NORTHWESTERN UNIVERSITY

Genetic and molecular mechanisms of phenotypic variation in the *Caenorhabditis elegans* species

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Biological Sciences

By

Stefan Zdraljevic

EVANSTON, ILLINOIS

September 2019

Abstract

Phenotypic variation is the functional unit that evolution acts upon and is the main contributor to the diversity of species. The phenotype of an individual is shaped by genetic and environmental factors [1]. Despite its necessity to all of life, the underlying genetic and environmental factors that contribute to phenotypic variation within a species also causes substantial human suffering when presented as a disease state or limit the efficacy of therapeutic interventions to treat disease. Therefore, to understand the mechanisms that underlie evolution and the factors that contribute to human health, it is critical to determine the effects of genetic and environmental factors on phenotypic variation. The development of high-throughput genomic sequencing has enabled the characterization of genetic variation on a massive scale. The characterization of genetic variation held promise to identify the genetic basis of many human diseases. Despite moderate success, the lofty aspirations of a comprehensive understanding of the genetic basis of human disease has been largely unsuccessful. A variety of reasons for this lack of success exist, including the underlying genetic complexity of disease phenotypes, genetic interactions, limited sample sizes, unaccounted for environmental variation, and gene-environment interactions. A particularly challenging phenotype to identify the genetic basis for in human populations is response to cancer chemotherapeutics. The main focus of my Ph.D. work has been to use *C. elegans* as a model system to identify and understand the effects of natural variation on chemotherapeutic responses.

Acknowledgments

It takes a village to raise a Ph.D. candidate, and this dissertation could not have been possible without the support I received from my village. First and foremost, I would like to acknowledge my advisor, Dr. Erik Andersen, who first inspired me to join his lab during the IBiS retreat. During my rotation in Erik's lab, I made a point of arriving early because I knew he was there, and we would talk for hours about a variety of topics. Beyond the research in Erik's lab, it was these conversations that motivated me to join his lab. Throughout my graduate education, Erik has provided constant support and advice on a wide range of topics, including science, interpersonal relationships, writing, navigating career paths, and the list goes on. Though our daily hour-long conversations tapered off as he became more busy, he always made himself available to talk when things came up. If such a thing exists, Erik is almost honest and open to a fault, but this is one of his characteristics that I admire most, because I am the same way. In fact, Erik and I have a few interesting similarities, including one pinky that can not be completely straightened, we are both Eagle Scouts, and we both probably work more than we should. It is his passion and enthusiasm for science that I will miss most in the lab.

For a lab to operate smoothly, a core team of individuals need to be present to keep the cogs oiled. Apart from Erik, Dr. Robyn Tanny is the glue that keeps the lab running smoothly and her role in keeping the lab functional can not be overstated. She has trained countless undergraduates and critical technicians that have enabled the lab to perform the amazing science that it does, including Sam Rosenberg, Sarah Giuliani, Kreena Patel, Briana Rodriguez, and Nicole Roberto. She has been unbelievably patient with me when I have been a bad lab citizen, which has only made me appreciate her role in the lab as the years went on. And, who doesn't like the dessert of their choosing on their birthday.

Dr. Daniel Cook joined the Andersen lab a few hours after I joined and we had a great relationship throughout graduate school. Dan was instrumental to getting the lab where it is today by laying the computational foundation on which the lab is built. Without Dan's constant help in all computational matters, I would not be nearly as computationally savvy as I am today. The amount of pestering that he had to deal with from me and others in the lab would have taken a toll on most people. Though I am sure it took its toll on Dan, he was always happy to take time out of his day to help, and for that I will be forever grateful. Despite all of the time he spent helping me and others, Dan was constantly exploring new avenues of research, to the point where he would have an idea one day and an answer within a week. He'd always joke that he was the "ideas guy", but his ability to take his ideas and put them into action made him much more than just the "ideas guy". Outside of the lab, we both shared a passion for biking. I'll never forget our trip together to Milwaukee, through rain and snow, getting lost a handful of times, and meeting then presidential candidate Bernie Sanders at a restaurant.

The first graduate student that I helped train in the lab was Dr. Shannon Brady during her rotation. Shannon really likes GWAS, a lot. It was immediately clear that Shannon and I have completely different personalities, where I am rough around the edges, a bit antisocial, and boisterous, my first impression of Shannon was that she was enthusiastic, engaging, and bubbly. At first, it seemed like our personality differences were going to make working together difficult for me. However, these initial impressions quickly dissipated as we worked and grew together. I was constantly in awe by Shannon's constant display of sheer grit as she kept hitting dead ends in her main dissertation project, which of course was supposed to be a "low hanging fruit" project. Apart from Erik, Shannon was the lab's "genetics guru", and I would often go to her for advice in all matters related to classical genetics. Her ability to clearly articulate difficult concepts have helped many people in the lab and beyond, and she has used this skill in various

outreach efforts. My graduate school experience would not have been nearly as enjoyable without Shannon in the lab, and our relationship has helped me become a more compassionate person.

Dr. Mostafa Zamanian was a postdoctoral researcher in the Andersen lab for my first two years. Though his time in the lab was short, we became quite close and I have learned a lot from him. Mostafa is one of the most intelligent people with whom I have ever had the pleasure to work, and I often came to him for advice on a wide range of topics. He was a second mentor to me while he was in the lab and has continued to advise me since he started his own lab. Mostafa has the unique ability to make people think differently, and anyone will be lucky to have him as a mentor. I look forward to keeping track of his progress as a PI and a baby daddy.

Dr. Steffen Hahnel was a postdoctoral researcher in the Andersen lab for a short stint. Steffen is one of the nicest and warmest people I have ever met. Despite being warned of my rancorous ways when he joined the lab, we quickly bonded. Steffen, Daehan (another postdoc that joined around the same time as Steffen), and myself went out for drinks the first opportunity we had. We all bonded instantly over science and life. I was fortunate enough to collaborate closely with Steffen on a project investigating the effect of natural genetic variation on *C. elegans* anthelmintic responses. This was a fruitful and fun collaboration, where I quickly learned that Steffen was a fount of knowledge on all matters related to parasitology. Our discussions helped shape my future scientific directions, and for that I am forever grateful. It was a sad day when Steffen left the lab, but I wish him all the best.

Two other postdocs in the lab with whom I have had the pleasure of collaborating are Dr. Daehan Lee and Dr. Timothy Crombie. Throughout our collaborations on separate projects, the three of us would often get into long and sometimes heated (Daehan, looking at you) discussions about science. When the conversation got heated, Tim would bring us back to reality and make us realize that Daehan and I were agreeing with each other, but getting lost in the nuance. During our 2017 lab collecting trip in Hawaii, Daehan and I bonded over philosophical topics after the day was over, while we sat by the moonlit ocean on Kauai. In 2018 Tim and I went on a separate collection trip in Hawaii, where we were with each other 24/7 for ten days. I've found that people can get on each other's nerves after spending that much time with each other, but that wasn't the case on our trip. Daehan and Tim's personalities are very different from each other, and having them both in the same room is always a treat. I wish Daehan the best of luck on his next journey, and look forward to seeing the product of Tim's hard work and dedication to science.

I would like to acknowledge other members of the Andersen Lab who supported the development of the projects in this dissertation. My long-time fellow graduate student Katie Evans Katie who is a talented scientist and has made me strongly reconsider Rmarkdown. My undergraduate mentees, Samuel Hamilton, Lily Kameny, and Thanda Kim. Clay Dilks joined the lab during my fifth year but has already impressed me with his work ethic and brings a calm, friendly energy to the lab. I had the pleasure of working with Gaotian Zhang, Ye Wang, Anita Huang, Ryan Chung, Kristen Laricchia, Josh Roberts, and Tyler Shimko during my time in graduate school, and they each contributed to the science that has come out of the Andersen Lab.

I am grateful to my committee members Dr. Richard Morimoto, Dr. Robert Holmgren, and Dr. Gregory Bietel who helped push my science forward during my qualifying exam and during each yearly meeting. In particular, I am grateful to Bob for answering every request for a letter of recommendation for fellowship and postdoctoral fellowships, for without his enthusiastic recommendations I would not be starting a new position in Dr. Leonid Kruglyak's lab in the fall. I would also like to thank Dr. Jason Brickner for letting me rotate in his lab, despite not being able to take a student that year. This rotation allowed me to test a hypothesis from my work as a research technician at the Molecular Sciences Institute (MSI). I am also

indebted to Jason for providing funding agencies and postdoctoral labs letters of recommendation on my behalf, and for always having an open door if I needed to discuss something.

I would likely not be in graduate school if it were not for Dr. Gustavo Pesce and Dr. Richard Yu, who were my first scientific mentors at MSI. After months of searching for a research position with no luck, I sent an email to Richard saying I would like to volunteer at MSI. During my very casual interview with Gustavo and Richard, they made it clear they wanted my full commitment and offered me a full-time paid research position. For the next three and a half years, I worked under the supervision of Gustavo who is as much of an intellectual as he is a smart ass, needless to say we got along swimmingly. It was during my time at MSI where I learned how to conduct scientific research, where I performed all aspects of the scientific process from cleaning test tubes, to making media, designing and implementing experiments, analyzing the data I generated, and writing. Critical to my successful transition from research technician to graduate student was Dr. Charles Denby, who was a postdoctoral researcher at MSI. During my graduate school application process, Charles took the time to go through my entire essay, line by line, and helped me find my voice when I had little scientific writing experience.

I obviously would not be where I am today without my parents, Mary and Sava Zdraljevic. Though they were not equipped to guide me in the intricacies of getting through graduate school, they provided me with a supportive upbringing and instilled in me the ideals that made me the person I am today.

Last, but certainly not least, graduate school would not have been possible without the love and support of my wife Tirzah Blanche. Tirzah and I met in college and quickly fell in love. For the last ten and a half years, Tirzah has helped me grow as a person in many ways. She put up with me during my rowdy college days and supported me when my the prospect of my finding a research position after undergraduate school looked bleak. Once I started my position at MSI, Tirzah was nothing but supportive until she realized how much I was willing to work because I felt the need to make a very good impression and do the best possible job I could. The amount of time that I put into work was both unhealthy and strained our relationship. I constantly insisted that I just wanted to get into graduate school, which is when my work load lighten. How wrong I was! Graduate school was just as demanding as my technician job, if not more so. It took a lot of persistence from Tirzah for me to establish a better work-life balance, which I am finally working toward. Despite my love of science and working toward discovery, there is more to life than work, and without Tirzah's encouragement, I probably would not have realized that. My gratitude for her support through the graduate process goes on, but I will just leave it at - she probably deserves 90% of the credit for my Ph.D..

List of Abbreviations

ABZ - Albendazole

ALT - Alternate allele

ANOVA - Analysis of variance statistical test

BCAA - Branched-chain amino acid

BCFA - Branched-chain fatty acid

BCKDH - Branched-chain α-keto acid dehydrogenase complex

BZ - Benzimidazole

BF - Bonferroni

C15ISO - 13-methyl myristic acid

C17ISO - 15-methyl hexadecanoic acid

C15SC - Pentadecanoic acid

C17SC - Heptadecanoic acid

CCA - Canonical correlation analysis

CeNDR - *C. elegans* Natural Diversity Resource

CRISPR - Clustered regularly interspaced short palindromic repeats

DALY - Disability-adjusted life year

DEL - Deletion

DUP - Duplication

EMMA - Efficient mixed model association

EMMAx - Efficient mixed model association expedited

eQTL - Expression quantitative trait locus

EXT - Optical density, extinction

FDR - False discovery rate

GFP - Green fluorescence protein

GWA - Genome-wide association

GWER - Genome-wide error rate

HDR - Homology-driven repair

HTA - High-throughput assay

Indel - Insertion or deletion

IQR - Interquartile range

INV - Inversion

Ka/Ks - The ratio of nonsynonymous to synonymous changes

KGDH - α-ketoglutarate dehydrogenase

L1-4 - Larval stages of *C. elegans*

LD - Linkage disequilibrium

LOD - Log of the odds

LoF - Loss of function

MAF - Minor allele frequency

MDA - Massive drug administration

mmBCFAs - Mono-methyl branched-chain fatty acids

NIL - Near-isogenic line

NHEJ - Non-homologous end joining

NTDs - Neglected tropical diseases

PC - Principal component

PCA - Principal component analysis

PCR - Polymerase chain reaction

PDH - pyruvate dehydrogenase

QTL - Quantitative trait locus

RIAIL - Recombinant inbred advanced intercross line

RIL - Recombinant inbred line

SKAT - Sequence kernel association test

SNP - Single-nucleotide polymorphism

SNV - Single-nucleotide variant

SV - Structural variant

TOF - Time of flight

VT - Variable threshold

UTR - Untranslated region

WHO - World health organization

Glossary

***ben-1* -** One of six C. elegans beta-tubulin orthologs

BIOSORT - A large particle flow cytometer developed by Union Biometrica

CB4856 - A wild isolate of *C. elegans* from Hawaii

CSV - Comma separated value data format

***dbt-1* -** Lipoamide acyltransferase component of branched-chain alpha-keto acid dehydrogenase complex

DL238 - A wild isolate of *C. elegans* from Hawaii

FASTA - A text-based format for representing genomic sequences

***H^2* -** Broad-sense heritability estimate

***h^2* -** Narrow-sense heritability estimate

JU258 - A wild isolate of *C. elegans* from Madeira

K medium - An enriched saline solution for *C. elegans* growth

M9 - Minimal medium for *C. elegans* growth

N2 - The canonical laboratory strain of *C. elegans* from Bristol, England

NGMA - Normal growth media with 1% agar and 0.7% agarose to prevent burrowing

***top-2* -** The gene that encodes the only *C. elegans* topoisomerase II protein

TSV - Tab-separated value data format

VCF - Variant caller format for representing sequence data

Dedication

For Mary and Sava Zdraljevic, and my loving partner Tirzah Blanche

Table of Contents

List of Tables

List of Figures

Chapter 1:

1. Introduction

Organismal fitness depends on adaptation to complex niches where chemical compounds and pathogens are omnipresent. These stresses can lead to the fixation of alleles in both xenobiotic responses and proliferative signaling pathways that promote survival in these niches. However, both xenobiotic responses and proliferative pathways vary within and among species. For example, genetic differences can accumulate within populations because xenobiotic exposures are not constant and selection is variable. Additionally, neutral genetic variation can accumulate in conserved proliferative pathway genes because these systems are robust to genetic perturbations given their essential roles in normal cell-fate specification. For these reasons, sensitizing mutations or chemical perturbations can disrupt pathways and reveal cryptic variation. With this fundamental view of how organisms respond to cytotoxic compounds and cryptic variation in conserved signaling pathways, it is not surprising that human patients have highly variable responses to chemotherapeutic compounds. These different responses result in the low FDA approval rates for chemotherapeutics and underscore the need for new approaches to understand these diseases and therapeutic interventions. Model organisms, especially the classic *Caenorhabditis elegans*, can be used to combine studies of natural variation across populations with responses to both xenobiotic compounds and chemotherapeutics targeted to conserved proliferative signaling pathways.

Responses to cancer chemotherapeutics vary across natural populations

In their natural habitats, metazoans are exposed to small molecules produced by bacteria, fungi, and plants as defense mechanisms to prevent predation. Modern medicinal chemistry has employed these cytotoxic small molecules to treat human diseases, so that approximately 70% of cancer chemotherapeutics (hereafter chemotherapeutics) developed from 1981-2010 were

derived originally from natural products [2]. Oftentimes, these small molecules disrupt essential cellular processes and can act as strong selective pressures that reduce genetic diversity [3]. By contrast, the combinations of small molecules in ecological niches change over time and can maintain genetic diversity within a species through balancing selection [4]. In addition to xenobiotic compounds, targeted chemotherapeutics specifically perturb the signaling pathways mutated in human cancers and are often lauded as great successes of personalized medicine. However, little evidence for their efficacy across a wide-range of genetically distinct patients exists because these proliferative signaling pathways evolved mechanisms to withstand the accumulation of genetic variation within populations [5]. Therefore, for both xenobiotic and targeted chemotherapeutics, it is not surprising that responses to chemotherapeutics are highly variable among the human population [6].

Variability in patient responses to chemotherapeutics can be caused by differences in the drug mechanism of action, absorption, metabolism, and elimination. Additionally, these processes can be impacted by germline variation, rare somatic mutations in the target tumor, environmental factors, and interactions among these factors and others [6]. This complexity results in a narrow range of concentrations that cause maximal tumor clearance among patients (defined as the therapeutic index). Also, chemotherapeutics are the most toxic drugs that are prescribed and cause severe and variable side effects among patient populations, thereby limiting the therapeutic index. In order to tailor treatments to individuals, drug responses must be correlated with genetic variants in specific patients. These data provide markers to broaden the therapeutic index for specific patients. The identification of genetic determinants that contribute to variable chemotherapeutic responses largely depends on the sample size of the patient population, the allele frequency and effect size of the causative variant(s), and the reliability of the responses being measured [7]. These factors are limited in clinical oncology because it is extremely difficult to acquire large cohorts of patients that undergo the same

therapeutic regimen [8], the high levels of genetic heterogeneity present in tumor [9] and patient populations [10], and the confounding effects of environmental variability [11,12]. As a result, only 6.4% of anti-cancer compounds in phase I clinical trials become FDA-approved chemotherapeutics, which is the lowest of any drug class [13]. Even if these limitations were resolved and genetic markers were associated with variable chemotherapeutic responses, the underlying mechanisms that are affected by the causal genetic variants would remain unknown, limiting clinical applications to recommendations based solely on genetic information.

Chemotherapeutic drug responses are conserved

The invertebrate model organism, *Caenorhabditis elegans*, has long facilitated the discoveries of molecular mechanisms associated with therapeutic responses [14]. This system enable the study of chemotherapeutic effects because xenobiotic-response pathways are highly conserved between invertebrates and humans [15], including cytochrome P450s [16,17], UDP-glucuronosyltransferases [3], and ABC transporters [18]. For example, two recent studies in *C. elegans* [19] and human cell lines [20] found that the widely administered chemotherapeutic cisplatin induces the same mutagenic profile in both species, suggesting that DNA repair mechanisms are shared. Additionally, the utility of *C. elegans* can be extended to chemotherapeutics that target cell proliferation pathways often constitutively activated in human cancers [21]. Because most of these pathways were discovered and characterized in studies of *C. elegans* vulval development and *D. melanogaster* compound eye development [22], the relevance of tractable models to understand conserved signaling pathways is long-standing. Cellular over proliferation associated with activating mutations in Ras pathway components have been shown to be conserved between *C. elegans* and humans [23]. For example, the severity of different activating mutations in the Ras pathway kinase, MEK1, and the suppressive effects of a MEK1 inhibitor have the same rank orders between invertebrates and vertebrates

[24]. Although this highlighted example and others are important for the understanding of cytotoxic and targeted chemotherapeutic responses, most studies have been performed only in a single genetic background without any consideration of natural genetic variation.

The effect of genetic background on chemotherapeutic responses

Individuals across populations harbor seemingly neutral genetic variation that causes phenotypic differences in the presence of chemical perturbations. This cryptic variation can cause large and divergent responses to chemotherapeutic regimens across cancer patient populations. Pharmacogenetics, pharmacogenomics, and genome-wide association studies of patient responses to chemotherapeutics focus on the identification and characterization of this genetic variation, but few broadly applicable results have been obtained [25]. Therefore, new approaches must be taken to understand how physiological responses to chemotherapeutics are affected by the genetic makeup of an individual without the difficulties associated with clinical oncology studies.

The *C. elegans* community has developed numerous strain resources with divergent genetic backgrounds, including wild isolates with whole-genome sequence data [26,27] and recombinant inbred lines (RILs) generated by crossing distinct genetic backgrounds [28,29]. WIthin *C. elegans*, drug responses generally affect fitness, including offspring production, growth rate, and viability. High-throughput assays have been developed to quantify these traits across a large number of individuals in tightly controlled environmental conditions. When applied to studying the effects of chemotherapeutics on diverse genetic backgrounds, these powerful assays enable the identification of genomic regions (quantitative trait loci or QTL) that vary across the population and are predictive of drug response [28,30–33] because environmental conditions are strictly controlled, drug responses from large numbers of divergent

individuals can be measured, and high levels of replication can be obtained. Additionally, the abundance of genome-editing tools available in *C. elegans* allow the functional validation and molecular characterization of genetic variants associated with chemotherapeutic responses [34]. Through these resources, assays, and genetic tools, investigators can rapidly go from a difference in drug response to the variant underlying that phenotypic difference.

Characterizing the effect of genetic variation on chemotherapeutic responses

To characterize the effect genetic variation has on phenotypic differences in a population, accurate genotype and phenotype information are required. Once accurate genotype and phenotype information are acquired, methods to identify meaningful associations between the two are required. In the following sections, I describe the methods we used to accurately define genotypes and phenotypes, and to identify associations between the two.

Genetic variation among *C. elegans* wild isolates

Over the last 50 years, the nematode *Caenorhabditis elegans* has been central to many important discoveries in the fields of developmental, cellular, and molecular biology. The vast majority of these insights came from the study of a single laboratory-adapted strain collected in Bristol, England known as N2 [35–42]. Recent sampling efforts have led to the identification of numerous wild *C. elegans* strains and enabled the study of genetic diversity and ecology of the species [43–50]. The earliest studies of *C. elegans* genetic variation showed that patterns of single-nucleotide variant (SNV) diversity were shared among most wild strains, with the exception of a Hawaiian strain, CB4856, which has distinct and high levels of variation relative to the other strains analyzed [51]. Subsequent analyses revealed that *C. elegans* has reduced levels of diversity relative to obligate outcrossing *Caenorhabditis* species and the facultative

selfer *C. briggsae* [52,53]. The most comprehensive analysis of *C. elegans* genetic diversity to date used data from thousands of genome fragments across a globally distributed collection of 97 genetically distinct individuals to show that recent selective sweeps have largely homogenized the genomes of a majority of individuals in the species [43]. The authors hypothesized that these selective sweeps might contain loci that facilitate human-assisted dispersal and/or increase fitness in human-associated habitats. Consistent with previous analyses, two Hawaiian strains, CB4856 and DL238, did not share patterns of reduced genetic diversity caused by the selective sweeps that affected the rest of the *C. elegans* population – a trend that has held true as the number of Hawaiian strains has increased [27,44,48,49]. Taken together, these studies suggest that the Hawaiian *C. elegans* population might be more representative of ancestral genetic diversity that existed prior to the selective pressures associated with recent human influence.

To date, the Andersen lab has amassed whole-genome sequence data for 330 distinct *C. elegans* strains. These data are available through the *C. elegans* Natural Diversity Resource (CeNDR) [26]. In addition to raw sequence data, CeNDR also contains genetic variant calls for the population of *C. elegans* wild isolates. The goal of CeNDR is to facilitate *C. elegans* researchers with less genomics experience to explore the effect of genetic variation on phenotypic differences among wild isolates. In Chapter 2, I describe the analysis of the whole-genome sequence data for the 330 wild isolates.

Phenotypic variation

To assess the effect of a particular chemotherapeutic perturbation on an individual *C. elegans* strain, we make use of a high-throughput fitness assay (HTA). The HTA depends on a large-particle flow cytometer (COPAS BIOSORT) that can accurately quantify nematode length, optical density, and brood size. We use these three animal phenotypes as proxies for drug

responses because as nematodes undergo developmental progression they increase in size and optical density and generate offspring. This high-throughput approach to accurately quantify drug effects in *C. elegans* is crucial to acquire the necessary strain phenotype data for association mapping. However, this assay is not limited to assessing the effects of drugs and has been applied to study growth rate [54] and dauer formation [48].

For a given HTA experiment, strains are passaged for four generations to reduce transgenerational effects from starvation or other stresses. Strains are then bleach-synchronized and aliquoted to 96-well microtiter plates at approximately one embryo per microliter in K medium [55]. Embryos are then hatched overnight to the L1 larval stage. The following day, hatched L1 animals are fed HB101 bacterial lysate at a final concentration of 5 mg/ml and grown to the L4 stage after two days at 20ºC. Three L4 larvae are then sorted using the COPAS BIOSORT into microtiter plates that contain HB101 lysate at 10 mg/ml, K medium, and either drug dissolved in 1% DMSO or 1% DMSO. The animals are then grown for four days at 20ºC. During this time, the animals will mature to adulthood and lay embryos that comprise the next generation. Prior to the measurement of fitness parameters from the population, animals are treated with sodium azide to straighten their bodies for more accurate length measurements. A figure representation of the assay is depicted in Figure 1-1.

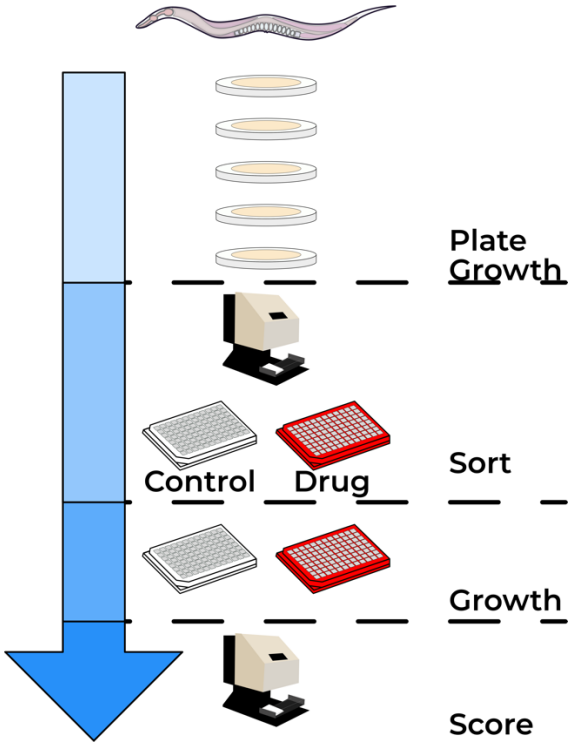

Figure 1-1 HTA workflow.
Individual strains are passaged for four generations on agar plates seeded with OP50 bacteria by transfer of five L4 larval stage animals to a fresh plate each generation every three days. Animals are bleach synchronized and aliquoted to 96-well microtiter plates in 50 µl of K medium at a concentration of one embryo per µL. Aliquoted embryos are incubated overnight at 20°C. The following day, 5 µl of 50 mg/ml HB101 lysate is added to each well. Animals are then grown for two days to the L4 larval stage. Then, three L4 animals are sorted to assay plates containing drug or DMSO using the BIOSORT. Four days later, 200 µl M9 plus 50 mM sodium azide is added to each well and strains are scored using the BIOSORT.

Genotype - Phenotype association

To perform genotype-phenotype association mapping, it is critical that genetically different strains have quantifiably different phenotypes. To assess how much genetic variation affects phenotypic differences among wild *C. elegans* strains, we estimate broad-sense heritability (H^2) for a small panel of strains prior to performing a large-scale assay. This estimate allows us to determine if there are reproducible phenotypic differences among phenotyped strains by accounting for within- and among-strain phenotypic variation. The level of phenotypic variation that can be explained by genetic factors influences the experimental design for a genome-wide mapping experiment, where lower levels of heritability require more strains to be phenotyped. In

general, a phenotype is a good candidate for genome-wide association mapping if more than 20% of the phenotypic variation in the heritability panel can be explained by genetic factors (Figure 1-2). The Andersen lab uses two independent genetic mapping approaches; linkage mapping and genome-wide association mapping. The goal of both of these approaches is to identify genomic loci that explain phenotypic differences among phenotyped strains, referred to as quantitative trait loci or QTL.

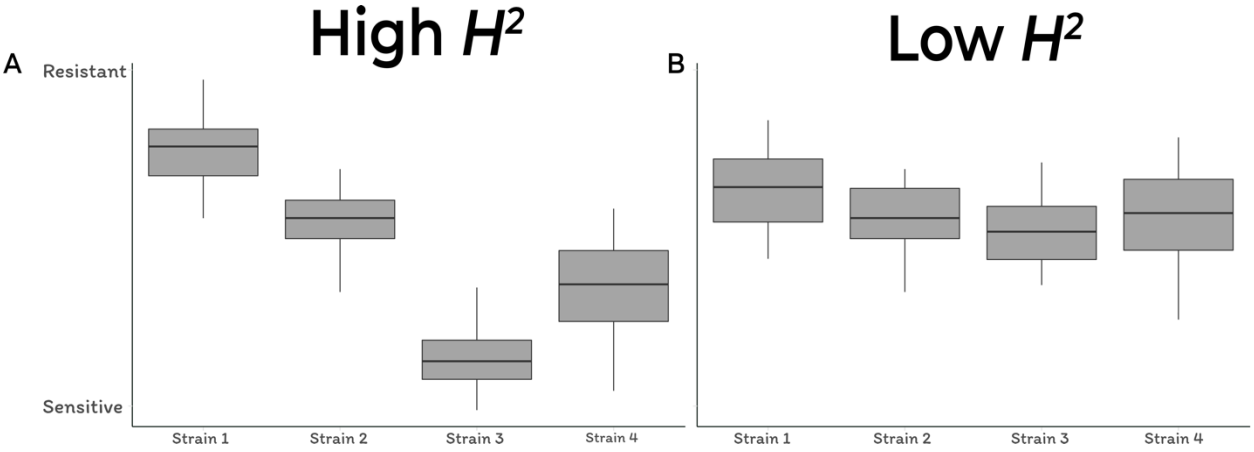

Figure 1-2 Heritability experiment.
In this example, four genetically divergent strains are phenotyped in the presence of a toxin. Each strain's phenotype is measured in multiple replicates and visualized using Tukey boxplots. The y-axis represents individual strain phenotypes in the presence of the drug. A) An illustrative example of trait with high broad-sense heritability (H^2) and B) an illustrative example of trait with low broad-sense heritability (H^2). The trait with high H^2 will be a better candidate for a large-scale genome-wide mapping experiment.

Linkage mapping

The first approach to identify QTL uses a panel of recombinant inbred advanced intercross lines (RIAILs) that have been generated from two diverged *C. elegans* strains, N2 and CB4856. These lines were generated by ten consecutive generations of random pair mating, starting with an F3 population generated from the three consecutive crosses 1) N2 (male) x CB4856 (hermaphrodite) and CB4856 (hermaphrodite) x N2 (male), 2) four possible F1 crosses, and 3) four possible F2 crosses [54]. Mating was followed by six generations of selfing to homozygose the genome. Generated RIAILs were then genotyped at 1400 markers across the genomes and cryopreserved to prevent the accumulation of new mutations. Once cryopreserved, the RIAILs

can be thawed and put through the HTA for any trait where the N2 and CB4856 strains have distinct phenotypes. Once RIAIL phenotypes are acquired, the log of the odds ratio (LOD score) is calculated for every genotyped marker using the formula LOD = $log(1\ cor(y, g)2)/2log(10)$, where y is the measured phenotype and g is the genotype information [23, 24].

Genome-wide association mapping

The second approach to identify QTL that is commonly used in the Andersen Lab is genome-wide association (GWA) mapping. In contrast to linkage mapping, the GWA mapping approaches attempts to identify genetic variants associated with phenotypic differences among all strains in the *C. elegans* species. All wild isolates are thawed prior to a mapping experiment using the HTA. Once phenotype data are generated and processed, we determine if the genotype at a given position in the genome explains phenotypic differences among individuals. Traditional tests for making this type of association include the t-test, the Wilcoxon rank-sum test, linear regression, or analysis of variance. However, these statistical tests often lead to spurious associations by assuming independence between the genetic marker being tested and the measured phenotype [56]. The independence assumption can be violated when studying populations of individuals because of population structure [57]. For example, if a subset of the assayed population is closely related and susceptible to a topoisomerase poison, then genomic regions corresponding to that subset's relatedness will be significantly associated with the susceptibility phenotype. This confounding effect can be greatly reduced by incorporating a strain x strain relatedness matrix, *K,* as a random effect in a linear mixed-model with equation $y = X + Zu + e$, where *y* is the measured phenotype, *X* are fixed effects such as the SNP to be tested for association, *Zu* is the phenotype *y* corrected for population structure *K*, and *e* are residual effects [56]. This methodology was developed for performing GWA on inbred populations, but should be applicable in our system as well. To verify that this is the case, I simulated QTL at every SNV in the genome and tested our ability to identify the QTL. The

EMMA algorithm developed by Kang *et al.* outperformed traditional methods such as the Wilcoxon rank-sum test and the t-test, and other more sophisticated mixed-model approaches like EMMAx and MLMM [58,59].

From QTL to causal variant

QTL often span large genomic regions that contain many genes that harbor genetic variation in the phenotyped population. As the goal of quantitative genetics is often to understand the genetic basis of phenotypic variation in a wild population, strategies are required to narrow the search space for the causal variant. One such strategy is to first isolate the QTL region in an clean genetic background and perform speed congenics [60,61]. This strategy relies on first generating near-isogenic lines (NILs), where the genomic region that overlaps an identified QTL of one strain is introgressed into another genetic background through a series of crosses (Figure 1-3A). Constructed NILs can then be phenotyped to determine if the introgressed genomic region recapitulates the effect of the QTL. If a NIL recapitulates the QTL effect, the NIL can be backcrossed to the parental strain and progeny can be screened for recombinants within the NIL interval. This approach will identify recombinant progeny that have smaller introgressed regions that can be phenotyped (Figure 1-3B). The result of this process is to narrow the search space for potential genetic variants contributing to phenotypic differences in the original population.

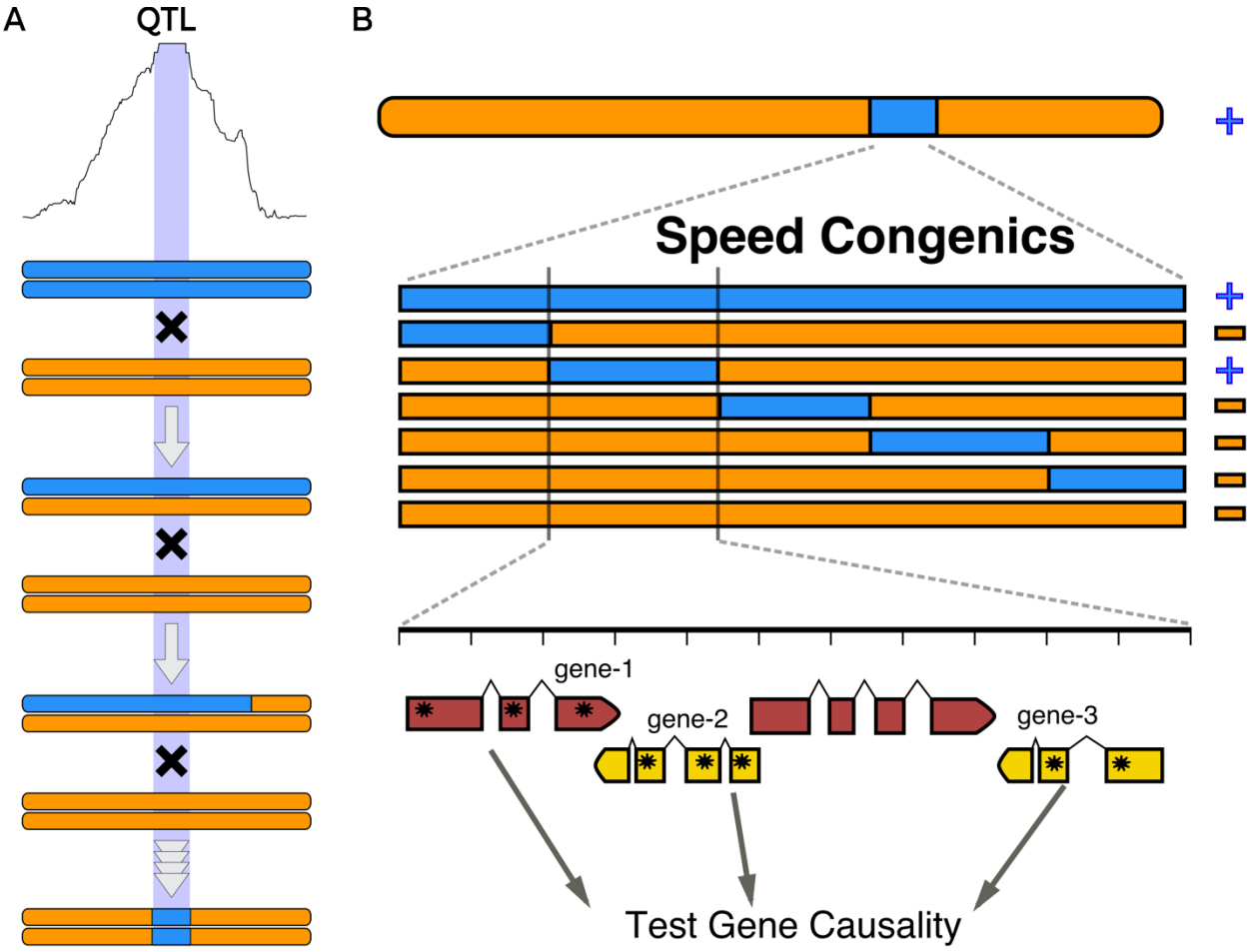

Figure 1-3 From QTL to causal gene.
A) Flow diagram for NIL construction. A QTL is identified on a chromosome with a corresponding confidence interval (purple rectangle). Two phenotypically different parental strains (orange and blue) are crossed to each other to generate a heterozygote. Insertion or deletion variants (indels) that differ between the parental strains and flank the QTL confidence interval can be used to verify that recombination did not occur within the QTL region being maintained. Six consecutive generations of back crossing to the orange parental strain will homogenize the genetic background such that the resulting strain will have the parental genotype in approximately 99% of the genome and the introgressed region of interest with the opposite genotype. Crosses will be followed by six consecutive generations of hermaphroditic selfing to homozygose any portion of the genome that was not converted to the orange genotype. Generated NILs are then phenotyped to confirm that the introgressed genomic region confers the expected phenotype. B) NILs are then broken up further using speed congenics and phenotyped. In the speed congenic approach, a NIL is backcrossed to the orange parent for two generations. Progeny are then distributed to a single well and allowed to grow. Genotyping is performed on a series of indels across the genomic interval to identify recombinants. Wells that contain recombinants are transferred from wells to agar plates. Individual animals are then distributed to plates and allowed to self and re-genotyped to identify the recombinant progeny. The confidence interval is narrowed down to the genomic region that confers the phenotype of interest, in this case + refers to resistance. Genes with variants in this region will then be tested for causality.

Once a QTL interval has been narrowed, targeted genome editing can be performed to causally connect a genetic variant with phenotypic differences. The RNA-dependent DNA endonuclease Cas9 of *Streptococcus pyogenes* has been adapted by researchers to generate targeted edits in a variety of laboratory model systems, including *C. elegans* [62,63]. The Cas9 protein can be targeted to a specific genomic position by a complementary strand of RNA, at which point it creates a double-stranded break [63]. If a homologous piece of DNA is provided homology-driven repair (HDR) will initiate to generate a precise edit [64–69]. With this approach, it has become straightforward to test the phenotypic effect of multiple genetic variants within a narrowed QTL.

The rate limiting step in the process of going from a phenotypic difference among genetically divergent individuals to a causal genetic variant is the identification of a causal genetic variant. In Chapters 3 - 6, I will discuss ways in which this process can be expedited.

2. Hyper-divergent genomic regions are prevalent throughout the *C. elegans* genome

Preface

Though the research I discuss in this chapter is the work I started latest and is still very much a work in progress, I present it first because a catalog of genetic variation is required for all quantitative genetics studies. I started this project after spending many hours scrolling through the alignment files of many genetically divergent *C. elegans* strains and noticed many regions with elevated levels of genetic diversity, relative to nearby regions. As Erik and a few members of the lab were discussing what the message of the next strain set publication was going to be, we decided that the characterization of these hyper-variable regions would be an interesting research project.

Abstract

Caenorhabditis elegans is a globally distributed bacterivore nematode species that inhabits a variety of ecological niches. In the present study, we characterized single-nucleotide, indel, and structural variation of 330 genetically distinct *C. elegans* strains. The genetic diversity of the worldwide *C. elegans* population is low, because of the recent chromosome-scale selective sweeps. However, many strains, including those that have undergone the selective sweep, have retained highly variable genomic regions. We find that approximately 70% of all single-nucleotide and indel variation present in the *C. elegans* species is localized to these genomic regions that, in total, span 3% percent of the genome. Interestingly, 30% of these hyper-divergent regions were present in more than 17 strains, which suggests that they are actively maintained in the population. Gene-enrichment analysis of the genetically divergent strains

present in at least 17 strains revealed significant enrichment of genes associated with environmental sensing, interaction with bacterial food, and response to pathogens. We hypothesize that these genetically divergent genomic regions might provide fitness advantages in specific niches.

Introduction

The nematode *Caenorhabditis elegans* has been central to many important discoveries in the fields of developmental, cellular, and molecular biology. Though the vast majority of these insights came from the study of a single laboratory-adapted strain collected in Bristol, England known as N2 [35–42], recent sampling efforts have led to the identification of hundreds of genetically distinct *C. elegans* strains [27,44,43]. Analysis of the genomic content of a subset of these strains revealed that recent selective sweeps have largely homogenized the genomes of individuals in the species [43]. The homogenization of the *C. elegans* species is a clear example of the effects of strong positive selection events [70]. However, many new *C. elegans* strains that have avoided this genomic homogenization have been identified since this initial study [27,44,71].

To better characterize the genetic diversity of the *C. elegans* species, we have isolated and analyzed the genomic content of 330 genetically distinct *C. elegans* strains. Many of observations from the previous analysis of 97 strains held true in this expanded strains set, including the presence of large-scale selective sweeps of a single haplotype. However, within our expanded collection there are many new strains that are remarkably divergent from the majority of the swept population. Furthermore, by acquiring and analyzing whole-genome sequence data, we have identified many genomic regions that have unusually high levels of variation within individual strains, relative to neighboring regions. We find that these genetically

divergent genomic regions are not unique to divergent strains, with 30% of the divergent regions being shared by more than five percent of the analyzed strains. Altogether, these genetically divergent regions cover approximately four percent of the entire *C. elegans* genome and contain nearly 70% of the genetic variation with the species. Though a majority of genetic variation within the species are in these divergent regions, we find that their presence does not have a large effect on the relatedness of individuals.

Closer inspection of the gene content within these divergent regions revealed significant enrichment of genes involved in environmental sensing and response to pathogens. These findings suggest that balancing selection is acting on these loci to maintain strain advantages in different environments. However, classical methods for detecting signatures of balancing selection do not recapitulate the same enrichment of environmental sensing genes. We hypothesize that this discrepancy is the result of long-term balancing selection acting on these loci, which classical tests of balancing selection have limited power to detect [72]. Therefore, our results highlight a novel method to detect long-term balancing selection and provide an important dataset that will enable a comprehensive understanding of the ecology and evolution of the *C. elegans* species.

Materials and Methods

Illumina library construction and whole-genome sequencing

To extract DNA, we transferred nematodes from two 10 cm NGMA plates spotted with OP50 *E. coli* into a 15 ml conical tube by washing with 10 mL of M9. We then used gravity to settle animals on the bottom of the conical tube, removed the supernatant, and added 10 mL of fresh M9. We repeated this wash method three times over the course of one hour to serially dilute the

E. coli in the M9 and allow the animals time to purge ingested *E. coli*. Genomic DNA was isolated from 100-300 µl nematode pellets using the Blood and Tissue DNA isolation kit cat# 69506 (QIAGEN, Valencia, CA) following established protocols [44]. The DNA concentration was determined for each sample with the Qubit dsDNA Broad Range Assay Kit cat# Q32850 (Invitrogen, Carlsbad, CA). The DNA samples were then submitted to the Duke Center for Genomic and Computational Biology per their requirements. The Illumina library construction and sequencing were performed at Duke University using KAPA Hyper Prep kits (Kapa Biosystems, Wilmington, MA) and the Illumina NovaSeq 6000 platform (paired-end 150 bp reads).

Small variant calling

To ensure reproducible data analysis, all genomic analyses were performed using pipelines generated in the Nextflow workflow management system framework [73]. Each Nextflow pipeline used in this study is briefly described below. All pipelines follow the "*pipeline name-nf*" naming convention and full descriptions can be found on the Andersen lab dry-guide website: (http://andersenlab.org/dry-guide/pipeline-overview/).

Raw sequencing reads were trimmed using *trimmomatic-nf*, which uses trimmomatic (v0.36) [74] to remove low-quality bases and adapter sequences. Following trimming, we used the *concordance-nf* pipeline to characterize *C. elegans* strains isolated in this study and previously described strains [27,44,49]. The *concordance-nf* pipeline calls single-nucleotide variants using the BCFtools (v.1.9) [75] variant calling software. The variants are filtered by: Depth (FORMAT/DP) ≥ 3; Mapping Quality (INFO/MQ) > 40; Variant quality (QUAL) > 30; (Allelic Depth (FORMAT/AD) / Num of high quality bases (FORMAT/DP)) ratio > 0.5. We determined the pairwise similarity of all strains by calculating the fraction of shared SNVs. Finally, we

classified two or more strains as the same isotype if they shared >99.9% SNVs. If a strain did not meet this criterion, we considered it as a unique isotype. Newly assigned isotypes were added to CeNDR [27].

After isotypes are assigned, we used *alignment-nf* with BWA (v0.7.17-r1188) [76,77] to align trimmed sequence data for distinct isotypes to the N2 reference genome (WS245) [78]. Next, we called single-nucleotide variants using GATK4 [79] and Strelka2 [80] [75]. The *wi-gatk* pipeline generates two population-wide VCFs that we refer to as the soft-filtered and hard-filtered VCFs. Variant calling for individual strains was performed with the GATK *HaplotypeCaller* command with the following parameters: *--emit-ref-confidence GVCF --genotyping-mode DISCOVERY --max-genotype-count 3000 --max-alternate-alleles 100*. Individual strain gVCF files were merged using the GATK *MergeVcfs* command. The merged gVCFs were then imported to a database using the GATK *GenomicsDBImport* command. Finally, the cohort variant database was genotyped using the GATK *GenotypeGVCFs* command. After variant calling, a soft-filtered VCF was generated for each sample by appending the following soft-filters to variant sites: Depth (FORMAT/DP) > 5; Mapping Quality (INFO/MQ) > 30; Variant quality (QUAL) > 30; (Allelic Depth (FORMAT/AD) / Number of high quality bases (FORMAT/DP)) ratio > 0.5, (INFO/ReadPosRankSum) < -5.0, (INFO/FS) > 50, (INFO/QD) < 50, (INFO/SOR) > 5. We refer to this VCF as the soft-filtered VCF. To construct the hard-filter VCF, we removed all variants that did not pass the filters described above.

We called small variants with Strelka2 using the *strelka-nf* pipeline. We used the *configureStrelkaGermlineWorkflow.py* script with default parameters to initialize the variant calling parameters, with ploidy across the genome set to one. Next, small variants were called using the *runWorkflow.py* script and individual sample VCFs were merged using the BCFtools *merge* command with the following parameters: *-m both --missing-to-ref*, and normalized using

the BCFtools *norm* command with the parameter: *-m -any*. Next, variants were recalled for individual strains using the *runWorkflow.py* script after the Manta workflow was configured with the *configureStrelkaGermlineWorkflow.py* script using the following parameters: *--forcedGT --ploidy 1*. The recalled strain VCFs were then merged using the BCFtools *merge* command with the same parameters as above. We used the default Strelka2 filter parameters to define high quality variants.

For each of the small variant cohort VCFs that was generated, we used SnpEff (v4.3t) to annotate the predicted effect of each structural variant [81]. We used the WS263 reference genome for SnpEff predictions. We ran SnpEff with the following parameters: *-no-downstream -no-intergenic -no-upstream -nodownload* and used the invertebrate codon table. Finally, we used the BCFtools *isec* command to identify variants that were found by both GATK4 and Strelka2. We refer to this VCF as the Intersection VCF.

Structural variant calling

Structural variants were called using three structural variant callers Delly2 (v0.8.1) [82], smoove (v0.2.3) [83], and Manta (v1.4.0) [84]. Prior to calling structural variants BAM alignment files for each strain were downsampled to 100X depths using sambama (v0.6.8) [85]. We used WS245 as the reference genome for calling structural variants [86].

Delly2 was run with the standard workflow. First, structural variants were called for individual strains using the *call* command with default parameters. Next, sample BCF files were merged using the *merge* command with the following parameters: *-m 100 -n 100000 -b 500 -r 0.5*. Next, structural variants were re-called using the *call* command and only passing structural variants were retained. Passing variants were identified using the default Delly2 parameters. Finally, the re-called BCF files were merged using the BCFtools (v1.9) *merge* command with *-m id [87]*.

Smoove was run with the standard workflow. First, structural variants were called for individual strains using the *call* command with default parameters and *-p 1*. Next, sample VCF files were merged using the *merge* command with the default parameters. Next, structural variants were re-called using the *genotype* command with the following parameters: *-d -x -p 1*. We only retained passing structural variants that were less than 100 Kb. Passing variants were identified using the default smoove parameters. Finally, the re-called VCF files were merged using the smoove *paste* command.

Manta was run with the recommended workflow. First, structural variants were called for individual strains using the *runWorkflow.py* script with default parameters. Next, we removed variants that were larger than 100 Kb using the BCFtools *filter* command and only retained passing structural variants. Passing variants were identified using the default Manta parameters. Next, individual sample VCFs were merged using the BCFtools *merge* command with *-m all*.

For each of the three cohort VCFs that were generated, discussed above, we used SnpEff (v4.3t) to annotate the predicted effect of each structural variant [81]. We used the WS263 reference genome for SnpEff predictions. We ran SnpEff with the following parameters: *-no-downstream -no-intergenic -no-upstream -nodownload* and used the invertebrate codon table.

To combine the results from Delly2, smoove, and Manta, we used SURVIVOR (v1.0.3) [88]. For each sample VCF, we used the SURVIVOR *merge* command with the following parameters: *1000 1 0 0 1 30*. We then converted the individual merged VCFs to BED files using BCFtools *query* and combined all sample BED files to form a cohort BED file [89].

Identification of divergent regions

We identified divergent regions for individual strains based on a variety of metrics defined by variant count, alignment depth, and evidence of structural variants.

For each strain, we counted the number of variants in 1000 bp windows using the bedtools *coverage* command with the *-count* parameter. Next, we identified genomic windows that have unusually high variant counts relative to neighboring genomic windows using the anomalize R package [90,91]. We used the *time_decompose* function with the following parameters: *method = 'stl' and trend = 100*, followed by the *anomalize* function with the following parameters: *method = "gesd", alpha = 0.005, max_anoms = 0.05, verbose = T*. We next used the variant count per 1000 bp window of all strains to select a cutoff for the maximum number of variants to be considered a non-divergent window. To do this, we combined all strain variant count data and defined the variant count threshold as the number of variants that 1% of all windows had (23 variants per 1000 bp window). We used this variant threshold to define masks, discussed below.

In parallel to finding local outliers, we calculated the coverage for each strain BAM file for every 1000 bp window using mosdpeth (v0.2.5). Next, we identified windows where the coverage was aberrant relative to surrounding windows using the anomalize R package. We used the *time_decompose* function with the following parameters: *method = 'stl' and trend = 50*, followed by the *anomalize* function with the following parameters: *method = "gesd", alpha = 0.005, max_anoms = 0.05, verbose = T*.

We used strain-level structural variant BED files to define the number of structural variants per 1000 bp window. We used the BEDtools *coverage* command to determine the number of

structural variants that were present in each window. For this analysis, we required structural variants to be identified by at least two callers, except for insertion structural variants, which are only reliable identified by Manta. Additionally, we excluded the translocation variants from our analysis.

Once we had variant count, depth, local outlier windows identified,and structural variant count for each 1000 bp window, we generated masks for individuals strains that were applied in sequential order. The first masks that were applied were defined by windows that were identified to be local outliers based on coverage. These mask names are called Masked_Low_Coverage and Masked_High_Coverage. Next, we applied a mask on windows that were local outliers based on small variant counts and contained more than 23 variants, as defined by the variant threshold described above (named Masked_Outlier). Next, we applied a mask on windows that were local outliers based on small variant counts and had structural variants cover more than 50% of the window (named Masked_SV). Next, we applied a mask on windows that had structural variants cover more than 50% of the window and contained more variants than 23 x fraction of the window that was covered by structural variants (named Masked_SV_count). Next, we applied a mask on windows that contained support for more than three structural variants and had an average depth of coverage of less than 5 reads (named Masked_SV_count_cov). Finally, we applied a mask to windows that had more than 23 variants, as defined by the variant threshold, described above. To fill gaps between neighboring windows, we defined the Masked_Two_Flank mask for windows that had two masked windows next to them. Additionally, we defined the Masked_One_Flank_Outlier for windows that were next to one masked window and the window was a local outlier based on variant counts. We iterated this final mask three additional times to generate Mask_P3-4.

For subsequent analysis of masked variants, we set strain genotypes to missing if they were identified by one of the masks described above. Unless otherwise stated, all subsequent analyses were performed using this masked VCF and the GATK-Strelka2 Intersection VCF.

Gene enrichment of divergent regions

We concatenated individual strain masked regions and calculated the frequency of each masked region across the population. Next, we split the masks into three categories based on the window frequency, windows with less than 1% frequency, 1-5% frequency, and greater than 5% frequency. The resulting window sets were then queried for protein coding genes within them using BEDtools *intersect* function. The resulting gene sets were then used to perform enrichment analysis using the clusterProfiler R package [92]. We used the *enrichGO* function with the following parameters: *ont = "MF", pAdjustMethod = "BH", keyType = 'ENSEMBL', pvalueCutoff = 0.05, qvalueCutoff = 0.05*, to identify molecular function GO-term enrichment. We used the *enrichGO* function with the following parameters: *ont = "MF", pAdjustMethod = "BH", keyType = 'ENSEMBL', pvalueCutoff = 0.05, qvalueCutoff = 0.05* to identify molecular function GO-term enrichment, and *ont = "BP", pAdjustMethod = "BH", keyType = 'ENSEMBL', pvalueCutoff = 0.05, qvalueCutoff = 0.05* to identify biological process GO-term enrichment. We used the *gseKEGG* with the parameters: *organism = 'cel', nPerm = 1000, keyType = "uniprot", pvalueCutoff = 0.05, verbose = FALSE*.

Admixture analysis

We performed admixture analysis using ADMIXTURE (v1.3.0) [93]. Prior to running ADMIXTURE, we LD-pruned the VCFs using PLINK (v1.9) [94,95] with the command *--indep-pairwise 50 10 0.1*. We also removed variants only present in one isotype. We ran ADMIXTURE

ten independent times for K sizes ranging from 5 to 8. We ran this analysis on the Masked and Intersection VCF and compared the results of both analyses, described below.

Principal component analysis

We performed principal component analysis using the Eigenstrat *smartpca* command (v6.1.4) [96]. Prior to running Eigenstrat, we LD-pruned the VCFs using PLINK (v1.9) [94,95] with the command *--indep-pairwise 50 10 0.95*. We also removed variants only present in one isotype. We ran *smartpca* two ways: with removing outlier strains and with retaining outlier strains. We ran this analysis on the Masked and Intersection VCF and compared the results of both analyses, described below.

Comparison of population structure

We compared the results from Admixture and PC analysis run on the Masked and Intersection VCFs. To compare the results from Admixture analysis, we used the *cc* function from the CCA R package to perform canonical correlation analysis [97]. To compare the results of the PCA, we used the *cor* R function with the parameter: *method = "spearman"*.

Results

Distribution of small genetic variants

The most comprehensive analysis of *C. elegans* genetic diversity to date used data from thousands of genome fragments across a globally distributed collection of 97 genetically distinct strains to show that recent selective sweeps have largely homogenized the genome [43]. Through the collaborative effort of many labs and citizen scientists across the globe, we have acquired 330 wild *C. elegans* strains from around the world [27,43,44]. These strains have been isolated from distinct substrates across diverse habitats. To get a more complete understanding

of genetic diversity across the *C. elegans* species, we acquired whole-genome sequence data of these 330 strains and characterized single-nucleotide variants (SNVs), small insertion and deletion variants, and structural variants.

Across all wild strains, we identified 390,833 indel variants and 2,369,213 SNVs. The distributions of these variant classes across the genome were similar, with lower levels of variation on the chromosome centers than the arms for the autosomes and a relatively even distribution of variation across chromosome X (Figure 2-1A). To understand the predicted effect of each variant, we ran SnpEff [81]. The results of this analysis show that a majority of variants do not affect protein coding genes (71.5% intergenic or intronic variants). However, we did identify 240,236 variants that encode for missense variants and 19,749 variants that have predicted high effects on gene function, such as stop-gained, frameshift, and splice site variants. In total, 8,776 genes were affected by variation with predicted high effects. However, we note that the SnpEff is not a haplotype-aware prediction algorithm and the effect of some predicted high-effect variants might be compensated for by nearby variants. Variants with predicted high effects on gene function are enriched on chromosomal arms (Figure 2-1B). Of the various classes of predicted high-effect variants, frameshift and stop-gained variants were the most prevalent classes of variant effects (Figure 2-1C).

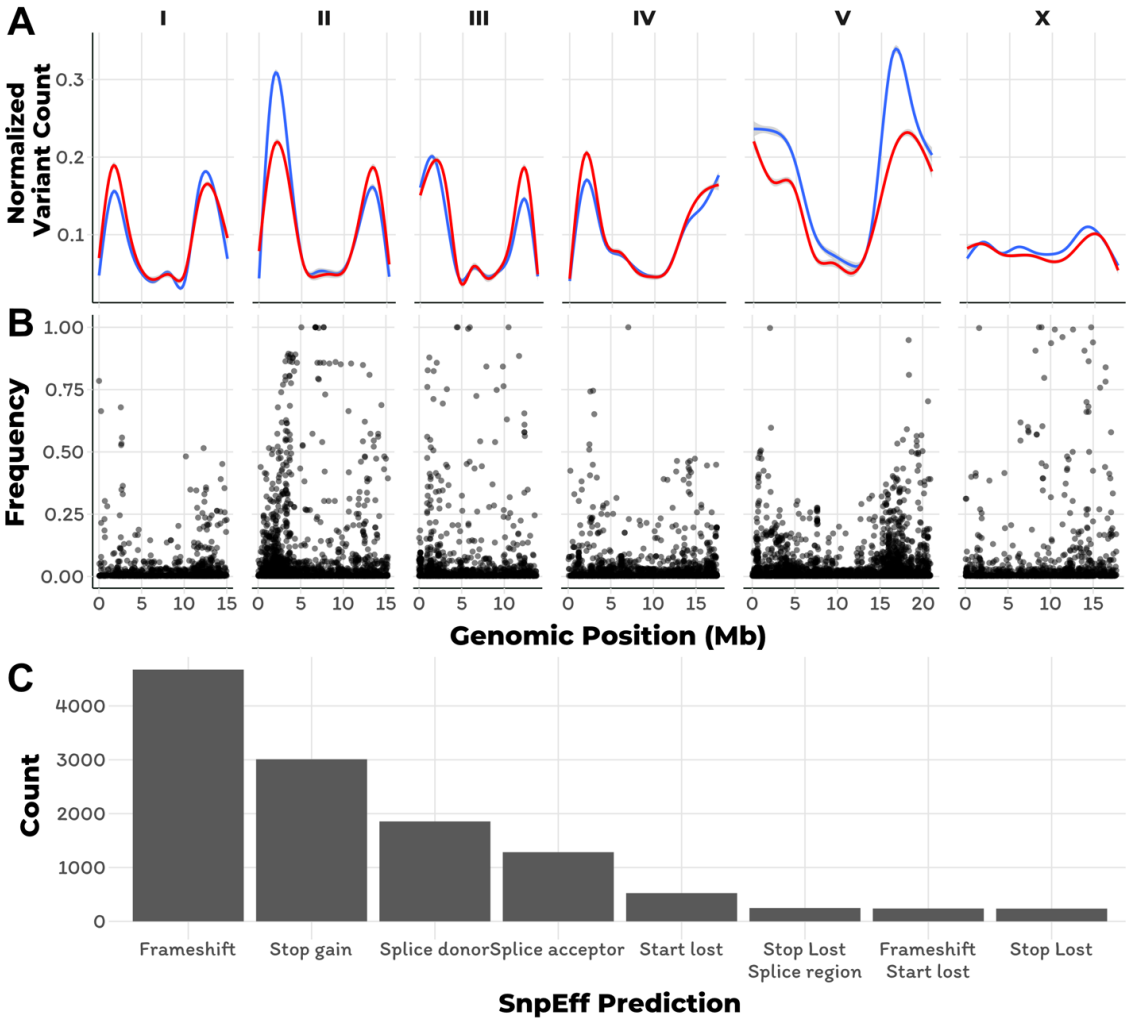

Figure 2-1 The genomic distribution of small variants and their predicted effect on gene function
A) Small variant counts across the six *C. elegans* chromosomes are shown (red = indel, blue = SNV). The y-axis represents the normalized variant counts across the genome on the x-axis, where each tick mark corresponds to 5 Mb. Variant counts were normalized to put them on the same scale. B) The genomic distribution of variants with predicted high effects on a gene's function is shown. The y-axis represents the variant's frequency in the *C. elegans* population. The x-axis represents the genomic location of the variants across each of the six *C. elegans* chromosomes, where each tick mark corresponds to 5 Mb C) The number of small variants with predicted high effects on a gene's function is shown. The y-axis represents the number of variants per predicted effect class on the x-axis. Only the eight effect classes with the most variants is shown.

We next looked at the variant distribution among individuals strains. Across all strains, the median number of variants per strain is 116,348 and the distribution of variants per strain is normally distributed (Figure 2-2A). Despite the normal distribution of variants per strain for most individuals, 23 strains have more variants than two IQR from the median. Of these divergent

strains, 18 were isolated on the Hawaiian islands (Figure 2-2B), which have previously been shown to harbor divergent *C. elegans* strains [27,43,44]. The strain with the most genetic variants is XZ1516, with 574,210 SNVs and indel variants, which is approximately five times more variation than the median of the entire population. This strain, along with the two strains with the next highest variation content, ECA701 (487,929 variants) and XZ1514 (429,117 variants), were all isolated from the Hawaiian island Kauai.

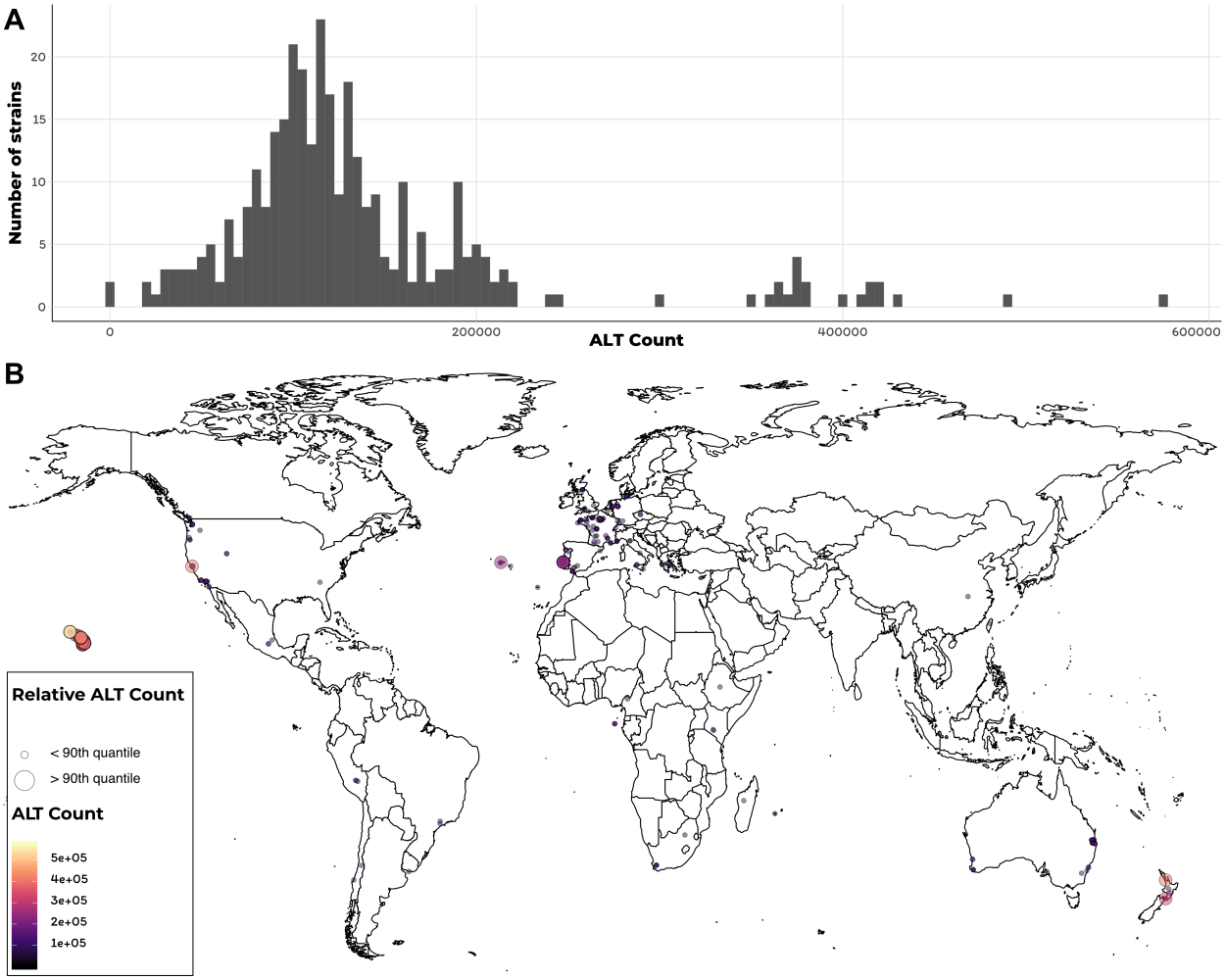

Figure 2-2 Alternate genotypes per strain and their global distribution
A) A histogram of the number of alternate alleles per strain is shown. The y-axis represents the number of strains for a given alternate allele count on the x-axis. B) The sampling location for the 330 *C. elegans* strains discussed in this chapter is shown. Each strain is represented by a dot that is colored by the number of alternate alleles that the strain contains. Larger dots correspond to the top 10% of strains based on number of alternate alleles in their genome.

Distribution of structural variants

In addition to SNV and small indel variants, we also called structural variants (SVs), including duplication, inversion, large insertion, and large deletion variants. We applied three SV calling algorithms to the *C. elegans* population and used the SURVIVOR software package to find overlapping SVs (see Materials and Methods). Across the *C. elegans* population, we identified a total of 20,217 structural variants that were identified by at least two SV calling algorithms. We identified 6,242 deletion, 6,311 insertion, 4,391 complex, 2,346 duplication, and 927 inversion variants. The genomic distribution of these variants was similar to the distribution of SNV and indel variants, with a larger proportion located on chromosome arms than chromosome centers (Figure 2-3A). For the most part, we found that the fraction of total SVs per chromosome was correlated with chromosome size, with chromosome V harboring the most SVs (average of 27% of all SV classes). We found no significant chromosomal enrichment for any SV class (Figure 2-3A**)**. Genome wide, the median size of each SV class was 768 bp for deletions, 1,295 bp for complex variants, 1,924 for inversions, and 3,170 for duplications (Figure 2-3B**)**. The average frequency of SVs in the population was 3% and the median was 0.3%, which held true across all variant classes (Figure 2-3C**)**.

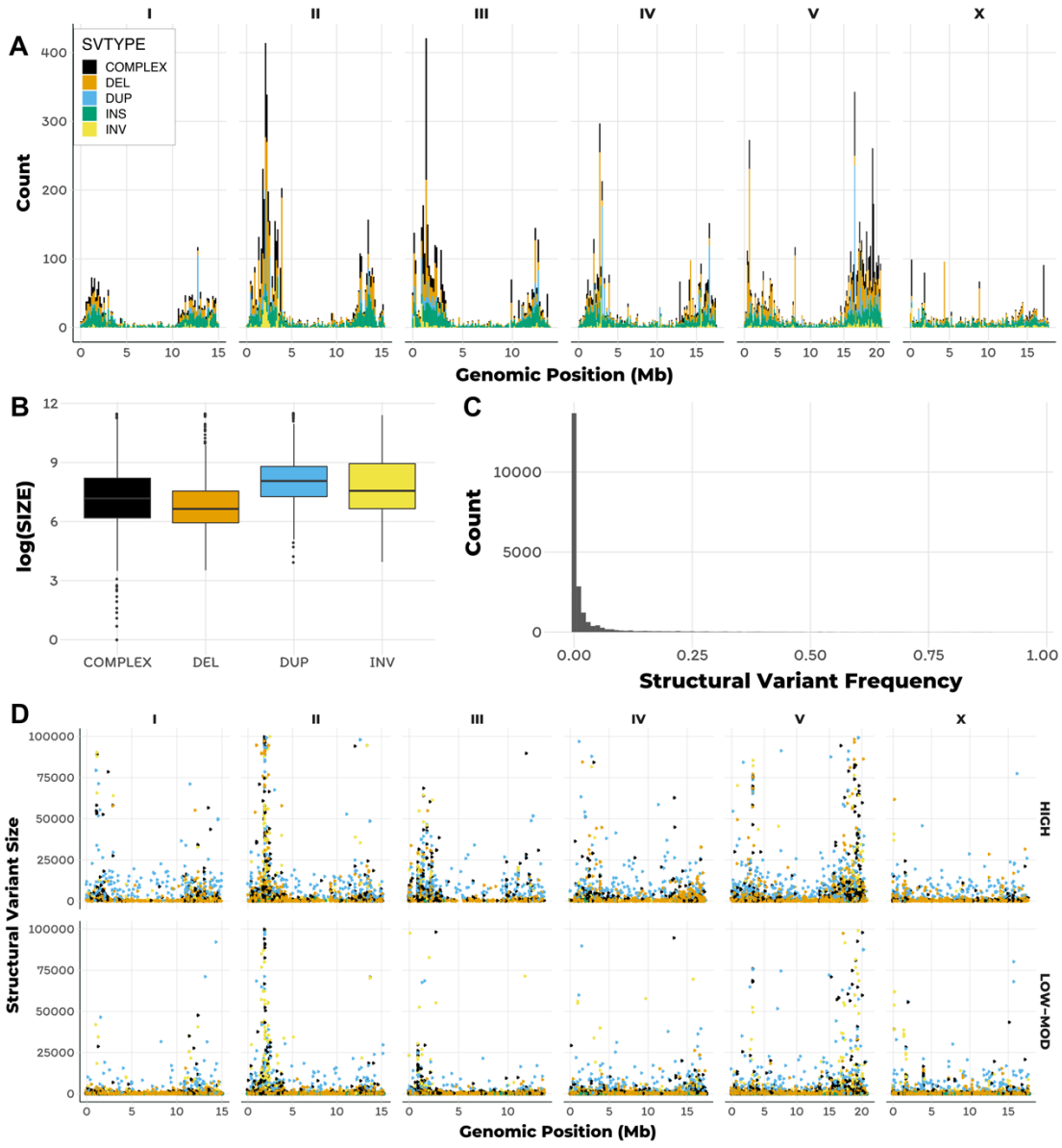

Figure 2-3 The genomic distribution, size, and frequency of structural variants
A) A histogram of structural variant (SV) counts across the six *C. elegans* chromosomes is shown. The colors of the bars correspond to the SV class (black = complex, orange = deletion, blue = duplication, green = insertion, yellow = inversion). The y-axis represents the variant counts across the genome on the x-axis, where each tick mark corresponds to 5 Mb. Variant counts were normalized to put them on the same scale. B) Tukey boxplots that represent the size of each detected structural variant is shown. The y-axis represents the log-scaled variant size for each SV class on the x-axis. C) A histogram of all detected SVs is shown. The y-axis represents the number of SVs and the x-axis represents the SV frequency. D) The genomic distribution of SVs across the six *C. elegans* chromosomes are shown. The y-axis represents the SV and the x-axis represents the genomic position, where each tick mark corresponds to 5 Mb. The top panel shows SVs that are predicted to have high effects on a gene's function and the bottom panel shows SVs that are not predicted to affect a gene's function.

We next looked at the distribution of SVs among genomic features. We found that 8,662 of SVs were located in non-intronic regions of genes, 7,411 were intergenic, and 4,144 were located in introns. Surprisingly, 6,115 of the SVs located in non-intronic regions of genes had high predicted effects on gene function, as annotated by SnpEff [81]. Of the SVs that had high predicted effects on gene function, 2,298 were deletions, 1,945 were complex variants, 1,123 were duplications, 420 were insertions, and 329 were inversions, which in total affected 2,180 unique genes (Figure 2-3D). Taken together, these results suggest that SVs likely have profound impacts on gene function in the *C. elegans* population.

Divergent regions are prevalent in wild *C. elegans* isolates

Our analysis of genetic variation within the *C. elegans* species revealed that certain genomic regions had elevated levels of variation, relative to neighboring regions (Figure 2-4A), we refer to these outlier regions as divergent regions. Inspection of the BAM alignment files on the CeNDR website confirmed that there were an unusually high number of high-quality variants in these regions, relative to surrounding regions. We therefore sought to systematically identify these regions for each *C. elegans* strain in our collection. To do this, we made use of small variant counts, structural variant counts, depth of coverage for each 1 kb window across the genome (see Materials and Methods). In total, the regions we identified as divergent contained 1,880,829 SNVs and 258,113 indel variants of 2,759,941 SNVs and 389,741 indel variants from the VCF we used as input for identification of divergent regions. The discrepancy between the variant counts described here and in the above section is because we included variant sites with missing genotype data to identify divergent regions. Using our approach, 3.983 Mb of the genome was classified as a divergent region. A vast majority of the high-frequency divergent regions are located on the arms of chromosomes, except for chromosome X where high-frequency divergent regions are evenly distributed (Figure 2-4B).

The localized distribution of 68.1% of the genetic variation to 4% of the genome suggested that a few hyper-divergent strains might be driving these results. However, this does not appear to be the case, because only 17% of the divergent regions were unique to a single strain, and 53% of the divergent regions were shared by less than five percent of the population (Figure 2-4C). Furthermore, we find that divergent regions that were identified by different masks (see Materials and Methods) have variable frequencies across the *C. elegans* population (Figure 2-4D). We find that the local outlier mask (Masked outlier, see Materials and methods) tend to have the highest frequencies in the population and the mask defined solely on variant count (Masked outlier, see Materials and methods) tend to be rare in the population. This observation is likely driven by a few hyper-divergent strains. These results demonstrate that a large fraction of the divergent regions are common in *C. elegans* population.

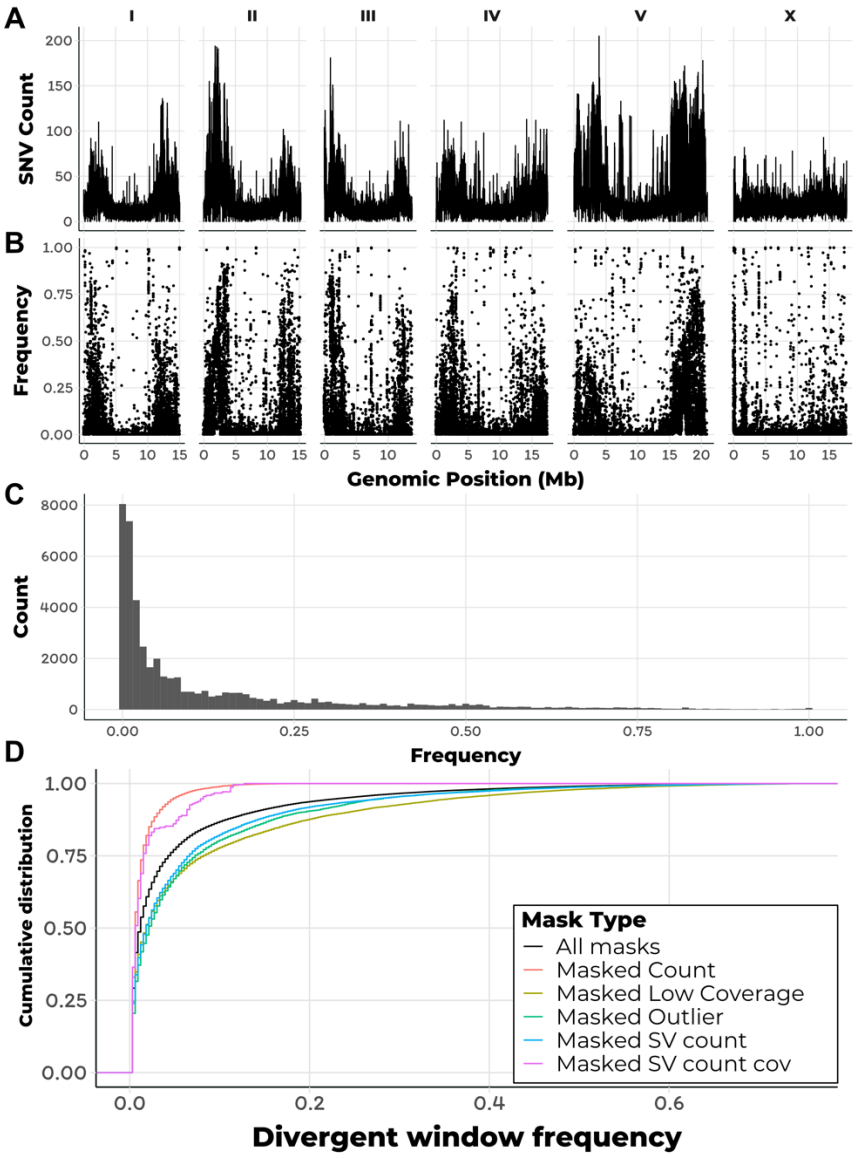

Figure 2-4 Variant counts per genomic window and the distribution of genetically divergent regions

A) A line plot of SNV counts per 1 kb window across the six *C. elegans* chromosomes is shown. The y-axis represents the variant counts across the genome on the x-axis, where each tick mark corresponds to 5 Mb. B) The genomic distribution of 1 kb windows that we identified to be genetically divergent. Each dot represents a single 1 kb window across the genome on the x-axis, where each tick mark corresponds to 5 Mb. The y-axis represents the frequency that window was identified as divergent in the *C. elegans* population. C) A histogram of the divergent window frequencies in the *C. elegans* population is shown. The y-axis represents the number of divergent windows per bin and the x-axis represents the frequency of the window in the population. D) The cumulative distribution of divergent windows separated by the mask that identified the window as divergent is shown (black = all divergent windows, salmon = count mask, olive = low coverage mask, green = local outlier mask, blue = number of SVs mask, pink = number of SVs and abnormal coverage mask). The y-axis represents the cumulative distribution of divergent window frequencies, which is shown on the x-axis.

Divergent regions do not affect population structure

Given that the divergent regions are so prevalent and common in the *C. elegans* population, we next asked if their presence has an effect on the population structure of the species. To this end, we explored the population structure within the species using two independent approaches, admixture and principal component analysis (PCA). We performed this analysis independently using the Intersection and Masked VCFs (see Materials and Methods).

To perform PCA, we ran *smartpca* from the Eigenstrat software package in two run modes; with outlier strain removal and with no strain removal [96]. In outlier strain removal mode, the first 14 principal components (PCs) generated from the Intersection and Masked VCFs had greater than 70% correlation, suggesting that the genetic relatedness among individual strains is largely unaffected by the presence of the divergent regions (Figure 2-5A). However, when we ran *smartpca* in outlier removal mode, where the 15 most genetically divergent strains are removed, we observed reduced correlation among the PCs generated from the Intersection and Masked VCFs (Figure 2-5B). These results suggest that the 15 most genetically divergent individuals in the *C. elegans* population contribute the most to the population structure in the species.

Because we masked divergent regions at the strain level, the Masked VCF contained a lot of missing genotype data. To determine the effect of this high level of missing data, we performed the same smartpca analysis, but only included sites with no missing genotype. The results from the smartpca procedure with no outlier removal was similar to that when we included sites with missing data (Figure 2-5C). In contrast to the smartpca analysis without outlier removal when missing genotype data was included, we observed higher correlation among the first 10 PCs

generated from the Masked and Intersection VCFs when sites with missing data were not considered (Figure 2-5D). These results show that high levels of missing data affect accurate classification of population structure and suggest that the divergent regions do not have a large impact on the overall population structure within the species.

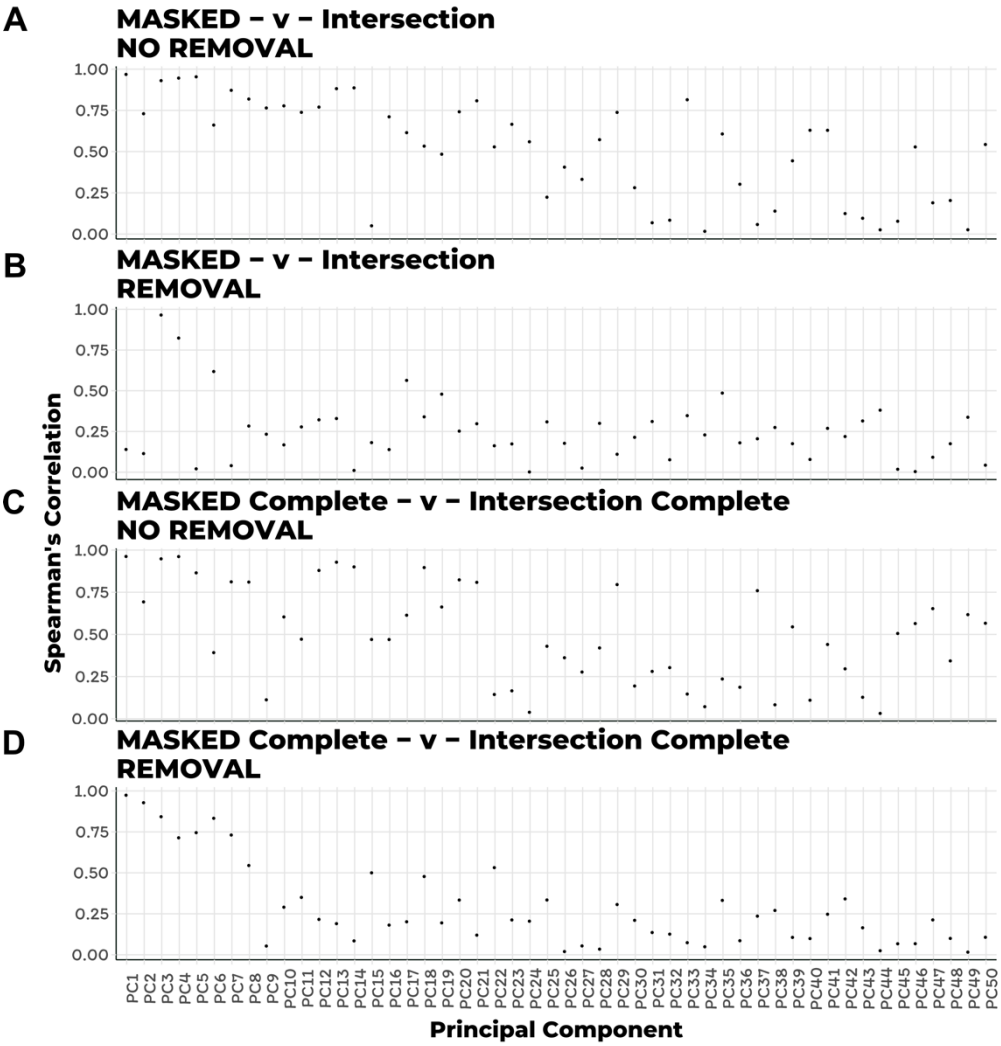

Figure 2-5 PCA comparison of population structure when divergent regions are masked
A-D) The Spearman's rank correlation coefficient comparing principal components calculated from the divergent region masked and GATK-Strelka2 intersection VCF is shown on the y-axis for the first 50 principal components (PCs) on the x-axis. The PCs in A) and C) were generated using the *smartpca* command without removing outlier strains. The PCs in B) and D) were generated using the *smartpca* command with removing the 15 most genetically divergent strains. The PCs in A) and B) were generated from VCF files that required sites to have less than 20% missing genotype information across all strains. The PCs in C) and D) were generated from VCF files that required sites 0% missing genotype information across all strains.

To independently confirm the results from the smartpca analysis, we performed admixture analysis on both the Masked and Intersection VCFs, and only considered variant sites with no missing genotype data. We used Admixture to perform this analysis on a subset of defined ancestral populations (K = 5 - 8) and compared the ancestral population assignments generated by these analyses using canonical correlation analysis [93]. The results from this approach to define population structure in the *C. elegans* species was in agreement with our observations from the PCA approach. Specifically, individual strains were assigned to the same ancestral population and varied slightly in their defined admixed proportions (Figure 2-6A-C). Furthermore, strain population assignments were more similar for divergent strains than strains with less genetic variation. Taken together, these results show that despite the high proportion of variants within the divergent genomic regions, the species-wide population structure is not affected by these regions.

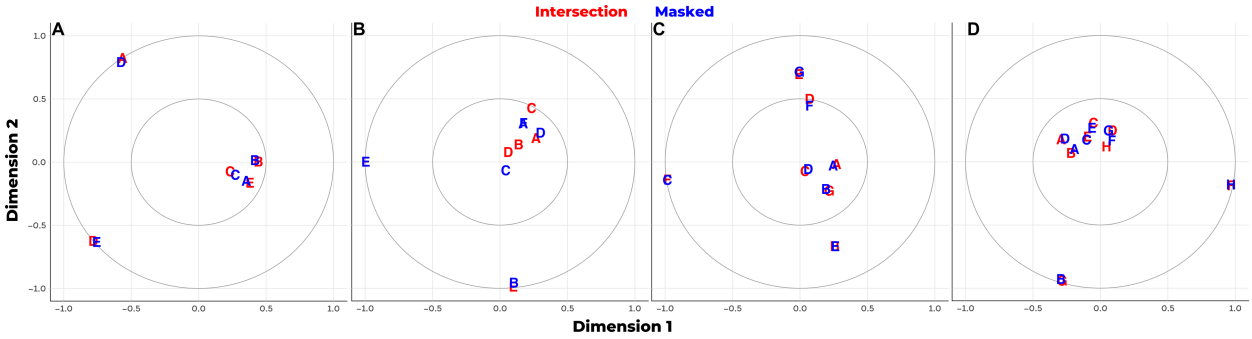

Figure 2-6 Admixture comparison of population structure when divergent regions are masked
The results of canonical correlation analysis performed on the strain population assignments generated by the ADMIXTURE analysis is shown for A) K=6, B) K=6, C) K=6, D) K=8. The x-axis represents the dimension that explains the most variance in the dataset and the y-axis represents the dimension that explains the second most variation in the data set. Each ancestral population is labeled with a colored letter, where red letters represents assignments calculated using the GATK-Strelka2 intersection VCF and blue letters represent assignments calculated using the divergent masked VCF. ADMIXTURE analysis was performed using variant sites with no missing genotype data across all strains. The closer two red and blue letters are on the plot means that the strains from both analysis procedures were assigned to the same ancestral population.

Divergent regions are enriched with environmental sensing genes

Given the unusually high levels of variation within the divergent regions we identified, we looked for the presence of gene enrichment within these regions. To do this, we performed GO term and KEGG pathway enrichment analysis on the genes within divergent regions. We subdivided the divergent regions into three groups based on the frequency of the region in the population; regions less than 1% (rare), regions between 1-5% (intermediate), and regions greater than 5% (common). For rare divergent regions, we found slightly significant enrichment of the GTPase binding and protein kinase activity molecular function GO terms, and a seven slightly enrichment biological process GO terms (Figure 2-7A-B). KEGG pathway enrichment of genes in rare divergent regions revealed several gene enrichments, the most significant of which were FoxO signaling, Wnt signaling, and other components of metabolism (Figure 2-7C). As compared to the low frequency regions, the intermediate frequency divergent regions had distinct molecular function GO term enrichment, which included olfactory receptor activity and GPCR activity (Figure 2-8A). Similarly, the most enriched biological process GO terms for intermediate frequency divergent regions were distinct from low frequency regions and all related to environmental sensing processes (Figure 2-8B). KEGG gene set enrichment of intermediate frequency divergent regions revealed FoxO signaling, Wnt signaling, and other components of metabolism enrichment, which were all enriched in low frequency regions (Figure 2-8C). Molecular function enrichment of common divergent was drastically different from that of low or intermediate frequency divergent regions, and included helicase activity, carbohydrate binding, GPCR activity, among other significantly enriched terms (Figure 2-9A). In contrast, common divergent regions enrichment overlapped with intermediate regions for biological processes involved in environmental sensing, but also contained several uniquely

enriched terms. Terms unique to common regions included response to xenobiotics, extracellular matrix biosynthetic, and telomere maintenance term enrichment. Similar pathways to those identified in low frequency divergent regions were enriched in common regions, including the FoxO, Wnt, mTOR, MAPk, Purine metabolism, and longevity pathways. However, genes involved in Ubiquitin mediated proteolysis, TGF-beta signaling, Protein processing in the endoplasmic reticulum, Autophagy, and the Splicesome were only enriched in common divergent regions. All of these pathways have previously been implicated in mediating responses to pathogenic and non-pathogenic bacteria in *C. elegans [98–105]*. Elevated levels of genetic diversity occurs within and among genes involved in pathogen response and environmental sensing [106–108]. Therefore, we asked if the common divergent regions we identified at the strain level are show signatures of balancing selection within the *C. elegans* population.

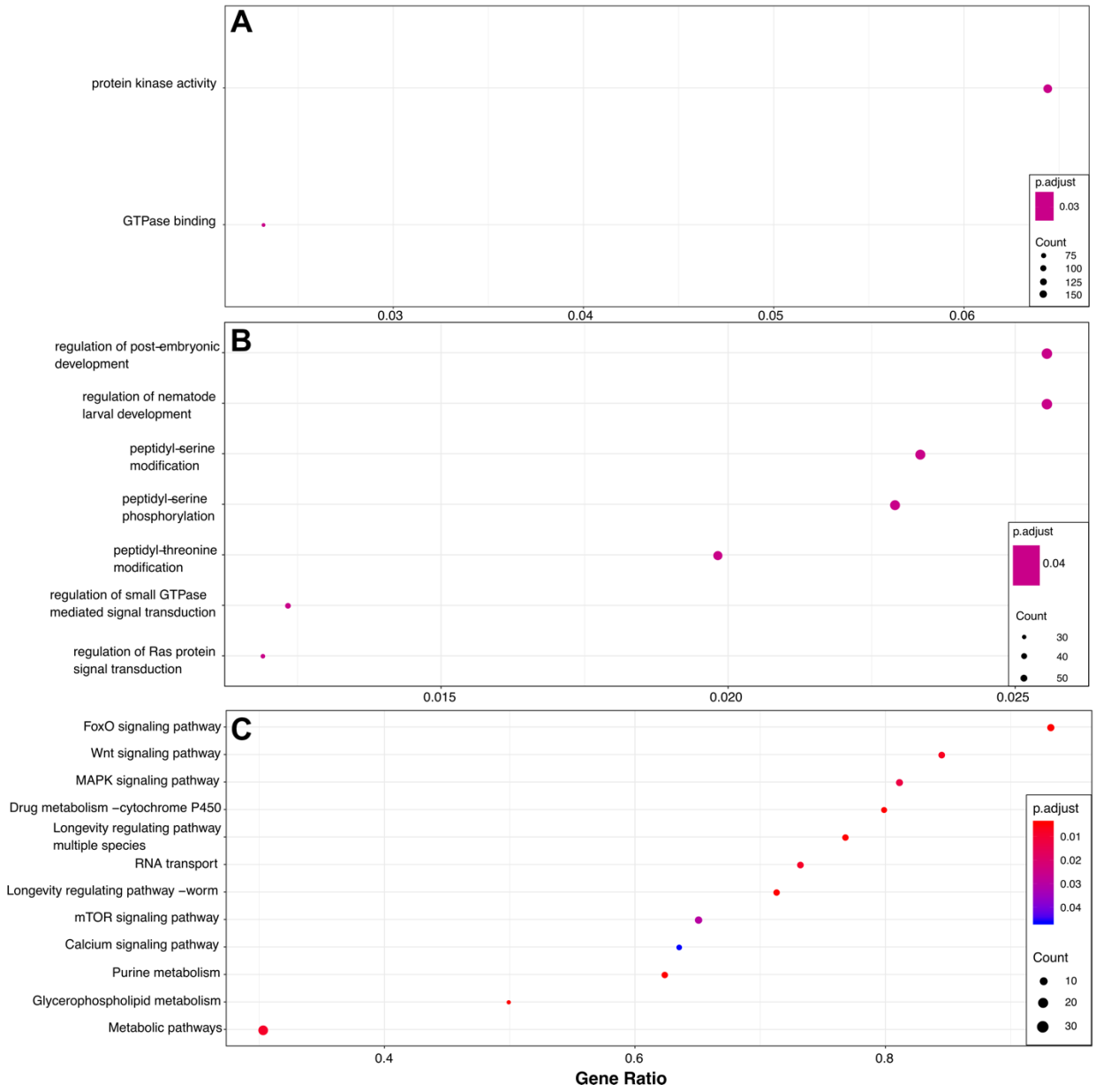

Figure 2-7 Low frequency divergent region enrichment analysis
A) The significantly enriched molecular function and B) biological process GO-terms are shown on the y-axis for divergent regions that are in less than 1% of the *C. elegans* population. C) The significantly enriched KEGG gene sets are shown on the y-axis. For all panels, dots are colored by the significance of enrichment and sized by the number of genes corresponding to the GO term or KEGG gene set. The x-axis represents the k/n ratio, where k is the size of the overlap of input genes with the specific GO term or gene set and n is the size of the overlap of the overlap of input genes with all the members of the collection of GO terms or gene.

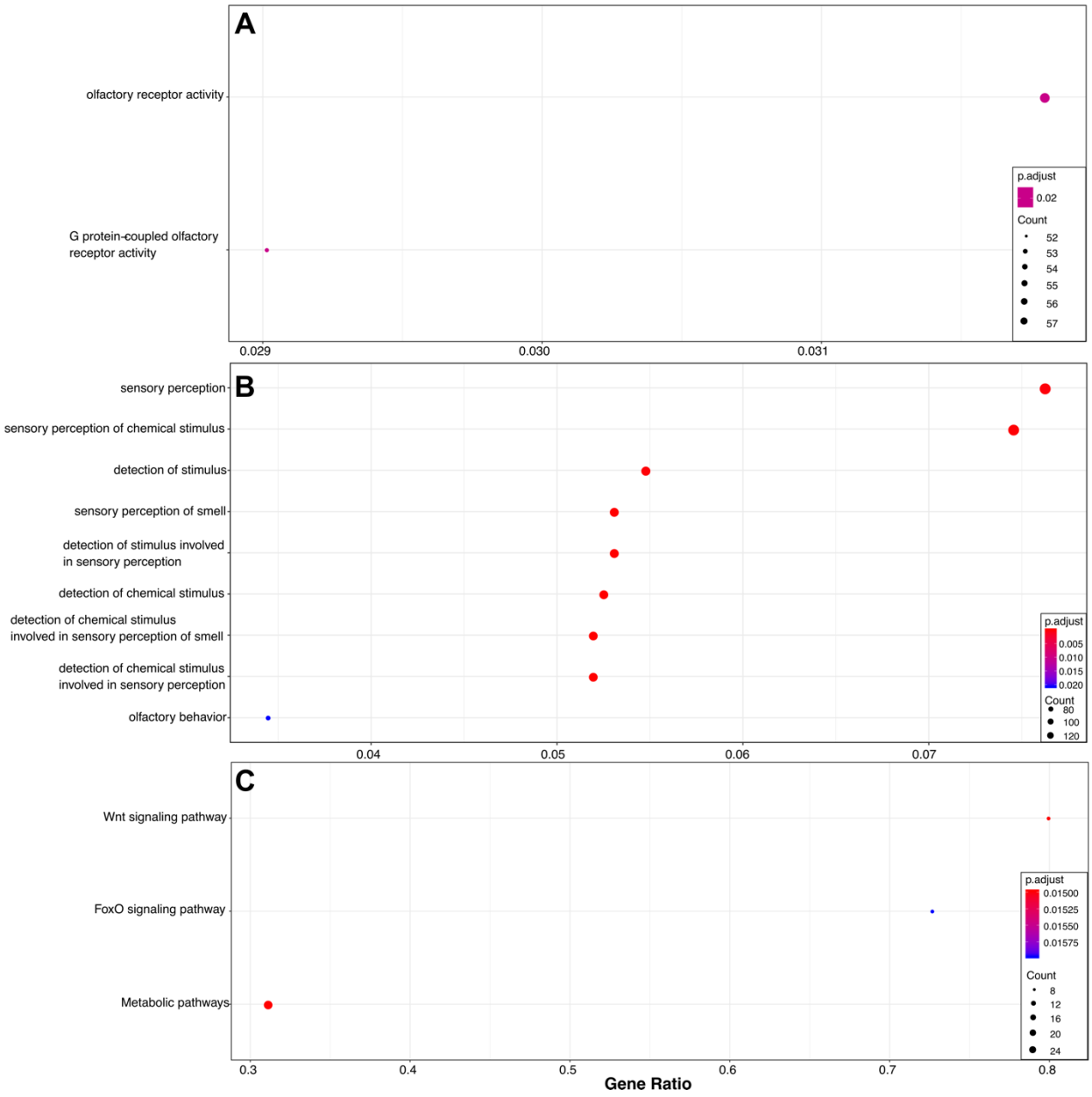

Figure 2-8 Intermediate frequency divergent region enrichment analysis
A) The significantly enriched molecular function and B) biological process GO-terms are shown on the y-axis for divergent regions that are between 1-5% of the *C. elegans* population. C) The significantly enriched KEGG gene sets are shown on the y-axis. For all panels, dots are colored by the significance of enrichment and sized by the number of genes corresponding to the GO term or KEGG gene set. The x-axis represents the k/n ratio, where k is the size of the overlap of input genes with the specific GO term or gene set and n is the size of the overlap of the overlap of input genes with all the members of the collection of GO terms or gene.

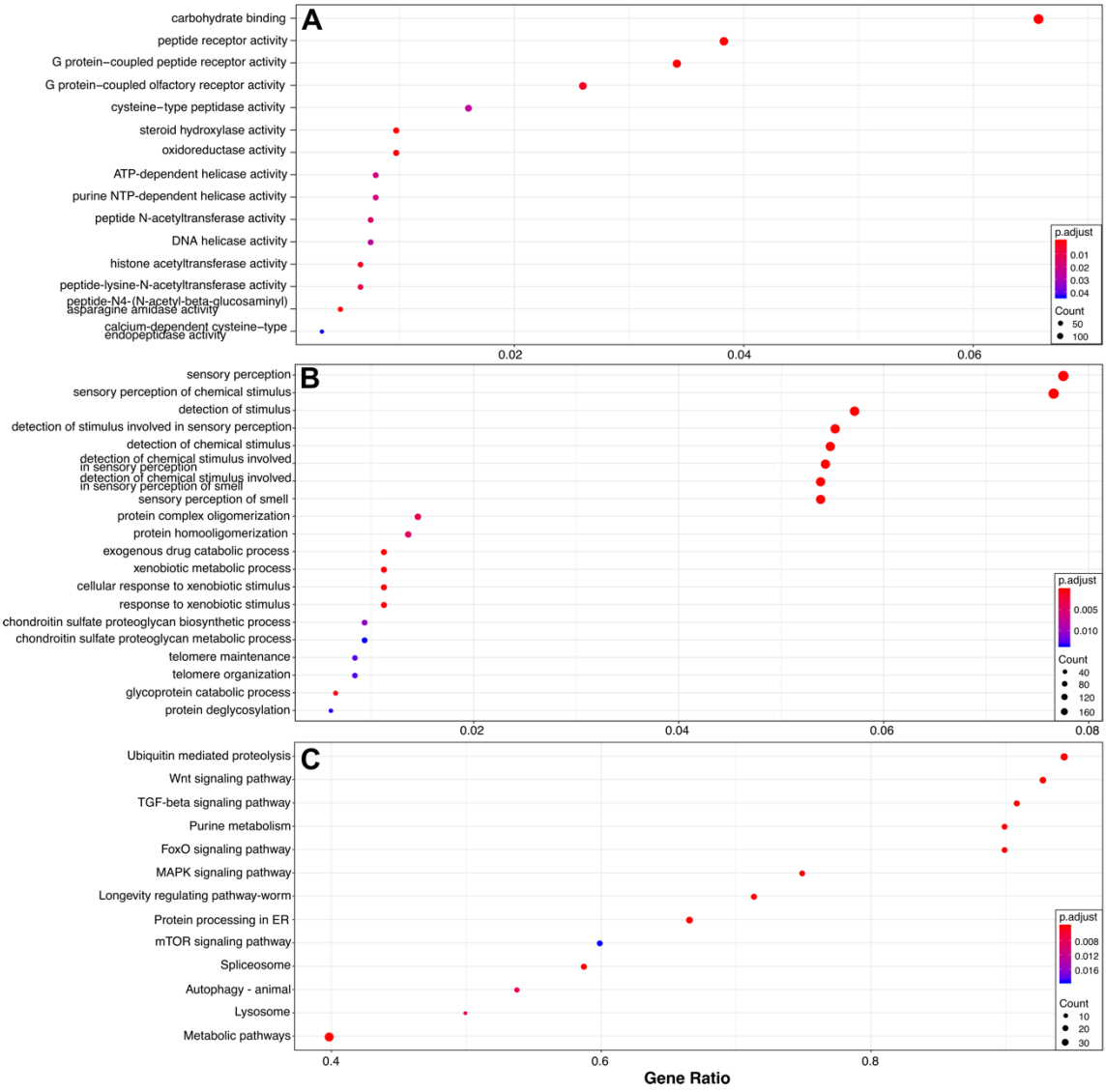

Figure 2-9 Common divergent region enrichment analysis
A) The significantly enriched molecular function and B) biological process GO-terms are shown on the y-axis for divergent regions that are in more than 5% of the *C. elegans* population. C) The significantly enriched KEGG gene sets are shown on the y-axis. For all panels, dots are colored by the significance of enrichment and sized by the number of genes corresponding to the GO term or KEGG gene set. The x-axis represents the k/n ratio, where k is the size of the overlap of input genes with the specific GO term or gene set and n is the size of the overlap of the overlap of input genes with all the members of the collection of GO terms or gene.

To determine if regions under balancing selection were enriched in the same GO terms we identified within the common divergent regions, we calculated Tajima's *D* across the genome for the *C. elegans* population [109]. We independently calculated Tajima's *D* across the genome using sliding windows and for every gene. We observed similar genome-wide patterns of

Tajima's *D* as were previously reported with a smaller strain set and fewer genomic markers (Figure 2-10A) [43]. Interestingly, we did not observe similar GO term enrichment for regions under balancing selection, as measured with Tajima's *D* statistic (*D* > 2) for the sliding window analysis and for the individual gene analysis (Figure 2-10B-C). In fact, we observed no KEGG pathway enrichment for regions under balancing selection. Taken together, these results suggest that the identification of divergent regions at the strain level, as opposed to the population level, can give strikingly different results for enrichment analysis.

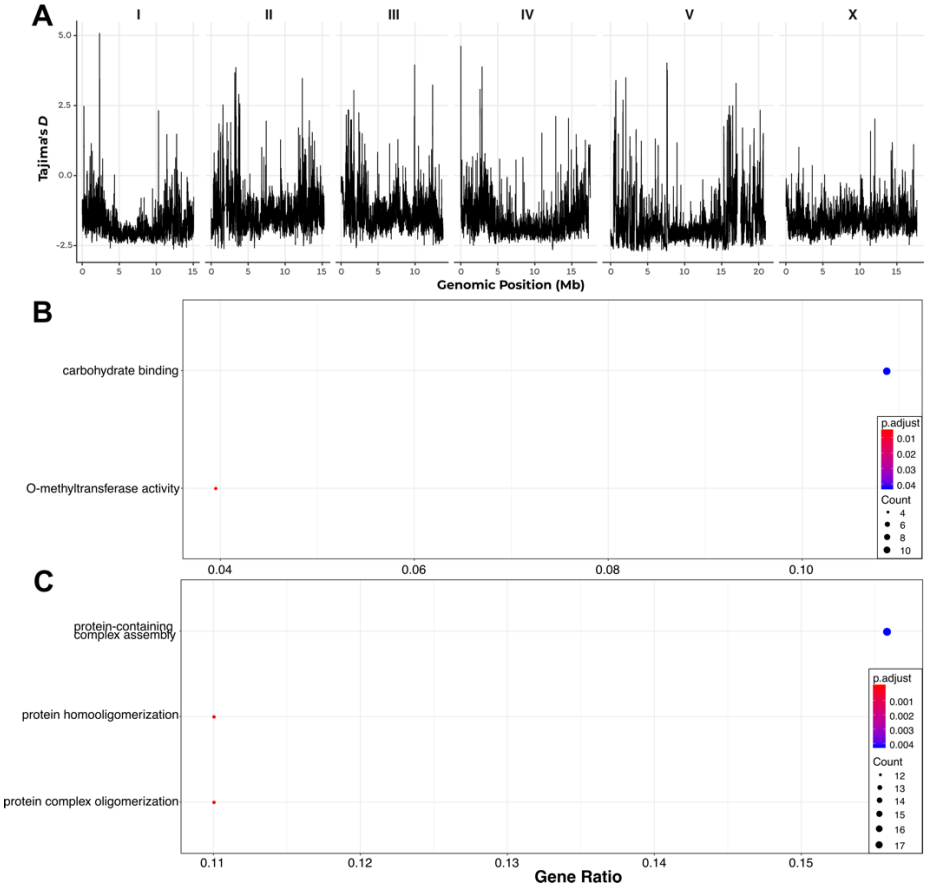

Figure 2-10 Genome-wide Tajima's D and enrichment analysis of regions under balancing selection

A) Divergence, as measured by Tajima's *D*, is shown on the y-axis across the six *C. elegans* chromosomes on the x-axis. Tajima's D was calculated for 10 kb windows across the genome, with 9 kb of overlap between windows. B) The significantly enriched molecular function and C) biological process GO-terms are shown on the y-axis for windows with a Tajima's *D* value greater than 2. The x-axis represents the k/n ratio, where k is the size of the overlap of input genes with the specific GO term or gene set and n is the size of the overlap of the overlap of input genes with all the members of the collection of GO terms or gene.

Discussion

This work represents the most comprehensive analysis of genetic diversity in the *C. elegans* species to date. We analyzed all classes of variation except translocations across 330 genetically distinct strains that have been collected around the world. We found that 9,486 genes have predicted high effects on gene function, though we note that we did not use a haplotype-aware prediction algorithm, which will likely reduce this number substantially. Nevertheless, the presence of high-impact variation in such a large fraction of the annotated genes in the *C. elegans* genome indicates that we still have much to learn about the genetic content of this species. It is well established that *C. elegans* has a large expansion of chemoreceptors relative to other related nematode species [106,110], and that many of these might be functionally redundant or provide a fitness advantage in specific niches.

An interesting observation of our scan of genomic variation within the species was the presence of highly divergent regions within the genome. We developed a novel approach to characterize divergent regions within individual strains based on the number small variants, structural variants, and alignment depth. These regions span 3% of the *C. elegans* genome, yet contain nearly 70% of the small genetic variants within the species. Given the high density of polymorphisms in these regions, we were surprised to find that they had little effect on the overall population structure within the species. This observation suggests that the variation within common divergent regions is consistent across individuals and that rare divergent regions show similar patterns of polymorphism as the rest of the genome of the individuals that harbor them.

Closer inspection of the genetically divergent genomic regions revealed that approximately 30% of these regions are present in at least 5% of the individuals within the species. This observation

suggests that these divergent regions are being maintained in the *C. elegans* populations. However, a classical test for maintenance of intermediate to common frequency alleles found no enrichment of genes typically thought to be under balancing selection, including immunity genes [108,111,112], genes that provide a fitness advantage in specific environmental niches [113,114], genes required for self-incompatibility [115,116], and others [117,118]. However, our analysis of genes within divergent regions identified by analyzing individual strains revealed strikingly different enrichment patterns than the population-level analysis. In accordance with what is known about genes expected to contain high levels of variation within a species, we find significant enrichment genes associated with environmental sensing, response to xenobiotics, metabolism, and response to pathogens. The genes involved in these biological processes were only enriched in divergent regions that have intermediate to high frequencies in the population. In contrast, KEGG gene-set enrichment revealed similarities between low frequency and common divergent regions, with a few interesting differences. A key difference is the enrichment of metabolic genes in common divergent regions, which suggests that individuals in the species have different metabolic requirements, a result consistent with my own and others' work [32,119]. A possible explanation of this result is that individual strains encounter diverse bacterial species with variable nutritional content in the environmental niches they inhabit, which is supported by recent studies of the *C. elegans* microbiome [120–123].

The second key difference between the KEGG gene-set enrichment of low-frequency and common divergent regions is the presence of enrichment of ubiquitin-mediated proteolysis genes in common divergent regions. Previous work has shown that ubiquitin ligases are under positive selection among nematodes and plants [107]. Subsequent analysis of the largely uncharacterized *pals* gene family has shown that this family of genes regulates the expression ubiquitin ligases and is involved in the intracellular parasite response, silencing of repetitive elements, and thermotolerance [99,100,124,125]. Taken together, our analysis of divergent

genomic regions within *C. elegans* complement previous work across nematode species that show ubiquitin ligases are under positive selection and suggest that selection at these loci contribute to the diversity we observe within the *C. elegans* population.

Future directions

As stated in the preface to this chapter, the work I discussed above is still its infancy. However, this work does provide a starting point to several interesting research directions.

One open question is whether or not the strains that contain divergent regions we identified contain functional genes. One approach to address this question is to perform RNAseq analysis and analyze the expression level and sequences of transcripts in these regions. The limitation of RNAseq analysis is that these genes might not be expressed in standard laboratory conditions. Unfortunately, this limitation is difficult to overcome because of the magnitude of largely unknown environmental factors that would have to be tested to induce the expression of genes not expressed in laboratory settings. A second limitation to the RNAseq approach is the level of diversity within these genes might make it difficult to detect their expression. To circumvent this issue, whole-genome assembly of divergent *C. elegans* strains will be required to properly annotate the genetic content within these regions, which the Andersen lab is currently working toward.

A second open question that I did not explore is whether the divergent regions we identified are enriched in specific environmental niches. My preliminary analysis into this question, which I did not present in this chapter, suggests that combinations of divergent regions on the left arm of chromosome II and the right arm of chromosome V are correlated with the substrate from which the strains were sampled. These results were obtained by performing PCA on a genotype

matrix that consisted of the identified divergent regions in the species. I found several principal components that were significantly correlated with isolation substrate and landscape and inspected the loadings of these correlated components. Further work is needed to explore if this result is driven by highly divergent strains on the Hawaiian islands, which tend to be isolated from leaf litter. Additionally, I performed this preliminary analysis on all classes of divergent regions (see Materials and Methods for mask types). Repeating this analysis with the divergent regions separated by the type of mask that identified the region or the frequency of the divergent region in the population might reveal additional associations between divergent regions and environmental niches. Establishing these associations will provide insights into how specific environmental factors contribute to the maintenance of genetic diversity within the species.

To date, a suitable outgroup species has not been identified to enable comprehensive comparative genomic analysis for *C. elegans*. This is a major limitation when attempting to determine the genomic factors that enable individual species to occupy specific environmental niches. However, recent whole-genome assemblies of ten new *Caenorhabditis* species can provide the framework for performing comparative genomic analysis of to establish genomic factors contributing to niche specificity [126]. Seven species pairs might be suitable for comparative genomics analyses, including *C. brenneri – C doughertyi, C. latens – C. remanei, C. nigoni – C. briggsa, C. afra – C. 63anziba, C. nouraguensis – C. becei, C. waitukubuli – C. panamensis, C. sinica – C. 63anzibari – C. tribulationis.* Comparative genomic analyses among these groups will be greatly facilitated by the identification and genomic characterization of individuals within each species.

3. Natural variation in a single amino acid underlies cellular responses to topoisomerase II poisons

Preface

In the spring quarter of 2014, I chose to complete my final first-year rotation in the Andersen Lab. As part of my rotation project I analyzed a large-scale phenotyping effort of 96 natural *C. elegans* isolates and 250 recombinant inbred lines (RILS) by members of the lab. This analysis included performing genome-wide association (GWA) and linkage mapping on the phenotyped populations to identify candidate quantitative trait loci (QTL) to further study during my Ph.D.. The following chapter describes my work chasing down one of the candidate QTL I identified during my rotation project, that explains phenotypic variation in response to the chemotherapeutic etoposide. This project could not have been completed without the help of excellent collaborators at the Broad Institute and at the University of Wisconsin Madison. Our findings were published in *PloS Genetics* in 2017 [31].

Abstract

Many chemotherapeutic drugs are differentially effective from one patient to the next. Understanding the causes of this variability is a critical step towards the development of personalized treatments and improvements to existing medications. Here, we investigate sensitivity to a group of anti-neoplastic drugs that target topoisomerase II using the model organism *Caenorhabditis elegans*. We show that wild strains of *C. elegans* vary in their sensitivity to these drugs, and we use an unbiased genetic approach to demonstrate that this natural variation is explained by a methionine-to-glutamine substitution in topoisomerase II (TOP-2). The presence of a non-polar methionine at this residue increases hydrophobic interactions between TOP-2 and its poison etoposide, as compared to a polar glutamine. We

hypothesize that this stabilizing interaction results in increased genomic instability in strains that contain a methionine residue. The residue affected by this substitution is conserved from yeast to humans and is one of the few differences between the two human topoisomerase II isoforms (methionine in hTOPIIα and glutamine in hTOPIIβ). We go on to show that this amino acid difference between the two human topoisomerase isoforms influences cytotoxicity of topoisomerase II poisons in human cell lines. These results explain why hTOPIIα and hTOPIIβ are differentially affected by various poisons and demonstrate the utility of *C. elegans* in understanding the genetics of drug responses.

Introduction

Antineoplastic regimens used to treat cancer are often associated with poor prognoses and severe side effects. Ideally, antineoplastic regimens can be tailored to an individual patient based on various genetic markers known to be associated with drug response to maximize therapeutic effectiveness and minimize unwanted side effects. Advances in sequencing technologies over the course of the past decade promised the discovery of many genetic variants that contribute to human health. Though large-scale sequencing projects have lead to the identification of many genetic variants associated with disease risk [127], relatively few variants have been identified that contribute to clinically relevant traits such as response to antineoplastic compounds. In fact, only 71 of over 500 FDA-approved antineoplastic compounds use genetic information to affect treatment efficacy (www.fda.gov). Unfortunately, the predictive power of these identified genetic variants can be inconsistent due to biases in the sampled population [128] and other key limitations of clinical genome-wide association (GWA) studies that attempt to link genetic variants with treatment outcomes. The major factor limiting the efficacy these studies is sample size because it is difficult to identify large numbers of individuals exposed to the same antineoplastic regimens. This limitation is compounded when

considering environmental [11,12] and tumor heterogeneity [9]. As a result, most variants discovered to be associated with outcomes in clinical GWA studies offer low predictive power for patient responses to treatment [129]. These limitations and others emphasize the need for novel approaches to identify variants that predict patient outcomes to antineoplastic compounds.

Studies of model organisms have greatly facilitated our understanding of basic cellular processes. In recent years, *Saccharomyces cerevisiae* and *Drosophila melanogaster* have been used to understand the physiological effects of small molecules and repurposed as screening platforms to identify new antineoplastic compounds [130–132]. The ability to generate extremely large numbers of recombinant yeast facilitates the identification of genomic regions that are predictive of drug response [133,134]. Furthermore, the specific genes and variants within regions can be identified and functionally validated in yeast [135–137]. By contrast, *D. melanogaster* studies offer the ability to study the physiological responses to drugs in the context of multiple tissue types, but functional validation of specific genes and variants associated with drug responses has been more limited [132]. The roundworm *Caenorhabditis elegans* has the advantages of both *S. cerevisiae* and *D. melanogaster* because large cross populations can be generated to study the physiological responses to drugs in a metazoan. These attributes have made *C. elegans* an important model for connecting differential drug responses with genetic variants present in the species [54,138].

Here, we take advantage of natural genetic variation present in *C. elegans* to identify the genetic basis underlying susceptibility to a panel of clinically relevant antineoplastic compounds that poison the activity of topoisomerase II enzymes. The inhibition of these enzymes by topoisomerase II poisons results in the accumulation of double-stranded breaks and genome instability [139–141]. Topoisomerase II enzymes are targeted by antineoplastic regimens

because proliferative cell populations require their enzymatic activity to relieve topological stress ahead of the replication fork [142]. Using two unbiased genetic mapping approaches, we show that divergent physiological responses to the topoisomerase II poison etoposide are determined by natural genetic variation in a *C. elegans* topoisomerase II enzyme. Furthermore, we show using CRISPR/Cas9-mediated genome editing that variation in a specific amino acid (Q797M) underlies the cytotoxic effects of etoposide. This residue is conserved in humans and is one of the few differences between the putative drug-binding pockets of the two topoisomerase II isoforms (M762 in hTOPIIα and Q778 in hTOPIIβ). Previous structural studies on hTOPIIβ implicated this glutamine residue in etoposide binding because of its proximity to the drug-binding pocket [143,144]. However, a study on hTOPIIα suggested that the corresponding methionine residue has no functional role in drug binding [145]. We present a mechanistic model to explain how variation at this residue underlies differential responses to etoposide and other topoisomerase II poisons. Finally, we use genome-edited human cell lines to show that this residue in hTOPIIα contributes to differential toxicity of various topoisomerase II poisons. These results demonstrate the power of using *C. elegans* natural genetic variation to identify mechanisms of drug susceptibility in human cells that could inform human-health decisions based on genetic information.

Materials and Methods

Strains

Animals were cultured at 20ºC with the bacterial strain OP50 on modified nematode growth medium (NGM), containing 1% agar and 0.7% agarose to prevent burrowing of the wild isolates. For each assay, strains were grown at least five generations with no strain entering starvation or encountering dauer-inducing conditions [146]. Wild *C. elegans* isolates used for genome-wide

association are described previously [27,44]. Recombinant inbred advanced intercross lines (RIAILs) used for linkage mapping were constructed previously *[54]*. Strains constructed for this manuscript are listed in Supplementary Information. Construction of individual strains is detailed in the corresponding sections below and listed in Table 2-1. Oligonucleotides used to construct strains are listed in Table 2-2.

Table 3-1 Strains used for experiments discussed in Chapter 3

Name	Allele Name	Genotype
ECA215	*eanIR135*	N2 (II:11.64 - 11.91 Mb)
ECA216	*eanIR136*	N2 (II:11.43 - 12.11 Mb)
ECA219	*eanIR139*	CB4856 (II:11.64 - 11.9 Mb)
ECA220	*eanIR140*	N2 (II:12.01 - 12.1 Mb)
ECA338	*K12D12.1(ok1930)*	(*K12D12.1(ok1930)/mIn1 [mIs14 dpy-10(e128)]* II)
VC1474	*K12D12.1(ok1930)*	(*K12D12.1(ok1930)/mIn1 [mIs14 dpy-10(e128)]* II)
EG7952	NA	[*oxTi207 [eft-3p::GFP::unc-54 3'UTR + hsp::peel-1 + NeoR + Cbr-unc-119(+)]]*
ECA401	top-2(*ean2*[Q797M]	N2 TOP2(Q797M)
ECA402	top-2(*ean3*[Q797M]	N2 TOP2(Q797M)
ECA547	top-2(*ean4*[M797Q])	CB4856 TOP2(M797Q)
ECA548	top-2(*ean5*[M797Q])	CB4856 TOP2(M797Q)
ECA549	top-2(*ean6*[M797Q])	CB4856 TOP2(M797Q)

Table 3-2 Oligonucleotides used for experiments discussed in Chapter 3

Name	Sequence	Description
oECA593	CCGGTGTTTCAGGGCAATTT	Generate NIL for etoposide interval 11470951-12115727 using RIAIL QX322
oECA596	GCTACGGAATGTGCTGCTAC	Generate NIL for etoposide interval 11470951-12115727 using RIAIL QX322
oECA601	GAAGTTTCGGGTCAATGTATCCA	Generate NIL for etoposide interval 11470951-12115727 using RIAIL QX327
oECA604	GCTACGGAATGTGCTGCTAC	Generate NIL for etoposide interval 11470951-12115727 using RIAIL QX327

oECA605	ATGCAACGTTTGACTGGCAT	Generate NIL for etoposide interval 11470951-12115727 using RIAIL QX103
oECA608	CCATTGAATTAGTTGGCGGC	Generate NIL for etoposide interval 11470951-12115727 using RIAIL QX103
oECA1087	GACGAGTACCATTGGAATAAT CGGG	Verification CRISPR allele swap at top-2 Q778M variant
oECA1124	GGGAGAAGAAGGACCGAAAG C	Verification CRISPR allele swap at *top-2* Q778M variant
oECA1003	TCCAATCAAAGGATTCGAGG	5' primer used to verify presence of *K12D12.1(ok1930)*
oECA1004	ATGTCCTGGCCTTCCTTTTT	3' primer used to verify presence of *K12D12.1(ok1930)*
dpy-10 repair	CACTTGAACTTCAATACGGCAA GATGAGAATGACTGGAAACCG TACCGCATGCGGTGCCTATGG TAGCGGAGCTTCACATGGCTT CAGACCAACAGCCTAT	Repair template to generate dpy-10(cn64) mutation
CB4856 Repair	ACAGCGGAAGGTTCTCTTCGC GTGCTTCAAGAGAGCAGACAA GCGTGAAGTCAAAGTAGCTCA ATTGGCTGGAGCTGTCGCTGA AATTTCTGCTTATCATCACGGA GAA***CAG***TCGCTTATGGGAACA ATTGTGAATCTCG	Repair template to generate CB4856 M778Q allele swap. Green represents the gRNA sequence, red represents the PAM site, purple is the mutated nucleotide of the PAM site, and bold-underline-italic encodes for the desired edit.
N2 Repair	ACAGCGGAAGGTTCTCTTCGC GTGCTTCAAGAGAGCAGACAA GCGTGAAGTCAAAGTAGCTCA ATTGGCTGGAGCTGTCGCTGA AATTTCTGCTTATCATCACGGA GAA***ATG***TCGCTTATGGGAACA ATTGTGAATCTCG	Repair template to generate N2 Q778M allele swap. Green represents the gRNA sequence, red represents the PAM site, purple is the mutated nucleotide of the PAM site, and bold-underline-italic encodes for the desired edit.
***top-2* crRNA CB4856**	TAAGCGACATTTCTCCGTGA	Guide sequence for crRNA to target *top-2* at M778 in CB4856
***top-2* crRNA N2**	TAAGCGACTGTTCTCCGTGA	Guide sequence for crRNA to target *top-2* at Q778 in N2
***dpy-10* crRNA**	GCTACCATAGGCACCACGAG	Guide sequence for crRNA to target dpy-10
TOP2A crRNA	GTCATCATTAGTGACATCTG	targets sense strand of human TOP2A
TOP2B crRNA	GCTTGCTATACAACAGAAGA	targets sense strand of human TOP2B
TOP2A M>Q (sense strand)	AGTAAAAGCCTCAGCTTAATGA ATCTTTTTTTCTCTACAGCAAT CACTAATGATGACCATTATCAA TTTGGCTCAGAATTTTG	Repair template sequences (edited bases are shown in red, with the amino-acid changing mutations bolded, and the sgRNA binding site underlined)

TOP2B M>Q (antisense strand)	CAAAGTTCTGAGCTAAGTTCAC AATAGTCATCATCAATGCCATC TATACAACAGAAGAAGACAGA ACATAACATTAATATTCT	Repair template sequences (edited bases shown in red, with the amino-acid changing mutations bolded, and the sgRNA binding site underlined)
TOP2A forward	TTGTGGAAAGGACGAAACACC GGTGAGGTTAAGTCATAATGTA TTTGT	PCR primers, first round. Underlined sequence binds to target DNA. Red nucleotides are sequence adaptors used for the second round of PCR
TOP2A reverse	TCTACTATTCTTTCCCCTGCAC TGTCCCCTGGCCTCTGCCACT AGAT	PCR primers, first round. Underlined sequence binds to target DNA. Red nucleotides are sequence adaptors used for the second round of PCR
TOP2B forward	TTGTGGAAAGGACGAAACACC GCTTTATTCTTCACTTGGATTT TAATTC	PCR primers, first round. Underlined sequence binds to target DNA. Red nucleotides are sequence adaptors used for the second round of PCR
TOP2B reverse	TCTACTATTCTTTCCCCTGCAC TGTCCACAGCTATAATTCCATC GAACA	PCR primers, first round. Underlined sequence binds to target DNA. Red nucleotides are sequence adaptors used for the second round of PCR
P5/forward prime	AATGATACGGCGACCACCGAG ATCTACACTCTTTCCCTACACG ACGCTCTTCCGATCT[s]TTGTG GAAAGGACGAAACACCG	Green: P5 & P7 sequences for attachment to Illumina flow cell Blue: Illumina sequencing primer binding sites [s]: Stagger sequence. To prevent monotemplate reads, a mixture of eight unique oligonucleotides are used that vary in their length in this region. The nucleotides used in this position are: [no additional bases, C, GC, AGC, CAAC, TGCACC, ACGCAAC, GAAGACCC] Red: Sequences that bind to round 1 PCR products NNNNNNNN: 8 nucleotide barcode used to identify each sample
P7/reverse primer	CAAGCAGAAGACGGCATACGA GATNNNNNNNNGTGACTGGAG TTCAGACGTGTGCTCTTCCGAT CTTCTACTATTCTTTCCCCTGC ACTGT	Same as above

High-throughput fitness assays

We used a modified version of the high-throughput fitness assay (HTA) described previously [54]. In short, strains are passaged for four generations to reduce transgenerational effects from starvation or other stresses. Strains are then bleach-synchronized and aliquoted to 96-well microtiter plates at approximately one embryo per microliter in K medium [55]. Embryos are then hatched overnight to the L1 larval stage. The following day, hatched L1 animals are fed HB101 bacterial lysate (Pennsylvania State University Shared Fermentation Facility, State College, PA) at a final concentration of 5 mg/ml and grown to the L4 stage after two days at 20ºC. Three L4 larvae are then sorted using a large-particle flow cytometer (COPAS BIOSORT, Union

Biometrica, Holliston, MA) into microtiter plates that contain HB101 lysate at 10 mg/ml, K medium, 31.25 μM kanamycin, and either drug dissolved in 1% DMSO or 1% DMSO. The animals are then grown for four days at 20ºC. During this time, the animals will mature to adulthood and lay embryos that encompass the next generation. Prior to the measurement of fitness parameters from the population, animals are treated with sodium azide (50 mM) to straighten their bodies for more accurate length measurements. Traits that are measured by the BIOSORT include brood size and animal length (time of flight or TOF).

Calculation of fitness traits for genetic mappings

Phenotype data generated using the BIOSORT were processed using the R package *easysorter*, which was specifically developed for processing this type of data set [147]. Briefly, the function *read_data*, reads in raw phenotype data, runs a support vector machine to identify and eliminate bubbles. Next, the *remove_contamination* function eliminates any wells that were contaminated prior to scoring population parameters for further analysis. Contamination is assessed by visual inspection. The *sumplate* function is then used to generate summary statistics of the measured parameters for each animal in each well. These summary statistics include the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} quantiles for TOF. Measured brood sizes are normalized by the number of animals that were originally sorted into the well. After summary statistics for each well are calculated, the *regress(assay=TRUE)* function in the *easysorter* package is used to fit a linear model with the formula (*phenotype ~ assay*) to account for any differences between assays. Next, outliers are eliminated using the *bamf_prune* function. This function eliminates strain values that are greater than two times the IQR plus the 75^{th} quantile or two times the IQR minus the 25^{th} quantile, unless at least 5% of the strains lie outside this range. Finally, drug-specific effects are calculated using the *regress(assay=FALSE)* function from *easysorter*, which fits a linear model with the formula (*phenotype ~ control phenotype*) to account for any differences in population parameters present in control DMSO-only conditions.

Topoisomerase II poisons dose-response assays

All dose-response experiments were performed on four genetically diverged strains (Bristol, Hawaii, DL238, and JU258) in technical quadruplicates prior to performing GWA and linkage mapping experiments. Animals were assayed using the HTA, and phenotypic analysis was performed as described above. Drug concentrations for GWA and linkage mapping experiments were chosen based on two criteria – an observable drug-specific effect and broad-sense heritability H^2. We aimed to use the first concentration for which a drug-specific effect with a maximum H^2 was observed. Broad-sense heritability estimates were calculated using the *lmer* function in the *lme4* package with the following model (*phenotype ~1 + (1|strain)*). Concentrations for each chemotherapeutic used in mapping experiments are; etoposide – 250μM, teniposide – 125 μM, amsacrine – 50 μM, dactinomycin – 15 μM, and XK469 – 1000 μM. All topoisomerase II poisons used in this study were purchased from Sigma (XK469 cat#X3628, etoposide cat#E1383, amsacrine cat#A9809, dactinomycin cat#A1410, and teniposide cat#SML0609).

Topoisomerase II poisons linkage mapping analysis

A total of 265 RIAILs were phenotyped in the HTA described previously for control and etoposide conditions. The phenotype data and genotype data were entered into R and scaled to have a mean of zero and a variance of one for linkage analysis. Quantitative trait loci (QTL) were detected by calculating 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 phenotype trait values [134]. The maximum LOD score for each chromosome for each trait was retained from three iterations of linkage mappings. We randomly permuted the phenotype values of each RIAIL while maintaining correlation structure among phenotypes 1000 times to estimate significance empirically. The ratio of expected peaks to

observed peaks was calculated to determine the genome-wide error rate of 5% of LOD 4.61. Broad-sense heritability was calculated as the fraction of phenotypic variance explained by strain from fit of a linear mixed-model of repeat phenotypic measures of the parents and RIAILs [148]. The total variance explained by each QTL was divided by the broad-sense heritability to determine how much of the heritability is explained by each QTL. Confidence intervals were defined as the regions contained within a 1.5 LOD drop from the maximum LOD score.

Topoisomerase II genome-wide association mapping

Genome-wide association (GWA) mapping was performed using 152 *C. elegans* isotypes. We used the *cegwas* R package for association mapping [27]. This package uses the EMMA algorithm for performing association mapping and correcting for population structure [56], which is implemented by the GWAS function in the *rrBLUP* package [149]. The kinship matrix used for association mapping was generated using a whole-genome high-quality single-nucleotide variant (SNV) set [44] and the *A.mat* function from the *rrBLUP* package. SNVs previously identified using RAD-seq [43] 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 $-log_{10}(p)$ value greater than the Bonferroni-corrected value were processed further using fine mapping. Tajima's D was calculated using the *tajimas_d* function in the *cegwas* package using default parameters (window size = 300 SNVs, sliding window distance = 100 SNVs, outgroup = N2).

Topoisomerase II QTL confidence interval mapping

Fine mapping was performed on variants from the whole-genome high-quality SNV set within a defined region of interest for all mappings that contained a significant QTL. Regions of interest surrounding a significant association were determined by simulating a QTL with 20% variance

explained at every RAD-seq SNV present in 5% of the phenotyped population. We then identified the most correlated SNV for each mapping. Next, we determined the number of SNVs away from the simulated QTL SNV position that captured 95% of the most correlated SNVs. A range of 50 SNVs upstream or downstream of the peak marker captured 95% of the most significant SNVs in the simulated mappings. We therefore used a region 50 SNVs from the last SNV above the Bonferroni-corrected *p*-value on the left side of the peak marker and 50 SNVs from the last SNV above the Bonferroni-corrected *p*-value on the right side of the peak marker.

The *snpeff* function from the *cegwas* package was used to identify SNVs from the whole-genome SNV set with high to moderate predicted functional effects present in a given region of interest [81]. The correlation between each variant in the region of interest and the kinship-corrected phenotype used in the GWA mapping was calculated using the *variant_correlation* function and processed using the *process_correlations* function in the *cegwas* package. ClustalX was used to perform the multiple sequence alignment between various topoisomerase II orthologs.

Topoisomerase II poison QTL near-isogenic line generation

NILs were generated by crossing N2xCB4856 RIAILs to each parental genotype. For each NIL, eight crosses were performed followed by six generations of selfing to homozygose the genome. Reagents used to generate NILs are detailed in Table 3-2. The NILs responses to 250 µM etoposide were quantified using the HTA fitness assay described above.

Toposiomerase II poison dominance test

Dominance experiments were performed using the fluorescent reporter strain EG7952 *oxTi207 [eft-3p::GFP::unc-54 3'UTR + hsp::peel-1 + NeoR + Cbr-unc-119(+)]*. Hermaphrodites of N2 and CB4856 were crossed to male EG7952 reporter strain, which expresses GFP, to ensure that we

could measure heterozygous cross progeny by the presence of GFP. Three GFP-positive progeny were manually transferred to a 96-well assay microtiter plate containing to 250 μM etoposide dissolved in 1% DMSO or 1% DMSO control, in addition to K medium, HB101 lysate at 10 mg/ml, and 31.25 μM kanamycin. Animals were grown for four days at 20ºC. The phenotypes of the progeny were scored using the BIOSORT as described above. Heterozygous progeny were computationally identified as those individuals that had fluorescence levels between the non-fluorescent and fluorescent parental strains.

Top-2 and *npp-3* complementation

To perform the complementation experiments, N2 and CB4856 males were both crossed to both VC1474 *top-2(ok1930)/mIn1 [mIs14 dpy-10(e128)]* and VC1505 *npp-3(ok1900)/mIn1 [mIs14 dpy-10(e128)]* hermaphrodites. Three non-GFP L4 hermaphrodite progeny were manually picked into experimental wells containing either 250 μM etoposide dissolved in 1% DMSO or 1% DMSO without etoposide, in addition to HB101 lysate at 10 mg/ml, K medium, and 31.25 μM kanamycin. Animals were grown for four days at 20ºC. The phenotypes of the progeny were scored using the BIOSORT as described above.

Top-2 reciprocal hemizygosity

VC1474 *top-2(ok1930)/mIn1 [mIs14 dpy-10(e128)]* was used for *top-2* complementation tests. *Top-2(ok1930)* and *mIn1[mIs14 dpy-10(e128)]* were individually introgressed into CB4856 for 10 generations. Once individual crosses were completed, CB4856 *mIn1 [mIs14 dpy-10(e128)]* was crossed to CB4856 *top-2(ok1930)* to generate ECA338, which contains a *mIn1*-balanced *top-2(ok1930)*. oECA1003 and oECA1004 were used to verify the presence of *top-2(ok1930)* during crosses.

To perform the reciprocal hemizygosity experiment, N2 and CB4856 males were both crossed to both VC1474 and ECA338 hermaphrodites. Three non-GFP L4 hermaphrodite progeny were manually picked into experimental wells containing either 250 µM etoposide dissolved in 1% DMSO or 1% DMSO without etoposide, in addition to HB101 lysate at 10 mg/ml, K medium, and 31.25 µM kanamycin. Animals were grown for four days at 20ºC. The phenotypes of the progeny were scored using the BIOSORT as described above.

Generation of top-2 allele replacement strains

All allele replacement strains were generated using CRISPR/Cas9-mediated genome engineering, using the co-CRISPR approach [150] with Cas9 ribonucleoprotein delivery [69]. Alt-R^{TM} crRNA and tracrRNA were purchased from IDT (Skokie, IL). tracrRNA (IDT, 1072532) was injected at a concentration of 13.6 µM. The *dpy-10* and the *top-2* crRNAs were injected at 4 µM and 9.6 µM, respectively. The *dpy-10* and the *top-2* single-stranded oligodeoxynucleotides (ssODN) repair templates were injected at 1.34 µM and 4 µM, respectively. Cas9 protein (IDT, 1074182) was injected at 23 uM. To generate injection mixes, the tracrRNA and crRNAs were incubated at 95ºC for 5 minutes and 10ºC for 10 minutes. Next, Cas9 protein was added and incubated for 5 minutes at room temperature. Finally, repair templates and nuclease-free water were added to the mixtures and loaded into pulled injection needles (1B100F-4, World Precision Instruments, Sarasota, FL). Individual injected P_0 animals were transferred to new 6 cm NGM plates approximately 18 hours after injections. Individual F_1 rollers were then transferred to new 6 cm plates and allowed to generate progeny. The region surrounding the desired Q797M (or M797Q) edit was then amplified from F_1 rollers using oECA1087 and oECA1124. The PCR products were digested using the *Hpy*CH4III restriction enzyme (R0618L, New England Biolabs, Ipswich, MA). Differential band patterns signified successfully edited strains because the N2 Q797, which is encoded by the CAG codon, creates an additional *Hpy*CH4III cut site. Non-Dpy,

non-Rol progeny from homozygous edited F_1 animals were propagated. If no homozygous edits were obtained, heterozygous F_1 progeny were propagated and screened for presence of the homozygous edits. F_1 and F_2 progeny were then Sanger sequenced to verify the presence of the proper edit. Allele swap strains responses to the topoisomerase II poisons were quantified using the HTA fitness ass described above.

TOP-2 molecular docking simulations

The *C. elegans* TOP-2 three-dimensional structure homology model was built by threading the *C. elegans* TOP-2 peptide to the human topoisomerase II beta structure (PDB accession code 3QX3; 59% identity, 77% similarity) using the Prime3.1 module implemented in Schrodinger software [151,152]. After building the model, a robust energy minimization was carried out in the Optimized Potentials for Liquid Simulations (OPLS) force field. The minimized structure was subjected to MolProbity analysis, and the MolProbity score suggested with greater than 95% confidence that the minimized structure model was a good high-resolution structure [153].

Next, the Prot-Prep wizard was used to prepare the TOP-2 homology model, which fixed the hydrogen in the hydrogen bond orientations, eliminated the irrelevant torsions, fixed the missing atoms, assigned the appropriate force field charges to the atoms [154]. After preparing the structure, the glutamine 797 was mutated to various rotamers of methionine (Q797M), which subsequently underwent minimization in the OPLS force field. The energy-minimized structure was used in the *in silico* experiments.

The structure data file of etoposide (DrugBank ID: DB00773) was obtained from PubMed and was subjected to ligand preparation panel of the Schrodinger software. Using the induced fit docking (IFD) module of Schrodinger and Suflex software, we carried out the docking of etoposide with the glutamine and methionine forms of the *C. elegans* TOP-2 homology model.

After the docking experiments, we analyzed the docked poses of the ligands bound to the TOP-2 homology models from both docking engines. Change in free energy (∆G) and the hydrophobicity parameter were calculated using Schrodinger.

Topoisomerase II CRISPR-Cas9 gene editing in human cells

Gene-editing experiments were performed in human 293T cells (ATCC) grown in DMEM with 10% FBS. On day zero, 500,000 cells were seeded per well in a six-well plate format. The following day, two master mixes were prepared: a) LT-1 transfection reagent (Mirus) was diluted 1:10 in Opti-MEM and incubated for 5 minutes; b) a DNA mix of 500 ng Cas9-sgRNA plasmid with 250 pmol repair template oligonucleotide (Table 3-2) was diluted in Opti-MEM in a final volume of 100 µL. 100 µL of the lipid mix was added to each of the DNA mixes and incubated at room temperature for 25 minutes. Following incubation, the full 200 µL volume of DNA and lipid mix was added drop-wise to the cells, and the cells were centrifuged at 1000xg for 30 min. Six hours post-transfection, the media on the cells was replaced, and the cells were passaged as needed. On day six, 5 million cells from each condition were pelleted to serve as an early time point for the editing efficiency, and 5 million cells were then passaged on the five drugs at two doses for 12 days, at which time all surviving cells were pelleted. Concentrations used for each small molecule are: etoposide – 500 nM, 100 nM; amasacrine – 500 nM, 100nM; teniposide – 20 nM, 4 nM; dactinomycin – 4 nM, 800 pM; and XK469 – 5 µM, 1µM.

Analysis of CRISPT-Cas9 topoisomerase II editing in human cells

gDNA was extracted from cell pellets using the QIAGEN (QIAGEN, Hilden, Germany) Midi or Mini Kits based on the size of the cell pellet (cat # 51183, 51104) according to the manufacturer's recommendations. TOP2A and B loci were first amplified with 17 cycles of PCR using a touchdown protocol and the NEBnext 2x master mix (New England Biolabs M0541).

The resulting product served as input to a second PCR, using primers that appended a sample-specific barcode and the necessary adaptors for Illumina sequencing. The resulting DNA was pooled, purified with SPRI beads (A63880, Beckman Coulter, Brea, CA), and sequenced on an Illumina MiSeq with a 300 nucleotide single-end read with an eight nucleotide index read. For each sample, the number of reads exactly matching the wild-type and edited TOP2A/B sequence were determined.

Results

A single major-effect locus explains variation in response to etoposide

We investigated etoposide sensitivity in *C. elegans* using a high-throughput fitness assay. In brief, animals were grown in liquid culture in presence of etoposide, and body lengths of progeny and offspring production were measured using a COPAS BIOSORT. In this assay, shorter body lengths are indicative of developmental delay. To identify an appropriate dose of etoposide for this assay, we performed dose-response experiments on four genetically diverged isolates of *C. elegans*: N2 (Bristol), CB4856 (Hawaii), JU258, and DL238. We chose 250 μM etoposide for further experiments because it was the lowest concentration at which we observed an etoposide-specific effect in all four strains tested, trait differences between the laboratory Bristol strain (N2) and a wild strain from Hawaii (CB4856) strains were maximized, and the median animal length was highly heritable.

When grown in etoposide, progeny of the Hawaii strain are on average 75 μm shorter than progeny of the Bristol strain. To map the genetic variants underlying this difference, we performed our high-throughput fitness assay on a panel of 265 recombinant inbred advanced intercross lines (RIAILs), generated between a Bristol derivative (QX1430) and Hawaii[54]. We

measured median animal length for each RIAIL strain grown in etoposide, and we corrected for assay-to-assay variability and effects of the drug carrier (DMSO) using a linear model. We used the resulting regressed median animal length trait (referred to as animal length) for quantitative trait locus (QTL) mapping. This mapping identified a major-effect QTL for etoposide resistance on chromosome II at 11.83 Mb (Figure 3-1A). This QTL explained 27% of the phenotypic variance among the recombinant lines. The QTL confidence interval spans from 11.67 to 11.91 Mb on chromosome II and contains 90 genes, 68 of which contain variation between the parental strains.

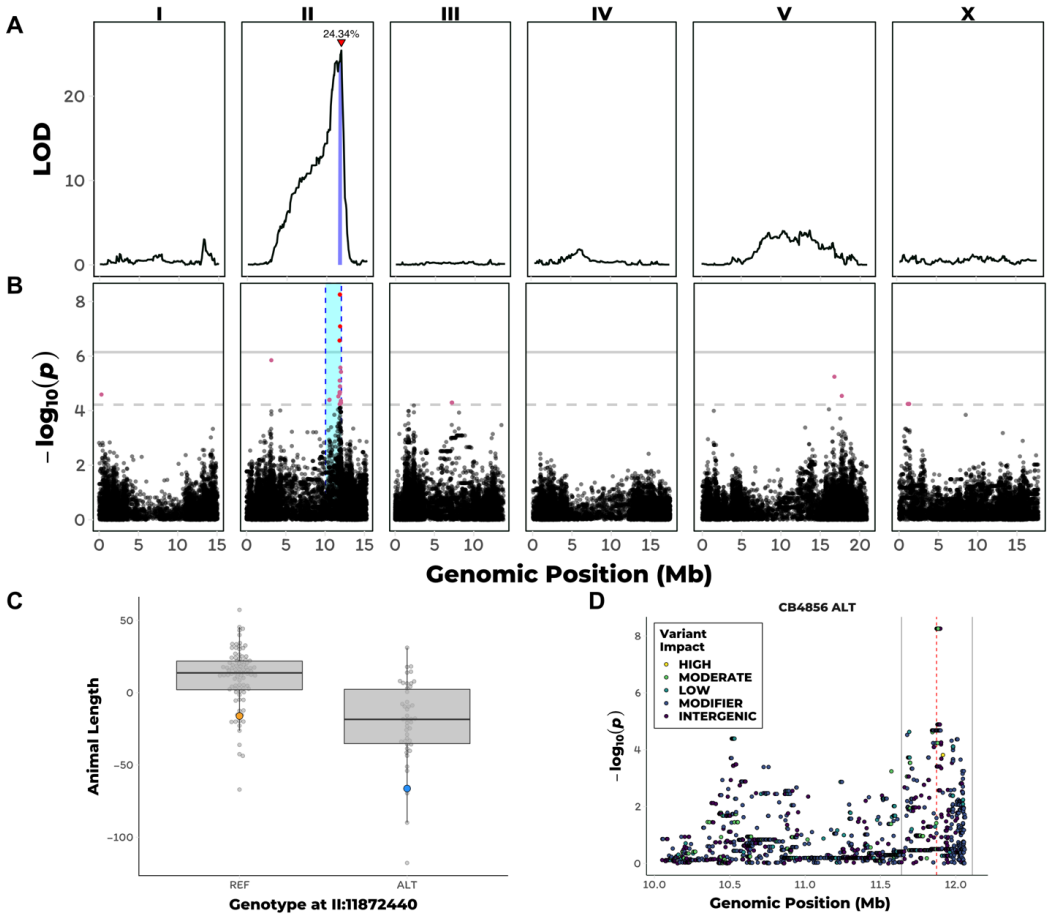

Figure 3-1 GWA and linkage mapping of etoposide response variation
A) Linkage mapping plot for regressed median animal length in the presence of etoposide is shown. The significance values (logarithm of odds, LOD, ratio) for 1454 markers between the Bristol and Hawaiian strain are on the y-axis, and the genomic position (Mb) separated by chromosome is plotted on the x-axis. Each tick on the x-axis corresponds to 5 Mb. The associated 1.5 LOD-drop confidence intervals are represented by blue bars. B) A manhattan plot regressed median animal size in the presence of etoposide is shown. Each dot represents an SNV that is present in at least 5% of the assayed wild

population. The genomic position in Mb, separated by chromosome, is plotted on the x-axis and the -log10(*p*) for each SNV is plotted on the y-axis. SNVs are colored red if they pass the genome-wide Bonferroni-corrected significance (BF) threshold, which is denoted by the gray horizontal line. SNVs are colored pink if they pass the genome-wide eigen-decomposition significance (ED) threshold, which is denoted by the dotted gray horizontal line. The genomic region of interests surrounding the QTL that pass the by cyan rectangles. C) Tukey box plots of phenotypes used for association mapping in (A) are shown. Each dot corresponds to the phenotype of an individual strain, which is plotted on the y-axis. Strains are grouped by their genotype at the peak QTL position (highest red SNV from panel B, chrII:11877529), where REF corresponds to the allele from the reference N2 strain. The N2 (orange) and CB4856 (blue) strains are highlighted. D) Fine mapping of the chromosome II region of interest (cyan region from panel A) is shown. Each dot represents an SNV present in the CB4856 strain. The association between the SNV and regressed median animal size in the presence of etoposide is shown is shown on the y-axis and the genomic position of the SNV is shown on the x-axis. Dots are colored by their SnpEff predicted effect.

We next sought to validate this QTL using homozygous reciprocal near-isogenic lines (NILs), which contain either the QTL confidence interval from the Bristol strain introgressed into the Hawaii strain or the interval from the Hawaii strain introgressed into the Bristol strain. NILs with the genomic interval derived from the Bristol strain have increased resistance to etoposide compared to the Hawaii strain. Similarly, NILs with the genomic interval derived from the Hawaii strain exhibited decreased resistance to etoposide. These results confirmed that genetic variation located on the right arm of chromosome II contributes to differential etoposide susceptibility.

The same locus on chromosome II explains variation in response to etoposide in a panel of wild *C. elegans* isolates

In the initial dose response experiments, we found that JU258 and DL238 had different responses to etoposide than the Bristol and Hawaii strains, suggesting that additional genetic variation present in the wild *C. elegans* population could also contribute to etoposide response. To identify this additional variation, we performed a genome-wide association (GWA) mapping of etoposide resistance in 138 wild *C. elegans* isolates. This analysis led to the identification of a QTL on the right arm of chromosome II with a peak position at 11.88 Mb (Figure 3-1B). This QTL has a genomic region of interest that spans from 11.70 to 12.15 Mb for which we found no

evidence of selection or geographic clustering of the resistant allele. In addition, this QTL overlaps with the QTL identified through linkage mapping described above. Of the 139 wild isolates assayed, including the Hawaiian strain, 46 have the alternate (non-Bristol) genotype at the peak position on chromosome II (Figure 3-1C). Similar to our observations using the recombinant lines, the 46 strains that contain the alternate genotype are more sensitive to etoposide than strains containing the Bristol genotype at the QTL peak marker. We hypothesized that variation shared between the Hawaiian strain and the other 45 alternate-genotype strains contributes to etoposide sensitivity because we detected overlapping QTL, with the same direction of effect, between GWA and linkage mapping experiments. This hypothesis suggested that we could condition a fine-mapping approach on variants found in the Hawaiian strain and shared across these 45 strains.

To fine-map the QTL, we focused on variants shared among wild isolates. Using data from the *C. elegans* whole-genome variation dataset[44] we calculated Spearman's *rho* correlations between animal length and each single-nucleotide variant (SNV) in the QTL confidence interval (Figure 3-1D). SNVs in only three genes, *npp-3*, *top-2*, and *ZK930.5*, were highly correlated with the etoposide response (*rho* > 0.45). Of these genes, the *top-2* gene encodes a topoisomerase II enzyme that is homologous to the two human isoforms of topoisomerase II. We prioritized *top-2* because topoisomerase II enzymes are the cellular targets for etoposide[139].

Genetic variation in *top-2* contributes to differential etoposide sensitivity

To determine if genetic variation present in the *top-2* gene contributes to differential etoposide sensitivity, we performed a reciprocal hemizygosity test [137]. Prior to this test, we determined that resistance to etoposide is dominant by measuring the lengths of F1 heterozygotes from a cross between the Bristol and Hawaii strains in the presence of etoposide. Additionally, we tested *npp-3* and *top-2* deletion alleles from the Bristol genetic background and found that only

loss of *top-2* contributes to etoposide sensitivity. To more definitively show a causal connection of *top-2* variation to etoposide sensitivity, we used a reciprocal hemizygosity test [137]. First, we introgressed the top-2(*ok1930*) deletion allele into the Hawaiian genetic background. The Bristol/Hawaii(Δ*top-2*) heterozygote that contains the Bristol *top-2* allele is more resistant to etoposide treatment than the Hawaii/Bristol(Δ*top-2*) heterozygote, which suggests that the Bristol top-2 allele underlies etoposide resistance (Figure 3-2A). The observed differences between the Hawaii/Bristol(Δtop-2) and Bristol/Hawaii(Δ*top-2*) heterozygotes confirmed that *top-2* variation underlies differential susceptibility to etoposide.

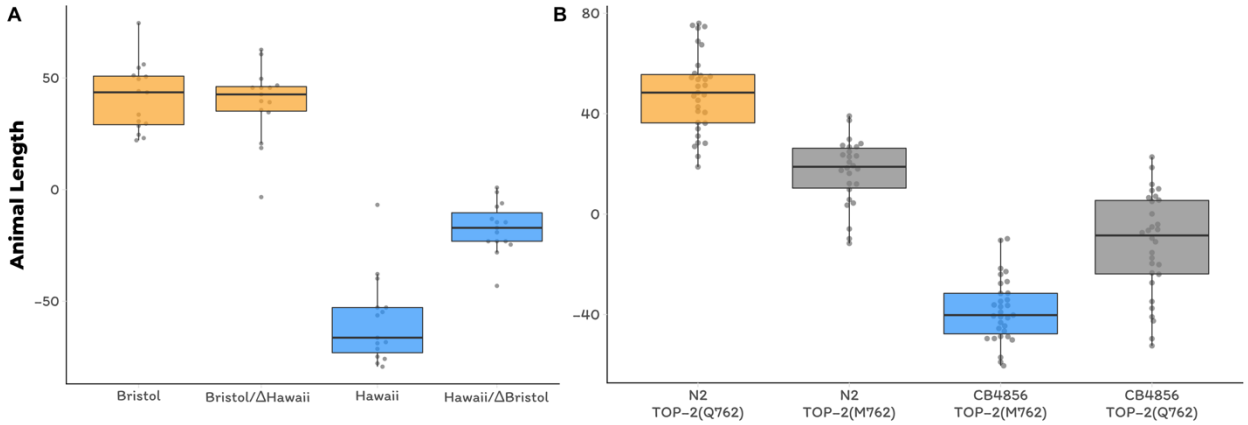

Figure 3-2 *top-2* reciprocal hemizygosity and validation of TOP-2 Q762M allele
A) Tukey box plots of the residual median animal length distribution of Bristol (orange) and Hawaii (blue) compared to two heterozygous top-2 deletion strains in response to etoposide are shown. The Bristol strain and the Bristol/Hawaii(*Δtop-2*) heterozygous strain are not significantly different from each other (Tukey's HSD, *p*-value 0.9990812), but all other comparisons are significant (Tukey's HSD *p*-value < 0.0001). B) Tukey box plots of residual median animal length after etoposide exposure are shown (Bristol, orange; Hawaii, blue; allele replacement strains, gray). Labels correspond to the genetic background and the corresponding residue at position 797 of TOP-2 (Q for glutamine, M for methionine). Every pair-wise strain comparison is significant (Tukey's HSD, Bristol—Hawaii, Bristol—Hawaii swap, Hawaii—Bristol swap *p*-value < 2.0E-14; Bristol—Bristol swap *p*-value = 1.2E-10; Bristol swap—Hawaii swap *p*-value = 3.4E-9; Hawaii—Hawaii swap *p*-value = 1.3E-8).

A glutamine-to-methionine variant in TOP-2 contributes to etoposide response

To identify genetic variants in *top-2* that contribute to etoposide resistance in the Bristol strain, we focused on genomic differences between the Bristol and Hawaii strains. Based on gene expression data between the Bristol and Hawaii strains [155], *top-2* is expressed at similar

levels. Therefore, we concluded that etoposide resistance in the Bristol strain is likely caused by coding variation. The *C. elegans top-2* gene contains 31 SNVs across the population-wide sample of 139 wild isolates. We narrowed our search to 16 variants present in the Hawaiian strain. Two of these variants are in the 3' UTR, three are in introns, and six are synonymous variants that likely do not contribute to etoposide resistance. The remaining five variants encode for amino acid changes in the TOP-2 enzyme. Of these five variants, four were highly correlated with etoposide sensitivity in the wild isolate panel: Q797M, I1206L, Q1217A, and D1387N. Multiple-sequence alignment of TOP-2 peptides across yeast, *D. melanogaster*, mice, and humans revealed that I1206L, Q1217A, and D1387N are in the highly variable C-terminal domain. By contrast, the Q797M variant is located in the conserved DNA binding and cleavage domain [156]. Structural data suggest that the TOP-2 Q797 residue lies within the putative etoposide-binding pocket [143], and the corresponding residue is a methionine (M762) in the hTOPOIIα and a glutamine (Q778) in hTOPOIIβ [145]. Additionally, the two human isoforms differ in one other residue within the putative etoposide-binding pocket (S800(α)/A816(β)). Therefore, the *C. elegans* glutamine-to-methionine TOP-2 variant mirrors one of two differences within the etoposide-binding pocket of the two human topoisomerase II enzyme isoforms. Crucially, hTOPOIIα forms a more stable DNA-TOPOII cleavage complex with etoposide than hTOPOIIβ [157]. We hypothesized that etoposide sensitivity in both *C. elegans* and the human isoforms is affected by this residue.

To test the effects of the Q797M variant on *C. elegans* response to etoposide, we used CRISPR/Cas9-mediated genome editing to change this residue. We replaced the glutamine residue in the Bristol strain with a methionine and the methionine residue in the Hawaii strain with a glutamine. We exposed the allele-replacement strains to etoposide and found that the methionine-containing Bristol animals were more sensitive than glutamine-containing Bristol animals (Figure 3-2B). Conversely, the glutamine-containing Hawaii animals were more

resistant to etoposide than the methionine-containing Hawaii animals (Figure 3-2B). These results confirm that this variant contributes to differential etoposide sensitivity between the Bristol and Hawaii strains.

Methionine mediates stronger hydrophobic interactions with etoposide than glutamine

We hypothesized that the non-polar functional group attached to the glycosidic bond of etoposide contributes to increased stability of the drug-enzyme complex by forming a more stable interaction with the methionine residue than with the glutamine residue. To test this hypothesis, we simulated etoposide docking into the putative drug-binding pocket of the TOP-2 homology model generated by threading the *C. elegans* peptide sequence into the hTOPOIIβ structure (RMSD = 1.564Å, PDB:3QX3; [143]). Upon etoposide binding, the free energy (ΔG) of the drug-binding pocket was -10.09 Kcal/mol for TOP-2 Q797 (Figure 3-3A, 3-3C) and -12.67 Kcal/mol for TOP-2 M797 (Figure 3-3B, 3-3C). This result suggests that etoposide interacts more favorably with TOP-2 M797 than with TOP-2 Q797, consistent with our results in live worms. A more favorable drug-enzyme interaction, as indicated by a more negative ΔG, likely causes increased stability of the TOP2 cleavage complexes, which has been shown to result in a greater number of double-stranded breaks throughout the genome [158]. Therefore, we expect *C. elegans* strains that contain a methionine at this residue to accumulate more genomic damage when exposed to etoposide. The resulting physiological effect of increased genomic damage likely delays development and causes the progeny of exposed individuals to be shorter.

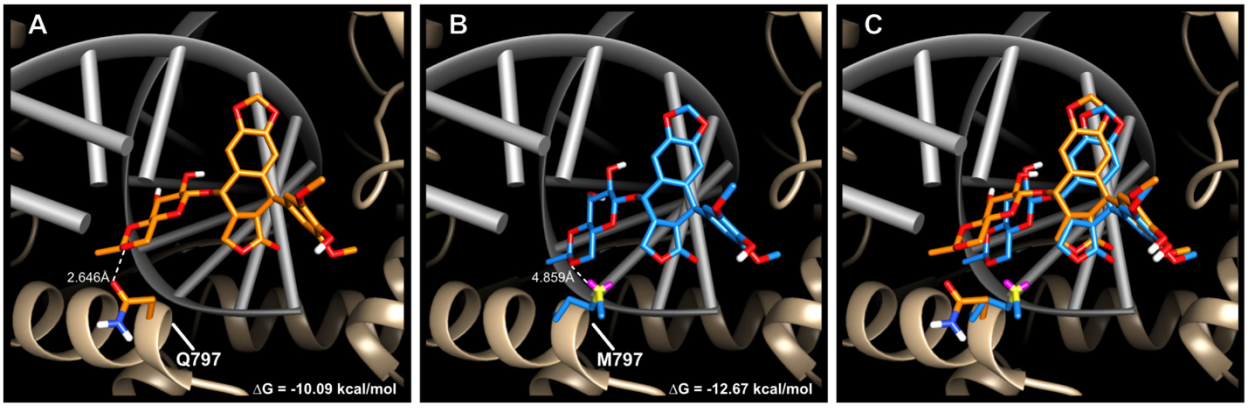

Figure 3-3 Molecular docking of etoposide in *C. elegans* TOP-2
Etoposide docked into the (A) glutamine-containing Bristol TOP-2 enzyme (*ΔG* = -10.09 kcal/mol), B) methionine-containing Hawaii TOP-2 enzyme (*ΔG* = -12.67 kcal/mol), and C) the overlay of both structures. Glutamine is colored in orange, and methionine is colored in blue. Etoposide docked into the glutamine-containing TOP-2 enzyme is orange, and etoposide docked into the methionine-containing TOP-2 enzyme is blue. DNA is colored in gray, and the ribbon representation of the TOP-2 protein is shown in tan.

TOP-2 variation causes allele-specific interactions with an expanded set of topoisomerase II poisons

Because the molecular docking simulations explain the observed physiological effects of etoposide exposure, we hypothesized that the 797 residue of TOP-2 would mediate differential interactions with additional topoisomerase II poisons based on their chemical structures. Like etoposide, teniposide, dactinomycin, and amsacrine each contain core cyclic rings that are thought to interfere with the re-ligation step of the topoisomerase II catalytic cycle through DNA interactions [139]. However, the functional groups attached to the core cyclic rings of each poison vary in their polarity and size, which could affect interactions with topoisomerase II enzymes. For example, the only difference between teniposide and etoposide is the presence of a thienyl or methyl group attached to the D-glucose derivative, respectively, but they share a similarly sized and hydrophobic functional group. We predicted that these two drugs would have comparable interactions with the TOP-2 alleles and elicit a similar physiological response. By contrast, the polar functional groups of dactinomycin likely have stronger interactions with the glutamine variant and induce increased cytotoxicity in animals that contain this allele. We

quantified the physiological responses of the TOP-2 allele-replacement strains exposed to these two drugs and found that each response matched our predictions (Figure 3-4A-B). Specifically, strains harboring the TOP-2 methionine allele were more sensitive to teniposide than those strains that contain the glutamine allele. Conversely, strains with the TOP-2 glutamine allele were more sensitive to dactinomycin than those strains with the methionine allele. Unlike etoposide, teniposide, or dactinomycin, the core cyclic rings of amsacrine do not have an equivalent functional group to interact with the TOP-2 797 residue, suggesting that variation at TOP-2 residue 797 will have no impact on amsacrine sensitivity. Although the Bristol and Hawaiian strains differed, we found that the allele status of TOP-2 had no quantifiable effect on amsacrine response (Figure 3-4C) and different genomic loci control response to this drug. These results support the hypothesis that the polarity of the putative drug-binding pocket determines the cytotoxic effects of multiple, but not all, topoisomerase II poisons. To further explore this hypothesis, we tested a drug (XK469) that has preferential hTOPOIIβ specificity [159]. Surprisingly, we found that the strains that contain the methionine allele (like hTOPOIIα) were more sensitive to XK469. This result indicates that an additional mechanism might contribute to XK469 specificity in human cells and underscores the importance of functional validation of specific residues that are thought to be involved in targeted drug binding.

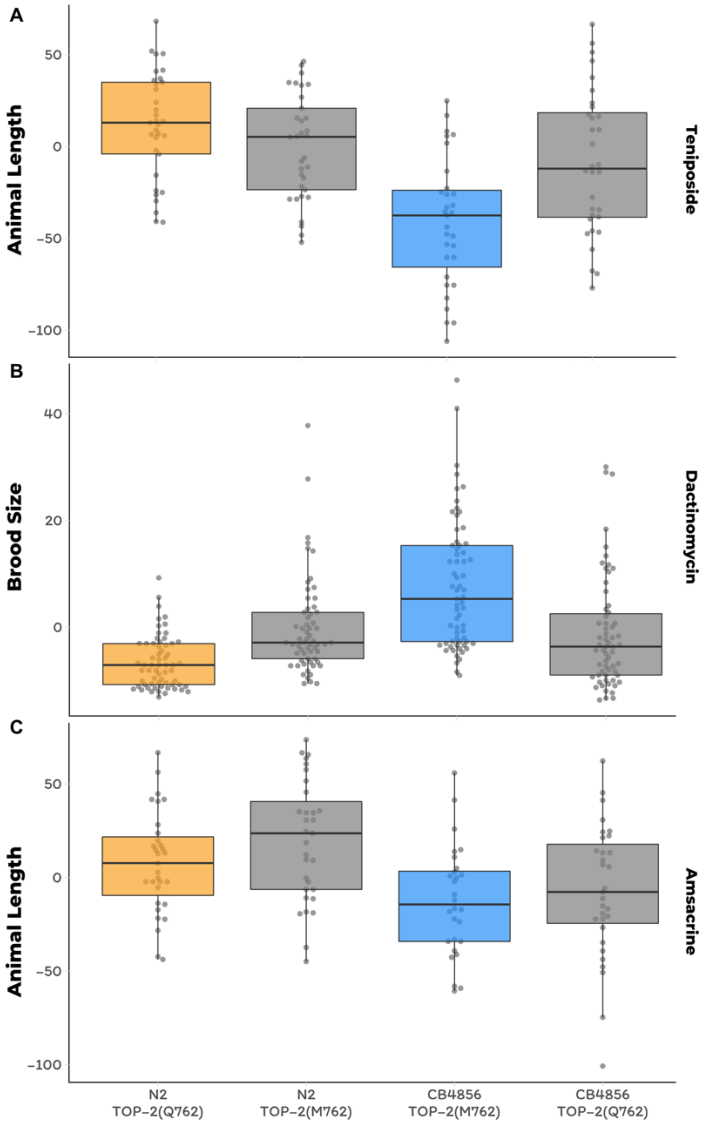

Figure 3-4 The Q762M allele affects variation to other topoisomerase II poisons
Tukey box plots of A) regressed brood size in response to teniposide. Mean phenotypes of the Bristol and Hawaiian allele swap strains are significantly different from the Bristol and Hawaiin parental strains (Tukey's HSD, Bristol—Bristol Swap *p*-value = 0.0009; Hawaii—Hawaiian Swap *p*-value < 1E-7). B) Tukey box plots of regressed brood size in response to dactinomycin. Mean phenotypes of the Bristol and Hawaiian allele swap strains are significantly different from the Bristol and Hawaiin parental strains (Tukey's HSD, Bristol—Bristol Swap *p*-value = 0.0002; Hawaii—Hawaiian Swap *p*-value = 3E-7). C) Tukey box plots of regressed animal length in response to amsacrine show allele swap strains are not significantly different from parental strains (Tukey's HSD, Bristol—Bristol Swap *p*-value = 0.925; Hawaii—Hawaiian Swap *p*-value = 0.414). Orange corresponds to the Bristol genetic background and blue to the Hawaii background. Labels correspond to the genetic background and the corresponding residue at position 797 of TOP-2 (Q for glutamine, M for methionine).

Variation in the equivalent site in topoisomerase II alpha causes differential susceptibility to diverse poisons in human cells

To determine if differences in the hydrophobicities of the two human topoisomerase II putative drug-binding pockets underlie etoposide sensitivity, we used CRISPR/Cas9 genome editing and a pooled-sequencing approach to create human embryonic kidney 293 cells (293T) that encode hTOPOIIα enzymes with a hTOPOIIβ-like drug-binding pocket. Cells were incubated with genome-editing machinery for six hours, allowed to recover for five days, and then split into two populations for etoposide exposure or no etoposide exposure (Figure 3-5A). Etoposide treatment provided a selective pressure that upon further passaging led to a greater than 160-fold enrichment of cells that contain the glutamine-edited hTOPOIIα allele as compared to populations of cells exposed to no drug (Figure 3-5B). These results show that cells with the glutamine-edited hTOPOIIα allele are more resistant to etoposide treatment than cells with the non-edited methionine hTOPOIIα allele. Notably, the rarity of genome editing events makes it unlikely that every copy of the hTOPOIIα gene in this diploid/polyploid cell line is edited. Because we see etoposide resistance in these incompletely edited cells, hTOPOIIα dimeric complexes likely contain one edited and one wild-type copy of hTOPOIIα and do not bind etoposide as well as causing less cytotoxicity. These data confirm both our dominance test and the two-drug model of etoposide binding ([160] in which both enzymes of the homodimer must be bound by poison to be completely inhibited. Additionally, we performed the reciprocal experiment to edit the glutamine-encoding hTOPOIIβ gene to a version that encodes methionine. If the methionine hTOPOIIβ allele is more sensitive to etoposide than the glutamine hTOPOIIβ, we would expect to observe a depletion of methionine-edited cells upon etoposide treatment. However, because glutamine-to-methionine editing occurred in less than 1% of the cells, it was difficult to detect further reductions in methionine allele frequencies. Overall, we

demonstrate that this residue underlies variation in etoposide response in both *C. elegans* and human cell lines.

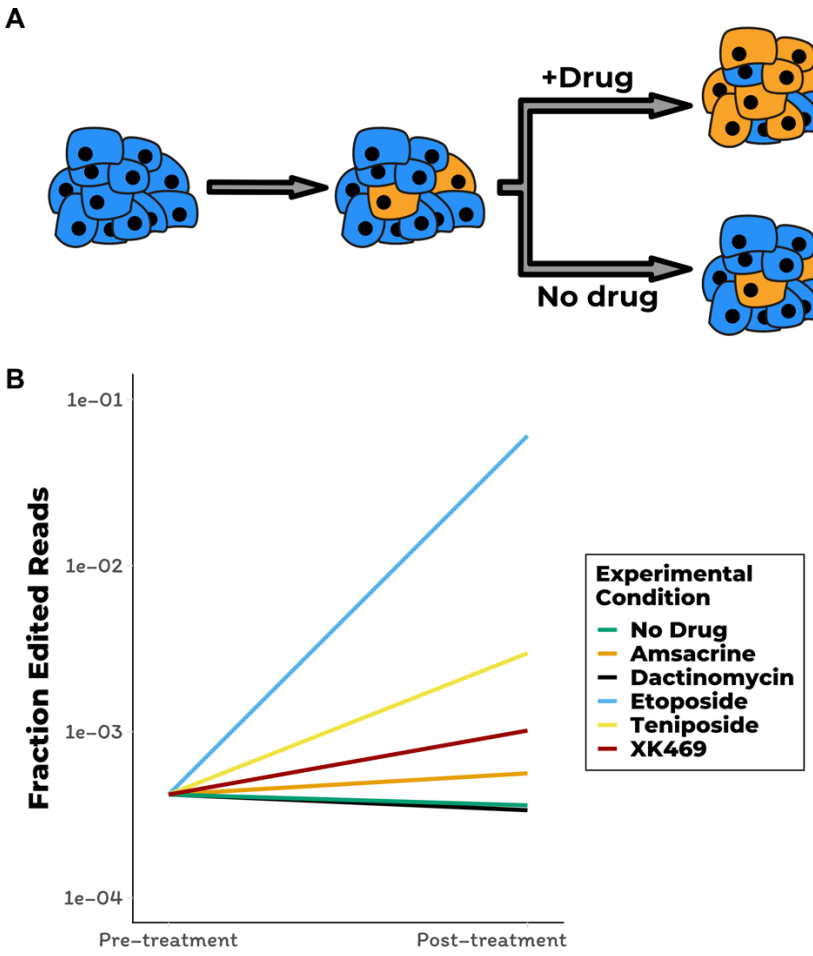

Figure 3-5 Human cell line validation of the Q762M allele
A) A cartoon depiction of human cell-line experiment is shown. A population of cells is incubated with CRISPR reagents to introduce the M762Q mutation in hTOPOIIα. A small fraction of cells are edited at this position and depicted in orange. Prior to splitting this population of cells into separate growth cultures with drug or control conditions, a subset of cells are prepared for sequencing to assess the fraction of edited cells. Split populations are then allowed to grow for two weeks and prepared for sequencing. In this example, blue cells contain the *wt* hTOPOIIα and the orange cells contain the edited M762Q hTOPOIIα, both strains contain the *wt* hTOPOIIβ. B) Log-transformed fraction of sequencing reads that contain the CRISPR-edited allele of hTOPOIIα is plotted on the y axis for pre-treatment and post-treatment cell populations. The fraction of cells that contain the hTOPOIIα glutamine allele increase by 168.2-, 8.2-, 2.8-, and 1.6-fold upon treatment with etoposide, teniposide, XK469, and amsacrine, respectively, when compared to the no-drug control. All four of these enrichments of the edited glutamine allele were significant by Fisher's exact test (etoposide and teniposide, *p*-value < 2.2E-16; amsacrine *p*-value = 0.025; XK469 *p*-value < 1.67E-13). The fraction of cells that contain the hTOPOIIα glutamine allele did not significantly increase in dactinomycin treated cells (Fisher's exact test, *p*-value = 0.7182).

Our *C. elegans* results using the genome-edited *top-2* strains show that variation at this residue underlies differences in some but not all topoisomerase II poisons. We exposed the edited human cell lines to these different poisons to test this hypothesis. We found that cells containing the glutamine-edited hTOPOIIα allele are less affected by both teniposide and XK469, as indicated by the respective 8.2- and 2.8-fold increase in edited cell frequency upon drug exposure as compared to cell populations with no drug added. These results mirror our observations that *C. elegans* strains with the methionine TOP-2 allele are more sensitive to these drugs. We observed moderate-to-no change in edited allele frequencies between cell populations exposed to amsacrine (~1.6-fold increase) or dactinomycin (~0.93-fold decrease) and those cells exposed to no-drug control conditions. These results indicate that the hTOPOIIα M762 residue does not interact with amsacrine. Our expectation was that dactinomycin would be more cytotoxic to cells that contain the glutamine hTOPOIIα allele, and therefore result in a depletion of edited cells. As mentioned above, we do not have the power to detect this depletion in response because of low CRISPR-editing efficiency. However, we did expect to see enrichment of hTOPOIIβ Q778M edited cells upon exposure to dactinomycin in the reciprocal experiment. Despite this expectation, we saw no change in hTOPOIIβ Q778M edited cell frequency between dactinomycin and the no drug control. Our inability to detect any enrichment of allele frequencies might result from dactinomycin having multiple cellular targets in addition to topoisomerase II [161]. Taken together, our findings testing a variety of poisons on human cell lines recapitulated the results from *C. elegans*.

Discussion

Few genetic markers have been identified that predict patient responses to chemotherapeutic regimens [162,163]. The goal of this study was to introduce new methods for the rapid and cost-effective identification of genetic variants that explain differences in chemotherapeutic response.

Our approach leveraged genetic and phenotypic variation present in the model organism *C. elegans* to identify a single amino acid variant (Q797M) in the topoisomerase II enzyme that underlies differences in etoposide response. Mechanistic insights into differential etoposide binding between the glutamine or methionine alleles gave us the power to predict the physiological responses to an expanded panel of topoisomerase II poisons. These results highlight how the combination of a highly sensitive phenotyping assay with classical and quantitative genetics approaches in *C. elegans* can rapidly identify the mechanistic underpinnings of phenotypic variability in response to a key class of antineoplastic compounds.

Our approach stands in stark contrast to previous underpowered human cell line [164] and clinical studies [8] that failed to identify any statistically significant associations between etoposide-induced cytotoxicity and genetic variation in the human population. However, the residue we identified in *C. elegans* does not vary in the human population [165], suggesting that GWA studies would not have identified this variant as a marker for etoposide sensitivity. Nevertheless, this residue is one of the few differences between the putative drug-binding pockets of the two human topoisomerase II isoforms (M762 in hTOPIIα and Q778 in hTOPIIβ), which allowed us to investigate the molecular underpinnings of drug binding. We verified that this single amino acid change in the human topoisomerase II isoforms results in profound differences in topoisomerase II poison-induced cytotoxicity using 293T cells. Though previous hTOPOIIβ structural studies have implicated this glutamine-methionine difference as functionally important for etoposide binding [143,144], studies involving hTOPOIIα have argued that this residue is not involved [145]. The results presented here unequivocally show that this residue contributes to differential topoisomerase II poison-induced cytotoxicity and have important implications for targeted drug design.

Although topoisomerase II poisons can bind and inhibit both hTOPOIIα and hTOPOIIβ, hTOPOIIα is the cellular target of poisons in most cancers because it is expressed in

proliferating cells [139]. However, recent evidence suggests that side effects associated with these treatments are caused by inhibition of hTOPOIIβ in differentiated cells [166]. For example, antineoplastic treatment regimens that contain the epipodophyllotoxins (*e.g.* etoposide or teniposide) are hypothesized to increase the risk of developing secondary malignancies caused by hTOPOIIβ-dependent 11q23 translocations [167–170]. Additionally, the most severe side effects associated with treatments that contain an alternative class of topoisomerase II poisons (anthracyclines, *e.g.* doxorubicin or daunorubicin) include dose-dependent cardiotoxicity and heart failure dependent on hTOPOIIβ [171–173]. Therefore, optimal topoisomerase II poisons will maximize interactions with hTOPOIIα to inhibit proliferating cells and minimize hTOPOIIβ interactions to reduce side effects. With this goal in mind, others have identified etoposide analogues with different isoform specificities but have not determined the mechanism of specificity [174]. Our study functionally validates a key residue determining isoform specificity and is critical to the improvement of this widely administered drug class. The importance of such functional characterization is underscored by our observation that *C. elegans* strains and human cells with the methionine TOP-2 allele are more sensitive to XK469, despite this drug being shown to be a β-specific poison [159]. Though XK469 has been shown to be a β-specific poison, no information regarding its drug binding pocket or the mechanism driving isoform specificity is currently known. Our results indicate that XK469 occupies a similar drug-binding pocket of TOP-2 as other topoisomerase II poisons and interacts with residue 797.

To date, no human genetic variants have been linked to topoisomerase II poison-induced cytotoxicity. Of the 291 and 279 respective SNVs in hTOPOIIα and hTOPOIIβ that encode for missense mutations, some are near the highly conserved DNA-binding domains or drug-binding pockets, which could affect drug response. However, the extent to which these variants impact responses to topoisomerase II poisons is unknown, so functional validation is required. The approach of editing human cells and following allele frequencies via sequencing represents a

scalable method to assess the functional role of these variants and avoids single-cell cloning. Importantly, differences in responses to topoisomerase II poisons might not be affected by variation in the topoisomerase II isoforms but instead mediated by variation in cellular import, metabolism, or export. Pharmacogenomic data available for many antineoplastic compounds [175,176], in combination with human variation data [165], can be used to prioritize and test variants in highly conserved regions of proteins known to be involved in these alternative processes. This biased approach focused on candidate variants is necessitated by the lack of power in clinical GWA studies and is not guaranteed to successfully connect variants to differences in drug response. For this reason, unbiased mapping approaches in model organisms combined with functional validation in genome-edited human cells will greatly expand our current understanding of how human genetic variation affects drug responses.

Future directions

The results described in this chapter the characterize of a single QTL that explains variation in *C. elegans* responses to etoposide and other topoisomerase II poisons. However, we identified multiple QTL that explain phenotypic variation to this class of cancer chemotherapeutic compounds. Therefore, a straightforward follow-up line of research would be to identify the causal genetic variants that underlie these QTL. To accomplish this, a similar line of experiments as described in this chapter can be used to narrow the QTL to a causal genetic variant or variants.

This research project was largely a proof-of-principle that we can leverage natural genetic variation within the *C. elegans* species to discover new biology related to cancer chemotherapeutics. However, most cancer chemotherapeutics are often administered as combination treatments. For example, etoposide and cisplatin are administered as a

combination therapy (called EP) to treat testicular cancers and thymomas. Despite being approved for 50 years, no genetic markers have been identified that contribute to variability in tumor remission and side effects associated with etoposide treatment among patients. Similarly, cisplatin has been FDA-approved for 38 years and has a single genetic test for variants present in thiopurine S-methyltransferase (TPMT) gene. However, the association between variation in TPMT and ototoxicity in the human population is not consistent, which suggests that other genetic or environmental factors play a role [128]. The lack of genetic markers associated with therapeutic interventions is compounded when trying to identify variants that affect responses to combination therapies like EP. As a result, only very few genetic variants have been identified that contribute to combination therapy, typically at the severe detriment to patients [177].

Given the limited genetic data available to predict combination cancer chemotherapeutic responses, performing combination treatment phenotyping assays in *C. elegans* is a logical next step for this project. To set this experiment up, a two dimensional dose response assay will be a first step to identify the proper dose for a large-scale experiment. This dose response assay can be performed using 12 divergent strains to identify individuals with the greatest phenotypic difference. Once an optimal combination dose is identified that maximizes broad-sense heritability, a large-scale experiment can be performed to identify QTL. This can be achieved by generating large pools of recombinant individuals that are a mixture of two parental genomes with differential sensitivity to the combined treatment. If no selective pressure is applied to this recombinant population, we would expect each parental genotype to be represented in ~50% of the recombinant lines at all positions of the genome. However if a selective pressure is applied to the recombinant population, the genomic regions that contain variants that affect response to the selective pressure will be enriched (diverged from the 50% expectation) for the parental genotype that performs better in the presence of the selective pressure. I will therefore subject recombinant populations to the selective pressure of the combined treatment treatment and

select the subset of individuals that are the least affected and most affected by the combined treatment. Genomic regions that contribute to the combined treatment response will be enriched or depleted based on the selected recombinant population, as compared to recombinants that have not undergone the combined treatment selection. In addition, individual drug treatments will have to be performed simultaneously to disentangle the effect of an individual drug QTL and combined treatment QTL. This approach is referred to bulk-segregant analysis (BSA) and has previously been validated in *C. elegans* [178]. Identifying the causal variants underlying each QTL can follow the same procedure as I described in this chapter. The BSA approach has the advantage of generating a large pool of recombinants that can be tested in a variety of conditions simultaneously.

One interesting observation from this study was the variant we identified to contribute to differential topoisomerase II poison responses in *C. elegans* is the same difference that exists between the two topoisomerase II isoforms in humans, which is likely because the DNA binding pocket of topoisomerase II enzymes is highly conserved. However, this also suggests that highly conserved enzymatic domains can tolerate certain mutations that drastically alter response to therapeutic interventions. Therefore, it will be interesting to explore the mutational landscape of such regions across species to identify residues that can tolerate mutation. This approach has the potential to identify target mutations that might arise in cancer cells that modify therapeutic interventions. Such a targeted approach can greatly reduce the search space for variants that modulate therapeutic interventions.

Contributions

The authors would like to thank Samuel Rosenberg for assistance on early mappings of drug sensitivities, Rama Mishra of the Center for Molecular Innovation and Drug Discovery core for

molecular dynamic simulations, Mudra Hegde of the Broad Institute for assistance with sequence analysis, and members of the Andersen laboratory for critical reading of this manuscript, and Joshua Bloom for editorial comments. Conceptualization: Stefan Zdraljevic, Christine Strand, John G. Doench, Erik C. Andersen. Data curation: Stefan Zdraljevic, Daniel E. Cook, John G. Doench, Erik C. Andersen. Formal analysis: Stefan Zdraljevic, John G. Doench, Erik C. Andersen. Funding acquisition: Erik C. Andersen. Investigation: Stefan Zdraljevic, Christine Strand, John G. Doench, Erik C. Andersen. Methodology: Stefan Zdraljevic, Christine Strand, Erik C. Andersen. Project administration: Erik C. Andersen. Resources: Stefan Zdraljevic, Hannah S. Seidel, Daniel E. Cook, Erik C. Andersen. Software: Stefan Zdraljevic, Daniel E. Cook, Erik C. Andersen. Supervision: Erik C. Andersen. Validation: Stefan Zdraljevic, Erik C. Andersen. Visualization: Stefan Zdraljevic, Erik C. Andersen. Writing – original draft: Stefan Zdraljevic, Christine Strand, Hannah S. Seidel, John G. Doench, Erik C. Andersen. Writing – review & editing: Stefan Zdraljevic, John G. Doench, Erik C. Andersen.

4. Natural variation in *C. elegans* arsenic toxicity is explained by differences in branched chain amino acid metabolism

Preface

The second project of my Ph.D. focused on a QTL I identified during my rotation project. This QTL explains variation in response to arsenic trioxide, which is a pervasive environmental toxin. As this project was a secondary focus of my graduate work, it took longer to complete. However, with the help of my lab, our collaborators at the Broad Institute, and at Cornell, I was able to complete this project and discover a new mechanism of arsenic toxicity. This work was published in *eLife* in 2019 [32].

Abstract

We find that variation in the *dbt-1* gene underlies natural differences in *Caenorhabditis elegans* responses to the toxin arsenic. This gene encodes the E2 subunit of the branched-chain α-keto acid dehydrogenase (BCKDH) complex, a core component of branched-chain amino acid (BCAA) metabolism. We causally linked a non-synonymous variant in the conserved lipoyl domain of DBT-1 to differential arsenic responses. Using targeted metabolomics and chemical supplementation, we demonstrate that differences in responses to arsenic are caused by variation in iso-branched chain fatty acids. Additionally, we show that levels of branched chain fatty acids in human cells are perturbed by arsenic treatment. This finding has broad implications for arsenic toxicity and for arsenic-focused chemotherapeutics across human populations. Our study implicates the BCKDH complex and BCAA metabolism in arsenic responses, demonstrating the power of *C. elegans* natural genetic diversity to identify novel mechanisms by which environmental toxins affect organismal physiology.

Introduction

An estimated 100 million people are currently at risk of chronic exposure to arsenic, a toxic metalloid that can be found in the environment [179]. The high prevalence of environmental arsenic and the severe toxicity associated with exposure has made it the number one priority for the United States Agency for Toxic Substances and Disease Registry (https://www.atsdr.cdc.gov/SPL/). Inorganic trivalent arsenic As(III) compounds, which include arsenic trioxide (As_2O_3), are the most toxic forms of environmental arsenic [180,181]. In humans, As(III) is detoxified by consecutive methylation events, forming dimethylarsenite (DMA) [182,183]. However, this methylation process also creates the highly toxic monomethylarsenite (MMA) intermediate, so ratios of DMA to MMA determine levels of arsenic toxicity. Both MMA and DMA are produced from As(III) via the arsenic methyltransferase (AS3MT) [184]. Interestingly, individuals from human subpopulations that inhabit high arsenic environments have higher DMA/MMA ratios than individuals from low-arsenic environments. The elevated DMA/MMA ratio in these individuals is associated with natural differences in the *AS3MT* gene [185–187], which shows signs of strong positive selection. These results suggest that a more active AS3MT enzyme in these human subpopulations makes more DMA and enables adaptation to elevated environmental arsenic levels [184]. Importantly, population-wide differences in responses to environmental arsenic cannot be explained solely by variation in *AS3MT*, indicating that other genes must impact arsenic toxicity.

Despite its toxicity, arsenic trioxide has been used as a therapeutic agent for hundreds of years. Most recently, it was introduced as a highly effective cancer chemotherapeutic for the treatment of acute promyelocytic leukemia (APL) [188–191]. Hematopoietic differentiation and apoptosis in APL patients is blocked at the level of promyelocytes by the Promyelocytic Leukemia/Retinoic Acid Receptor alpha fusion protein caused by a t(15;17) chromosomal translocation [192,193].

Arsenic trioxide has been shown to directly bind a cysteine-rich region of the RING-B box coiled-coil domain of PML-RARα, which causes the degradation of the oncogenic fusion protein [194,195]. The success of arsenic trioxide (Trisenox®) has spurred its use in over a hundred clinical trials in the past decade [196]. Despite these successes, individual differences in the responses to arsenic-based treatments, including patient-specific dosing regimens and side effects, limit the full therapeutic benefit of this compound [197]. Medical practitioners require knowledge of the molecular mechanisms for how arsenic causes toxicity to provide the best individual-based therapeutic benefits.

Studies of the free-living roundworm *Caenorhabditis elegans* have greatly facilitated our understanding of basic cellular processes [198–201], including a number of studies that show that the effects of arsenic are similar to what is observed in mammalian model systems and humans. These effects include mitochondrial toxicity [202,203], the generation of reactive oxygen species (ROS) [204], genotoxicity [205], genome-wide shifts in chromatin structure [206], reduced lifespan [204], and the induction of the heat-shock response [207]. However, these studies were all performed in the genetic background of the standard *C. elegans* laboratory strain (N2). To date, 152 *C. elegans* strains have been isolated from various locations around the world [27,43,44], which contain a largely unexplored pool of genetic diversity much of which could underlie adaptive responses to environmental perturbations [208].

We used two quantitative genetic mapping approaches to show that a major source of variation in *C. elegans* responses to arsenic trioxide is caused by natural variation in the *dbt-1* gene, which encodes an essential component of the highly conserved branched-chain α-keto acid dehydrogenase (BCKDH) complex [209]. The BCKDH complex is a core component of branched-chain amino acid (BCAA) catabolism, which has not been previously implicated in arsenic responses. Furthermore, we show that a single missense variant in DBT-1(S78C),

located in the highly conserved lipoyl-binding domain, underlies phenotypic variation in response to arsenic. Using targeted and untargeted metabolomics and chemical rescue experiments, we show that differences in wild isolate responses to arsenic trioxide are caused by differential synthesis of mono-methyl branched chain fatty acids (mmBCFA), metabolites with a central role in development [198]. These results demonstrate the power of using the natural genetic diversity across the *C. elegans* species to identify mechanisms by which environmental toxins affect physiology.

Materials and Methods

Strains

Animals were fed the bacterial strain OP50 and grown at 20°C on modified nematode growth medium (NGM), containing 1% agar and 0.7% agarose to prevent burrowing of the wild isolates [55]. For each assay, strains were grown for five generations with no strain entering starvation or encountering dauer-inducing conditions [146]. Wild *C. elegans* isolates used for genome-wide association and recombinant inbred advanced intercross lines (RIAILs) used for linkage mapping have been described previously [27,44,54]. Strains constructed for this manuscript are listed above in **Table 3-1**. Oligonucleotides used to construct strains are shown **in Table 3-2**.

Table 4-1 Strains used for experiments discussed in Chapter 4

Strain Name	Genetic background	Genotype	Description	Experiments
ECA581	N2	N2::dbt-1(ean15[C78S])	dbt-1 allele swap strain in the N2 genetic background	Allele swap, Rescue
ECA590	CB4856	CB4856::dbt-1(ean34[S78C])	dbt-1 allele swap strain in the CB4856 genetic background	Allele swap, Rescue
ECA414	N2	eanIR188[ChrII:5.75 - 8.02Mb]	left interval NIL, CB4856 introgressed into the N2 genetic background	NIL
ECA434	N2	eanIR208[ChrII:7.83 - 9.66Mb]	right interval NIL, CB4856 introgressed into the N2 genetic background	NIL

Table 4-2 Oligonucleotides used for experiments discussed in Chapter 4

Name	Sequence	Use
oECA609	tttcacacaaaccatgcgct	generate NIL from RIAIL QX103 - Forward primer, primes at 6765038 bp

oECA610	actcgtctgctgggtattct	generate NIL from RIAIL QX103 - Reverse primer, primes at 6765480 bp
oECA611	tgtcttcgcacctttactcg	generate NIL from RIAIL QX103 - Forward primer, primes at 8737946 bp
oECA612	cattcaagtcagggcgtatcc	generate NIL from RIAIL QX103 - Reverse primer, primes at 8738299 bp
oECA1163	GAAGGAATTGCCGAAGTT CAGGTTAAG	amplify region surrounding C78S variant
oECA1165	CCGTCATCTCCACAAAAA GCTTTATCTCTC	amplify region surrounding C78S variant
dbt-1 gRNA	CCATCTCCTGTAGATACG AC	target Cas9 to generate allele swap strains
N2 Repair	CTTCCAGGTACGTGAAAG AAGGAGATACGATTTCGC AGTTCGATAAAGTCTGTGA AGTGCAAAGTGATAAAGC AGCAGTAACCATCTCCAG TAGATACGACGGAATTGT CAAAAAATTGTAAGTTTCT TCCTAA	REPAIR TEMPLATE - MAKE SENSITIVE N2 into RESISTANT CB4856
CB4856 Repair	TTAGGAAGAAACTTACAAT TTTTTGACAATTCCGTCGT ATCTACAGGAGATGGTTA CTGCTGCTTTATCGCTTTG CACTTCACAGACTTTATCG AACTGCGAAATCGTATCTC CTTCTTTCACGTACCTGGA AG	REPAIR TEMPLATE - MAKE Resistant CB4856 to sensitive N2
dpy-10 repair	CACTTGAACTTCAATACGG CAAGATGAGAATGACTGG AAACCGTACCGCATGCGG TGCCTATGGTAGCGGAGC TT CACATGGCTTCAGACCAA CAGCCTAT	Repair template to generate dpy-10(cn64) mutation

High-throughput arsenic-response assay

We used the high-throughput fitness assay (HTA) described previously [54]. In short, strains are passaged for four generations before bleach-synchronization and aliquoted to 96-well microtiter plates at approximately one embryo per microliter in K medium [55]. The final concentration of NaCl in the K medium for the genome-wide association (GWA) and linkage mapping assays was 51 mM. For all subsequent experiments the final NaCl concentration was 10.2 mM. The following day, hatched and synchronized L1 animals were fed HB101 bacterial lysate (Pennsylvania State University Shared Fermentation Facility, State College, PA, [210]) at a final concentration of 5 mg/ml and grown for two days at 20ºC. Next, three L4 larvae were sorted using a large-particle flow cytometer (COPAS BIOSORT, Union Biometrica, Holliston, MA) into microtiter plates that contain HB101 lysate at 10 mg/ml, K medium, 31.25 μM kanamycin, and

either arsenic trioxide dissolved in 1% water or 1% water alone. The animals were then grown for four days at 20ºC. For linkage mapping and GWA mapping experiments, we added polychromatic fluorescent beads (Polysciences, cat. #19507-5) to each well for five minutes. The populations were treated with sodium azide (50 mM) prior to being measured with the BIOSORT. To reduce experimental costs, the polychromatic fluorescent beads were not added to follow-up experiments. For all experiments, we report the results for four independently quantified traits): the normalized brood size (norm.n), mean progeny length per well (mean.TOF), the mean optical density normalized by animal length per well (mean.norm.EXT), and the mean fluorescence normalized by animal length per well (mean.norm.yellow). All raw experimental data can be found on FigShare (https://doi.org/10.6084/m9.figshare.7458980.v2).

Arsenic-response trait calculations

Phenotype data generated using the BIOSORT were processed using the R package *easysorter*, which was specifically developed for processing this type of data [147]. Briefly, the function *read_data*, reads in raw phenotype data and runs a support vector machine to identify and eliminate bubbles. Next, the *remove_contamination* function eliminates any wells that were identified as contaminated prior to scoring population parameters. This analysis results in processed BIOSORT data where each observation is for a given strain corresponds to the measurements for an individual animal. However, the phenotypes we used for mapping and follow-up experiments are summarized statistics of populations of animals in each well of a 96-well plate. The *sumplate* function was used to generate summary statistics of the measured parameters for each animal in each well. These summary statistics include the 10th, 25th, 50th, 75th, and 90th quantiles for time of flight (TOF), animal extinction (EXT), and three fluorescence channels (Green, Yellow, and Red), which correspond to animal length, optical density, and ability to pump fluorescent beads, respectively. Measured brood sizes (n) are normalized by the

number of animals that were originally sorted into each well (norm.n). For mapping experiments, a single well replicate for each strain is summarized using the *sumplate* function. For follow-up experiments, multiple replicates for each strain indicated by a unique plate, well, and column were used. After summary statistics for each well are calculated, we accounted for differences between assays using the *regress(assay=TRUE)* function in the *easysorter* package. Outliers in the GWA and linkage mapping experiments were identified and eliminated using the *bamf_prune* function in *easysorter*. For follow-up experiments that contained multiple replicates for each strain, we eliminated strain replicates that were more than two standard deviations from the strain mean for each condition tested. Finally, arsenic-specific effects were calculated using the *regress(assay=FALSE)* function from *easysorter*, which accounts for strain-specific differences in growth parameters present in control conditions.

Principal component analysis of processed BIOSORT measured traits

The COPAS BIOSORT measures individual animal length (TOF), optical density (EXT), fluorescence (green, yellow, and red). We use these data to calculate the total number of animals in a well and then normalize by the number of animals initially sorted into the well (brood size). All of these measurements were then summarized using the *easysorter* package to generate various summary statistics of each measured parameter, including five distribution quantiles and measures of dispersion [54]. We used four independently quantified traits as inputs to principal component analysis (PCA): the normalized brood size (norm.n), mean progeny length per well (mean.TOF), the mean optical density normalized by animal length per well (mean.norm.EXT), and the mean fluorescence normalized by animal length per well (mean.norm.yellow). Though we only used fluorescent beads in the GWA and linkage mapping experiments, we found that fluorescence-based traits exhibited an arsenic-specific effect that correlated with strain sickness. Prior to principal component analysis (PCA), HTA phenotypes were scaled to have a mean of zero and a standard deviation of one using the *scale* function in

R. PCA was performed using the *prcomp* function in R [211]. Eigenvectors were subsequently extracted from the object returned by the *prcomp* function.

Arsenic dose-response assays

All dose-response experiments were performed on four genetically diverged strains (N2, CB4856, DL238, and JU775) in technical quadruplicates prior to performing GWA and linkage mapping experiments. Animals were assayed using the HTA, and phenotypic analyses were performed as described above. The arsenic trioxide concentration for GWA and linkage mapping experiments was chosen based on an observable effect for animal length and brood size phenotypes in the presence of arsenic.

Linkage mapping

A total of 262 RIAILs were phenotyped in the HTA described previously for control and arsenic trioxide conditions [31,54]. The phenotype and genotype data were entered into R and scaled to have a mean of zero and a variance of one for linkage analysis. Quantitative trait loci (QTL) were detected by calculating 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 phenotype trait values [134]. The maximum LOD score for each chromosome for each trait was retained from three iterations of linkage mappings. We randomly permuted the phenotype values of each RIAIL while maintaining correlation structure among phenotypes 1000 times to estimate the significance threshold empirically. The significance threshold was set using a genome-wide error rate of 5%. Confidence intervals were defined as the regions contained within a 1.5 LOD drop from the maximum LOD score [212].

Principal component analysis of RIAILs

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 zero and a standard deviation of one. The *princomp* function within the *stats* package in R [91] was used to run a PCA 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, totaling 97 PCs across all toxins. We additionally performed a two-dimensional genome scan using the *scantwo* function in the *qtl* package [212] for all 47 significantly mapped PCs. Significant interactions were determined by permuting the phenotype data for each PC 1,000 times and determining the 5% genome-wide error rate.

Heritability estimates

For dose response experiments, broad-sense heritability (H^2) estimates were calculated using the *lmer* function in the lme4 package with the following linear mixed model (phenotype ~1 + (1|strain)) [213]. $H^2$was then calculated as the fraction of the total variance that can be explained by the random component (strain) of the mixed model. Prior to estimating H^2, we removed outlier replicates that we defined as replicates with values greater than two standard deviations away from the mean phenotype. Outliers were defined on a per-trait and per-strain basis.

Heritability estimates for the linkage mapping experiment were calculated using two approaches. In both approaches, we used the previously described RIAIL genotype matrix to compute relatedness matrices [54]. In the first approach, a variance component model using the R package regress was used to estimate the fraction of phenotypic variation explained by

additive and epistatic genetic factors, H^2, or just additive genetic factors, h^2 [214,215], using the formula (y ~ 1, ~ZA+ZAA), where y is a vector of RIAIL phenotypes, ZA is the additive relatedness matrix, and ZAA is the pairwise-interaction relatedness matrix. The additive relatedness matrix was calculated as the correlation of marker genotypes between each pair of strains. In addition, a two-component variance 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.

The second approach utilized a linear mixed model and the realized additive and epistatic relatedness matrices [149,216–218]. We used the *mmer* function in the sommer package with the formula (y ~ A+E) to estimate variance components, where y is a vector of RIAIL phenotypes, A is the realized additive relatedness matrix, and E is the epistatic relatedness matrix. This same approach was used to estimate heritability for the GWA mapping phenotype data, with the only difference being that we used the wild isolate genotype matrix described below. Heritability estimates for RIAIL and wild isolate data are in.

Effect size calculations for dose response assay

We first fit a linear model with the formula (phenotype ~ strain) for all measured and principal component traits for each concentration of arsenic trioxide using the *lm* R function. Next, we extracted effect sizes using the *anova_stats* function from the sjstats R package [219].

Generation of NILs

NILs were generated by crossing N2xCB4856 RIAILs to each parental genotype. For each NIL, eight crosses were performed followed by six generations of selfing to homozygose the genome. Reagents used to generate NILs are detailed in the Key Resources Table. The NILs

responses to 1000 μM arsenic trioxide were quantified using the HTA described above. NIL whole-genome sequencing and analysis was performed as described previously [220].

Genome-wide association mapping

Genome-wide association (GWA) mapping was performed using phenotype data from 86 *C. elegans* isotypes. Genotype data were acquired from the latest VCF release (Release 20180527) from CeNDR that was imputed as described previously [27]. We used BCFtools [87] to filter variants that had any missing genotype calls and variants that were below 5% minor allele frequency. We used PLINK v1.9 [94,95] to LD-prune the genotypes at a threshold of $r^2 < 0.8$, using *--indep-pairwise 50 10 0.8.* This resulting genotype set consisted of 59,241 markers that were used to generate the realized additive kinship matrix using the *A.mat* function in the *rrBLUP* R package [149]. These markers were also used for genome-wide mapping. However, because these markers still have substantial LD within this genotype set, we performed eigen decomposition of the correlation matrix of the genotype matrix using *eigs_sym* function in Rspectra package [221]. The correlation matrix was generated using the *cor* function in the correlateR R package [222]. We set any eigenvalue greater than one from this analysis to one and summed all of the resulting eigenvalues [223]. This number was 500.761, which corresponds to the number of independent tests within the genotype matrix. We used the *GWAS* function in the rrBLUP package to perform genome-wide mapping with the following command: *rrBLUP::GWAS(pheno = PC1, geno = Pruned_Markers, K = KINSHIP, min.MAF = 0.05, n.core = 1, P3D = FALSE, plot = FALSE)*. To perform fine-mapping, we defined confidence intervals from the genome-wide mapping as +/- 100 SNVs from the rightmost and leftmost markers above the Bonferroni significance threshold. We then generated a QTL region of interest genotype matrix that was filtered as described above, with the one exception that we did not perform LD pruning. We used PLINK v1.9 to extract the LD between the markers used for fine mapping and the peak QTL marker identified from the genome-wide scan. We used the

same command as above to perform fine mapping, but with the reduced variant set. The workflow for performing GWA mapping can be found on https://github.com/AndersenLab/cegwas2-nf. All trait mapping results can be found on FigShare (https://doi.org/10.6084/m9.figshare.7828706.v1).

Generation of *dbt-1* allele replacement strains

Allele replacement strains were generated using CRISPR-Cas9-mediated genome editing, using the co-CRISPR approach [150] with Cas9 ribonucleoprotein delivery [69]. Alt-R™ crRNA and tracrRNA were purchased from IDT (Skokie, IL). tracrRNA (IDT, 1072532) was injected at a concentration of 13.6 µM. The *dpy-10* and the *dbt-1* crRNAs were injected at 4 µM and 9.6 µM, respectively. The *dpy-10* and the *dbt-1* single-stranded oligodeoxynucleotides (ssODN) repair templates were injected at 1.34 µM and 4 µM, respectively. Cas9 protein (IDT, 1074182) was injected at 23 µM. To generate injection mixes, the tracrRNA and crRNAs were incubated at 95ºC for five minutes and 10ºC for 10 minutes. Next, Cas9 protein was added and incubated for five minutes at room temperature. Finally, repair templates and nuclease-free water were added to the mixtures and loaded into pulled injection needles (1B100F-4, World Precision Instruments, Sarasota, FL). Individual injected P_0 animals were transferred to new 6 cm NGM plates approximately 18 hours after injections. Individual F_1 rollers were then transferred to new 6 cm plates to generate self-progeny. The region surrounding the desired S78C (or C78S) edit was then amplified from F_1 rollers using primers oECA1163 and oECA1165. The PCR products were digested using the *Sfc*I restriction enzyme (R0561S, New England Biolabs, Ipswich, MA). Differential band patterns signified successfully edited strains because the N2 S78C, which is encoded by the CAG codon, creates an additional *Sfc*I cut site. Non-Dpy, non-Rol progeny from homozygous edited F_1 animals were propagated. If no homozygous edits were obtained, heterozygous F_1 progeny were propagated and screened for the presence of the homozygous

edits. F_1 and F_2 progeny were then Sanger sequenced to verify the presence of the proper edited sequence. The phenotypes of allele replacement strains in control and arsenic trioxide conditions were measured using the HTA described above.

Rescue with 13-methyltetradeconoic acid

Strains were grown as described for a standard HTA experiment. In addition to adding arsenic trioixde to experimental wells, we also added a range of C15iso (13-methyltetradecanoic acid, Matreya Catalog # 1605) concentrations to assay rescue of arsenic effects.

Growth conditions for metabolite profiling

For L1 larval stage assays, chunks (~1 cm) were taken from starved plates and placed on multiple fresh 10 cm plates. Prior to starvation, animals were washed off of the plates using M9, and embryos were prepared by bleach synchronization. Approximately 40,000 embryos were resuspended in 25 ml of K medium and allowed to hatch overnight at 20ºC. L1 larvae were fed 15 mg/ml of HB101 lysate the following morning and allowed to grow at 20ºC for 72 hours. We harvested 100,000 embryos from gravid adults by bleaching. These embryos were hatched overnight in 50 ml of K medium in a 125 ml flask. The following day, we added arsenic trioxide to a final concentration of 100 µM and incubated the cultures for 24 hours. After 24 hours, we added HB101 bacterial lysate (2 mg/ml) to each culture. Finally, we transferred the cultures to 50 ml conical tubes, centrifuged the cultures at 3000 RPM for three minutes to separate the pellet and supernatant. The supernatant and pellets from the cultures were frozen at -80ºC and prepared for analysis. For young adult stage assays, 45,000 animals per culture were prepared as described above but in S medium, at a density of three animals per microliter, and fed HB101 lysate (5 mg/mL). These cultures were shaken at 200 RPM, 20ºC in 50 mL Erlenmeyer flasks for 62 h. For harvesting, we settled 15 mL of cultures for 15 minutes at room temperature and

then pipetted the top 12 mL of solution off of the culture. The remaining 3 mL of animal pellet was washed with 10 mL of M9, centrifuged at 1000 g for one minute, and then the supernatant removed. This wash was repeated once more with M9 and again with water. The final nematode pellet was snap frozen in liquid nitrogen.

Nematode metabolite extractions

Pellets were lyophilized 18-24 hours using a VirTis BenchTop 4K Freeze Dryer until a chalky consistency was achieved. Dried pellets were transferred to 1.5 mL microfuge tubes and dry pellet weight recorded. Pellets were disrupted in a Spex 1600 MiniG tissue grinder after the addition of three stainless steel grinding balls to each sample. Microfuge tubes were placed in a Cryoblock (Model 1660) cooled in liquid nitrogen, and samples were disrupted at 1100 RPM for two cycles of 30 seconds. Each sample was individually dragged across a microfuge tube rack eight times, inverted, and flicked five times to prevent clumping. This process was repeated two additional rounds for a total of six disruptions. Pellets were transferred to 4 mL glass vials in 3 mL 100% ethanol. Samples were sonicated for 20 minutes (on/off pulse cycles of two seconds at power 90 A) using a Qsonica Ultrasonic Processor (Model Q700) with a water bath cup horn adaptor (Model 431C2). Following sonication, glass vials were centrifuged at 2750 RCF for five minutes in an Eppendorf 5702 Centrifuge using rotor F-35-30-17. The resulting supernatant was transferred to a clean 4 mL glass vial and concentrated to dryness in an SC250EXP Speedvac Concentrator coupled to an RVT5105 Refrigerated Vapor Trap (Thermo Scientific). The resulting powder was suspended in 100% ethanol according to its original dry pellet weight: 0.01 mL 100% ethanol per mg of material. The suspension was sonicated for 10 minutes (pulse cycles of two seconds on and three seconds off at power 90 A) followed by centrifugation at 20,817 RCF in a refrigerated Eppendorf centrifuge 5417R at 4°C. The resulting supernatant was transferred to an HPLC vial containing a Phenomenex insert (cat #AR0-4521-12) and centrifuged at 2750 RCF for five minutes in an Eppendorf 5702 centrifuge. The resulting

supernatant was transferred to a clean HPLC vial insert and stored at -20°C or analyzed immediately.

Mass spectrometric analysis

Reversed-phase chromatography was performed using a Dionex Ultimate 3000 Series LC system (HPG-3400 RS High Pressure pump, TCC-3000RS column compartment, WPS-3000TRS autosampler, DAD-3000 Diode Array Detector) controlled by Chromeleon Software (ThermoFisher Scientific) and coupled to an Orbitrap Q-Exactive mass spectrometer controlled by Xcalibur software (ThermoFisher Scientific). Metabolites were separated on a Kinetex EVO C18 column, 150 mm x 2.1 mm, particle size 1.7 µm, maintained at 40°C with a flow rate of 0.5 mL/min. Solvent A: 0.1% ammonium acetate in water; solvent B: acetonitrile (ACN). A/B gradient started at 5% B for 30 seconds, followed by a linear gradient to 95% B over 13.5 minutes, then a linear gradient to 100% B over three minutes. 100% B was maintained for one minute. Column was washed after each run with 5:1 isopropanol:ACN, flow rate of 0.12 mL/min for five minutes, followed by 100% ACN for 2.9 minutes, a linear gradient to 95:5 water:ACN over 0.1 minutes, and then 95:5 water:ACN for two minutes with a flow rate of 0.5 mL/min. A heated electrospray ionization source (HESI-II) was used for the ionization with the following mass spectrometer parameters: spray voltage: 3 kV; capillary temperature: 320°C; probe heater temperature: 300°C; sheath gas: 70 AU; auxiliary gas flow: 2 AU; resolution: 240,000 FWHM at m/z 200; AGC target: 5e6; maximum injection time: 300 ms. Each sample was analyzed in negative and positive modes with m/z range 200-800. Fatty acids and most ascarosides were detected as $[M-H]^-$ ions in negative ionization mode. Peaks of known abundant ascarosides and fatty acids were used to monitor mass accuracy, chromatographic peak shape, and instrument sensitivity for each sample [224].

Statistical analyses

All *p*-values testing the differences of strain phenotypes in the NIL, allele-replacement, and C15ISO experiments were performed in R using the *TukeyHSD* function with an ANOVA model with the formula (*phenotype* ~ *strain*). *p*-values of individual pairwise strain comparisons are reported in each figure legend.

CRISPR-Cas9 gene editing in human cells

The 293T cells were sourced from CCLE. Identity authenticated by SNP profiling. Cells were regularly tested for Mycoplasma (~bimonthly). Gene-editing experiments were performed in a single parallel culture experiment using human 293T cells (ATCC) grown in DMEM with 10% FBS. On day zero, 300,000 cells were seeded per well in a six-well plate format. The following day, two master mixes were prepared: a) LT-1 transfection reagent (Mirus) was diluted 1:10 in Opti-MEM and incubated for five minutes; b) a DNA mix of 500 ng Cas9-sgRNA plasmid with 250 pmol repair template oligonucleotide was diluted in Opti-MEM in a final volume of 100 µL. 250 µL of the lipid mix was added to each of the DNA mixes and incubated at room temperature for 25 minutes. Following incubation, the full 350 µL volume of DNA and lipid mix was added dropwise to the cells. These six-well plates were then centrifuged at 1000 x g for 30 minutes. After six hours, the media on the cells was replaced. For the next six days, cells were expanded and passaged as needed. On day seven, one million cells were taken from each set of edited and unedited cells and placed into separate T75s with either media-only or 5 µM arsenic-containing media. Days seven to fourteen, arsenic and media-only conditions were maintained at healthy cell densities. Days fourteen to eighteen, arsenic exposed cell populations were maintained off arsenic to allow the populations to recover prior to sequencing. Media-only conditions were maintained in parallel. On day eighteen, all arsenic and media-only conditions were pelleted for genomic DNA extraction.

Analysis of CRISPR-Cas9 editing in human cells

Genomic DNA was extracted from cell pellets using the QIAGEN (QIAGEN, Hilden, Germany) Midi or Mini Kits based on the size of the cell pellet (51183, 51104) according to the manufacturer's recommendations. DBT1 loci were first amplified with 17 cycles of PCR using a touchdown protocol and the NEBnext 2x master mix (New England Biolabs M0541). The resulting product served as input to a second PCR, using primers that appended a sample-specific barcode and the necessary adaptors for Illumina sequencing. The resulting DNA was pooled, purified with SPRI beads (A63880, Beckman Coulter, Brea, CA), and sequenced on an Illumina MiSeq with a 300-nucleotide single-end read with an eight nucleotide index read. For each sample, the number of reads exactly matching the wild-type and edited DBT1 sequence were determined.

Preparing human cells for mass spectroscopy

Mass spectroscopy experiments used human 293T cells (ATCC) grown in DMEM with 10% FBS. On day zero, 150,000 cells were seeded into 15 cm tissue cultures dishes with 15 mL of either 2.5 μM arsenic or no arsenic media. Each condition had five replicates. On day three, the no arsenic cells were approaching confluence and required passaging. Arsenic conditions were at ~30% confluence and received a media change. On day seven, both conditions were near confluence, media was removed, and plates were rinsed with ice cold PBS, remaining liquid removed. Plates were frozen at -80ºC before processing for mass spectrometric analysis. Cells were scraped off the plates with PBS and pelleted in microfuge tubes. Cell pellets were lyophilized 18-24 hours using a VirTis BenchTop 4K Freeze Dryer and extracted in 100% ethanol using the same sonication program as described for nematode extraction. Following sonication, samples were centrifuged at 20,817 RCF in a refrigerated Eppendorf centrifuge 5417R at 4 °C. Clarified supernatant was aliquoted to a new tube and concentrated to dryness

in an SC250EXP Speedvac Concentrator coupled to an RVT5105 Refrigerated Vapor Trap (Thermo Scientific). The resulting material was suspended in .1 mL 100% ethanol and analyzed by LC-MS as described. Metabolite measurements can be found in **Figure 5-source data 3**.

Tajima's D calculation

We used the VCF corresponding to CeNDR release 20160408 (https://elegansvariation.org/data/release/20160408) to calculate Tajima's D. Tajima's D was calculated using the *tajimas_d* function in the *cegwas* package using default parameters (window size = 500 SNVs, sliding window distance = 50 SNVs, outgroup = N2). Isolation locations of strains can be found in.

Results

Natural variation of chromosome II underlies differences in arsenic responses

We quantified arsenic trioxide sensitivity in *C. elegans* using a high-throughput fitness assay that utilizes the COPAS BIOSORT [31,54]. In this assay, three L4 larvae from each strain were sorted into arsenic trioxide or control conditions. After four days of growth, we quantified various attributes of populations that relate to the ability of *C. elegans* to grow in the presence of arsenic trioxide or control conditions (see Materials and Methods). To determine an appropriate concentration of arsenic trioxide for mapping experiments, we performed dose-response experiments on four genetically diverged isolates of *C. elegans*: N2, CB4856, JU775, and DL238. To assess arsenic-induced toxicity, we studied four independently measured traits: brood size, animal length, optical density, and fluorescence (see Materials and Methods). When compared to control conditions, all four strains produced fewer progeny at all arsenic trioxide

concentrations, and the lowest concentration at which we observed a significant reduction in brood size for all strains was 1 mM. We used statistical summaries of measurements of individual animals as replicated identical genotypes to estimate broad-sense heritability (H^2) across the four strains, but this analysis might not represent the effects of arsenic on individuals within natural populations. For the brood size trait in 1 mM arsenic trioxide, we calculated H^2 to be 0.65 and the strain effect to be 0.48 (partial omega squared, ${\omega_p}^2$), indicating that this trait has a large genetic component and a large strain effect. In addition to brood size effects, we observed that the progeny of animals exposed to arsenic trioxide were shorter in length than the progeny of animals grown in control conditions, which indicates an arsenic-induced developmental delay (animal length (mean.TOF) H^2= 0.13 and ${\omega_p}^2$= 0.09). As *C. elegans* develop, the animals increase in optical density. Therefore, it is not surprising that we found an arsenic-induced decrease in optical density of progeny populations, which is further support for an arsenic-induced developmental delay (H^2= 0.19; ${\omega_p}^2$= 0.07). We also observed an arsenic-induced effect on yellow autofluorescence (H^2= 0.50 and ${\omega_p}^2$= 0.21). Overall, the CB4856 strain was less affected by arsenic – that strain produced approximately 16% more offspring that were on average 20% larger than the other three strains when treated with 1 mM arsenic trioxide. These results suggest that the CB4856 strain was more resistant to arsenic trioxide than the other three strains. In addition to the BIOSORT-quantified traits, we generated a synthetic principal component (PC) trait using the four quantified traits described above (see Materials and Methods). For 1 mM arsenic trioxide, we estimated the broad-sense heritability (H^2) of the first principal component to be 0.37 with an effect size of 0.001 (${\omega_p}^2$). The first principal component explained a large fraction (0.75) of the total phenotypic variance within the experiment, which is likely because the four input traits correlate with arsenic-induced toxicity. We noted that the first principal component (PC1) was most strongly influenced by the optical density trait, as indicated by the loadings, suggesting that PC1 is a biologically relevant trait.

Furthermore, because we observe a large range of effect sizes and broad-sense heritability estimates across measured traits, we focused our analyses on the PC1 trait derived from the four BIOSORT-quantified described above for all subsequent experiments (Materials and Methods).

The increased arsenic trioxide resistance of CB4856 compared to N2 motivated us to perform linkage mapping experiments with a panel of recombinant inbred advanced intercross lines (RIAILs) that were previously constructed through ten generations of intercrossing between an N2 derivative (QX1430) and CB4856 [54]. To capture arsenic trioxide-induced phenotypic differences, we exposed a panel of 252 RIAILs to 1 mM arsenic trioxide and corrected for growth differences among RIAILs in control conditions and assay-to-assay variability using linear regression (see Materials and Methods). We performed linkage mapping on processed traits and the eigenvector-transformed traits (principal components or PCs) obtained from PCA that explained 90% of the variance in the processed trait set (Materials and Methods). The rationale of this approach was to minimize trait fluctuations that could be caused by only measuring the phenotypes of one replicate per RIAIL strain, and PC1 captured overall arsenic-induced toxicity. In agreement with our observations from the dose-response experiment, we found that PC1 captures 69.5% of the total measured trait variance and is strongly influenced by animal length traits. Linkage mapping analysis of the PC1 trait revealed that arsenic trioxide-induced phenotypic variation is significantly associated with genetic variation on the center of chromosome II (Figure 3-1A). An additional quantitative trait locus (QTL) was significantly associated with variation in arsenic responses on chromosome X (Figure 3-1A). Consistent with the loadings of PC1, we determined that PC1 is highly correlated with the both brood size and animal length traits (Figure 3-1B), suggesting that PC1 captures RIAIL variation in these traits. To further support this relationship to interpretable biological significance, we found that the four traits used as input for PCA all map to the same region on the center of chromosome II. The

QTL on the center of chromosome II explains 35.4% of the total RIAIL phenotypic variation for the PC1 trait, which accounts for 63.4% of the total phenotypic variation that can be explained by genetic factors (H^2= 0.56). Taken together, the two QTL identified by mapping the PC1 trait account for 40% of the total RIAIL variation, corresponding to 71.6% of the total phenotypic variation that can be explained by genetic factors. However, we did not account for errors in genomic heritability estimates. In addition to the two QTL that explain variation of the PC1 trait, we identified a QTL on chromosome I for the brood size and optical density traits, and a QTL on chromosome V that explained variation in animal length and optical density upon arsenic exposure. The PC1 QTL confidence interval spans from 7.04 to 8.87 Mb on chromosome II. This QTL overlaps with the brood size (6.18-9.31 Mb) and animal length (6.92-8.70 Mb) QTL confidence intervals and is identical to the optical density and fluorescence QTL. However, each of these QTL confidence intervals span genomic regions greater than 1.5 megabases and contain hundreds of genes that vary between the N2 and CB4856 strains.

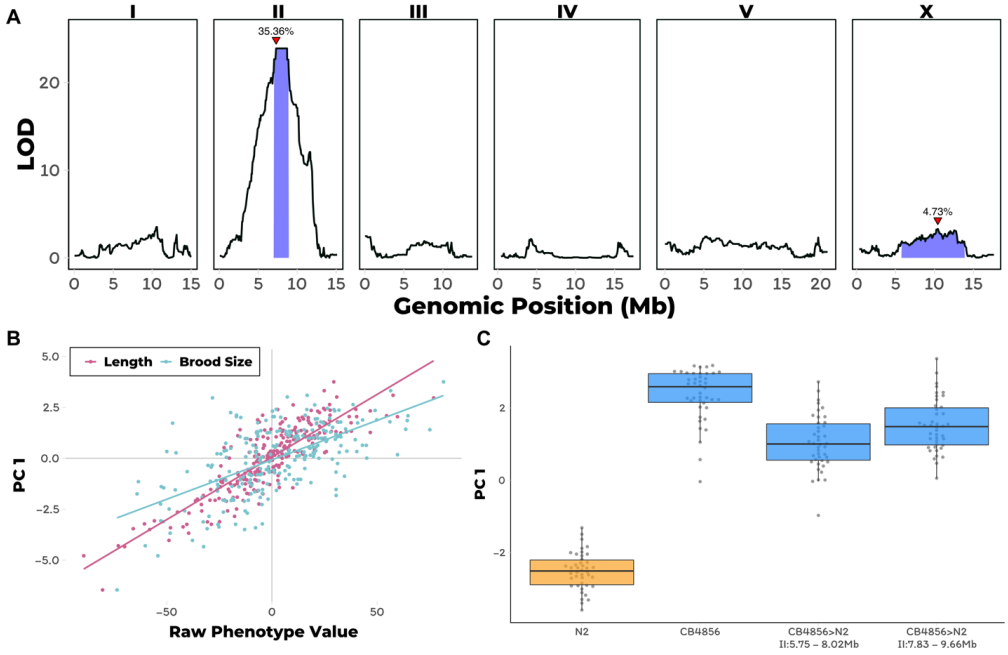

Figure 4-1 Linkage mapping of arsenic response variation and NIL QTL validation
A) Linkage mapping plots for the first principal component trait in the presence of 1000 μM arsenic trioxide is shown. The significance values (logarithm of odds, LOD, ratio) for 1454 markers between the N2 and CB4856 strains are on the y-axis, and the genomic position (Mb) separated by chromosome is plotted on the x-axis. The associated 1.5 LOD-drop confidence intervals are represented by blue boxes.

The phenotypic variance explained by each QTL is shown above the peak QTL marker, which is marked by red triangles. B) The correlation between brood size (blue; r2 = 0.38, *p*-value=1.65E-27) or animal length (pink; r2 = 0.74, *p*-value=3.16E-74) with the first principal component trait. Each dot represents an individual RIAIL's phenotype, with the animal length and brood size phenotype values on the x-axis and the first principal component phenotype on the y-axis. C) Tukey box plots of near-isogenic line (NIL) phenotype values for the first principal component trait in the presence of 1000 μM arsenic trioxide is shown. NIL genotypes are indicated below the plot as genomic ranges. The N2 trait is significantly different than the CB4856 and NIL traits (Tukey HSD *p*-value<1E-5).

Next, we constructed near-isogenic lines (NILs) to isolate and narrow the chromosome II QTL in a controlled genetic background. We introgressed genomic regions from the CB4856 strain on the left and right halves of the confidence interval into the N2 genetic background. In the presence of arsenic trioxide, both of these NILs recapitulated the parental CB4856 PC1 phenotype (Figure 3-1C) and had similar trait values for the four traits used as inputs into the PCA. Furthermore, we showed that similar to the RIAIL phenotypes, the measured traits were correlated and contributed similarly to the PC1 trait. Furthermore, the PC1 trait was highly correlated with the four input traits. The phenotypic similarity of these NILs to the CB4856 parental strain suggested that the two NILs might share an introgressed region of the CB4856 genome. To identify this shared introgressed region, we performed low-coverage whole-genome sequencing of the NIL strains and defined the left and right bounds of the CB4856 genomic introgression to be from 5.75 to 8.02 Mb and 7.83 to 9.66 Mb in the left and right NILs, respectively. The left and right NILs recapitulate 70.6% and 81.9% of the effect size difference between N2 and CB4856 as measured by Cohen's F, respectively [225], which exceeds our observations the linkage mapping results where the QTL on chromosome II explained 63.4% of the total phenotypic variation in the RIAIL population. This discrepancy was observed likely because the NILs are a more homogenous genetic background and the experiment was performed at higher replication than the linkage mapping. We observed similar levels of phenotypic recapitulation for the four traits used as inputs for the PCA (brood size: 57.8 and 87.4%, animal length: 98.5 and 100%; 69.5 and 68.1%; fluorescence: 58.5 and 64.2% for the left and right NILs). Taken together, these results suggested that genetic differences between

N2 and CB4856 within 7.83 to 8.02 Mb on chromosome II conferred resistance to arsenic trioxide.

In parallel to the linkage-mapping approach described above, we performed a genome-wide association (GWA) mapping experiment by quantifying the responses to arsenic trioxide for 86 wild *C. elegans* strains [43]. Consistent with previous experiments, the PC1 trait was influenced less by the brood size trait, as indicated by the loadings. In agreement with the results from the linkage mapping approach, PC1 differences among the wild isolates mapped to a QTL on the center of chromosome II that spans from 7.6 Mb to 8.21 Mb (Figure 3-2A). However, we noted that the brood size trait did not map to a significant QTL with the GWA mapping approach, which is most likely due to the lower statistical power of this approach. Interestingly, the genomic estimates of broad- and narrow-sense heritability (H^2; h^2) were low for all of the wild isolates measured and principal component traits (**Figure 2-figure supplement 2; Figure 2-source data 6**), which could be because the center of chromosome II has not experienced the chromosome-scale selective sweeps [43] that contribute to much of the population structure within the species. The marker found to be most correlated with the PC1 trait from GWA mapping (II:7,931,252), explains 84.6% of the total heritable phenotypic variation. In addition to the PC1 trait, three of the four measured traits also mapped to significant QTL on the center of chromosome II. (**Figure 2-figure supplement 3; Figure 2-source data 7**). Notably, the CB4856 strain, which was one of the parents used to construct the RIAIL panel used for linkage mapping, had the non-reference genotype at the marker most correlated with PC1 (**Figure 3-2B**), suggesting that the same genetic variant(s) might be contributing to the differential arsenic trioxide response between the RIAIL and wild isolate populations.

To fine map the PC1 QTL, we focused on variants from the *C. elegans* whole-genome variation dataset [44] that are shared among at least five percent of the 86 wild isolates exposed to

arsenic trioxide. Under the assumption that the linkage and GWA mapping QTL are caused by the same genetic variation, we only considered variants present in the CB4856 strain. Eight markers within the QTL region are in complete linkage disequilibrium with each other and are most correlated with the PC1 trait ((**Figure 3-2C**)**-figure supplement 4; Figure 2-source data 8**). Only one of these markers is located within an annotated gene (*dbt-1*) and is predicted to encode a cysteine-to-serine variant at position 78 (C78S). Although it is possible that the causal variant underlying differential arsenic trioxide response in the *C. elegans* population is an intergenic variant, we focused on the DBT-1(C78S) variant as a candidate to test for an effect on arsenic response.

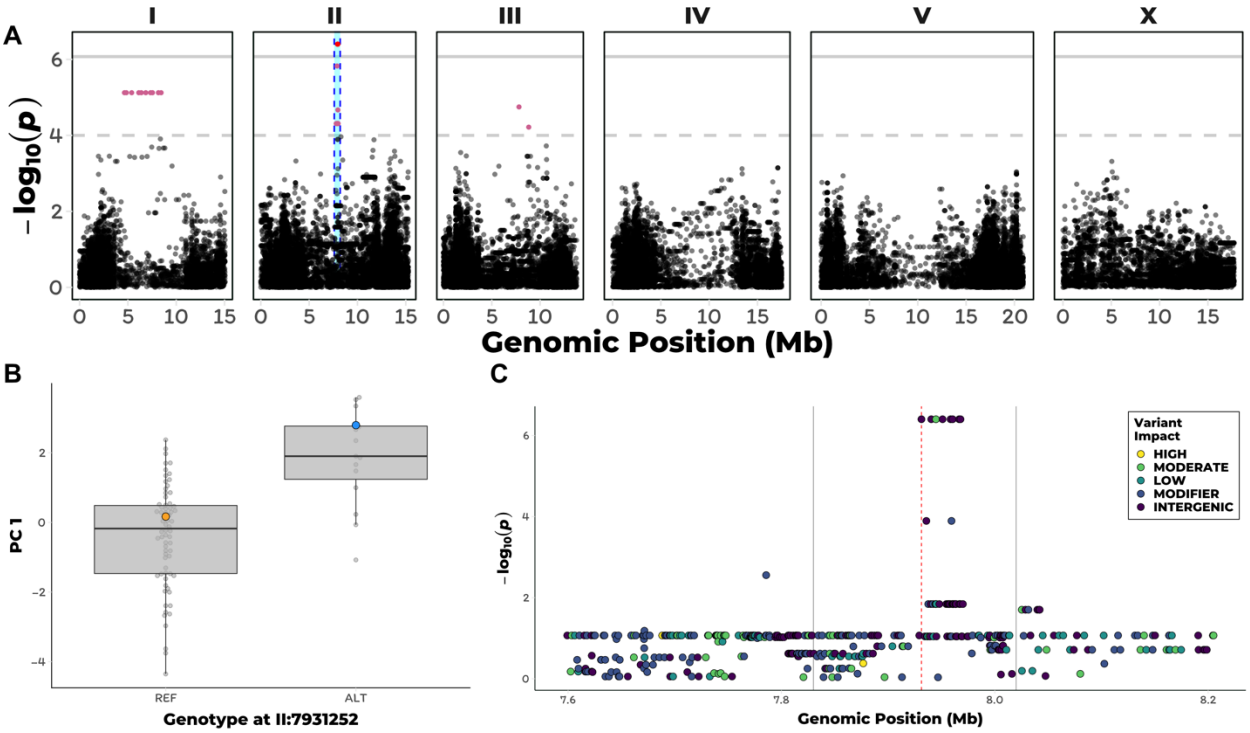

Figure 4-2 GWA mapping of arsenic response variation
A) A manhattan plot for the first principal component in the presence of 1000 µM arsenic trioxide is shown. Each dot represents an SNV that is present in at least 5% of the assayed wild population. The genomic position in Mb, separated by chromosome, is plotted on the x-axis and the -log10(*p*) for each SNV is plotted on the y-axis. SNVs are colored red if they pass the genome-wide Bonferroni-corrected significance (BF) threshold, which is denoted by the gray horizontal line. SNVs are colored pink if they pass the genome-wide eigen-decomposition significance (ED) threshold, which is denoted by the dotted gray horizontal line. The genomic region of interests surrounding the QTL that pass the BF and ED thresholds are represented by cyan and pink rectangles, respectively. B) Tukey box plots of phenotypes used for association mapping in A) are shown. Each dot corresponds to the phenotype of an individual strain, which is plotted on the y-axis. Strains are grouped by their genotype at the peak QTL position (red

SNV from panel A, ChrII:7,931,252), where REF corresponds to the allele from the reference N2 strain. The N2 (orange) and CB4856 (blue) strains are highlighted. C) Fine mapping of the chromosome II region of interest (cyan region from panel A, 7.60–8.21 Mb) is shown. Each dot represents an SNV present in the CB4856 strain. The association between the SNV and first principal component is shown on the y-axis and the genomic position of the SNV is shown on the x-axis. Dots are colored by their SnpEff predicted effect.

A cysteine-to-serine variant in DBT-1 contributes to arsenic response variation

The *C. elegans dbt-1* gene encodes the E2 component of the branched-chain α-keto acid dehydrogenase complex (BCKDH) [209]. The BCKDH complex is a core component of branched-chain amino acid (BCAA) catabolism and catalyzes the irreversible oxidative decarboxylation of amino acid derived branched-chain α-ketoacids [226]. The BCKDH complex belongs to a family of α-ketoacid dehydrogenases that include pyruvate dehydrogenase (PDH) and α-ketoglutarate dehydrogenase (KGDH) [227]. All three of these large enzymatic complexes include a central E2 component that is lipoylated at one critical lysine residue (two residues in PDH). The function of these enzymatic complexes depends on the lipoylation of these lysine residues [227,228]. In *C. elegans*, the putative lipoylated lysine residue is located at amino acid position 71 of DBT-1, which is in close proximity to the C78S residue that we found to be highly correlated with arsenic trioxide resistance.

To confirm that the C78S variant in DBT-1 contributes to differential arsenic trioxide responses, we used CRISPR-Cas9-mediated genome editing to generate allele-replacement strains by changing the C78 residue in the N2 strain to a serine and the S78 residue in the CB4856 strain to a cysteine. When treated with arsenic trioxide, the N2 DBT-1(S78) allele-replacement strain recapitulated 56.4% of the phenotypic difference between the CB4856 and N2 strains as measured with the first principal component (Cohen's F) [225] (Figure 3-3). Similarly, the

CB4856 DBT-1(C78) allele-replacement strain recapitulated 64.8% of the total phenotypic difference between the two parental strains. The degree to which the allele-replacement strains recapitulated the difference in the PC1 trait between the N2 and CB4856 strains matched our observations from the linkage mapping experiment, where the chromosome II QTL explained 63.4% of the total phenotypic variation in the RIAIL population. This result suggested that the majority of heritable variation in arsenic trioxide response was explained by the DBT-1(C78S) allele. We obtained similar results for the BIOSORT-quantified traits, suggesting that overall animal physiology is affected by arsenic exposure. However, when considering brood size, the N2 DBT-1(C78S) allele-replacement strain produced an intermediate number of progeny in the presence of arsenic trioxide relative to the parental N2 and CB4856 strains. And the CB4856 DBT-1(S78C) allele-replacement strain produced fewer offspring than both parental strains. These results suggested that additional genetic variants between the N2 and CB4856 strains might interact with the DBT-1(C78S) allele to affect different aspects of physiology. Nevertheless, these results functionally validated that the DBT-1 C78S variant underlies differences in physiological responses to arsenic trioxide.

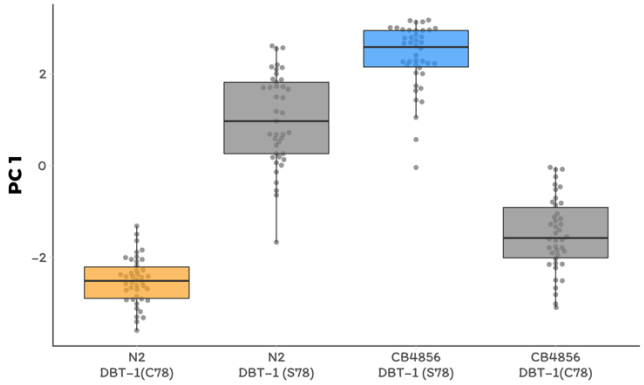

Figure 4-3 Functional validation of the DBT-1 C78S allele
A) Tukey box plots of the first principal component generated by PCA on allele-replacement strain henotypes measured by the COPAS BIOSORT 1000 µM arsenic trioxide exposure are shown (N2,orange; CB4856, blue; allele replacement strains, gray). Labels correspond to the genetic background and the corresponding residue at position 78 of DBT-1 (C for cysteine, S for serine). All pair-wise comparisons are significantly different (Tukey HSD, *p*-value < 1E-7).

Arsenic trioxide inhibits the DBT-1 C78 allele

Mono-methyl branched chain fatty acids (mmBCFA) are an important class of molecules that are produced via BCAA catabolism [198,209,229,230]. The production of mmBCFA requires the BCKDH, fatty acid synthase (FASN-1), acetyl-CoA carboxylase (POD-2), fatty acyl elongases (ELO-5/6), β-ketoacyl dehydratase (LET-767), and acyl CoA synthetase (ACS-1) [198,209,229,231–233]. Strains that lack functional *elo-5*, *elo-6*, or *dbt-1* produce less C15ISO and C17ISO mmBCFAs, arrest at the L1 larval stage, and can be rescued by supplementing the growth media with C15ISO or C17ISO [198,209,229] (Figure 3-4A).

Because DBT-1 is involved in BCAA catabolism, we hypothesized that the DBT-1(C78S)-dependent difference in progeny length between the N2 and CB4856 strains after arsenic trioxide treatment might be caused by differential larval arrest through depletion of downstream mmBCFAs. To test this hypothesis, we quantified the abundance of the monomethyl-branched (ISO) and straight-chain (SC) forms of C15 and C17 in the N2, CB4856, and allele-replacement genetic backgrounds. We measured the metabolite levels in staged L1 animals and normalized the detected amounts of C15ISO and C17ISO relative to the abundances of C15SC and C17SC, respectively, to mitigate the confounding effects of differences in developmental rates that could be caused by genetic background differences after arsenic trioxide exposure. Generally, the ratios of C15ISO/C15SC and C17ISO/C17SC were reduced in arsenic-treated animals relative to controls (Figure 3-4B). However, arsenic trioxide treatment had a 7.6-fold stronger effect on the C15ISO/C15SC ratio in N2, which naturally has the C78 allele, than on the N2 DBT-1(S78) allele replacement strain. This difference suggests that the DBT-1(C78) allele is more strongly inhibited by arsenic trioxide (0.04 to 0.004, Tukey HSD p-value = 0.0358, n = 6). Similarly, we observed a 6.6-fold arsenic-induced reduction in the C17ISO/C17SC ratio when comparing the N2 DBT-1(C78) and N2 DBT-1(S78) strains (Tukey HSD p-value =

0.003747, n = 6). When comparing the CB4856 DBT-1(S78) and CB4856 DBT-1(C78) strains, we observed a 2.8-fold lower C15ISO/C15SC ratio (Tukey HSD *p*-value = 0.0427733, n = 3) and 1.5-fold lower C17ISO/C17SC ratio (Tukey HSD *p*-value = 0.164721, n = 3) in the in the CB4856 DBT-1(C78) strain. We noted that the C17ISO/straight-chain ratio difference was not significantly different between the two CB4856 genetic background strains. However, we observed a significant arsenic-induced decrease in raw C17ISO production in the CB4856 DBT-1(C78) strain (Tukey HSD *p*-value = 0.029) and no significant difference in the CB4856 DBT-1(S78) strain (Tukey HSD *p*-value = 0.1). Importantly, these DBT-1(C78S) allele-specific reductions in ISO/straight-chain ratios were not driven by arsenic-induced differences in straight-chain fatty acids. These results explained the majority of the physiological differences between the N2 and CB4856 strains in the presence of arsenic trioxide (Figure 3-3) and suggested that the DBT-1(C78) allele was inhibited by arsenic trioxide more strongly than DBT-1(S78). Taken together, the differential reduction in branched-chain fatty acids likely underlies the majority of physiological differences between the sensitive and resistant *C. elegans* strains.

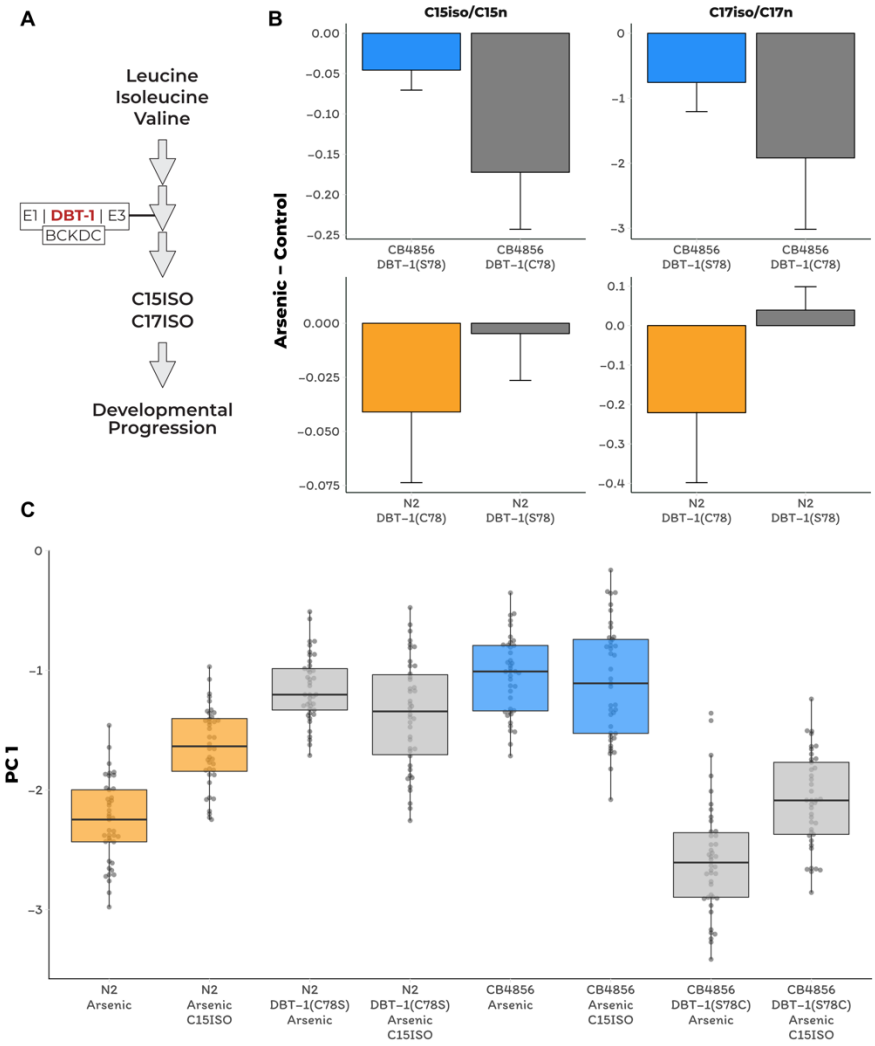

Figure 4-4 Branched-chain fatty acid (BCFA) production in arsenic and BCFA rescue of arsenic sensitivity

A) A simplified model of BCAA catabolism in *C. elegans*. The BCKDH complex, which consists of DBT-1, catalyzes the irreversible oxidative decarboxylation of branched-chain ketoacids. The products of these breakdowns can then serve as building blocks for the mmBCFAs that are required for developmental progression. B) The difference in the C15ISO/C15SC (left panel) or C17ISO/C17SC (right panel) ratios between 100 µM arsenic trioxide and control conditions is plotted on the y-axis for three independent replicates of the CB4856 and CB4856 allele replacement strains and six independent replicates of the N2and N2 allele replacement strains. The difference between the C15 ratio for the CB4856-CB4856 allele replacement comparison is significant (Tukey HSD *p*-value = 0.0427733), but the difference between the C17 ratios for these two strains is not (Tukey HSD *p*-value = 0.164721). The difference between the C15and C17 ratios for the N2-N2 allele replacement comparisons are both significant (C15: Tukey HSD *p*-value = 0.0358; C17: Tukey HSD *p*-value = 0.003747). C) Tukey box plots median animal length after arsenic trioxide or arsenic trioxide and 0.64 µM C15ISO exposure are shown (N2, orange; CB4856, blue; allele replacement strains, gray). Labels correspond to the genetic background and the corresponding residue at position 78 of DBT-1 (C for cysteine, S for serine). Every pair-wise strain comparison is significant except for the N2 DBT-1(S78) - CB4856 comparisons (Tukey's HSD *p*-value < 1.43E-6).

In addition to arsenic-induced differences in branched chain fatty acid production, we observed significant differences in branched/straight-chain ratios between the parental and allele replacement strains when L1 larval animals were grown in control conditions. Strains with the DBT-1(C78) had higher ISO/SC ratios relative to strains with the DBT-1(S78) for the C17 (CB4856 DBT-1(C78): Tukey HSD *p*-value = 0.0342525, n = 3; N2 DBT-1(C78): Tukey HSD *p*-value = 0.0342525, n = 6) and C15 ratios (CB4856: Tukey HSD *p*-value = 0.0168749, n = 3; N2: Tukey HSD *p*-value = 0.1239674, n = 6). We noted that the C15ISO/straight-chain ratio was not significantly different when comparing the N2 and N2 allele replacement strain, but the direction of effect matched our other observations, and we saw significant differences in C15ISO levels (N2-C15ISO DBT-1(C78): Tukey HSD *p*-value = 0.0265059, n = 6). Importantly, the DBT-1 allele-specific differences in the fatty acid ratio and ISO measurements were not driven by differences in straight-chain fatty acids. However, we did not observe the same effect of the DBT-1(C78S) allele at the young adult life stage. Taken together, these results suggest that the DBT-1(C78) allele produces more branched chain fatty acids than the DBT-1(S78) allele, but this effect was dependent on the developmental stage of the animals.

To test the hypothesis that differential arsenic-induced depletion of branched-chain fatty acids in strains with the DBT-1(C78S) causes physiological differences in growth, we tested if mmBCFA supplementation could rescue the effects of arsenic trioxide-induced toxicity. We exposed the parental and the DBT-1 allele-replacement strains to media containing arsenic trioxide alone, C15ISO alone, or a combination of arsenic trioxide and C15ISO. In agreement with previous experiments, the PC1 trait was more strongly correlated with the animal length, optical density, and fluorescence traits than the brood size trait. C15ISO supplementation of the arsenic growth media caused a 53.5% rescue of the allele-specific effect in the N2 genetic background (Figure 3-4C). Similarly, when arsenic-exposed CB4856 DBT-1(C78) animals were supplemented with C15ISO, the allele-specific PC1 phenotypic difference was reduced by 25.6% when compared

to the difference between the CB4856 DBT-1(C78) and CB4856 DBT-1(S78) strains in arsenic trioxide alone. By contrast, CB4856 DBT-1(S78) and N2 DBT-1(S78) phenotypes were unaffected by C15ISO supplementation in arsenic trioxide media. We observed similar trends for the animal length, optical density, and fluorescence traits that we used as inputs for PCA but not for brood size. Collectively, these data support the hypothesis that the cysteine/serine variant in DBT-1 contributes to arsenic sensitivity in *C. elegans* by reducing ISO fatty acid biosynthesis, and the DBT-1(C78) variant can be partially rescued by supplementation with mmBCFAs.

Arsenic exposure increases mmBCFA production and favors a cysteine allele in human DBT1

To test whether our results from *C. elegans* translate to human variation in arsenic sensitivity, we tested the role of DBT1 variation on arsenic trioxide responses and mmBCFA biosynthesis in human cells. The human DBT1 enzyme contains a serine at position 112 that corresponds to the C78 residue in *C. elegans* DBT-1 (Figure 3-5A). Using CRISPR-Cas9, we edited batch cultures of 293T cells to generate a subset of cells with DBT1(S112C). These cells were exposed to arsenic trioxide or control conditions, and we monitored changes in the fraction of cells carrying the DBT1(C112) allele. We found that arsenic exposure caused a 4% increase in the fraction of cells that contained the DBT1(C112) allele (Figure 3-5B, Fisher's exact test, *p*-value < 1.9E-16). Though the human DBT1 does not vary within the human population at S112, two residues in close spatial proximity to S112 do vary among individuals in the population (R113C and W84C) (Figure 3-5A)[234]. To test the effects of these residues on arsenic trioxide sensitivity, we performed the same editing and arsenic selection procedure described above. Over the course of the selection experiment, cells with DBT1(W84C) and DBT1(R113C) increased by 2% and 1%, respectively (Figure 3-5B). Therefore, it appears that all three missense variants caused a slight increase in fitness in batch-edited human cell cultures

exposed to arsenic – the opposite result we found in *C. elegans*. To determine if branched-chain fatty acid production was altered by arsenic exposure, as we found in *C. elegans*, we measured mmBCFA production in unedited 293T cells in arsenic and mock-treated cultures. We found that overall fatty acid production was markedly reduced in arsenic-treated cultures. In contrast to our observations in *C. elegans*, straight-chain fatty acids were more drastically reduced than ISO fatty acids, suggesting pleiotropic effects and a general perturbation of fatty acid metabolism.

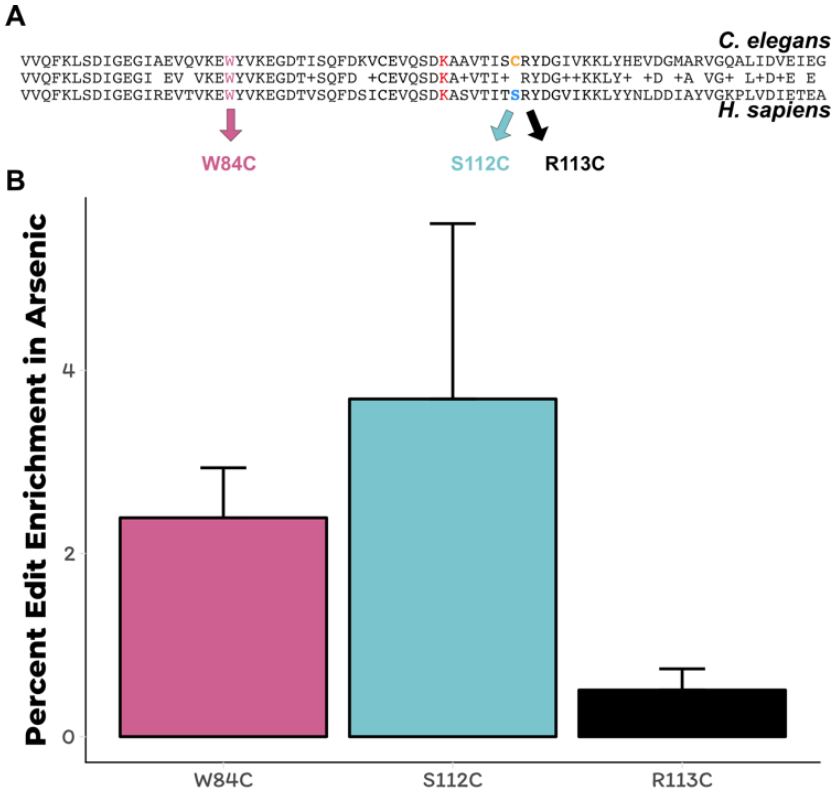

Figure 4-5 Human cell lines with DBT-1 cysteine allele are less sensitive to arsenic
A) Alignment of C. elegans DBT-1 and H. sapiens DBT1. The residues tested for an arsenic-specific effect are indicated with arrows - W84C (pink), S112C (blue), and R113C (black). The lysine that is post-translationally modified with a lipoid acid is highlighted in red. (B) The percent increase of edited human cells that contain the W84C, S112C, or R113C amino acid change in DBT1 in the presence 5 μM arsenic trioxide relative to control conditions are shown. The number of reads in 5 μM arsenic trioxide for all replicates are significantly different from control conditions (Fisher's exact test, *p*-value<0.011).

Discussion

In this study, we characterized the effects of *C. elegans* natural genetic variation on physiological responses to the pervasive environmental toxin arsenic trioxide. Though the

effects of this toxin have been extensively studied in a variety of systems [180,181,227,235,236], recent evidence from human population studies have revealed local adaptations within region-specific subpopulations [184,186,187,237]. Our investigation into the natural variation in *C. elegans* responses to arsenic trioxide led to the discovery of a novel mechanism by which this compound could elicit toxicity. We show that arsenic trioxide differentially inhibits two natural alleles of the E2 domain of the BCKDH complex, which is encoded by the *dbt-1* gene. Specifically, strains with the DBT-1(C78) allele are more sensitive to arsenic trioxide than strains carrying the DBT-1(S78). Furthermore, we show that the increased sensitivity of the DBT-1(C78) allele is largely explained by differences in the production of mmBCFAs (Figure 3-4B-C), which are critical molecules for developmental progression beyond the first larval stage. Arsenic is thought to inhibit the activity of both the pyruvate dehydrogenase (PDH) and the α-ketoglutarate (KGDH) dehydrogenase complexes through interactions with the reduced form of lipoate [227], which is the cofactor for the E2 domain of these complexes. Like the PDH and KGDH complexes, the E2 domain of BCKDH complex requires the cofactor lipoate to perform its enzymatic function [238–240]. The inhibition of DBT-1 by arsenic trioxide could involve three-point coordination of arsenic by the C78 thiol and the reduced thiol groups of the nearby lipoate. However, based on the crystal structure (PDB:1Y8N), the atomic distance between the C78 thiol group and the thiol groups from the lipoylated lysine is ~32 Å, which might be too large a distance for coordinating arsenic [241]. Alternatively, arsenic trioxide could inhibit DBT-1(C78) through coordination between the thiol groups of C78 and C65 (~8.5 Å). Analogous thiol-dependent mechanisms have been proposed for the inhibition of other enzymes by arsenic [236]. Despite structural similarities and a shared cofactor, no evidence in the literature indicates that BCKDH is inhibited by arsenic trioxide, so these results demonstrate the first connection of arsenic toxicity to BCKDH E2 subunit inhibition.

Multiple sequence alignments show that cysteine residues C65 and C78 of *C. elegans* DBT-1 correspond to residues S112 and C99 of human DBT1 (Figure 3-5A). Though DBT1 does not vary at position 112 within the human population, two residues (R113C and W84C) in close spatial proximity do [234]. We hypothesized that cysteine variants in DBT1 would sensitize human cells to arsenic trioxide. However, we found that the cysteine variants (W84C, S112C, and R113C) proliferated slightly more rapidly than the parental cells in the presence of arsenic. Notably, a growing body of evidence suggests that certain cancer cells upregulate components involved in BCAA metabolism, and this upregulation promotes tumor growth [242,243]. Perhaps the increased proliferation of human cell lines that contain the DBT1 C112 allele (Figure 3-5) is caused by increased activity of the BCKDH complex. It is worth noting that human cell lines grown in culture do not have the same strict requirements for mmBCFA, and the requirements for different fatty acids are variable among diverse immortalized cell lines [244,245]. Furthermore, in *C. elegans*, the developmental defects associated with *dbt-1* loss-of-function can be rescued by neuronal-specific expression of *dbt-1* [209], suggesting that the physiological requirements of mmBCFA in *C. elegans* depend on the coordination of multiple tissues that cannot be recapitulated with cell-culture experiments. These results further highlight the complexity of arsenic toxicity, as well as the physiological requirements for BCAA within and across species and could explain the discrepancy between the physiological effects we observed in *C. elegans* and human cell-line experiments. Given that arsenic trioxide has become the standard-of-care for treating APL [246] and is gaining traction in treating other leukemias, it might be important to further explore the effects of arsenic on BCAA metabolism and cancer growth.

The C78 allele of DBT-1 is likely the derived allele in the *C. elegans* population because all other organisms have a serine at the corresponding position. The loss of the serine allele in the *C. elegans* population might have been caused by relaxed selection at this locus, though this

hypothesis is difficult to test because of the effects of linked selection and decreased recombination in chromosome centers. It is hypothesized that the *C. elegans* species originated from the Pacific Rim and that the ancestral state more closely resembles the CB5846 strain than the N2 strain [43,247]. The CB4856 strain was isolated from the Hawaiian island of O'ahu [248], where environments could have elevated levels of arsenic in the soil from volcanic activity, the farming of cane sugar, former construction material (canec) production facilities, or wood treatment plants (Hawaii.gov). It is possible that as the *C. elegans* species spread across the globe into areas with lower levels of arsenic in the soil and water, the selective pressure to maintain high arsenic tolerance was relaxed and the cysteine allele appeared. Alternatively, higher mmBCFA levels at the L1 larval stage in strains with the DBT-1(C78) allele (Figure 4-figure supplement 3-4) might cause faster development in certain conditions, although we did not observe allele-specific growth differences in laboratory conditions. Despite these clues suggesting selection in local environments, the genomic region surrounding the *dbt-1* locus does not show a signature of selection as measured by Tajima's *D* [109], and the strains with the DBT-1 S78 allele show no signs of geographic clustering. Nevertheless, our study suggests that *C. elegans* is a powerful model to investigate the molecular mechanisms for how populations respond to environmental toxins.

Future directions

A key finding of this research project was that the cysteine variant we identified in *C. elegans* is not identified in any other species's reference genome. We found higher mmBCFA levels at the L1 larval stage in strains with the DBT-1(C78) allele, which suggests that this allele provides individual with a fitness advantage in certain environmental niches. Though we did not observe a signature of selection at this locus, the complex and largely unknown demographic history of the *C. elegans* species likely skews these population genetic statistics. Nevertheless, these

results strongly suggest that diverse *C. elegans* strains have different metabolic requirements, which has been further supported by recent work in the N2 and CB4856 strains [119]. My own work, discussed in Chapter 2, provides further support for this hypothesis, where I found enrichment of metabolic genes in genetically divergent loci of individual strains. Further exploration of the intersection of the metabolite QTL identified by Gao *et al.* and the genetically divergent regions of the genome I discussed in Chapter 1 might provide clues to the metabolic requirements for individual *C. elegans* strains to inhabit a specific environmental niche. One approach to achieve this goal is to perform comparative metabolomic analysis in individuals with the DBT-1 cysteine versus individuals with the serine allele. The results described by Gao *et al.* suggest that there is substantial differences in fatty acid metabolism among the N2 (cysteine) and CB4856 (serine) strains, but fully characterizing the effects of altered fatty acid production on the overall metabolic network will reveal how metabolic networks adjust to perturbation and reveal insights into the evolution and functional constraints of these networks. The DBT-1 allele replacement strains we constructed in this study provide an excellent platform to perform these exploratory experiments.

In addition to further work in *C. elegans*, our results showing that fatty acid production is affected by arsenic exposure in human cells opens a line of scientific inquiry into the effects fatty acid production in on chemotherapeutic interventions. As It has been established that tumors alter their metabolic networks, including altered fatty acid metabolism [249], investigating the effects of cancer chemotherapeutics on cancer metabolism might pave the way for novel chemotherapeutic combinations to increase the efficacy of tumor clearance.

Contributions

Stefan Zdraljevic, Conceptualization, Data curation, Formal analysis, Funding acquisition, Validation, Investigation, Visualization, Methodology, Writing—original draft, Project administration, Writing— review and editing; Bennett William Fox, Data curation, Formal analysis, Investigation, Methodology, Writing—review and editing; Christine Strand, Data curation, Formal analysis, Validation, Investigation, Writing—review and editing; Oishika Panda, Data curation, Formal analysis, Investigation; Francisco J Tenjo, Investigation; Shannon C Brady, Resources, Writing—review and editing; Tim A Crombie, Investigation, Writing—review and editing; John G Doench, Resources, Supervision, Methodology, Project administration, Writing—review and editing; Frank C Schroeder, Supervision, Funding acquisition, Methodology, Project administration, Writing—review and editing; Erik C Andersen, Conceptualization, Resources, Supervision, Funding acquisition, Methodology, Project administration, Writing—review and editing

5. Extreme allelic heterogeneity at a *Caenorhabditis elegans* beta-tubulin locus explains natural resistance to benzimidazoles

Preface

Steffen Hahnel, Ph.D., joined the Andersen lab in 2016 and was interested in studying the effects of natural variation in *C. elegans* responses to anthelmintic compounds. With the help of Briana Rodriguez, Steffen performed exposed 249 wild *C. elegans* strains to a variety of anthelmintic compounds. Because Steffen did not intend to stay in the Andersen lab for an extended period of time, I was brought into the project to perform the heavy lifting for the data analysis. During our collaboration, Steffen and I became friends and learned a lot from each other. I exposed him to coding and data analysis, and he exposed me to the parasitology field and the ins and outs of anthelmintics. This fruitful collaboration culminated in a publication at *PLoS Pathogens* in 2018 [71].

Abstract

Benzimidazoles (BZ) are essential components of the limited chemotherapeutic arsenal available to control the global burden of parasitic nematodes. The emerging threat of BZ resistance among multiple nematode species necessitates the development of novel strategies to identify genetic and molecular mechanisms underlying this resistance. All detection of parasitic helminth resistance to BZ is focused on the genotyping of three variant sites in the orthologs of the β-tubulin gene found to confer resistance in the free-living nematode *Caenorhabditis elegans*. Because of the limitations of laboratory and field experiments in parasitic nematodes, it is difficult to look beyond these three sites to identify additional mechanisms that might contribute to BZ resistance in the field. Here, we took an unbiased

genome-wide mapping approach in the free-living nematode species *C. elegans* to identify the genetic underpinnings of natural resistance to the commonly used BZ, albendazole (ABZ). We found a wide range of natural variation in ABZ resistance in natural *C. elegans* populations. In agreement with known mechanisms of BZ resistance in parasites, we find that a majority of the variation in ABZ resistance among wild *C. elegans* strains is caused by variation in the β-tubulin gene *ben-1*. This result shows empirically that resistance to ABZ naturally exists and segregates within the *C. elegans* population, suggesting that selection in natural niches could enrich for resistant alleles. We identified 25 distinct *ben-1* alleles that are segregating at low frequencies within the *C. elegans* population, including many novel molecular variants. Population genetic analyses indicate that *ben-1* variation arose multiple times during the evolutionary history of *C. elegans* and provide evidence that these alleles likely occurred recently because of local selective pressures. Additionally, we find purifying selection at all five β-tubulin genes, despite predicted loss-of-function variants in *ben-1*, indicating that BZ resistance in natural niches is a stronger selective pressure than loss of one β-tubulin gene. Furthermore, we use genome-editing to show that the most common parasitic nematode β-tubulin allele that confers BZ resistance, F200Y, confers resistance in *C. elegans*. Importantly, we identified a novel genomic region that is correlated with ABZ resistance in the *C. elegans* population but independent of *ben-1* and the other β-tubulin loci, suggesting that there are multiple mechanisms underlying BZ resistance. Taken together, our results establish a population-level resource of nematode natural diversity as an important model for the study of mechanisms that give rise to BZ resistance.

Introduction

Parasitic nematodes have a tremendous impact on global health and socio-economic development, especially in the developing world [250]. They are among the most widespread

human pathogens, and almost two billion people are estimated to suffer from infection of one or multiple nematode species [250,251]. The main endemic areas of nematode infections are highly correlated with tropical and subtropical regions worldwide. Because of their detrimental impact on human health, several nematode infections belong to a class of diseases designated by the World Health Organization (WHO) as Neglected Tropical Diseases (NTDs). Altogether, the loss of disability-adjusted life years (DALY) caused by parasitic nematodes is conservatively estimated to be 10 million DALYs per year, which ranks them among the top of all NTDs [250]. Apart from this drastic impact on human health, several nematode species infect a variety of key crops and livestock causing substantial economic losses throughout the world [252].

Global control of nematode infections relies on the efficacy of a limited repertoire of anthelmintic drugs, including benzimidazoles (BZ). BZs are frequently used in mass drug administration (MDA) programs to treat parasitic nematode infections in endemic regions. However, the long-term success of these MDA programs is limited by the persistence and re-emergence of nematode infections in these regions. The most significant of these factors is the high reinfection rate caused by long-term parasitic nematode reservoirs, which necessitates frequent deworming of affected communities [251]. As a consequence of continuous drug-pressure and relaxation cycles, parasite populations have the potential to develop anthelmintic resistance. This resistance is already a significant problem in veterinary medicine [253], and resistance in the parasitic nematodes that infect humans is becoming increasingly evident. For example, reduced cure rates have been observed for soil-transmitted helminthiases, which might be attributable to BZ resistance [254–256]. To face this growing threat, detailed knowledge of anthelmintic resistance mechanisms is required to improve diagnostic tools for field surveys and educate drug-treatment strategies in MDA programs.

In vitro studies have shown that BZs inhibit the polymerization of microtubules [257–259]. Mutagenesis screens that selected for BZ resistance in *Saccharomyces cerevisiae* [260] and *Caenorhabditis elegans* [261] have independently identified numerous β-tubulin mutant alleles. These findings were then translated to veterinary-relevant nematodes, where three major single-nucleotide variants (SNVs) in parasitic nematode β-tubulin genes were found to be highly correlated with BZ resistance. The most common variant causes a phenylalanine to tyrosine amino acid change at position 200 (F200Y) [262]. Other SNVs linked to BZ resistance have been described at positions 167 (F167Y) and 198 (E198A) in several nematode species [263,264]. Although all three of these missense variants have been shown to reduce BZ binding affinity to purified β-tubulins *in vitro* [265,266], the BZ β-tubulin binding site remains experimentally uncharacterized. Furthermore, these missense variants do not explain all of the BZ resistance observed in the field [253,267]. This discrepancy might indicate that additional, non-target related components such as drug efflux pumps and detoxification enzymes might contribute to BZ resistance [253,268,269].

Additional genetic variants associated with natural BZ resistance can be identified using quantitative genetic approaches that consider genotypic and phenotypic variation present within a wild population of parasitic nematodes. However, parasitic nematodes are not easily amenable to quantitative genetic approaches because of their complicated life cycles, their poorly annotated reference genomes, and limited molecular and genetic tools [270–272]. Recent successes in the parasite species *Onchocerca volvulus* and *Haemonchus contortus* found genomic intervals that underlie ivermectin resistance [270,273,274], but similar approaches have not been applied to BZ resistance yet. By contrast, the free-living nematode species *C. elegans* has a short life cycle, a well annotated reference genome [27,275,276], and an abundance of molecular and genetic tools for the characterization of BZ responses [138,277,278]. A recent example that applied quantitative genetic approaches to investigated

BZ responses in *C. elegans* used a mapping population generated between two genetically divergent strains to identify a quantitative trait locus (QTL) linked to BZ resistance [278]. Importantly, the QTL identified in this study did not overlap with β-tubulin genes, suggesting that quantitative genetic approaches in *C. elegans* can be used to discover novel mechanisms associated with BZ resistance.

In the present study, we leveraged the power of *C. elegans* natural diversity to perform genome-wide association (GWA) mappings for BZ resistance. We used a set of 249 wild *C. elegans* isolates available through the *C. elegans* Natural Diversity Resource (CeNDR) [27] to identify genomic regions that contribute to albendazole (ABZ) resistance. ABZ is a broadly administered BZ used to treat parasitic nematode infections in humans and livestock [4,30]. We show that a major source of ABZ resistance in the *C. elegans* population is driven by putative loss-of-function (LoF) variants in the *ben-1* locus. Notably, we found 25 distinct *ben-1* alleles with low minor allele frequencies (MAF) that contribute to ABZ resistance. We show that these putative *ben-1* LoF variants arose independently during the evolutionary history of the *C. elegans* species, which suggests that local BZ selective pressures might have contributed to the extreme allelic heterogeneity at this locus. We next made use of the extensive molecular toolkit available in *C. elegans* to verify that the introduction of the *ben-1* LoF alleles do not result in any detectable fitness consequences in standard laboratory conditions, but does confer ABZ resistance. Taken together, these results suggest that *C. elegans* and likely free-living stages of parasitic species encounter natural compounds that promote BZ resistance through selection for standing or *de novo* variation in conserved nematode-specific β-tubulin genes.

Materials and Methods

Strains

Animals were cultured at 20°C on modified nematode growth medium (NGM), containing 1% agar and 0.7% agarose [146] and seeded with OP50 *E. coli*. Prior to each assay, strains were passaged for at least four generations without entering starvation or encountering dauer-inducing conditions [146]. For the genome-wide association (GWA) studies, 249 wild isolates from CeNDR (version 20170531) were used as described previously [27,31]. Construction of *ben-1* allele-replacement and *ben-1* deletion strains are described in the corresponding section. All strain information can be found in **S1 Table**.

Table 5-1 Strains used for experiments discussed in Chapter 5

Strain Name	Genetic background	Genotype	Description	Experiments
PTM229	N2	dpy-10(kah81)	strain contains a barcode signature in the dpy-10 gene	the strain used as reference in competition assay
ECA880	N2	ben-1 (ean62[588 C>G])	the strain only contains the silent mutation in the Cas9 protospacer-adjacent motif (PAM) which was used to generate the F200Y allele swap in ben-1	used in HTA as control for the F200Y allele swap
ECA881	N2	ben-1 (ean63[588 C>G])	the strain only contains the silent mutation in the Cas9 protospacer-adjacent motif (PAM) which was used to generate the F200Y allele swap in ben-1	used in HTA as control for the F200Y allele swap
ECA882	N2	ben-1 (ean64)	the strain contains a _ 1 kb deletion in the ben-1 ORF	used in the HTA
ECA883	N2	ben-1 (ean65)	the strain contains a _ 1 kb deletion in the ben-1 ORF	used in the HTA
ECA884	N2	ben-1 (ean66)	the strain contains a _ 1 kb deletion in the ben-1 ORF	used in the HTA and competiton assay
ECA917	N2	ben-1 (ean98[599 T>A, 588 C>G])	the strain contains the silent mutation in the Cas9 protospacer-adjacent motif (PAM) and the F200Y allele swap in ben-1	used in the HTA and competiton assay
ECA918	N2	ben-1 (ean99[599 T>A, 588 C>G])	the strain contains the silent mutation in the Cas9 protospacer-adjacent motif (PAM) and the F200Y allele swap in ben-1	used in the HTA
ECA919	N2	ben-1 (ean100[599 T>A, 588 C>G])	the strain contains the silent mutation in the Cas9 protospacer-adjacent motif (PAM) and the F200Y allele swap in ben-1	used in the HTA
ECA920	N2	ben-1 (ean101[599 T>A, 588 C>G])	the strain contains the silent mutation in the Cas9 protospacer-adjacent motif (PAM) and the F200Y allele swap in ben-1	used in the HTA
ECA921	N2	ben-1 (ean102[599 T>A, 588 C>G])	the strain contains the silent mutation in the Cas9 protospacer-adjacent motif (PAM) and the F200Y allele swap in ben-1	used in the HTA

High-throughput albendazole-response assay

The high-throughput fitness assays (HTA) were performed as described previously [31] with the exception that each strain was assayed in four technical replicates that consisted of four

independent bleach synchronization steps across two days with two independent drug and control preparations. In short, strains were propagated for four generations on agar plates, followed by bleach synchronization. The embryos were titered to 96-well microtiter plates at a final concentration of approximately one embryo per microliter of K medium [55] with modified salt concentrations (10.2 mM NaCl, 32 mM KCl, 3 mM $CaCl_2$, 3 mM $MgSO_4$). After overnight incubation, hatched L1 larvae were fed with 5 mg/mL HB101 bacterial lysate (Pennsylvania State University Shared Fermentation Facility, State College, PA) and cultured for two days until they reached the L4 larval stage. Using the large particle flow cytometer COPAS BIOSORT (Union Biometrica, Holliston MA), three L4 larvae per well were sorted into new microtiter plates containing modified K medium, 10 mg/mL HB101 lysate, 50 µM kanamycin, and either 12.5 µM albendazole (ABZ) dissolved in 1% DMSO or 1% DMSO alone as a control. During the following four-day incubation, animals were allowed to mature and to produce offspring. Fitness parameters, including traits for animal length and brood size, were measured for each population under drug and control conditions using the COPAS BIOSORT platform after 96 hours. To facilitate body straightening for more accurate length determinations, animals were treated with sodium azide (50 mM) immediately before measurement.

Albendazole-response trait calculations

Fitness parameters measured by the COPAS BIOSORT platform were processed using the R package *easysorter* [147] with some modifications to incorporate technical replicates into the analysis. The *easysorter* package is specifically developed for this type of data and includes reading and modifying of the raw data, pruning of anomalous data points, and regression of control and experimental phenotypes. Measured parameters include time-of-flight (animal length), extinction (optical density), and total object count (brood size) for each well. In short, the function *read_data* reads in raw phenotype data and runs a support vector machine to identify

and eliminate air bubbles, which can be confused with nematodes. In the next step, the function *remove_contamination* removes data obtained from microtiter wells that were manually identified to contain bacterial or fungal contamination. To generate summarized statistics for each well, the function *sum_plate* calculates the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} quantiles for all fitness parameters obtained. In this process, brood size is normalized by the number of animals sorted originally in each well. Next, biologically impossible data points passing certain cut-offs (n > 1000, n < 5, norm.n > 350) are eliminated (function *bio_prune*). We removed outlier replicates if they were outside 1.8 times the standard deviation of that strain's median phenotype of a particular trait. To adjust differences among replicates of each assay, a linear model was applied using the formula (phenotype ~ experiment + assay), which replaced the *easysorter* function *regress* (*assay* = *TRUE*). The experiment component of the linear model corresponds to two independent drug preparations of the same strains and the assay component corresponds to blocks of different strains performed across multiple weeks. In addition, the *bamf_prune* function of the *easysorter* package was skipped, because previous outlier removal based on technical replicates made this step unnecessary. Finally, drug-specific phenotype data is calculated using the *regress* (*assay* = *FALSE*) function of the *easysorter* package. This function fits a linear model with the formula (phenotype ~ control phenotype) to account for any differences in population parameters present in the control DMSO-only conditions.

Albendazole dose-response experiments

Albendazole (Sigma Aldrich Product #A4673) (ABZ) dose-response experiments were performed on a set of four genetically divergent *C. elegans* strains, including the laboratory strain N2 and three different wild isolates (CB4856, JU775, DL238), to determine suitable drug concentrations for subsequent GWA experiments. To this end, strain-specific drug responses were measured in four technical replicates using the HTA as described above at concentrations of 3.125 µM, 6.25 µM, 12.5 µM, and 25 µM ABZ. A suitable drug concentration for subsequent

GWA experiments was selected based on the lowest concentration in which we observed a significant difference in ABZ response among strains for brood size and animal length (S1-2 Data**)**.

Albendazole genome-wide association mappings

The GWA mappings were performed on the processed ABZ HTA phenotype data of 209 *C. elegans* wild isolates. In short, the *easysorter* processed phenotype data was analysed using the *cegwas* R package for association mapping [27]. This package uses the EMMA algorithm for performing association mapping and correcting for population structure [56], which is implemented by the *GWAS* function in the rrBLUP package [149]. In detail, the GWAS function was used with the following command: rrBLUP::GWAS (min.MAF = 0.05, P3D = FALSE). The kinship matrix used for association mapping was generated using whole-genome high-quality single-nucleotide variants (SNVs) [44] and the *A.mat* function from the rrBLUP package. All SNVs included in the marker set for GWA mapping had a minimum 5% minor allele frequency in the 240 isotype set [43]. Quantitative trait loci (QTL) were defined by at least one SNV that passed the Bonferroni-corrected threshold and were processed further using fine mapping, as described previously [31]. Computational fine mapping of the genomic regions of interest was performed as described previously [31]. We used the following linear model to correct for the presence of a putative *ben-1* LoF variant: $lm(animal\ length \sim (ben-1\ LoF))$. The list of strains that were considered to have putative LoF variants and the *ben-1*-corrected phenotype data are presented (S16 Data). We did not include CB4856 as a strain with a putative LoF variant because it has a 9 bp in-frame deletion near the end of the gene. We performed single-marker mappings as described above and gene-burden mappings as described below using these regressed phenotypes (S17-21 Data).

Albendazole burden mapping

Burden test analyses were performed using RVtests [279] and the variable-threshold method [280]. We called SNV using bcftools [87] with settings previously described [27,44,279]. We next performed imputation using BEAGLE v4.1 [281] with *window* set to 8000, *overlap* set to 3000, and *ne* set to 17500. Within RVtests, we set the minor allele frequency range from 0.003 to 0.05 for burden testing.

Generation of *ben-1* allele replacement and deletion strains

All *ben-1* edited strains assayed in this study were generated in the N2 background by CRISPR/Cas9-mediated genome editing, using a co-CRISPR approach [150] and Cas9 ribonucleoprotein (RNP) delivery [69]. For the *ben-1* F200Y allele replacement strains, sgRNAs for *ben-1* and *dpy-10* were ordered from Synthego (Redwood City, CA) and injected at final concentrations of 5 μM and 1 μM, respectively. Single-stranded oligodeoxynucleotides (ssODN) templates for homology-directed repair (HDR) of *ben-1* and *dpy-10* (IDT, Skokie IL) were used at final concentrations of 6 μM and 0.5 μM, respectively. Cas9 protein (IDT, Product #1074182) was added to the injection mixture at a concentration of 5 μM and incubated with all other components for ten minutes at room temperature prior to injection. All concentrations used for the sgRNA-mediated allele replacement were adapted from the work of Prior and colleagues [282]. N2 *ben-1* deletion strains were generated using two *ben-1* specific crRNAs synthesized by IDT (Skokie, IL), which targeted exon 2 and exon 4. For the injection mixture, *ben-1* crRNAs were used at final concentration of 8.3 μM each, mixed with *dpy-10* crRNA and tracrRNA (IDT, Product #1072532) at final concentrations of 1.2 μM and 17.6 μM, respectively, and incubated at 95°C for five minutes. After cooling to room temperature, Cas9 protein was added at a final concentration of 15.25 μM (IDT Product #1074181) and incubated for five minutes before *dpy-10* ssODN was added to a concentration of 5 μM.

RNP injection mixtures were microinjected into the germline of young adult hermaphrodites (P0) and injected animals were singled to fresh 6 cm NGM plates 18 hours after injection. Two days later, F1 progeny were screened, and animals expressing a Rol phenotype were transferred to new plates and allowed to generate progeny (F2). Afterwards, F1 animals were genotyped by PCR. For the *ben-1*(F200Y) allele replacement, ssODN repair templates contained the desired edit and a conservative change of the PAM site to prevent sgRNA:Cas9 cleavage. In addition, the PAM change introduced a new restriction site. PCRs were performed using the primers oECA1297 and oECA1298, and PCR products were incubated with BTsCI restriction enzyme (R0647S, New England Biolabs, Ipswich, MA) to identify successfully edited animals by differential band patterns. For genotyping of *ben-1* deletion strains, the primers oECA1301 and oECA1302 were used and successful deletions were identified by shorter PCR products. Non-Rol progeny (F2) of F1 animals positive for the desired edits were propagated on separate plates to generate homozygous progeny. F2 animals were genotyped afterwards, and PCR products were Sanger sequenced for verification. Generated *ben-1*(F200Y) replacement strains and *ben-1* deletion strains were phenotyped for ABZ response using the HTA as described above. All oligonucleotide sequences are listed in the supplement (**Table 4-2**).

Table 5-2 Oligonucleotides used for experiments discussed in Chapter 5

Name	Sequence	Use
crEC36	GCTACCATAGGCACCACGAG	crRNA to generate dpy-10(cn64) mutation
crECA38	GCTACCATAGGCACCACGAG	sgRNA to generate dpy-10(cn64) mutation
dpy-10 repair	CACTTGAACTTCAATACGGCAAGATG AGAATGACTGGAAACCGTACCGCAT GCGGTGCCTATGGTAGCGGAGCTT CACATGGCTTCAGACCAACAGCCTAT	Repair template to generate dpy-10(cn64) mutation
crECA39	GTCAATGCAGAAAGTCTCAT	sgRNA to generate the ben-1 F200Y allele swap in N2
crECA41	AAACTAATAATTAAAATTCAGGTCTCG GACACAGTCGTCGAGCCATACAACG CTACTCTTTCTGTCCACCAGCTCGTT GAAAATACGGATGAGACTTACTGCAT TGACAACGAGGCTCTTTATGATA	repair template to generate the ben-1 F200Y allele swap in N2
crECA44	AGATCGACGAGAACAGCGCG	crRNA to generate a ben-1 deletion. Used in combination with crECA45
crECA45	CAGGTACTATTATGGAGACG	crRNA to generate a ben-1 deletion. Used in combination with crECA44
oECA1301	TGGAGCACGTTTCCAGCAAC	forward primer to confirm ben-1 deletion
oECA1302	ATTCGCGCTCTATAGCAACCA	reverse primer to confirm ben-1 deletion
ddPCR-forward	GGCAAGATGAGAATGACTGGAAAC	used for ddPCR
ddPCR-reverse	GTGAAGCTCCGCTACCATAGG	used for ddPCR
Taqman probe	CACCACGAGCGGTACG	taqman probe used to quantify allele frequencies in the competition

VIC		assay
Taqman probe FAM	CACCACGAGCAGTACG	taqman probe used to quantify allele frequencies in the competition assay

Competition assays

Pairwise multi-generation competition assays were performed between a *ben-1* wild-type strain PTM229 and the *ben-1* edited strains, containing either the F200Y allele replacement or a *ben-1* deletion. All strains were generated in the N2 background. For the assay, strains were bleach-synchronized and embryos were transferred to 10 cm NGM plates. 48 hours later, seven L4 larvae per strain were transferred to 50 fresh 6 cm NGM plates containing either 1.25 μM ABZ for drug selection or DMSO as control. The ABZ concentration was determined by dose-response assays beforehand as the lowest concentration that caused a developmental delay in PTM229 and N2 when added to NGM. For each strain combination, 50 plates were grown, representing 10 technical replicates of five independent populations. Plates were grown for one week until starvation. Animals were transferred to fresh plates by cutting out a 0.5 cm^3 agar chunk. After every culture transfer, starved animals were washed off the plates with M9, and DNA was collected using the Qiagen DNeasy Kit (Catalog #69506). Allele frequencies of PTM229 compared to wild-type N2 or the *ben-1* edit strains in each replicate populations were measured using Taqman analysis in a Bio-Rad QX200 digital droplet PCR system. Digital PCR was performed following the standard protocol provided by Bio-Rad with the absolute quantification method. To calculate the relative fitness w of the competitive strains, we used linear regression to fit the relative allele frequencies at the first, third, fifth, and seventh weeks into a one-locus genic selection model [283]. All oligonucleotide sequences are listed in the supplement (S2 Table).

Computational modeling of *ben-1* variants

We obtained the predicted BEN-1 peptide sequence from WormBase [275]. We used the online utility PHYRE2 to generate a homology model of the predicted BEN-1 peptide [284]. We specified the intensive homology search option for running PHYRE2. Visualization of the BEN-1 homology model and highlighting of the variants of interest was performed using PyMol [285].

Statistical analyses

Phenotype data are shown as Tukey box plots and p-values were used to assess significant differences of strain phenotypes in allele-replacement experiments. All calculations were performed in R using the *TukeyHSD* function on an ANOVA model with the formula (*phenotype ~ strain*). P-values less than 0.05 after Bonferroni correction for multiple testing were considered to be significant.

Population genetics

Sliding window analysis of population genetic statistics was performed using the PopGenome package in R [286,287]. All sliding window analyses were performed using the imputed SNV VCF available on the CeNDR website with the most diverged isotype XZ1516 set to the outgroup [27,281,288]. Window size was set to 100 SNVs with a slide distance of one SNV. The Tajima's *D* calculation, using all of the manually curated variants in the *ben-1* locus, was performed using modified code from the developers of the LDhat package [289]. Estimates of Ka/Ks were performed using an online service (http://services.cbu.uib.no/tools/kaks). Multiple sequence alignments were performed using MUSCLE [290] through the online service (https://www.ebi.ac.uk/Tools/msa/muscle/). The β-tubulin neighbor-joining phylogeny construction was performed using Jalview 2 [291]. Linkage disequilibrium (LD) of QTL markers was calculated using the *genetics* [292] package in R. LD calculations are reported as $r =$

$-D \,/\, sqrt(\, p(A) \,*\, p(a) \,*\, p(B) \,*\, p(b)\,)$where $D \;=\; p(AB) \;-\; p(A) * p(B)$.*r = -D / sqrt(p(A) * p(a) * p(B) * p(b))*

Results

Genetically distinct *C. elegans* natural isolates respond differently to ABZ

We investigated the differences in responses across a panel of *C. elegans* wild strains to one of the most commonly used BZ compounds (albendazole, ABZ) in human and veterinary medicine [27,44,253]. To test the effects of ABZ treatment on *C. elegans*, we exposed four genetically divergent *C. elegans* isolates to various ABZ concentrations and measured the number and length of progeny the animals produced. We used these two traits because they are both reliable indicators for drug efficacy against nematodes. While brood size reflects the reproductive success under ABZ treatment, animal length indicates to influence of ABZ on the morphological development of the progeny. After four days of exposure to ABZ, we detected a decrease in brood size and progeny length for all four assayed strains (**S1 Fig; S1-2 Data**). Additionally, this assay revealed differential ABZ sensitivity among these wild strains, as measured by brood size and progeny length. For subsequent GWA mapping experiments, we used an ABZ concentration of 12.5 µM, which is a concentration that induced robust differences in ABZ responses among the four assayed wild strains.

Natural variation in *C. elegans* ABZ responses maps to multiple genomic regions, including the *ben-1* locus

To identify genomic loci that underlie strain-specific ABZ responses, we performed GWA mappings using HTA fitness data obtained from exposing 209 *C. elegans* wild isolates to either DMSO (control) or ABZ and DMSO conditions. We employed two statistical methods to perform

GWA mappings using the regressed animal length (q90.TOF) and normalized brood size (norm.n) traits, a single-marker and a gene-burden approach [279,280]. Using the single-marker approach [31,44,56,149], we identified three distinct QTL that explained variation in animal length among wild isolates exposed to ABZ, whereas the brood-size trait did not map to any significant genomic loci. Two of the animal-length QTL we identified are located on chromosome II, and the third is located on chromosome V (Figure 5-1A). The QTL on the left arm of chromosome II spans from 25 kb to 3.9 Mb and has a peak-marker position at 458 kb. The second QTL on chromosome II has a peak-marker position at 11 Mb and spans from 10.17 Mb to 11.6 Mb. The QTL on the right arm of chromosome V spans from from 18.04 Mb to 18.68 Mb, with the peak marker located at 18.35 Mb. Remarkably, the β-tubulin gene *ben-1*, shown to be the major determinant for ABZ resistance in the *C. elegans* laboratory strain N2 [261,262], did not map using this single marker-based approach (Figure 5-1A).

To complement the single-marker mapping described above, we performed a gene-burden approach to map BZ resistance [279,280]. Compared to the single-marker approach that includes SNVs with a minimum 5% minor allele frequency among all strains, the gene-burden approach incorporates all rare strain-specific variation within a gene for association testing [279,293]. With this alternative mapping strategy, we identified the same two QTL on chromosome II that we identified with the single-marker approach. In addition to the chromosome II QTL, we found a significant association between variation at the *ben-1* locus on chromosome III and animal-length variation in response to ABZ (Figure 5-1B). These results suggest that no variants are shared above 5% minor allele frequency among the wild isolates at the *ben-1* locus (chromosome III, 3,537,688-3,541,628 bp). An additional significant gene was detected on the right arm of chromosome I.

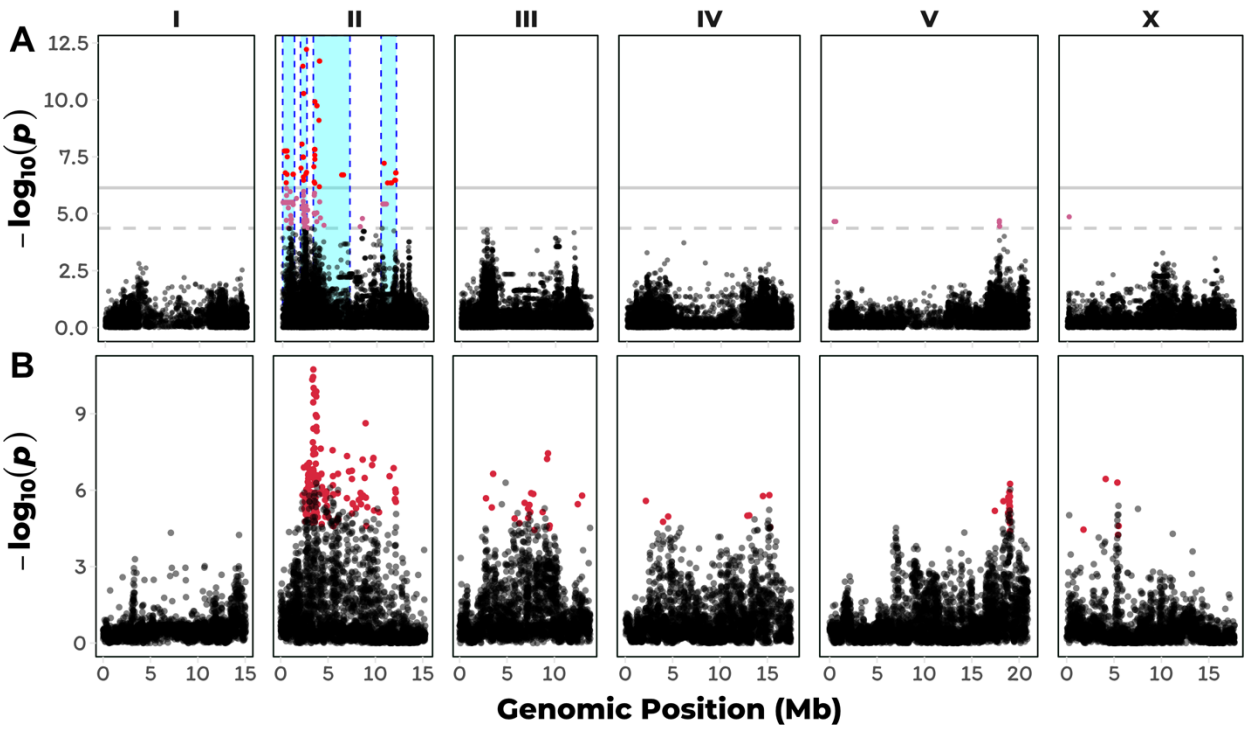

Figure 5-1 GWA and burden mapping of albendazole response variation
A) The single-marker based Manhattan plot for animal length (q90.TOF) in the presence of ABZ is shown. Each dot represents an SNV that is present in at least 5% of the assayed population. The genomic location of each SNV is plotted on the x-axis, and the statistical significance of the correlation between genotype and phenotype is plotted on the y-axis. SNVs are colored red if they pass the genome-wide Bonferroni-corrected significance (BF) threshold, which is denoted by the gray horizontal line. SNVs are colored pink if they pass the genome-wide eigen-decomposition significance (ED) threshold, which is denoted by the dotted gray horizontal line. The genomic region of interests surrounding the QTL that pass the BF and ED thresholds are represented by cyan and pink rectangles, respectively. In total, four QTL were identified, located on chr II. B) A genome-wide Manhattan plot of animal length (q90.TOF) after ABZ exposure based on the gene-burden mapping approach is shown. Each dot represents a single *C. elegans* gene with its genomic location plotted on the x-axis, and the statistical significance of the correlation between genotype and phenotype plotted on the y-axis. Genes passing the burden test statistical significance threshold are colored in red. The burden test analysis identified genes significantly correlated with ABZ resistance on chr II, chr III, chr IV, chr IV, and chr X.

C. elegans ABZ resistance correlates with extreme allelic heterogeneity at the *ben-1* locus

To explain the differences between the single-marker and gene-burden based GWA mapping results with respect to *ben-1*, we investigated the natural variation found at this genomic locus in more detail (Figure 5-2). We first used the *snpeff* function of the *cegwas* package [27,81] to look for SNVs in *ben-1* with predicted moderate-to-high effects on gene function [27,81]. These

moderate-to-high impact variants include missense variants, splice donor and acceptor variants, and alternative start and stop codons that might disrupt the open reading frame of *ben-1*.

In the set of 209 strains used for the GWA mapping, we identified 18 strains with moderate-to-high impact variants in the *ben-1* locus, as predicted by *snpEff* [81], including twelve strains with amino-acid substitutions. Amino-acid substitutions in β-tubulin genes are important markers for BZ resistance in parasitic nematodes and are hypothesized to reduce binding affinity of BZs to β-tubulin [265,266]. In the *C. elegans* panel of wild strains, we identified six strains that have the F200Y mutation in BEN-1, which is known to be the most common BZ resistance marker in livestock parasites [262]. Additionally, we detected novel amino-acid substitutions in BEN-1, which have not previously been associated with ABZ resistance in nematodes. These missense variants include an A185P substitution variant, found in two strains, as well as the E69G, Q131L, S145F, M257I, and D404N substitution variants, all of which were unique to single strains. In contrast to the amino-acid variants F200Y, E198A, and F167Y found in parasitic nematodes, which cluster at the putative ABZ binding site [294–296], these novel variants are more distributed throughout the protein structure. We classified a strain as resistant to ABZ if its animal-length phenotype after ABZ exposure was greater than 50% of the difference between the most and least ABZ-resistant strains (Figure 5-2A). Eleven of the twelve strains with amino-acid substitutions are more resistant to ABZ than strains with no variation at the *ben-1* locus. Similarly, strains with predicted high-impact variants in *ben-1* are resistant to ABZ treatment. The strains with high-impact variants include five strains with unique premature stop codons and one strain with a predicted splice-donor variant at the end of exon 1. However, many ABZ-resistant strains in the *C. elegans* population do not contain any of these rare and common variants. Therefore, these strains either have other types of deleterious variants at the *ben-1* locus or are resistant to ABZ through a distinct mechanism. To differentiate these two possibilities, we manually curated all of the strains raw sequence read alignment files (BAM

files), publicly available through the CeNDR website [27]. This in-depth investigation revealed extreme allelic heterogeneity at the *ben-1* locus in the natural *C. elegans* population. Twenty-six strains had either rare deletions or insertions in *ben-1* (20 deletions, 4 insertions; S9 Data). Interestingly, a subset of individuals with missense variants described above were found to be in perfect linkage disequilibrium (LD) with nearby deletion alleles. All six strains with the F200Y allele share the same deletion of approximately 160 bp that partially removes exons 3 and 4. The close proximity of the annotated F200Y to this 160 bp deletion suggested that the F200Y variant is actually an error in read alignment, which we confirmed by manual review of the BAM files of corresponding strains (JU751, JU830, JU2581, JU2587, JU2593, JU2829) (genome browser on CeNDR) [27]. The majority of paired-end reads at this genomic position that lead to the call of the F200Y allele in *ben-1*, map with their mate reads to the locus of the β-tubulin gene *tbb-2*, which is in relative close genomic distance (chromosome III, 4,015,769 - 4,017,643 bp). Additionally, the amino acid substitution (E69G), which is only observed in a single isotype (KR314), co-occurs with a deletion in exon 2 that is predicted to cause a frameshift in the *ben-1* open reading frame.

Of all the strains with *ben-1* SNVs (12 strains) and indels (24 strains), only three individuals were not classified as resistant to ABZ treatment. The least resistant of these three strains was the CB4856 strain, which contains a nine base pair deletion near the end of the *ben-1* coding sequence. Because this in-frame deletion variant is at the end of the coding sequence, we hypothesized that CB4856 still contains a functional copy of the *ben-1* gene and this variant likely does not confer resistance to ABZ. The next least resistant strain is QG2075, which shares a four base pair deletion in exon 1 of *ben-1* with the ABZ-resistant isotype WN2033. During the growth phase of our HTA assay, we noted that the QG2075 strain has a slow-growth phenotype in normal growth conditions, which is likely confounding our phenotypic measurements in ABZ. Finally, the JU2862 strain has a unique four base pair deletion in exon 4

of *ben-1*, that is predicted to cause a frameshift in the *ben-1* open reading frame. We note that JU2862 is right at our arbitrary ABZ-resistance threshold and upon re-phenotyping with higher replication might be classified as ABZ resistant.

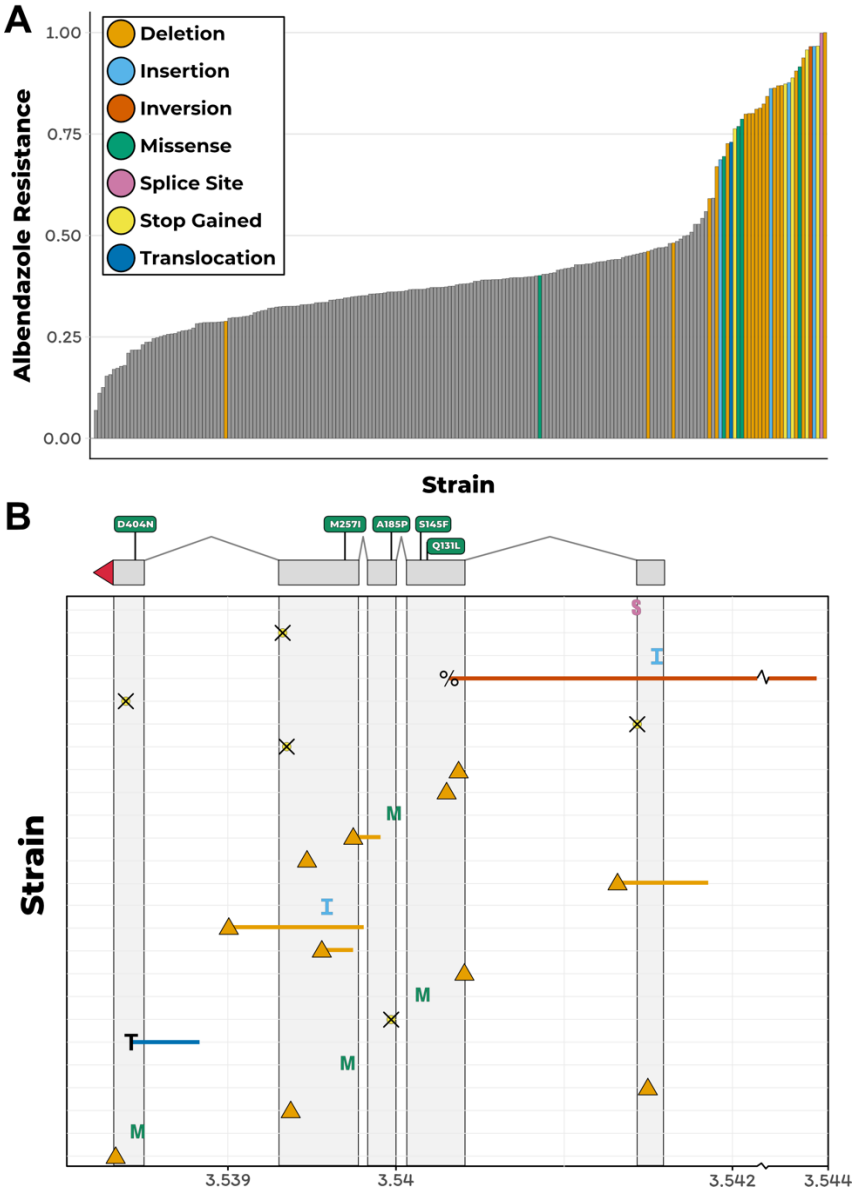

Figure 5-2 *C. elegans* albendazole response phenotype and the distribution of *ben-1* alleles
A) A bar plot for the phenotypic response of *C. elegans* strains under ABZ exposure is shown. Each bar represents a single wild strain included in the HTA. Strains are sorted by their relative resistance to ABZ based on the mean animal length (q90.TOF) from four replicate measures. All strains that were found to have variants with predicted moderate-to-high impact and/or structural variation are colored by their specific type of variant. Strains similar to the N2 reference genome with respect to the *ben-1* locus are shown in grey. (B) An overview of variants found in the *ben-1* locus among C. elegans wild strains is shown. The genomic position of each variant is shown on the x-axis (S: splice donor variant; M: missense variant leading to amino acid substitution; crossed circle (⊗): alternative stop codon; triangle: deletion; I: insertion; percent sign: inversion; T: transposon insertion. Colors correspond to colors in (A). Each line on

the y-axis corresponds to a strain(s) with a *ben-1* in increasing order by their length in ABZ. For variants shared by two or more strains, the mean phenotype value of all corresponding strains determines their placement on the y-axis. A representation of the *ben-1* gene model is shown above the variant summary panel, including 5' UTRs (red triangle), five exons (gray rectangles), and four introns (thin lines). Additionally, missense variants leading to amino acid substitutions are highlighted with their exact position and the corresponding amino acid exchange in the single letter code.

Among the remaining ABZ-resistant strains, we identified a putative transposon insertion in exon 5 of *ben-1* in the JU3125 strain. The genomic origin of this putative transposon insertion is from position 17.07 Mb on chromosome X and corresponds to a cut and paste DNA transposon Tc5B, which is part of the TcMar-Tc4 transposon superfamily [297,298]. Finally, the MY518 strain is resistant to ABZ but did not contain any of the above classes of variation. However, we did identify a 1 kb inversion that spans exon 1 and the promoter region of *ben-1*. Remarkably, all structural variants present in *ben-1* are only present in one or few wild strains and are mostly predicted to cause loss of *ben-1* gene function by coding for a non-functional β-tubulin (Figure 5-2). Altogether, the putative loss-of-function variants described above explain 73.8% of the phenotypic variation present in the natural *C. elegans* population.

Within-species selective pressures at the *ben-1* locus

The presence of multiple low-frequency *ben-1* alleles that confer resistance to ABZ treatment suggests that selection might have acted on this locus within the *C. elegans* population. To determine if this hypothesis is possible, we calculated the Ka/Ks ratio between *ben-1* genes from *C. elegans* and from two diverged nematode species *C. remanei* and *C. briggsae* [299]. The Ka/Ks ratio between the *ben-1* coding sequences of *C. elegans* and these two diverged species is ~0.008, which indicates that the evolution of this locus is constrained across species. However, among the 249 wild isolates within the *C. elegans* species [27], we identified 10 synonymous and 22 nonsynonymous, stop-gained, or splice-site variants in the *ben-1* locus when compared to the N2 reference genome [275]. We note that the synonymous variants identified in *ben-1* are primarily found in wild strains that contain putative loss-of-function alleles.

These results indicate that, despite the evolutionary constraint we observe at the *ben-1* locus across nematode species, an excess of potentially adaptive mutations (~2.2X nonsynonymous variants) have arisen within the *C. elegans* species. These results are consistent with our estimates of Tajima's *D* [109] at the *ben-1* locus. When we considered all variant types across coding and non-coding regions of *ben-1*, we found Tajima's *D* to be -2.02. When we only consider putative loss-of-function variants in the *ben-1* locus, the Tajima's *D* estimate is -2.59. This large negative value of Tajima's *D* likely reflects the high number of rare *ben-1* alleles present in the *C. elegans* population. We next calculated Tajima's *D* for the genomic region surrounding *ben-1* (Figure 5-3A). These results show a strong negative dip in Tajima's *D* (< -1.5) between 3.50 and 3.55 Mb, which is the region directly surrounding *ben-1* (3.537 - 3.541 Mb). This observation indicates that the *C. elegans* population has undergone a population expansion or that the region surrounding *ben-1* may have been subject to selective pressures. To differentiate between these two possibilities, we calculated Fay and Wu's *H* [300] and Zeng's *E* [301] for *ben-1* (Figure 5-3A). We noticed that these two statistics showed peaks around the *ben-1* locus, which do not correspond to our observations of multiple low minor allele frequency alleles. However, when we consider *ben-1* coding variation and non-coding variation separately, we found the H_{coding} statistic to be 0.34 and the $H_{noncoding}$ statistic to be -5.5, which indicate that when considering high frequency derived alleles, the coding sequence of this locus is evolving neutrally and that there are many high-frequency derived alleles in the intronic and UTR regions. By contrast, we found the E_{coding} statistic, which considers low and high frequency alleles, to be -1.7 and the $E_{noncoding}$ to be 3.4. Taken together, these results indicate that recent selective pressures have acted on the *ben-1* locus.

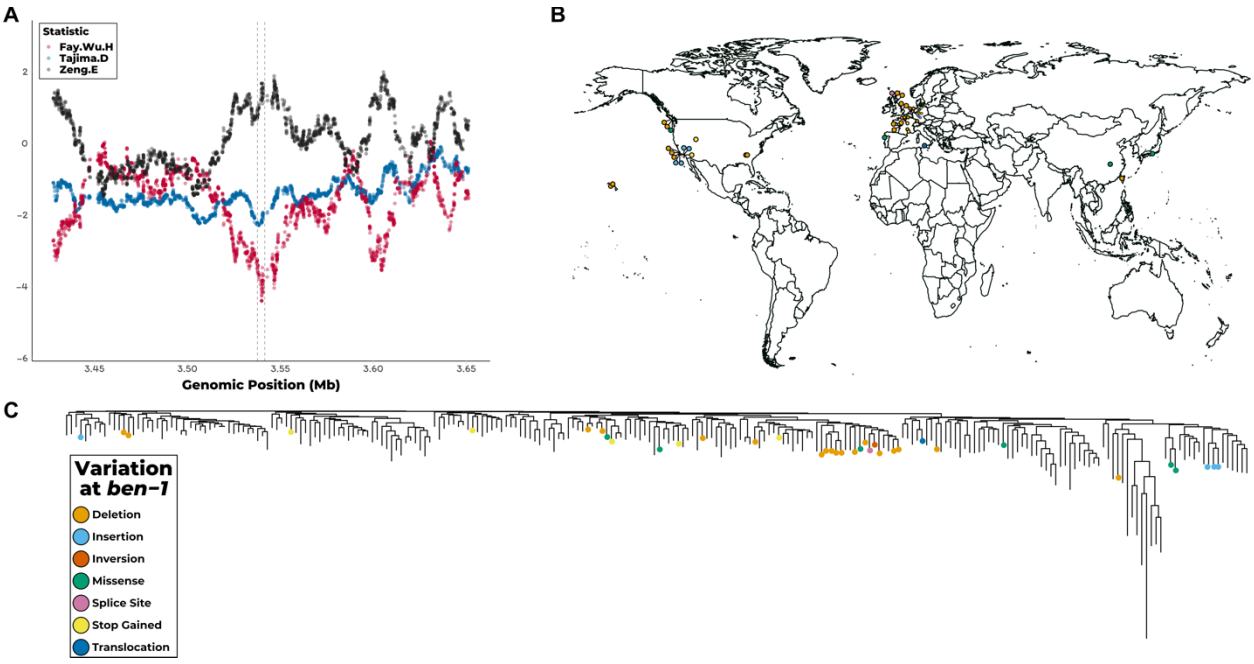

Figure 5-3 Recent selection at the *ben-1* locus, the global and phylogenetic distribution of *ben-1* alleles
A) Fay and Wu's *H* (red), Tajima's *D* (blue), and Zeng's *E* (black) for the genomic region surrounding *ben-1*. Genomic position is shown on the x-axis, and the value for each displayed statistic is shown on the y-axis. The neutrality statistics were calculated using a sliding window approach (100 SNVs window size and 1 SNV slide distance) using only SNV data. (B) The global distribution of strains that contain moderate-to-high predicted variation in *ben-1*. Each dot corresponds to the sampling location of an individual strain and is colored by the type of variant discovered in the *ben-1* locus. (C) The genome-wide phylogeny of 249 *C. elegans* strains showing that variation in the *ben-1* locus occurred independently multiple times during the evolutionary history of the species. The dots on individual branch nodes correspond to strains with variation in *ben-1* and have the same color code as in panel B.

If localized selective pressures, possibly caused by increased levels of environmental BZ, drove the loss of *ben-1* function in a subset of wild strains, then we could expect to see geographic clustering of resistant strains. When we looked at the distribution of strains with putative *ben-1* loss-of-function variants, we observed no trend in the sampling locations of these individuals (Figure 5-3B). However, we do note that highly diverse strains sampled from various locations on the Pacific Rim [27,43,44] do not harbor any putative *ben-1* loss-of-function alleles (Figure 5-3B). This diverse set of strains is hypothesized to represent the ancestral state of *C. elegans*, because they do not show signs of chromosome-scale selective sweeps throughout their genomes [43]. The lack of *ben-1* loss-of-function variants in these ancestral strains suggests

that variation in *ben-1* arose after individuals in the species spread throughout the world. We next considered that strains with putative *ben-1* loss-of-function alleles might have been isolated from similar local environments and substrates. Strains that are predicted to have a functional *ben-1* gene have been isolated from a greater diversity of environmental sampling locations and sampling substrates than strains with predicted loss-of-function variants in *ben-1*. However, a hypergeometric test for enrichment within specific locations and substrates showed no signs of significant enrichment. This lack of enrichment might be caused by biases in global coverage of *C. elegans* sampling and low sampling density. Nevertheless, the observation that the diversity of *ben-1* alleles arose independently on various branches of the *C. elegans* phylogeny lends support to the hypothesis that local selective pressures have acted on the *ben-1* locus (Figure 5-3C).

ben-1 natural variants confer BZ resistance to sensitive *C. elegans* strains

The majority of the variants we observe at the *ben-1* locus are predicted to result in the loss of gene function. Our findings in *C. elegans* are in contrast to findings in parasitic nematode populations that describe the F200Y and other missense variants as the major variants contributing to BZ resistance [262]. To test whether the putative loss-of-function variants in *ben-1* that we observe in the *C. elegans* population confer the same level of BZ resistance as the known F200Y allele, we introduced a *ben-1* deletion in the N2 strain using CRISPR/Cas9 (referred to as Del). In addition to the Del strain, we introduced the F200Y *ben-1* allele into the N2 strain (referred to as F200Y) to test whether this allele confers BZ resistance in *C. elegans* and to compare the levels of BZ resistance between the loss-of-function and missense alleles. We exposed the N2 parental, Del, and F200Y strains to 12.5 µM ABZ using the HTA described above. Both the Del and F200Y variants conferred ABZ resistance to the otherwise sensitive N2 parental strain (Figure 5-4A). Interestingly, we found no significant difference in BZ resistance between the Del and F200Y strains (*p*-value = 0.99, TukeyHSD). These results show that

putative loss-of-function variants of *ben-1* and known parasitic BZ resistance alleles confer BZ resistance in otherwise isogenic *C. elegans* genetic backgrounds.

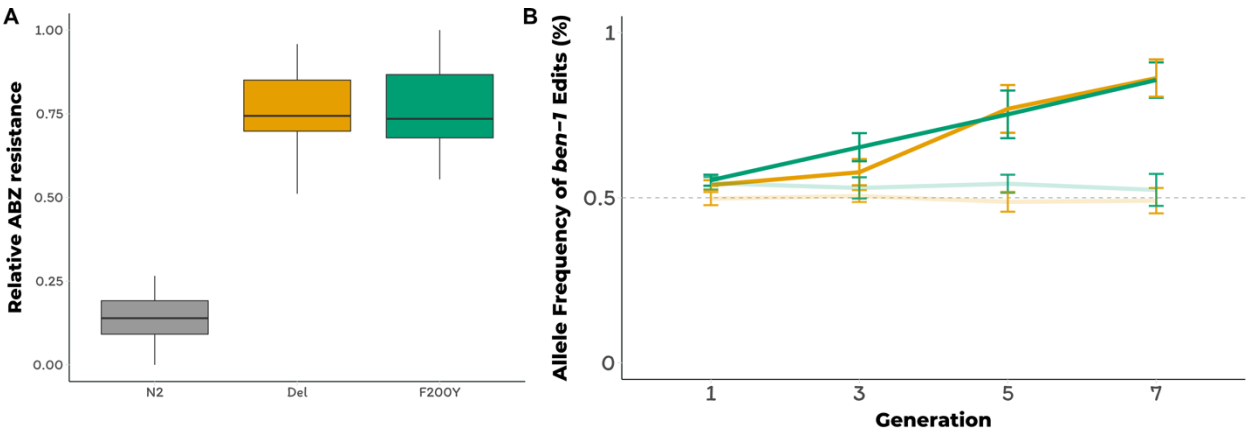

Figure 5-4 F200Y allele replacement and *ben-1* deletion confer resistance to the ABZ sensitive N2 strain
A) Tukey box plots of the animal length phenotypes for at least 100 replicates of the generated *ben-1* allele-replacement strains after ABZ treatment is shown. The y-axis represents the animal length phenotype after correcting for growth in DMSO conditions. Both *ben-1* allele strains are significantly more resistant to ABZ treatment than the N2 strain (*p*-value < 1E-10, TukeyHSD). B) The results from a multi-generation competition experiment between a barcoded N2 strain and the *ben-1* allele-replacement strains are shown. The generation number is shown on the x-axis, and the allele frequency of the ben-1 allele-replacement strains is shown on the y-axis.

To complement the results from the liquid-based HTA experiments, we performed a plate-based competition assay. In the competition experiments, we individually competed the F200Y, Del, and parental N2 strains against an N2 strain that contains a barcode sequence (PTM229). We quantified the relative allele frequencies of the barcoded strain for the first, third, fifth, and seventh generations of the competition assay (Figure 5-4B). Throughout the competition assay, and for all strains tested, the relative frequencies of the barcoded strain did not significantly deviate from the initial frequency when grown on DMSO plates. These results suggest that in standard laboratory conditions, the BEN-1 F200Y and *ben-1* deletion alleles do not have fitness consequences. We observed the same trend when we competed the N2 and the barcoded N2 strains on ABZ plates. However, the allele frequencies of the barcoded strain dropped to ~20% when competed against strains that contain either of the two *ben-1* alleles on ABZ plates (relative fitness w = ~ 1.3, both). These two independent assays show that the two *ben-1* alleles

confer BZ resistance with no negative fitness consequence under standard laboratory growth conditions.

Additional genomic intervals contribute to ABZ resistance in the *C. elegans* population

Above, we showed that extreme allelic heterogeneity at the *ben-1* locus explains 73.8% of the phenotypic variation in response to ABZ treatment. To identify potential genomic loci that contribute to the remaining 26.2% of the observed BZ response variation, we statistically corrected the animal length phenotype in response to ABZ treatment by using the presence of a putative loss-of-function variant at the *ben-1* locus in a strain as a covariate for linear regression. After correcting for variation at the *ben-1* locus, we performed GWA mapping using both the single-marker (Figure 5-5) and gene-burden approaches. Both of these approaches resulted in the disappearance of the two QTL on chromosome II and the QTL on chromosome V (Figure 5-5). This result suggests that the cumulative variation at the *ben-1* locus is in complex interchromosomal linkage disequilibrium (LD) with these three loci, despite the loci on chromosomes II and V not being in strong LD with each other. Interestingly, using the single-marker GWA-mapping approach, we found additional genomic loci on chromosomes III and X that were significantly associated with ABZ resistance (Figure 5-5). However, the gene-burden based approach did not result in significant associations between ABZ resistance and rare variation within any genes. These results suggest that ABZ resistance in the *C. elegans* population is caused by at least two loci because additional genetic variation present in the *C. elegans* population contributes to phenotypic variation in response to ABZ treatment.

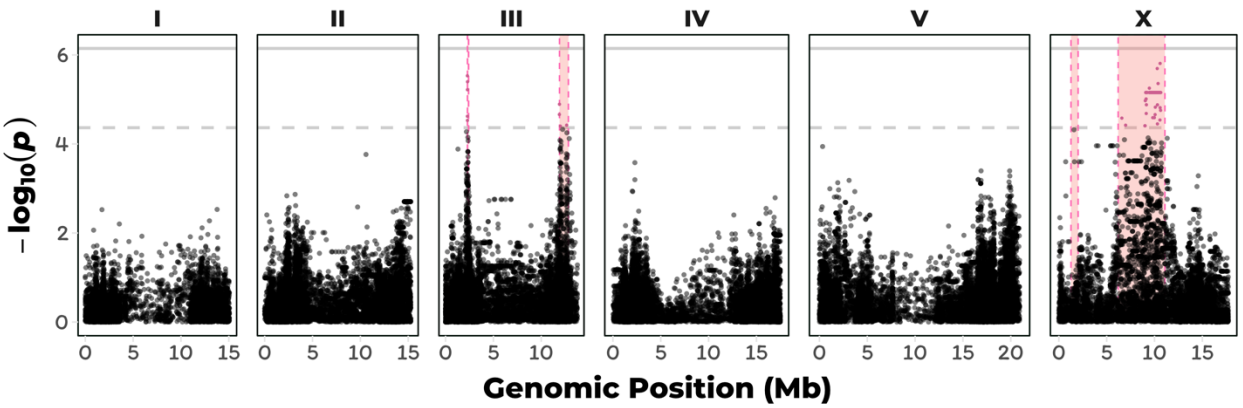

Figure 5-5 Regression of the putative *ben-1* LoF variants identifies novel QTL
(A) Manhattan plot of the single-marker based GWA mapping showing nominally significant QTL on chromosomes III and X are associated with residual phenotypic variation in response to ABZ treatment. Each dot represents an SNV that is present in at least 5% of the assayed population. The genomic location of each SNV is plotted on the x-axis, and the statistical significance is plotted on the y-axis. SNVs are colored pink if they pass the eigen-corrected significance threshold, which is shown by a gray horizontal dashed line. The genomic region of interest is represented by pink rectangle surrounding QTL.

Discussion

ABZ is a broadly administered BZ used to treat parasitic nematode infections in humans and livestock [253,302]. Already a significant problem in veterinary medicine, the heavy reliance of ABZ and other BZ compounds in MDA programs, the small repertoire of other anthelmintic compounds, and high rates of re-infection of parasitic nematodes in endemic regions around the world have raised the fear of the emergence of BZ resistance among parasitic nematode populations that infect humans [253,303]. Though the BZ-resistance alleles found in model organisms through classical genetic screens have had little direct translatability to parasitic nematodes, these classic alleles have been instrumental toward the elucidation of the mechanism of action of BZs [253,304]. Recent advances in sequencing technologies have enabled researchers to take a quantitative genetics approach to search for novel mechanisms of anthelmintic resistance in natural parasitic and non-parasitic nematodes [27,138,206,278,305,306]. In the present study, we leveraged genetic diversity within the *C. elegans* population to study the genetic basis of ABZ resistance within this species.

To identify the genetic basis of ABZ resistance in *C. elegans*, we used two unbiased genome-wide association approaches, single-marker [56,149] and gene-burden [280] based mappings. We found that the *C. elegans* population harbors extreme allelic heterogeneity at the *ben-1* locus, and that this variation contributes to differential ABZ resistance among wild isolates (Figure 5-1B, Figure 5-2A). The variants in the *ben-1* locus that contribute to ABZ resistance are present at low allele frequencies (<0.05 MAF) within the global *C. elegans* population and arose independently during the evolutionary history of the species (Figure 5-3). As a result of this complex demographic history, common variants (>0.05 MAF) used as genetic markers in the mappings near the *ben-1* locus are not in LD with the *ben-1* alleles that confer ABZ resistance. Therefore, we were unable to detect an association between the *ben-1* locus and ABZ response with the single-marker based mapping approach. We note that burden-based mapping approaches are only possible in species with well annotated genomes and genome-wide genetic variation, such as *C. elegans* and other model organisms. To extend this type of analysis to parasitic species, more effort will need to be made toward improving parasitic nematode reference genomes and the accumulation of genome-wide variation resources [86,307]. In addition to the strong association between ABZ responses and variants in the *ben-1* locus, we identified a novel ABZ-response QTL on chromosome X after accounting for variants in the *ben-1* locus (Figure 5-5). The identification of this QTL suggests that mechanisms independent of β-tubulin could account for BZ resistance in the *C. elegans* and in parasitic nematode populations.

The majority (22/25) of *ben-1* alleles that correlate with ABZ resistance are predicted to result in the loss of *ben-1* function (Figure 5-2). These results suggest that the function of the *ben-1* gene is not essential to the survival of *C. elegans* in natural habitats. This observation is in close agreement with previous work using the laboratory-derived N2 strain, which showed that

putative *ben-1* loss-of-function alleles do not confer observable defects in standard laboratory growth conditions [261]. However, our analysis of Ka/Ks between the *C. elegans ben-1* and the distantly related *C. briggsae* and *C. remanei* β-tubulin coding sequences suggest that evolution of this gene is highly constrained. A possible explanation for this discrepancy is that the *ben-1* gene has a highly specialized function in natural settings and that survival in the presence of a strong selective pressure, such as the presence of increased concentrations of environmental BZs, outweighs the necessity of this specialization. A possible source of environmental BZs are microbes that *C. elegans* might contact in their natural habitat. For example, 5,6-dimethylbenzimidazole is a BZ derivative that is produced naturally by prokaryotes [308]. However, whether these BZ derivatives bind to and inhibit β-tubulins remains unclear. Alternatively, natural selection on *C. elegans ben-1* might occur by contamination of nematode niches by synthetic BZs produced by humans. Initially developed as fungicides in the 1960s, several BZs, including thiabendazole, benomyl, and carbendazim are extensively used to treat crops, fruits, and grains [309]. In this context, BZ are broadly discussed as potential pollutants of soil and freshwater [310–312]. Degradation of some BZs is inhibited by their weak solubility in water and their sorption to soil particles, which can lead to a long persistence in the environment. Although depending on multiple factors like soil pH, the half time of carbendazim in soil can reach between 30 to 180 days [312]. Thiabendazole, another environmentally stable BZ, is intensively used in the post-harvest treatment of fruits and can reach half times of more than one year [311]. It was detected in biological waste samples in concentrations of up to 600 µg/kg [313]. A third source of environmental BZs might be runoff from livestock farms because they are used extensively to prevent parasite infections in these animals [310–312,314]. The observation that putative *ben-1* loss-of-function alleles are only found at low frequencies in the *C. elegans* population might suggest that individuals with these alleles are eventually removed from the population, though we see no evidence for a decrease in fitness in laboratory competition experiments, which might not recapitulate natural settings (Figure 5-4). Taken

together, our observations indicate that the independent putative *ben-1* loss-of-function alleles might have arose recently within the *C. elegans* population. Considering these findings in retrospect, it is extremely fortunate that N2 was used for initial studies of BZ sensitivity. Rapid progress on the mechanism of action for BZ compounds from the free-living *C. elegans* model to parasites might have not occurred otherwise, strongly illustrating the pitfalls associated with the study of a single genetic background to elucidate anthelmintic resistance.

The remaining (3/25) *ben-1* alleles we found to be correlated with ABZ resistance in the *C. elegans* population result in missense variants. These missense variants are S145F, A185P, and M257I and are found in one, two, and one *C. elegans* strain, respectively. Remarkably, substitutions of the amino acid residues A185 (A185S) and M257 (M257L) were previously described to confer BZ resistances in *Tapesia yallundae* and *Aspergillus nidulans*, respectively [315,316]. Of these two previously identified residues, M257 is postulated to directly interact with BZs, because it is in close three-dimensional proximity to the known BZ interacting F200 residue [296]. The S145F allele is in close proximity to the highly conserved GGGTGS motif of the GTP binding and hydrolysis site [317]. A fourth missense mutation Q131L, which is present in a strain that was not phenotyped in our mapping study because it grows slowly under normal laboratory conditions, is located near the β-tubulin/α-tubulin interaction interface [318,319]. Upon re-phenotyping this isotype with the Q131L variant, we found that it is resistant to ABZ treatment. A final missense variant (D404N) present in the *C. elegans* population was not correlated with ABZ resistance (Figure 5-2). Though theoretical structure-based evidence suggests that most of the missense variants in the *C. elegans* population associated with ABZ resistance cause a non-functional β-tubulin, further experiments are necessary to confirm these results.

The high prevalence of putative *ben-1* loss-of-function variants in the natural *C. elegans* population stands in stark contrast to the relative paucity of allelic diversity found in orthologous β-tubulin genes in parasitic nematodes. Perhaps parasitic nematode species do not have similarly high levels of functional redundancy of β-tubulins as has been observed in *C. elegans*. The main *C. elegans* β-tubulins are *tbb-1* and *tbb-2*, which are both ubiquitously highly expressed, functionally redundant in laboratory conditions, and contain a tyrosine at position 200 [320–322]. The co-expression of *tbb-1*, *tbb-2*, and *ben-1* in the nervous system [323] suggests that only one β-tubulin with F200 is required to confer sensitivity to BZs. Two other genes that encode for β-tubulins in *C. elegans* are *mec-7* and *tbb-4*. Both genes are only expressed in a subset of neuronal cells involved in chemo- and mechanosensation, and both have a phenylalanine at amino acid position 200 [322,324–326]. *tbb-6* encodes for a highly diverged *C. elegans* β-tubulin that could be a stress-resistant tubulin because it is highly expressed during the unfolded protein response [327]. Despite having phenylalanine at amino acid position 200, these tubulins (TBB-4, MEC-7, and TBB-6) are likely not involved in BZ response because we see no difference in ABZ-resistance between control and ABZ conditions for BZ-resistant strains. Of the six β-tubulin genes described above, we only observe variation in *ben-1* and *tbb-6*, both of which have low-frequency variants with high predicted functional effects. These observations are also shown by estimates of Tajima's *D* at these loci.

The β-tubulin repertoire among parasitic nematode species is highly diverse likely because of multiple gene duplication events [328]. For example, *H. contortus* has four known β-tubulin genes**,** of which *Hco-tbb-iso-1* and *Hco-tbb-iso-2* are thought to be the targets of BZ, and *Hco-tbb-iso-3* and *Hco-tbb-iso-4* are closely related in sequence and expression pattern to *C. elegans tbb-4* and *mec-7 [328]*. Despite the presence of a phenylalanine at amino acid position 200 in all four of the β-tubulins, the focus of the parasitology community is on missense variants present in *Hco-tbb-iso-1*. However, our data argue that it is important to consider the

high level of sequence similarity among the β-tubulin genes. We see that nearby structural variants, like deletions, can alter sequence read alignments and misannotate orthologous β-tubulin gene sequences to artifactually create these variant sites. Careful read alignments and high read depth will be necessary to interpret variation across highly divergent parasites with orthologous β-tubulin genes. Additionally, similar to our observations of *ben-1* in *C. elegans*, deletion alleles of *H. contortus Hco-tbb-iso-2* have been observed in few field studies, but their importance for BZ resistance remains unclear [329,330]. This observation led Kwa and colleagues [329,331] to hypothesize that BZ resistance requires two steps. First, mutation of F200 to tyrosine in one β-tubulin isoform, and second, deletion of the second β-tubulin that contains F200. This hypothesis aligns well with our observation that the single F200Y change in *C. elegans* BEN-1 confers BZ resistance in the presence of TBB-1 (Y200) and TBB-2 (Y200). An additional layer of complexity comes from recent evidence in *Trichuris trichiura* that non-BEN-1 β-tubulins might also play a role in BZ resistance in diverse parasitic species [332]. Our results showing that BZ resistance is a complex trait within the *C. elegans* population (**Figure 5-5**) and others' similar observations within parasitic nematode populations necessitates further investigation into the genetic and molecular mechanisms that underlie this trait.

Future Directions

An exciting result of these research project was the identification of a beta-tubulin-independent QTL on chromosomes III and X. This novel finding suggests that additional mechanisms of benzimidazole resistance exist in the *C. elegans* population, and likely exist in parasitic nematode populations. Further characterization of these QTL will reveal a BEN-1-independent mechanism of benzimidazole resistance, that can be used to inform the parasitology community of potential resistance loci in parasitic populations.

This study established *C. elegans* as a powerful system to test the effect of alleles predicted to confer ABZ resistance in parasitic populations. Therefore, we can test the effect of the four other alleles predicted to contribute to ABZ resistance in parasites - E198V, E198L, E198A, and F167Y. The same approach we used to validate the F200Y allele in the N2 background can be used to test the effect of these alleles. Furthermore, we identified five previously uncharacterized *ben-1* alleles (Q131L, S145F, A185P, M257I, and D404N) associated with albendazole resistance in the natural *C. elegans* population. However, the strains these alleles were discovered in represent a highly diverse set of genetic backgrounds and thus causal studies of these alleles must be performed to functionally validate their role in ABZ response. To validate these alleles, they must be introduced into a controlled genetic background and tested for ABZ resistance. The characterization of these alleles might provide insights into albendazole-tubulin interactions, which can be further supported through structural studies of the *C. elegans* BEN-1 protein.

Recent results from another group have found that the transcription factor SKN-1 and the metabolic gene *ugt-22* modulate albendazole resistance in the N2 genetic background [333]. My own inspection of the *ugt-22* locus in the natural *C. elegans* population revealed substantial variation at this locus, including a variant with high predicted effect in ECA670. Though candidate gene approaches do not consistently work, testing the effect of the *ugt-22* variation present in natural *C. elegans* might reveal that this locus modulates the effect of albendazole in the species.

Though I did not discuss the additional results in this Chapter, we also phenotyped the wild *C. elegans* population to several classes of anthelmintics, including a variety of benzimidazoles. We have identified a number of QTL for each of these classes of compounds that are independent of loci known to be involved in drug resistance. Therefore, there is great potential

to discover novel mechanisms of anthelmintic resistance by exploring the underlying variants within these QTL. A current Andersen lab graduate student, Clayton Dilks, is currently exploring a number of these QTL.

Finally, our observation that *ben-1* is highly conserved among related nematode species and highly variable within *C. elegans* suggests that we have a lot to learn about the evolution and function of *ben-1* and related beta-tubulins. I discussed potential explanations for the high levels of *ben-1* variation in *C. elegans* in the Discussion section of this chapter. However, another interesting avenue of research will be to explore the levels of *ben-1* variation among related free-living nematode species. Based on our observations in *C. elegans*, it is likely that other free-living nematodes also have elevated levels of variation at the *ben-1* locus. If this were the case, it might allow us to disentangle the two possible scenarios we discussed that could have caused elevated levels of *ben-1* variation in *C. elegans*. In contrast, if other free-living nematode species do not contain elevated levels of *ben-1* variation, we might gain insights into the evolution of the beta-tubulin gene family across nematodes.

Contributions

We thank Shannon C. Brady, Kyle Siegel, and members of the Andersen lab for editing the manuscript for flow and content. We thank members of the Andersen lab for making reagents used in the experiments presented in the manuscript. Additionally, we would also like to thank WormBase for providing an essential resource to study C. elegans biology. Conceptualization: Steffen R. Hahnel, Erik C. Andersen. Formal analysis: Steffen R. Hahnel, Stefan Zdraljevic. Funding acquisition: Patrick T. McGrath, Erik C. Andersen. Investigation: Steffen R. Hahnel, Stefan Zdraljevic, Briana C. Rodriguez, Yuehui Zhao, Patrick T. McGrath, Erik C. Andersen.

Project administration: Erik C. Andersen. Supervision: Erik C. Andersen. Writing – original draft: Steffen R. Hahnel, Stefan Zdraljevic. Writing – review & editing: Erik C. Andersen.

7. Discussion

The majority of my graduate work was focused on characterizing genetic variation within the *C. elegans* species and the effect this variation has on individual strain drug responses. When I joined the lab, I knew nothing about commonly used quantitative genetic approaches to tackle these questions. At that time, the two approaches we used to characterize the genetic basis of variable drug responses were GWA and linkage mapping, which I described in the Introduction of this dissertation. As I have progressed through graduate school, I have learned the benefits and caveats of these approaches and learned about new techniques that I have not covered in this dissertation. In this section, I will not include a discussion or future perspective on my work discussed in Chapters 2 - 5 because I included that in the Future Directions of each of those Chapters. Instead, I will discuss what I have learned about my experiences performing quantitative genetics in *C. elegans* and provide an outlook for future graduate students to expedite the discovery of genetic variants that contribute to phenotypic differences in the species. The identification of a causal variant underlying phenotypic differences in the *C. elegans* population was the first step toward uncovering the molecular mechanisms associated with these differences. Though characterizing the molecular basis of phenotypic variation is only a small fraction of the field of quantitative genetics, working toward unraveling these mechanisms was the most exciting part of my graduate work because we were figuring out and contextualize new biology.

Improvements to characterizing genetic variation

The variant calling I discussed in Chapters 2 - 5 is all based on the N2 reference genome, where individual strain read data is aligned to the N2 reference FASTA file. Support for a genetic variant exists at any position of the genome where strain sequence reads do not match

the reference genome. However, the nature of this approach is reference biased, which can affect our ability to accurately characterize the true natural diversity with in the *C. elegans* genome. An excellent example of this bias was shown for a divergent *C. elegans* strain, CB4856 [247]. Using short-read assembly methods, the authors identified nearly 5 Mb of genomic sequence in CB4856 that was not present in the N2 reference genome, and because this large amount of genomic content did not align to the reference genome, the natural diversity contained within this 5 Mb of genome could not be characterized. This early work was recently extended by using PacBio long-read sequence technology [334], where the authors identified several megabases of structural variants that affected nearly 3,000 annotated genes. More recently, a new build of the N2 genome was generated using long-read sequence data that uncovered 1.8 Mb of new genomic sequence that was previously uncharacterized [335]. These studies underscore the reference bias of using short-read sequence data to identify genetic variants in a species with genetically divergent individuals. Therefore, a necessary next step to comprehensively characterize the genetic diversity within the *C. elegans* genome is to assemble the genomes several genetically divergent individuals. Following whole-genome assembly of divergent strains, a graph-based variant calling method can be used to more accurately identify genetic variation within the species [336].

The power of combined mapping approaches

I was fortunate to be the first graduate student to join the Andersen lab. By the time I joined the lab, Erik and others had phenotyped 96 wild C. elegans isolates and 250 RIAILs in the presence of 50 different environmental perturbations. I was tasked with processing these data to generate a candidate list of QTL that could be followed up to identify the underlying genetic variants that contribute to phenotypic variation within the species. Fortunately, one of the parental strains that was used to generate the panel of RIAILs, CB4856, was also one of the 96 wild isolates

phenotyped. Having this strain in both phenotyped populations was critical to the progression of a majority of my graduate work. The approach I took when identifying a target list of QTL was to look for overlapping QTL identified by the GWA and linkage mapping approaches. The idea behind this approach was fairly simple, if QTL identified by these independent mapping methods had confidence intervals that overlapped, there was a chance the same underlying causal variant was contributing to phenotypic variation in both strain sets. Once I identified overlapping QTL, I filtered the variants within the GWA QTL based on their presence in the CB4856 strain. For both the etoposide and arsenic projects, this approach reduced the number of potential candidate variants within the GWA QTL from thousands to three and one, respectively. Despite the success of this approach for these two projects, the vast majority of QTL between the two data sets did not overlap, which begs the question - how does one identify the causal variant contributing to phenotypic variation in a population when complementary mapping approaches do not identify overlapping QTL?

For QTL identified by linkage mapping, the approach to identify a causal genetic variant can be straightforward. One could construct near-isogenic lines (NILs) and perform speed congenics [61], which I described in the Introduction. Overall, the resolution of a QTL identified by linkage mapping in *C. elegans* is limited by the number of recombinants phenotyped, which will hinder a researcher's ability to identify causal genetic variants, if that is the goal. For example, *C. elegans* has very limited recombination in the centers of chromosomes [29], which typically causes the size of a QTL in these regions to span multiple megabases of the genome. Furthermore, two-parent recombinant panels do not enable fine-mapping of QTL intervals, which makes follow-up QTL narrowing a necessary step to identify causal variants. However, other methods exist to circumvent this limitation of two-parent recombinant panels. A powerful approach is to construct a set of recombinant panels using a round-robin cross design, which was recently applied in yeast [337]. The round-robin cross design, which involves the

construction of multiple two-parent recombinant panels with multiple parents and each parent contributing to two independent recombinant panels, enabled the researchers to map QTL to single causal genes. Though the round-robin approach is relatively straightforward in yeast because of the ease of generating recombinant yeast strains, it should be feasible in *C. elegans* as well, though the parental lines should be chosen with care. A similar approach to using individuals that span the genetic diversity of the species, as implemented in Bloom *et al.*, should be strongly considered to maximize genetic diversity within the mapping populations, while accounting for known incompatibility loci with the *C. elegans* species [116,338].

In the above paragraphs I discuss combining single-marker mapping methods to expedite the discovery of variants contributing to phenotypic variation. However, as discussed in Chapter 5, these methods can often be misleading because of the complex LD structure in the *C. elegans* genome, which can occur between different chromosomes [43]. A powerful complementary approach to single-marker mapping is burden-based mapping. In Chapter 5, we applied the sequence kernel association test (SKAT) to the albendazole phenotype data [339], which looks for association between the cumulative variation in a gene and the phenotype of interest. This approach identified the same QTL as the single-marker mapping, but also identified the *ben-1* locus as the gene that was most significantly associated with albendazole responses in the natural *C. elegans* population. Though we used SKAT for the albendazole study, many other burden-based mapping approaches exist with variable power to detect QTL based on effect sizes, heritability, minor allele frequency, and polygenicity of the trait [340].

The albendazole study was pushed forward by prior functional knowledge of BEN-1. If we did not have this prior knowledge or solely relied on the single-marker mapping approach, our research direction might have led us to the QTL on chromosome II. Our study discussed in Chapter 5 highlights the known issue of linkage disequilibrium associated with GWA mapping.

The hermaphroditic life cycle and complex demographic history of *C. elegans* makes linkage disequilibrium a substantial issue when performing association studies in the species. For this reason, it is critical to carefully consider patterns of linkage disequilibrium among discovered QTL. In Chapter 5, we show that accounting for variation at the *ben-1* locus causes the other major QTL on chromosome II to disappear, suggesting that complex inter-chromosomal linkage disequilibrium is a major issue for QTL studies in *C. elegans*. Accounting for LD among markers can be performed prior to mapping by LD pruning with PLINK [94], though LD clumping is the preferred method [341], which identifies and groups correlated markers across the genome. A GWA mapping method that applies LD clumping has been developed and out performs existing linear mixed model methods, such as EMMA, in simulations studies and in Arabidopsis [342]. Unfortunately, if associated markers are in perfect linkage disequilibrium within the phenotyped population, no method will be able to prioritize one over the other. However, I remain hopeful that as the Andersen lab continues its extensive collection efforts to uncover more genetically divergent strains the LD structure surrounding causal genomic loci will break down, allowing for the expedited discovery of causal genetic variants. This expanded set of divergent strains will enable the lab to utilize mapping approaches analogous to “trans-ethnic studies” in humans that attempt to break LD structure by analyzing genetically distinct populations [343]. However, this approach will only work if the variants contributing to phenotypic variation are the same in both populations. Nevertheless, implementing these newer mapping approaches will be a critical step as the Andersen lab continues to perform genome-wide association mapping.

Many approaches based on statistical genetics exist for fine-mapping QTL identified by GWA mapping that I have not explored or implemented during my graduate work. The goal of all such approaches is to minimize the amount of follow-up work required to functionally validate a causal genetic variant. Looking to the future, these approaches should be introduced in the standard workflow of the Andersen lab to expedite the molecular characterization of phenotypic

variation in the *C. elegans* species. One such approach makes use of association mapping summary statistics and patterns of LD within a QTL to prioritize a set of candidate variants for functional validation [344]. In the case of *C. elegans*, where inter-chromosomal LD is prevalent, this approach can be easily modified to prioritize candidate markers. Other approaches make use of existing functional data of genomic features within QTL to prioritize candidate variants [345]. However, function-based approaches will be limited when the function of many genes in the *C. elegans* genome remain unknown and the effects of cryptic genetic variation can not be accounted for [346]. Uncharacterized complex genetic interaction networks will further the efficacy of function-based approaches. One way around the issue of cryptic genetic variation in complex genetic networks is to integrate transcriptomic data with QTL mapping data through a process known as mediation analysis [347,348], which attempts to find correlations between phenotypic and transcriptomic datasets.

Alternate methods for QTL mapping

Key to the *C. elegans* community's success to identify QTL and variants underlying phenotypic variation is the ability to rapidly build large populations of recombinant inbred lines. As discussed, these RIL panels can be repeatedly assayed to identify QTL [29,33,60,138,146,155,206,349–351]. However, a large amount of effort is required to generate, prepare for long-term storage, and phenotype a collection of RILs. This effort scales with the cross design used to generate the RILs and the desired number of recombinant individuals. As a result, the limited number RIL panels that exist were generated with very few parental genetic backgrounds, and therefore are likely missing genetic variants that affect the quantified phenotypes of interest. Though the generation of RIL panels has proven useful for the community, new methods must be applied to probe the effects of a larger sampling of genetic diversity present in the *C. elegans* population. Recent efforts in the *C. elegans* community have

implemented the classical genetic technique bulk-segregant analysis (BSA) [352] toward the discovery of QTL [178].

In contrast to constructing, cryopreserving, sequencing, and phenotyping individuals for a RIL panels, the BSA approach does not require isolating, cryopreserving, or phenotyping individual recombinants. Instead, pools of recombinant individuals (referred to as the bulk population) are generated from parents that vary in one or more phenotypes of interest. The ease of generating large numbers of recombinant *C. elegans* strains makes this an ideal metazoan model for such an approach. Once generated, the bulk population will be phenotyped. A common phenotyping approach for BSA is to apply a selective pressure to the bulk population. If no selective pressure is applied to this population, we would expect each parental genotype to be represented in ~50% of the recombinant lines at all positions of the genome. However if a selective pressure is applied to the recombinant population, the genomic regions that contain variants that affect response to the selective pressure will be enriched (diverged from the 50% expectation) for the parental genotype that performs better in the presence of the selective pressure. Therefore, one could subject the bulk population to a selective pressure and select a subset of individuals at one end of the phenotypic distribution for sequencing. Statistical power can be gained if both ends of the phenotypic distribution are isolated for sequencing and if independently generated bulk populations are subjected to the same selective pressure. Upon sequencing, genomic regions that contribute to phenotypic variation among the parental lines will be enriched or depleted in the selected populations. An approach that is similar to BSA is selection with recurrent backcrossing [353], where two individuals with different phenotypes are crossed and the progeny are put under selection. Progeny with the desired phenotype are then backcrossed to the parent with the opposite phenotype. Multiple rounds of selection and backcrossing will isolate the loci that contribute to phenotypic differences between the parental lines.

8. References

1. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet. 2008;9: 255–266. doi:10.1038/nrg2322

2. Newman DJ, Cragg GM. Natural products as sources of new drugs over the 30 years from 1981 to 2010. J Nat Prod. 2012;75: 311–335. doi:10.1021/np200906s

3. Bock KW. The UDP-glycosyltransferase (UGT) superfamily expressed in humans, insects and plants: Animal-plant arms-race and co-evolution. Biochem Pharmacol. 2016;99: 11–17. doi:10.1016/j.bcp.2015.10.001

4. Russell RJ, Scott C, Jackson CJ, Pandey R, Pandey G, Taylor MC, et al. The evolution of new enzyme function: lessons from xenobiotic metabolizing bacteria versus insecticide-resistant insects. Evol Appl. 2011;4: 225–248. doi:10.1111/j.1752-4571.2010.00175.x

5. Prasad V, Fojo T, Brada M. Precision oncology: origins, optimism, and potential. Lancet Oncol. 2016;17: e81–6. doi:10.1016/S1470-2045(15)00620-8

6. Turner RM, Park BK, Pirmohamed M. Parsing interindividual drug variability: an emerging role for systems pharmacology. Wiley Interdiscip Rev Syst Biol Med. 2015;7: 221–241. doi:10.1002/wsbm.1302

7. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nature Publishing Group. Nature Publishing Group; 2014;15: 335–346. doi:10.1038/nrg3706

8. Low S-K, Chung S, Takahashi A, Zembutsu H, Mushiroda T, Kubo M, et al. Genome-wide association study of chemotherapeutic agent-induced severe neutropenia/leucopenia for patients in Biobank Japan. Cancer Sci. 2013;104: 1074–1082. doi:10.1111/cas.12186

9. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell. 2013;155: 27–38. doi:10.1016/j.cell.2013.09.006

10. McClellan J, King M-C. Genetic heterogeneity in human disease. Cell. 2010;141: 210–217. doi:10.1016/j.cell.2010.03.032

11. Liu J, Huang J, Zhang Y, Lan Q, Rothman N, Zheng T, et al. Identification of gene-environment interactions in cancer studies using penalization. Genomics. 2013;102: 189–194. doi:10.1016/j.ygeno.2013.08.006

12. Hunter DJ. Gene-environment interactions in human diseases. Nat Rev Genet. 2005;6: 287–298. doi:10.1038/nrg1578

13. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014;32: 40–51. doi:10.1038/nbt.2786

14. Kaletta T, Hengartner MO. Finding function in novel targets: *C. elegans* as a model organism. Nat Rev Drug Discov. 2006;5: 387–398. doi:10.1038/nrd2031

15. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov. 2017;16: 19–34. doi:10.1038/nrd.2016.230

16. Kawashima A, Satta Y. Substrate-dependent evolution of cytochrome P450: rapid turnover of the detoxification-type and conservation of the biosynthesis-type. PLoS One. 2014;9: e100059. doi:10.1371/journal.pone.0100059

17. Harlow PH, Perry SJ, Widdison S, Daniels S, Bondo E, Lamberth C, et al. The nematode Caenorhabditis elegans as a tool to predict chemical activity on mammalian development and identify mechanisms influencing toxicological outcome. Sci Rep. ncbi.nlm.nih.gov; 2016;6: 22965. doi:10.1038/srep22965

18. Xiong J, Feng J, Yuan D, Zhou J, Miao W. Tracing the structural evolution of eukaryotic ATP binding cassette transporter superfamily. Sci Rep. 2015;5: 16724. doi:10.1038/srep16724

19. Meier B, Cooke SL, Weiss J, Bailly AP, Alexandrov LB, Marshall J, et al. C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. Genome Res. 2014;24: 1624–1636. doi:10.1101/gr.175547.114

20. Szikriszt B, Póti Á, Pipek O, Krzystanek M, Kanu N, Molnár J, et al. A comprehensive survey of the mutagenic impact of common cancer cytotoxics. Genome Biol. 2016;17: 99. doi:10.1186/s13059-016-0963-7

21. Aaron Hobbs G, Der CJ, Rossman KL. RAS isoforms and mutations in cancer at a glance. Journal of Cell Science. 2016; doi:10.1242/jcs.182873

22. Shaye DD, Greenwald I. OrthoList: a compendium of C. elegans genes with human orthologs. PLoS One. 2011;6: e20085. doi:10.1371/journal.pone.0020085

23. Reiner DJ, Lundquist EA. Small GTPases. WormBook. 2016; doi:10.1895/wormbook.1.67.2

24. Jindal GA, Goyal Y, Yamaya K, Futran AS, Kountouridis I, Balgobin CA, et al. In vivo severity ranking of Ras pathway mutations associated with developmental disorders. Proceedings of the National Academy of Sciences. 2017;114: 510–515. doi:10.1073/pnas.1615651114

25. Relling MV, Evans WE. Pharmacogenomics in the clinic. Nature. 2015;526: 343–350. doi:10.1038/nature15817

26. Cook DE, Zdraljevic S, Roberts JP, Andersen EC. CeNDR, the Caenorhabditis elegans natural diversity resource. Nucleic Acids Res. 2016; doi:10.1093/nar/gkw893

27. Cook DE, Zdraljevic S, Roberts JP, Andersen EC. CeNDR, the Caenorhabditis elegans natural diversity resource. Nucleic Acids Res. 2017;45: D650–D657. doi:10.1093/nar/gkw893

28. Andersen EC, Shimko TC, Crissman JR, Ghosh R, Bloom JS, Seidel HS, et al. A Powerful New Quantitative Genetics Platform, Combining *Caenorhabditis elegans* High-Throughput Fitness Assays with a Large Collection of Recombinant Strains. G3 . Genetics Society of America; 2015;5: g3.115.017178–920. doi:10.1534/g3.115.017178

29. Rockman MV, Kruglyak L. Recombinational landscape and population genomics of Caenorhabditis elegans. PLoS Genet. 2009;5: e1000419. doi:10.1371/journal.pgen.1000419

30. Mondal S, Hegarty E, Martin C, Gökçe SK, Ghorashian N, Ben-Yakar A. Large-scale microfluidics providing high-resolution and high-throughput screening of Caenorhabditis elegans poly-glutamine aggregation model. Nat Commun. 2016;7: 13023. doi:10.1038/ncomms13023

31. Zdraljevic S, Strand C, Seidel HS, Cook DE, Doench JG, Andersen EC. Natural variation in a single amino acid substitution underlies physiological responses to topoisomerase II poisons. PLoS Genet. 2017;13: e1006891. doi:10.1371/journal.pgen.1006891

32. Zdraljevic S, Fox BW, Strand C, Panda O, Tenjo FJ, Brady SC, et al. Natural variation in C. elegans arsenic toxicity is explained by differences in branched chain amino acid metabolism. Elife. 2019;8. doi:10.7554/eLife.40260

33. Evans KS, Brady SC, Bloom JS, Tanny RE, Cook DE, Giuliani SE, et al. Shared Genomic Regions Underlie Natural Variation in Diverse Toxin Responses. Genetics. 2018;210: 1509–1525. doi:10.1534/genetics.118.301311

34. Paix A, Folkmann A, Rasoloson D, Seydoux G. High Efficiency, Homology-Directed Genome Editing in Caenorhabditis elegans Using CRISPR-Cas9 Ribonucleoprotein Complexes. Genetics. 2015;201: 47–54. doi:10.1534/genetics.115.179382

35. Hodgkin JA, Brenner S. Mutations causing transformation of sexual phenotype in the nematode Caenorhabditis elegans. Genetics. 1977;86: 275–287. Available: https://www.ncbi.nlm.nih.gov/pubmed/560330

36. Brenner S. The genetics of Caenorhabditis elegans. Genetics. 1974;77: 71–94. Available: https://www.ncbi.nlm.nih.gov/pubmed/4366476

37. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature. 1998;391: 806–811. doi:10.1038/35888

38. Grishok A, Tabara H, Mello CC. Genetic requirements for inheritance of RNAi in C. elegans. Science. 2000;287: 2494–2497. Available: https://www.ncbi.nlm.nih.gov/pubmed/10741970

39. Consortium TCES. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science. 1998; 1–8.

40. Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC. Green fluorescent protein as a marker for gene expression. Science. 1994;263: 802–805. Available: https://www.ncbi.nlm.nih.gov/pubmed/8303295

41. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75: 843–854. Available: https://www.ncbi.nlm.nih.gov/pubmed/8252621

42. Sulston JE, Schierenberg E, White JG, Thomson JN. The embryonic cell lineage of the

nematode Caenorhabditis elegans. Dev Biol. 1983;100: 64–119. doi:10.1016/0012-1606(83)90201-4

43. Andersen EC, Gerke JP, Shapiro JA, Crissman JR, Ghosh R, Bloom JS, et al. Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. Nat Genet. 2012;44: 285–290. doi:10.1038/ng.1050

44. Cook DE, Zdraljevic S, Tanny RE, Seo B, Riccardi DD, Noble LM, et al. The Genetic Basis of Natural Variation in Caenorhabditis elegans Telomere Length. Genetics. 2016;204: 371–383. doi:10.1534/genetics.116.191148

45. Richaud A, Zhang G, Lee D, Lee J, Félix M-A. The Local Coexistence Pattern of Selfing Genotypes in Caenorhabditis elegans Natural Metapopulations. Genetics. 2018;208: 807–821. doi:10.1534/genetics.117.300564

46. Barrière A, Félix M-A. Isolation of C. elegans and related nematodes. WormBook. 2014; 1–19. doi:10.1895/wormbook.1.115.2

47. Félix M-A, Duveau F. Population dynamics and habitat sharing of natural populations of Caenorhabditis elegans and C. briggsae. BMC Biol. bmcbiol.biomedcentral.com; 2012;10: 59. doi:10.1186/1741-7007-10-59

48. Lee D, Zdraljevic S, Cook DE, Frézal L, Hsu J-C, Sterken MG, et al. Selection and gene flow shape niche-associated copy-number variation of pheromone receptor genes [Internet]. bioRxiv. 2019. p. 580803. doi:10.1101/580803

49. Hahnel SR, Zdraljevic S, Rodriguez BC, Zhao Y, McGrath PT, Andersen EC. Extreme allelic heterogeneity at a Caenorhabditis elegans beta-tubulin locus explains natural resistance to benzimidazoles [Internet]. bioRxiv. 2018. p. 372623. doi:10.1101/372623

50. Ferrari C, Salle R, Callemeyn-Torre N, Jovelin R, Cutter AD, Braendle C. Ephemeral-habitat colonization and neotropical species richness of Caenorhabditis nematodes. BMC Ecol. 2017;17: 43. doi:10.1186/s12898-017-0150-z

51. Koch R, van Luenen HG, van der Horst M, Thijssen KL, Plasterk RH. Single nucleotide polymorphisms in wild isolates of Caenorhabditis elegans. Genome Res. 2000;10: 1690–1696. Available: http://www.ncbi.nlm.nih.gov/pubmed/11076854

52. Thomas CG, Wang W, Jovelin R, Ghosh R, Lomasko T, Trinh Q, et al. Full-genome evolutionary histories of selfing, splitting, and selection in Caenorhabditis. Genome Res. 2015;25: 667–678. doi:10.1101/gr.187237.114

53. Dey A, Chan CKW, Thomas CG, Cutter AD. Molecular hyperdiversity defines populations of the nematode Caenorhabditis brenneri. Proc Natl Acad Sci U S A. 2013;110: 11056–11060. doi:10.1073/pnas.1303057110

54. Andersen EC, Shimko TC, Crissman JR, Ghosh R, Bloom JS, Seidel HS, et al. A Powerful New Quantitative Genetics Platform, Combining Caenorhabditis elegans High-Throughput Fitness Assays with a Large Collection of Recombinant Strains. G3 . 2015;5: 911–920. doi:10.1534/g3.115.017178

55. Boyd WA, Smith MV, Freedman JH. Caenorhabditis elegans as a model in developmental

toxicology. Methods Mol Biol. Totowa, NJ: Humana Press; 2012;889: 15–24. doi:10.1007/978-1-61779-867-2_3

56. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient Control of Population Structure in Model Organism Association Mapping. Genetics. 2008;178: 1709–1723. doi:10.1534/genetics.107.080101

57. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2: e190. doi:10.1371/journal.pgen.0020190

58. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. Nature Publishing Group; 2010;42: 348–354. doi:10.1038/ng.548

59. Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet. Nature Publishing Group; 2012;44: 825–830. doi:10.1038/ng.2314

60. Doroszuk A, Snoek LB, Fradin E, Riksen J, Kammenga J. A genome-wide library of CB4856/N2 introgression lines of Caenorhabditis elegans. Nucleic Acids Res. 2009;37: e110. doi:10.1093/nar/gkp528

61. Wong GT. Speed congenics: applications for transgenic and knock-out mouse strains. Neuropeptides. 2002;36: 230–236. Available: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=12359513&retmode=ref&cmd=prlinks

62. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007;315: 1709–1712. doi:10.1126/science.1138140

63. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337: 816–821. doi:10.1126/science.1225829

64. Katic I, Großhans H. Targeted heritable mutation and gene conversion by Cas9-CRISPR in Caenorhabditis elegans. Genetics. 2013;195: 1173–1176. doi:10.1534/genetics.113.155754

65. Lo T-W, Pickle CS, Lin S, Ralston EJ, Gurling M, Schartner CM, et al. Precise and heritable genome editing in evolutionarily diverse nematodes using TALENs and CRISPR/Cas9 to engineer insertions and deletions. Genetics. 2013;195: 331–348. doi:10.1534/genetics.113.155382

66. Arribere JA, Bell RT, Fu BXH, Artiles KL, Hartman PS, Fire AZ. Efficient marker-free recovery of custom genetic modifications with CRISPR/Cas9 in Caenorhabditis elegans. Genetics. 2014;198: 837–846. doi:10.1534/genetics.114.169730

67. Paix A, Wang Y, Smith HE, Lee C-YS, Calidas D, Lu T, et al. Scalable and versatile genome editing using linear DNAs with microhomology to Cas9 Sites in Caenorhabditis elegans. Genetics. 2014;198: 1347–1356. doi:10.1534/genetics.114.170423

68. Dickinson DJ, Goldstein B. CRISPR-Based Methods for Caenorhabditis elegans Genome Engineering. Genetics. Genetics; 2016;202: 885–901. doi:10.1534/genetics.115.182162

69. Paix A, Folkmann A, Rasoloson D, Seydoux G. High Efficiency, Homology-Directed Genome Editing in Caenorhabditis elegans Using CRISPR-Cas9 Ribonucleoprotein Complexes. Genetics. 2015;201: 47–54. doi:10.1534/genetics.115.179382

70. Stephan W. Selective Sweeps. Genetics. 2019;211: 5–13. doi:10.1534/genetics.118.301319

71. Hahnel SR, Zdraljevic S, Rodriguez BC, Zhao Y, McGrath PT, Andersen EC. Extreme allelic heterogeneity at a Caenorhabditis elegans beta-tubulin locus explains natural resistance to benzimidazoles. PLoS Pathog. 2018;14: e1007226. doi:10.1371/journal.ppat.1007226

72. Cheng X, DeGiorgio M. Detection of Shared Balancing Selection in the Absence of Trans-Species Polymorphism. Mol Biol Evol. 2019;36: 177–199. doi:10.1093/molbev/msy202

73. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. nature.com; 2017;35: 316–319. doi:10.1038/nbt.3820

74. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170

75. Danecek P, Schiffels S, Durbin R. Multiallelic calling model in bcftools (-m). 2014; 10–11.

76. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

77. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. arXiv [q-bio.GN]. 2013. p. 3. doi:arXiv:1303.3997 [q-bio.GN]

78. Lee RYN, Howe KL, Harris TW, Arnaboldi V, Cain S, Chan J, et al. WormBase 2017: molting into a new stage. Nucleic Acids Res. 2018;46: D869–D874. doi:10.1093/nar/gkx998

79. De Summa S, Malerba G, Pinto R, Mori A, Mijatovic V, Tommasi S. GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. BMC Bioinformatics. 2017;18: 119. doi:10.1186/s12859-017-1537-8

80. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;15: 591–594. doi:10.1038/s41592-018-0051-x

81. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, 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 . 2012;6: 80–92. doi:10.4161/fly.19695

82. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28: i333–i339. doi:10.1093/bioinformatics/bts378

83. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15: R84. doi:10.1186/gb-2014-15-6-r84

84. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32: 1220–1222. doi:10.1093/bioinformatics/btv710

85. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31: 2032–2034. doi:10.1093/bioinformatics/btv098

86. Howe KL, Bolt BJ, Cain S, Chan J, Chen WJ, Davis P, et al. WormBase 2016: expanding to enable helminth genomic research. Nucleic Acids Res. 2016;44: D774–80. doi:10.1093/nar/gkv1217

87. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27: 2987–2993. doi:10.1093/bioinformatics/btr509

88. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun. 2017;8: 14061. doi:10.1038/ncomms14061

89. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842. doi:10.1093/bioinformatics/btq033

90. anomalize [Internet]. Github; Available: https://github.com/business-science/anomalize

91. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. R Foundation for Statistical Computing; Available: http://www.R-project.org/

92. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16: 284–287. doi:10.1089/omi.2011.0118

93. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19: 1655–1664. doi:10.1101/gr.094052.109

94. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–575. doi:10.1086/519795

95. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4: 7. doi:10.1186/s13742-015-0047-8

96. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38: 904–909. doi:10.1038/ng1847

97. González I, Déjean S, Martin P, Baccini A. CCA: An R Package to Extend Canonical Correlation Analysis. Journal of Statistical Software, Articles. 2008;23: 1–14. doi:10.18637/jss.v023.i12

98. Berg M, Monnin D, Cho J, Nelson L, Crits-Christoph A, Shapira M. TGFβ/BMP immune

signaling affects abundance and function of C. elegans gut commensals. Nat Commun. 2019;10: 604. doi:10.1038/s41467-019-08379-8

99. Panek J, Reddy KC, Luallen RJ, Fulzele A, Bennett EJ, Troemel ER. A newly defined cullin-RING ubiquitin ligase promotes thermotolerance as part of the Intracellular Pathogen Response in C. elegans [Internet]. bioRxiv. 2019. p. 586834. doi:10.1101/586834

100. Reddy KC, Dror T, Sowa JN, Panek J, Chen K, Lim ES, et al. An Intracellular Pathogen Response Pathway Promotes Proteostasis in C. elegans. Curr Biol. 2017;27: 3544–3553.e5. doi:10.1016/j.cub.2017.10.009

101. Zou C-G, Ma Y-C, Dai L-L, Zhang K-Q. Autophagy protects C. elegans against necrosis during Pseudomonas aeruginosa infection. Proc Natl Acad Sci U S A. 2014;111: 12480–12485. doi:10.1073/pnas.1405032111

102. Kuo C-J, Hansen M, Troemel E. Autophagy and innate immunity: Insights from invertebrate model organisms. Autophagy. 2018;14: 233–242. doi:10.1080/15548627.2017.1389824

103. De Arras L, Laws R, Leach SM, Pontis K, Freedman JH, Schwartz DA, et al. Comparative genomics RNAi screen identifies Eftud2 as a novel regulator of innate immunity. Genetics. 2014;197: 485–496. doi:10.1534/genetics.113.160499

104. Doherty MF, Adelmant G, Cecchetelli AD, Marto JA, Cram EJ. Proteomic analysis reveals CACN-1 is a component of the spliceosome in Caenorhabditis elegans. G3 . 2014;4: 1555–1564. doi:10.1534/g3.114.012013

105. De Arras L, Alper S. Limiting of the innate immune response by SF3A-dependent control of MyD88 alternative mRNA splicing. PLoS Genet. 2013;9: e1003855. doi:10.1371/journal.pgen.1003855

106. Thomas JH, Robertson HM. The Caenorhabditis chemoreceptor gene families. BMC Biol. 2008;6: 42. doi:10.1186/1741-7007-6-42

107. Thomas JH. Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. Genome Res. 2006;16: 1017–1030. doi:10.1101/gr.5089806

108. Koenig D, Hagmann J, Li R, Bemm F, Slotte T, Neuffer B, et al. Long-term balancing selection drives evolution of immunity genes in Capsella. Elife. 2019;8. doi:10.7554/eLife.43606

109. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123: 585–595. Available: https://www.ncbi.nlm.nih.gov/pubmed/2513255

110. Vidal B, Aghayeva U, Sun H, Wang C, Glenwinkel L, Bayer EA, et al. An atlas of Caenorhabditis elegans chemoreceptor expression. PLoS Biol. 2018;16: e2004218. doi:10.1371/journal.pbio.2004218

111. McLaughlin RN Jr, Malik HS. Genetic conflicts: the usual suspects and beyond. J Exp Biol. 2017;220: 6–17. doi:10.1242/jeb.148148

112. Levin TC, Malik HS. Rapidly Evolving Toll-3/4 Genes Encode Male-Specific Toll-Like

Receptors in Drosophila. Mol Biol Evol. 2017;34: 2307–2323. doi:10.1093/molbev/msx168

113. O'Sullivan O, O'Callaghan J, Sangrador-Vegas A, McAuliffe O, Slattery L, Kaleta P, et al. Comparative genomics of lactic acid bacteria reveals a niche-specific gene set. BMC Microbiol. 2009;9: 50. doi:10.1186/1471-2180-9-50

114. Wu Q, Han T-S, Chen X, Chen J-F, Zou Y-P, Li Z-W, et al. Long-term balancing selection contributes to adaptation in Arabidopsis and its relatives. Genome Biol. 2017;18: 217. doi:10.1186/s13059-017-1342-8

115. Goldberg EE, Kohn JR, Lande R, Robertson KA, Smith SA, Igić B. Species selection maintains self-incompatibility. Science. 2010;330: 493–495. doi:10.1126/science.1194513

116. Seidel HS, Ailion M, Li J, van Oudenaarden A, Rockman MV, Kruglyak L. A novel sperm-delivered toxin causes late-stage embryo lethality and transmission ratio distortion in C. elegans. PLoS Biol. 2011;9: e1001115. doi:10.1371/journal.pbio.1001115

117. Charlesworth D. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet. 2006;2: e64. doi:10.1371/journal.pgen.0020064

118. Brisson D. Negative Frequency-Dependent Selection Is Frequently Confounding. Frontiers in Ecology and Evolution. 2018;6: 10. doi:10.3389/fevo.2018.00010

119. Gao AW, Sterken MG, Uit de Bos J, van Creij J, Kamble R, Snoek BL, et al. Natural genetic variation in C. elegans identified genomic loci controlling metabolite levels. Genome Res. 2018;28: 1296–1308. doi:10.1101/gr.232322.117

120. Dirksen P, Marsh SA, Braker I, Heitland N, Wagner S, Nakad R, et al. The native microbiome of the nematode Caenorhabditis elegans: gateway to a new host-microbiome model. BMC Biol. 2016;14: 38. doi:10.1186/s12915-016-0258-1

121. Samuel BS, Rowedder H, Braendle C, Félix M-A, Ruvkun G. Caenorhabditis elegans responses to bacteria from its natural habitats. Proc Natl Acad Sci U S A. 2016;113: E3941–9. doi:10.1073/pnas.1607183113

122. Zhang F, Berg M, Dierking K, Félix M-A, Shapira M, Samuel BS, et al. Caenorhabditis elegans as a Model for Microbiome Research. Front Microbiol. 2017;8: 485. doi:10.3389/fmicb.2017.00485

123. Schulenburg H, Félix M-A. The Natural Biotic Environment of Caenorhabditis elegans. Genetics. 2017;206: 55–86. doi:10.1534/genetics.116.195511

124. Reddy KC, Dror T, Underwood RS, Osman GA, Elder CR, Desjardins CA, et al. Antagonistic paralogs control a switch between growth and pathogen resistance in C. elegans. PLoS Pathog. 2019;15: e1007528. doi:10.1371/journal.ppat.1007528

125. Leyva-Díaz E, Stefanakis N, Carrera I, Glenwinkel L, Wang G, Driscoll M, et al. Silencing of Repetitive DNA Is Controlled by a Member of an Unusual Caenorhabditis elegans Gene Family. Genetics. 2017;207: 529–545. doi:10.1534/genetics.117.300134

126. Stevens L, Félix M-A, Beltran T, Braendle C, Caurcel C, Fausett S, et al. Comparative genomics of 10 new Caenorhabditis species. Evolution Letters. 2019;3: 217–236. doi:10.1002/evl3.110

127. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90: 7–24. doi:10.1016/j.ajhg.2011.11.029

128. Boddy AV. Genetics of cisplatin ototoxicity: confirming the unexplained? Clin Pharmacol Ther. 2013;94: 198–200. doi:10.1038/clpt.2013.116

129. Park J-H, Gail MH, Greene MH, Chatterjee N. Potential usefulness of single nucleotide polymorphisms to identify persons at high cancer risk: an evaluation of seven common cancers. J Clin Oncol. American Society of Clinical Oncology; 2012;30: 2157–2162. doi:10.1200/JCO.2011.40.1943

130. Willoughby LF, Schlosser T, Manning SA, Parisot JP, Street IP, Richardson HE, et al. An in vivo large-scale chemical screening platform using Drosophila for anti-cancer drug discovery. Dis Model Mech. 2013;6: 521–529. doi:10.1242/dmm.009985

131. Perlstein EO, Ruderfer DM, Roberts DC, Schreiber SL, Kruglyak L. Genetic basis of individual differences in the response to small-molecule drugs in yeast. Nat Genet. 2007;39: 496–502. doi:10.1038/ng1991

132. King EG, Kislukhin G, Walters KN, Long AD. Using Drosophila melanogaster to identify chemotherapy toxicity genes. Genetics. 2014;198: 31–43. doi:10.1534/genetics.114.161968

133. Ehrenreich IM, Torabi N, Jia Y, Kent J, Martis S, Shapiro JA, et al. Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature. 2010;464: 1039–1042. doi:10.1038/nature08923

134. Bloom JS, Ehrenreich IM, Loo WT, Lite T-LV, Kruglyak L. Finding the sources of missing heritability in a yeast cross. Nature. 2013;494: 234–237. doi:10.1038/nature11867

135. Demogines A, Smith E, Kruglyak L, Alani E. Identification and dissection of a complex DNA repair sensitivity phenotype in Baker's yeast. PLoS Genet. 2008;4: e1000123. doi:10.1371/journal.pgen.1000123

136. Liti G, Louis EJ. Advances in quantitative trait analysis in yeast. PLoS Genet. 2012;8: e1002912. doi:10.1371/journal.pgen.1002912

137. Stern DL. Identification of loci that cause phenotypic variation in diverse species with the reciprocal hemizygosity test. Trends Genet. 2014;30: 547–554. doi:10.1016/j.tig.2014.09.006

138. Ghosh R, Andersen EC, Shapiro JA, Gerke JP, Kruglyak L. Natural variation in a chloride channel subunit confers avermectin resistance in C. elegans. Science. 2012;335: 574–578. doi:10.1126/science.1214318

139. Pommier Y, Leo E, Zhang H, Marchand C. DNA Topoisomerases and Their Poisoning by Anticancer and Antibacterial Drugs. Chem Biol. Elsevier Ltd; 2010;17: 421–433. doi:10.1016/j.chembiol.2010.04.012

140. Pommier Y, Schwartz RE, Kohn KW, Zwelling LA. Formation and rejoining of deoxyribonucleic acid double-strand breaks induced in isolated cell nuclei by antineoplastic intercalating agents. Biochemistry. 1984;23: 3194–3201. Available:

https://www.ncbi.nlm.nih.gov/pubmed/6087890

141. Gómez-Herreros F, Romero-Granados R, Zeng Z, Álvarez-Quilón A, Quintero C, Ju L, et al. TDP2-dependent non-homologous end-joining protects against topoisomerase II-induced DNA breaks and genome instability in cells and in vivo. PLoS Genet. 2013;9: e1003226. doi:10.1371/journal.pgen.1003226

142. Nitiss JL. Targeting DNA topoisomerase II in cancer chemotherapy. Nat Rev Cancer. 2009;9: 338–350. doi:10.1038/nrc2607

143. Wu C-C, Li T-K, Farth L, Lin L-Y, Lin T-S, Yu Y-J, et al. Structural Basis of Type II Topoisomerase Inhibition by the Anticancer Drug Etoposide . Science. 2011;333: 456–459. doi:10.1126/science.1203963

144. Wu C-C, Li Y-C, Wang Y-R, Li T-K, Chan N-L. On the structural basis and design guidelines for type II topoisomerase-targeting anticancer drugs. Nucleic Acids Res. Oxford University Press; 2013;41: 10630–10640. doi:10.1093/nar/gkt828

145. Wendorff TJ, Schmidt BH, Heslop P, Austin CA, Berger JM. The structure of DNA-bound human topoisomerase II alpha: conformational mechanisms for coordinating inter-subunit interactions with DNA cleavage. J Mol Biol. 2012;424: 109–124. doi:10.1016/j.jmb.2012.07.014

146. Andersen EC, Bloom JS, Gerke JP, Kruglyak L. A variant in the neuropeptide receptor npr-1 is a major determinant of Caenorhabditis elegans growth and physiology. PLoS Genet. 2014;10: e1004156. doi:10.1371/journal.pgen.1004156

147. Shimko TC, Andersen EC. COPASutils: an R package for reading, processing, and visualizing data from COPAS large-particle flow cytometers. PLoS One. 2014;9: e111090. doi:10.1371/journal.pone.0111090

148. Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc Natl Acad Sci U S A. 2005;102: 1572–1577. doi:10.1073/pnas.0408709102

149. Endelman JB. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome Journal. Crop Science Society of America; 2011;4: 250–256. doi:10.3835/plantgenome2011.08.0024

150. Kim H, Ishidate T, Ghanta KS, Seth M, Conte D, Shirayama M, et al. A co-CRISPR strategy for efficient genome editing in Caenorhabditis elegans. Genetics. 2014;197: 1069–1080. doi:10.1534/genetics.114.166389

151. Jacobson MP, Friesner RA, Xiang Z, Honig B. On the role of the crystal environment in determining protein side-chain conformations. J Mol Biol. 2002;320: 597–608. doi:10.1016/S0022-2836(02)00470-9

152. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, et al. A hierarchical approach to all-atom protein loop prediction. Proteins. 2004;55: 351–367. doi:10.1002/prot.10613

153. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, et al. MolProbity: all-

atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. 2007;35: W375–83. doi:10.1093/nar/gkm216

154. Sastry GM, Adzhigirey M, Day T, Annabhimoju R, Sherman W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. J Comput Aided Mol Des. 2013;27: 221–234. doi:10.1007/s10822-013-9644-8

155. Rockman MV, Skrovanek SS, Kruglyak L. Selection at linked sites shapes heritable phenotypic variation in C. elegans. Science. 2010;330: 372–376. doi:10.1126/science.1194208

156. Schmidt BH, Osheroff N, Berger JM. Structure of a topoisomerase II-DNA-nucleotide complex reveals a new control mechanism for ATPase activity. Nat Struct Mol Biol. 2012;19: 1147–1154. doi:10.1038/nsmb.2388

157. Bandele OJ, Osheroff N. The efficacy of topoisomerase II-targeted anticancer agents reflects the persistence of drug-induced cleavage complexes in cells. Biochemistry. American Chemical Society; 2008;47: 11900–11908. doi:10.1021/bi800981j

158. Deweese JE, Osheroff N. The DNA cleavage reaction of topoisomerase II: wolf in sheep's clothing. Nucleic Acids Res. Oxford University Press; 2009;37: 738–748. doi:10.1093/nar/gkn937

159. Gao H, Huang K-C, Yamasaki EF, Chan KK, Chohan L, Snapka RM. XK469, a selective topoisomerase IIβ poison. Proceedings of the National Academy of Sciences. 1999;96: 12168–12173. doi:10.1073/pnas.96.21.12168

160. Bromberg KD, Burgin AB, Osheroff N. A Two-drug Model for Etoposide Action against Human Topoisomerase IIα. J Biol Chem. 2003;278: 7406–7412. doi:10.1074/jbc.M212056200

161. Koba M, Konopa J. [Actinomycin D and its mechanisms of action]. Postepy Hig Med Dosw . 2005;59: 290–298. Available: https://www.ncbi.nlm.nih.gov/pubmed/15995596

162. Moen EL, Godley LA, Zhang W, Dolan ME. Pharmacogenomics of chemotherapeutic susceptibility and toxicity. Genome Med. 2012;4: 90. doi:10.1186/gm391

163. Giacomini KM, Brett CM, Altman RB, Benowitz NL, Dolan ME, Flockhart DA, et al. The pharmacogenetics research network: from SNP discovery to clinical drug response. Clin Pharmacol Ther. 2007;81: 328–345. doi:10.1038/sj.clpt.6100087

164. Huang RS, Duan S, Bleibel WK, Kistner EO, Zhang W, Clark TA, et al. A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. National Acad Sciences; 2007;104: 9758–9763. doi:10.1073/pnas.0703736104

165. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536: 285–291. doi:10.1038/nature19057

166. Chen SH, Chan N-L, Hsieh T-S. New mechanistic and functional insights into DNA topoisomerases. Annu Rev Biochem. 2013;82: 139–170. doi:10.1146/annurev-biochem-061809-100002

167. Felix CA, Kolaris CP, Osheroff N. Topoisomerase II and the etiology of chromosomal translocations. DNA Repair . 2006;5: 1093–1108. doi:10.1016/j.dnarep.2006.05.031

168. Cowell IG, Sondka Z, Smith K, Lee KC, Manville CM, Sidorczuk-Lesthuruge M, et al. Model for MLL translocations in therapy-related leukemia involving topoisomerase IIβ-mediated DNA strand breaks and gene proximity. Proc Natl Acad Sci U S A. 2012;109: 8989–8994. doi:10.1073/pnas.1204406109

169. Azarova AM, Lyu YL, Lin C-P, Tsai Y-C, Lau JY-N, Wang JC, et al. Roles of DNA topoisomerase II isozymes in chemotherapy and secondary malignancies. Proc Natl Acad Sci U S A. National Acad Sciences; 2007;104: 11014–11019. doi:10.1073/pnas.0704002104

170. Ratain MJ, Kaminer LS, Bitran JD, Larson RA, Le Beau MM, Skosey C, et al. Acute nonlymphocytic leukemia following etoposide and cisplatin combination chemotherapy for advanced non-small-cell carcinoma of the lung. Blood. 1987;70: 1412–1417. Available: https://www.ncbi.nlm.nih.gov/pubmed/2822173

171. Zhang S, Liu X, Bawa-Khalfe T, Lu L-S, Lyu YL, Liu LF, et al. Identification of the molecular basis of doxorubicin-induced cardiotoxicity. Nat Med. 2012;18: 1639–1642. doi:10.1038/nm.2919

172. Vejpongsa P, Yeh ETH. Topoisomerase 2β: A Promising Molecular Target for Primary Prevention of Anthracycline-Induced Cardiotoxicity. Clinical Pharmacology & Therapeutics. Wiley Online Library; 2014;95: 45–52. Available: http://onlinelibrary.wiley.com/doi/10.1038/clpt.2013.201/full

173. Yeh ETH, Bickford CL. Cardiovascular complications of cancer therapy: incidence, pathogenesis, diagnosis, and management. J Am Coll Cardiol. 2009;53: 2231–2247. doi:10.1016/j.jacc.2009.02.050

174. Mariani A, Bartoli A, Atwal M, Lee KC, Austin CA, Rodriguez R. Differential Targeting of Human Topoisomerase II Isoforms with Small Molecules. J Med Chem. 2015;58: 4851–4856. doi:10.1021/acs.jmedchem.5b00473

175. Yang J, Bogni A, Schuetz EG, Ratain M, Dolan ME, McLeod H, et al. Etoposide pathway. Pharmacogenet Genomics. 2009;19: 552–553. doi:10.1097/FPC.0b013e32832e0e7f

176. Sim SC, Altman RB, Ingelman-Sundberg M. Databases in the area of pharmacogenetics. Hum Mutat. 2011;32: 526–531. doi:10.1002/humu.21454

177. Scripture CD, Figg WD. Drug interactions in cancer therapy. Nat Rev Cancer. 2006;6: 546–558. doi:10.1038/nrc1887

178. 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. 2019;10: 2680. doi:10.1038/s41467-019-10636-9

179. Ravenscroft P, Brammer H, Richards K. Arsenic pollution : a global synthesis [Internet]. Chichester, U.K.: Wiley-Blackwell; 2009. Available: http://www.worldcat.org/oclc/214285927?referer=xid

180. Ratnaike RN. Acute and chronic arsenic toxicity. Postgrad Med J. 2003;79: 391–396. doi:10.1136/pmj.79.933.391

181. Mandal BK, Suzuki KT. Arsenic round the world: a review. Talanta. 2002;58: 201–235. doi:10.1016/S0039-9140(02)00268-0

182. Khairul I, Wang QQ, Jiang YH, Wang C, Naranmandura H. Metabolism, toxicity and anticancer activities of arsenic compounds. Oncotarget. 2017;8: 23905–23926. doi:10.18632/oncotarget.14733

183. Stýblo M, Drobná Z, Jaspers I, Lin S, Thomas DJ. The role of biomethylation in toxicity and carcinogenicity of arsenic: a research update. Environ Health Perspect. 2002;110 Suppl 5: 767–771. Available: https://www.ncbi.nlm.nih.gov/pubmed/12426129

184. Schlebusch CM, Gattepaille LM, Engström K, Vahter M, Jakobsson M, Broberg K. Human adaptation to arsenic-rich environments. Mol Biol Evol. 2015;32: 1544–1555. doi:10.1093/molbev/msv046

185. Chung C-J, Hsueh Y-M, Bai C-H, Huang Y-K, Huang Y-L, Yang M-H, et al. Polymorphisms in arsenic metabolism genes, urinary arsenic methylation profile and cancer. Cancer Causes Control. 2009;20: 1653–1661. doi:10.1007/s10552-009-9413-0

186. Fujihara J, Fujii Y, Agusa T, Kunito T, Yasuda T, Moritani T, et al. Ethnic differences in five intronic polymorphisms associated with arsenic metabolism within human arsenic (+3 oxidation state) methyltransferase (AS3MT) gene. Toxicol Appl Pharmacol. 2009;234: 41–46. doi:10.1016/j.taap.2008.09.026

187. Gomez-Rubio P, Meza-Montenegro MM, Cantu-Soto E, Klimecki WT. Genetic association between intronic variants in AS3MT and arsenic methylation efficiency is focused on a large linkage disequilibrium cluster in chromosome 10. J Appl Toxicol. 2010;30: 260–270. doi:10.1002/jat.1492

188. Chen GQ, Shi XG, Tang W, Xiong SM, Zhu J, Cai X, et al. Use of arsenic trioxide (As2O3) in the treatment of acute promyelocytic leukemia (APL): I. As2O3 exerts dose-dependent dual effects on APL cells. Blood. 1997;89: 3345–3353. Available: https://www.ncbi.nlm.nih.gov/pubmed/9129041

189. Antman KH. Introduction: the history of arsenic trioxide in cancer therapy. Oncologist. 2001;6 Suppl 2: 1–2. doi:10.1634/theoncologist.6-suppl_2-1

190. Murgo AJ. Clinical trials of arsenic trioxide in hematologic and solid tumors: overview of the National Cancer Institute Cooperative Research and Development Studies. Oncologist. 2001;6 Suppl 2: 22–28. doi:10.1634/theoncologist.6-suppl_2-22

191. Emi N. Arsenic Trioxide: Clinical Pharmacology and Therapeutic Results. Chemotherapy for Leukemia. Springer, Singapore; 2017. pp. 221–238. doi:10.1007/978-981-10-3332-2_13

192. de Thé H, Chomienne C, Lanotte M, Degos L, Dejean A. The t(15;17) translocation of acute promyelocytic leukaemia fuses the retinoic acid receptor alpha gene to a novel transcribed locus. Nature. 1990;347: 558–561. doi:10.1038/347558a0

193. Grignani F, Valtieri M, Gabbianelli M, Gelmetti V, Botta R, Luchetti L, et al. PML/RAR

alpha fusion protein expression in normal human hematopoietic progenitors dictates myeloid commitment and the promyelocytic phenotype. Blood. 2000;96: 1531–1537. Available: https://www.ncbi.nlm.nih.gov/pubmed/10942402

194. Zhang X-W, Yan X-J, Zhou Z-R, Yang F-F, Wu Z-Y, Sun H-B, et al. Arsenic trioxide controls the fate of the PML-RARalpha oncoprotein by directly binding PML. Science. 2010;328: 240–243. doi:10.1126/science.1183424

195. Tomita A, Kiyoi H, Naoe T. Mechanisms of action and resistance to all-trans retinoic acid (ATRA) and arsenic trioxide (As2O 3) in acute promyelocytic leukemia. Int J Hematol. 2013;97: 717–725. doi:10.1007/s12185-013-1354-4

196. Hoonjan M, Jadhav V, Bhatt P. Arsenic trioxide: insights into its evolution to an anticancer agent. J Biol Inorg Chem. 2018;23: 313–329. doi:10.1007/s00775-018-1537-9

197. Zeidan AM, Gore SD. New strategies in acute promyelocytic leukemia: moving to an entirely oral, chemotherapy-free upfront management approach. Clin Cancer Res. 2014;20: 4985–4993. doi:10.1158/1078-0432.CCR-13-2725

198. Kniazeva M, Crawford QT, Seiber M, Wang C-Y, Han M. Monomethyl branched-chain fatty acids play an essential role in Caenorhabditis elegans development. PLoS Biol. 2004;2: E257. doi:10.1371/journal.pbio.0020257

199. Luz AL, Godebo TR, Smith LL, Leuthner TC, Maurer LL, Meyer JN. Deficiencies in mitochondrial dynamics sensitize Caenorhabditis elegans to arsenite and other mitochondrial toxicants by reducing mitochondrial adaptability. Toxicology. 2017;387: 81–94. doi:10.1016/j.tox.2017.05.018

200. Spracklin G, Fields B, Wan G, Becker D, Wallig A, Shukla A, et al. The RNAi Inheritance Machinery of Caenorhabditis elegans. Genetics. 2017;206: 1403–1416. doi:10.1534/genetics.116.198812

201. Watson E, MacNeil LT, Arda HE, Zhu LJ, Walhout AJM. Integration of Metabolic and Gene Regulatory Networks Modulates the C. elegans Dietary Response. Cell. Elsevier; 2013;153: 253–266. doi:10.1016/j.cell.2013.02.050

202. Luz AL, Meyer JN. Effects of reduced mitochondrial DNA content on secondary mitochondrial toxicant exposure in Caenorhabditis elegans. Mitochondrion. 2016;30: 255–264. doi:10.1016/j.mito.2016.08.014

203. Luz AL, Godebo TR, Bhatt DP, Ilkayeva OR, Maurer LL, Hirschey MD, et al. From the Cover: Arsenite Uncouples Mitochondrial Respiration and Induces a Warburg-like Effect in Caenorhabditis elegans. Toxicol Sci. 2016;152: 349–362. doi:10.1093/toxsci/kfw093

204. Schmeisser S, Schmeisser K, Weimer S, Groth M, Priebe S, Fazius E, et al. Mitochondrial hormesis links low-dose arsenite exposure to lifespan extension. Aging Cell. 2013;12: 508–517. doi:10.1111/acel.12076

205. Wyatt LH, Luz AL, Cao X, Maurer LL, Blawas AM, Aballay A, et al. Effects of methyl and inorganic mercury exposure on genome homeostasis and mitochondrial function in Caenorhabditis elegans. DNA Repair . 2017;52: 31–48. doi:10.1016/j.dnarep.2017.02.005

206. Large EE, Xu W, Zhao Y, Brady SC, Long L, Butcher RA, et al. Selection on a Subunit of the NURF Chromatin Remodeler Modifies Life History Traits in a Domesticated Strain of Caenorhabditis elegans. PLoS Genet. Public Library of Science; 2016;12: e1006219. doi:10.1371/journal.pgen.1006219

207. Wang Y, Ezemaduka AN, Li Z, Chen Z, Song C. Joint Toxicity of Arsenic, Copper and Glyphosate on Behavior, Reproduction and Heat Shock Protein Response in Caenorhabditis elegans. Bull Environ Contam Toxicol. 2017;98: 465–471. doi:10.1007/s00128-017-2042-5

208. Zdraljevic S, Andersen EC. Natural diversity facilitates the discovery of conserved chemotherapeutic response mechanisms. Curr Opin Genet Dev. 2017;47: 41–47. doi:10.1016/j.gde.2017.08.002

209. Jia F, Cui M, Than MT, Han M. Developmental Defects of Caenorhabditis elegans Lacking Branched-chain α-Ketoacid Dehydrogenase Are Mainly Caused by Monomethyl Branched-chain Fatty Acid Deficiency. J Biol Chem. 2016;291: 2967–2973. doi:10.1074/jbc.M115.676650

210. García-González AP, Ritter AD, Shrestha S, Andersen EC, Yilmaz LS, Walhout AJM. Bacterial Metabolism Affects the C. elegans Response to Cancer Chemotherapeutics. Cell. 2017;169: 431–441.e8. doi:10.1016/j.cell.2017.03.046

211. Team RC. R: A language and environment for statistical computing [Internet]. Vienna, Austria; 2014. 2017.

212. Broman KW, Wu H, Sen S, Churchill GA. R/qtl: QTL mapping in experimental crosses. Bioinformatics. 2003;19: 889–890. Available: https://www.ncbi.nlm.nih.gov/pubmed/12724300

213. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models using lme4 [Internet]. arXiv [stat.CO]. 2014. Available: http://arxiv.org/abs/1406.5823

214. Bloom JS, Kotenko I, Sadhu MJ, Treusch S, Albert FW, Kruglyak L. Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. Nat Commun. 2015;6: 8712. doi:10.1038/ncomms9712

215. David Clifford PM. The regress function. R News. 2006: 6–10. Available: https://cran.r-project.org/web/packages/regress/regress.pdf

216. Covarrubias-Pazaran G. Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. PLoS One. 2016;11: e0156744. doi:10.1371/journal.pone.0156744

217. Su G, Christensen OF, Ostersen T, Henryon M, Lund MS. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS One. 2012;7: e45293. doi:10.1371/journal.pone.0045293

218. Endelman JB, Jannink J-L. Shrinkage estimation of the realized relationship matrix. G3 . 2012;2: 1405–1413. doi:10.1534/g3.112.004259

219. Lüdecke D. sjstats: Statistical Functions for Regression Models [Internet]. 2018.

doi:10.5281/zenodo.1442812

220. Brady S, Evans K, Bloom J, Tanny R, Cook D, Giuliani S, et al. Common loci underlie natural variation in diverse toxin responses [Internet]. bioRxiv. 2018. p. 325399. doi:10.1101/325399

221. Qiu Y. RSpectra [Internet]. Github; Available: https://github.com/yixuan/RSpectra

222. Bilgrau AE. correlateR [Internet]. Github; 2018. Available: https://github.com/AEBilgrau/correlateR

223. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity . 2005;95: 221–227. doi:10.1038/sj.hdy.6800717

224. Artyukhin AB, Zhang YK, Akagi AE, Panda O, Sternberg PW, Schroeder FC. Metabolomic “Dark Matter” Dependent on Peroxisomal β-Oxidation in Caenorhabditis elegans. J Am Chem Soc. 2018;140: 2841–2852. doi:10.1021/jacs.7b11811

225. Cohen J. Statistical Power Analysis for the Behavioral Sciences [Internet]. Routledge; 2013. Available: https://market.android.com/details?id=book-2v9zDAsLvA0C

226. Adeva-Andany MM, López-Maside L, Donapetry-García C, Fernández-Fernández C, Sixto-Leal C. Enzymes involved in branched-chain amino acid metabolism in humans. Amino Acids. 2017;49: 1005–1028. doi:10.1007/s00726-017-2412-7

227. Bergquist ER, Fischer RJ, Sugden KD, Martin BD. Inhibition by methylated organo-arsenicals of the respiratory 2-oxo-acid dehydrogenases. J Organomet Chem. 2009;694: 973–980. doi:10.1016/j.jorganchem.2008.12.028

228. Reed LJ, Hackert ML. Structure-function relationships in dihydrolipoamide acyltransferases. J Biol Chem. 1990;265: 8971–8974. Available: https://www.ncbi.nlm.nih.gov/pubmed/2188967

229. Kniazeva M, Euler T, Han M. A branched-chain fatty acid is involved in post-embryonic growth control in parallel to the insulin receptor pathway and its biosynthesis is feedback-regulated in C. elegans. Genes Dev. 2008;22: 2102–2110. doi:10.1101/gad.1692008

230. Baugh LR. To Grow or Not to Grow: Nutritional Control of Development During Caenorhabditis elegans L1 Arrest. Genetics. 2013;194: 539–555. doi:10.1534/genetics.113.150847

231. Watts JL, Ristow M. Lipid and Carbohydrate Metabolism in Caenorhabditis elegans. Genetics. 2017;207: 413–446. doi:10.1534/genetics.117.300106

232. Entchev EV, Schwudke D, Zagoriy V, Matyash V, Bogdanova A, Habermann B, et al. LET-767 is required for the production of branched chain and long chain fatty acids in Caenorhabditis elegans. J Biol Chem. 2008;283: 17550–17560. doi:10.1074/jbc.M800965200

233. Zhu H, Shen H, Sewell AK, Kniazeva M, Han M. A novel sphingolipid-TORC1 pathway critically promotes postembryonic development in Caenorhabditis elegans. Elife. 2013;2: e00429. doi:10.7554/eLife.00429

234. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet. 2008;Chapter 10: Unit 10.11. doi:10.1002/0471142905.hg1011s57

235. Paul S, Banerjee N, Chatterjee A, Sau TJ, Das JK, Mishra PK, et al. Arsenic-induced promoter hypomethylation and over-expression of ERCC2 reduces DNA repair capacity in humans by non-disjunction of the ERCC2-Cdk7 complex. Metallomics. 2014;6: 864–873. doi:10.1039/c3mt00328k

236. Shen S, Li X-F, Cullen WR, Weinfeld M, Le XC. Arsenic binding to proteins. Chem Rev. 2013;113: 7769–7792. doi:10.1021/cr300015c

237. Li J, Packianathan C, Rossman TG, Rosen BP. Nonsynonymous Polymorphisms in the Human AS3MT Arsenic Methylation Gene: Implications for Arsenic Toxicity. Chem Res Toxicol. 2017;30: 1481–1491. doi:10.1021/acs.chemrestox.7b00113

238. Pettit FH, Yeaman SJ, Reed LJ. Purification and characterization of branched chain alpha-keto acid dehydrogenase complex of bovine kidney. Proc Natl Acad Sci U S A. 1978;75: 4881–4885. Available: https://www.ncbi.nlm.nih.gov/pubmed/283398

239. Heffelfinger SC, Sewell ET, Danner DJ. Identification of specific subunits of highly purified bovine liver branched-chain ketoacid dehydrogenase. Biochemistry. 1983;22: 5519–5522. doi:10.1021/bi00293a011

240. Yeaman SJ. The 2-oxo acid dehydrogenase complexes: recent advances. Biochem J. 1989;257: 625–632. Available: https://www.ncbi.nlm.nih.gov/pubmed/2649080

241. Kato M, Chuang JL, Tso S-C, Wynn RM, Chuang DT. Crystal structure of pyruvate dehydrogenase kinase 3 bound to lipoyl domain 2 of human pyruvate dehydrogenase complex. EMBO J. 2005;24: 1763–1774. doi:10.1038/sj.emboj.7600663

242. Burrage LC, Nagamani SCS, Campeau PM, Lee BH. Branched-chain amino acid metabolism: from rare Mendelian diseases to more common disorders. Hum Mol Genet. 2014;23: R1–8. doi:10.1093/hmg/ddu123

243. Tönjes M, Barbus S, Park YJ, Wang W, Schlotter M, Lindroth AM, et al. BCAT1 promotes cell proliferation through amino acid catabolism in gliomas carrying wild-type IDH1. Nat Med. 2013;19: 901–908. doi:10.1038/nm.3217

244. Hughes-Fulford M, Chen Y, Tjandrawinata RR. Fatty acid regulates gene expression and growth of human prostate cancer PC-3 cells. Carcinogenesis. 2001;22: 701–707. doi:10.1093/carcin/22.5.701

245. Agostini M, Silva SD, Zecchin KG, Coletta RD, Jorge J, Loda M, et al. Fatty acid synthase is required for the proliferation of human oral squamous carcinoma cells. Oral Oncol. 2004;40: 728–735. doi:10.1016/j.oraloncology.2004.01.011

246. Coombs CC, Tavakkoli M, Tallman MS. Acute promyelocytic leukemia: where did we start, where are we now, and the future. Blood Cancer J. 2015;5: e304. doi:10.1038/bcj.2015.25

247. Thompson OA, Snoek LB, Nijveen H, Sterken MG, Volkers RJM, Brenchley R, et al.

Remarkably Divergent Regions Punctuate the Genome Assembly of the Caenorhabditis elegans Hawaiian Strain CB4856. Genetics. 2015;200: 975–989. doi:10.1534/genetics.115.175950

248. J. Hodgkin TD. Natural Variation and Copulatory Plug Formation in Caenorhabditis Elegans. Genetics. Genetics Society of America; 1997;146: 149. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1207933/

249. DeBerardinis RJ, Chandel NS. Fundamentals of cancer metabolism. Sci Adv. 2016;2: e1600200. doi:10.1126/sciadv.1600200

250. Hotez PJ, Alvarado M, Basáñez M-G, Bolliger I, Bourne R, Boussinesq M, et al. The global burden of disease study 2010: interpretation and implications for the neglected tropical diseases. PLoS Negl Trop Dis. 2014;8: e2865. doi:10.1371/journal.pntd.0002865

251. Lustigman S, Prichard RK, Gazzinelli A, Grant WN, Boatin BA, McCarthy JS, et al. A research agenda for helminth diseases of humans: the problem of helminthiases. PLoS Negl Trop Dis. 2012;6: e1582. doi:10.1371/journal.pntd.0001582

252. Charlier J, van der Voort M, Kenyon F, Skuce P, Vercruysse J. Chasing helminths and their economic impact on farmed ruminants. Trends Parasitol. 2014;30: 361–367. doi:10.1016/j.pt.2014.04.009

253. Kotze AC, Hunt PW, Skuce P, von Samson-Himmelstjerna G, Martin RJ, Sager H, et al. Recent advances in candidate-gene and whole-genome approaches to the discovery of anthelmintic resistance markers and the description of drug/receptor interactions. Int J Parasitol Drugs Drug Resist. 2014;4: 164–184. doi:10.1016/j.ijpddr.2014.07.007

254. De Clercq D, Sacko M, Behnke J, Gilbert F, Dorny P, Vercruysse J. Failure of mebendazole in treatment of human hookworm infections in the southern region of Mali. Am J Trop Med Hyg. 1997;57: 25–30. Available: https://www.ncbi.nlm.nih.gov/pubmed/9242313

255. Albonico M, Bickle Q, Ramsan M, Montresor A, Savioli L, Taylor M. Efficacy of mebendazole and levamisole alone or in combination against intestinal nematode infections after repeated targeted mebendazole treatment in Zanzibar. Bull World Health Organ. 2003;81: 343–352. Available: https://www.ncbi.nlm.nih.gov/pubmed/12856052

256. Humphries D, Simms BT, Davey D, Otchere J, Quagraine J, Terryah S, et al. Hookworm infection among school age children in Kintampo north municipality, Ghana: nutritional risk factors and response to albendazole treatment. Am J Trop Med Hyg. 2013;89: 540–548. doi:10.4269/ajtmh.12-0605

257. Neff NF, Thomas JH, Grisafi P, Botstein D. Isolation of the beta-tubulin gene from yeast and demonstration of its essential function in vivo. Cell. 1983;33: 211–219. Available: https://www.ncbi.nlm.nih.gov/pubmed/6380751

258. Laclette JP, Guerra G, Zetina C. Inhibition of tubulin polymerization by mebendazole. Biochem Biophys Res Commun. 1980;92: 417–423. doi:10.1016/0006-291X(80)90349-6

259. Ireland CM, Gull K, Gutteridge WE, Pogson CI. The interaction of benzimidazole carbamates with mammalian microtobule protein. Biochem Pharmacol. 1979;28: 2680–

2682. doi:10.1016/0006-2952(79)90049-2

260. Thomas JH, Neff NF, Botstein D. Isolation and characterization of mutations in the beta-tubulin gene of Saccharomyces cerevisiae. Genetics. 1985;111: 715–734. Available: https://www.ncbi.nlm.nih.gov/pubmed/2998923

261. Driscoll M, Dean E, Reilly E, Bergholz E, Chalfie M. Genetic and molecular analysis of a Caenorhabditis elegans beta-tubulin that conveys benzimidazole sensitivity. J Cell Biol. 1989;109: 2993–3003. Available: https://www.ncbi.nlm.nih.gov/pubmed/2592410

262. Kwa MS, Veenstra JG, Roos MH. Benzimidazole resistance in Haemonchus contortus is correlated with a conserved mutation at amino acid 200 in beta-tubulin isotype 1. Mol Biochem Parasitol. 1994;63: 299–303. Available: https://www.ncbi.nlm.nih.gov/pubmed/7911975

263. Silvestre A, Cabaret J. Mutation in position 167 of isotype 1 beta-tubulin gene of Trichostrongylid nematodes: role in benzimidazole resistance? Mol Biochem Parasitol. 2002;120: 297–300. Available: https://www.ncbi.nlm.nih.gov/pubmed/11897135

264. Ghisi M, Kaminsky R, Mäser P. Phenotyping and genotyping of Haemonchus contortus isolates reveals a new putative candidate mutation for benzimidazole resistance in nematodes. Vet Parasitol. 2007;144: 313–320. doi:10.1016/j.vetpar.2006.10.003

265. Lacey E, Gill JH. Biochemistry of benzimidazole resistance. Acta Trop. 1994;56: 245–262. Available: https://www.ncbi.nlm.nih.gov/pubmed/8203306

266. Lubega GW, Prichard RK. Specific interaction of benzimidazole anthelmintics with tubulin: high-affinity binding and benzimidazole resistance in Haemonchus contortus. Mol Biochem Parasitol. 1990;38: 221–232. Available: https://www.ncbi.nlm.nih.gov/pubmed/2325707

267. Krücken J, Fraundorfer K, Mugisha JC, Ramünke S, Sifft KC, Geus D, et al. Reduced efficacy of albendazole against Ascaris lumbricoides in Rwandan schoolchildren. Int J Parasitol Drugs Drug Resist. 2017;7: 262–271. doi:10.1016/j.ijpddr.2017.06.001

268. Yilmaz E, Ramünke S, Demeler J, Krücken J. Comparison of constitutive and thiabendazole-induced expression of five cytochrome P450 genes in fourth-stage larvae of Haemonchus contortus isolates with different drug susceptibility identifies one gene with high constitutive expression in a multi-resistant isolate. Int J Parasitol Drugs Drug Resist. 2017;7: 362–369. doi:10.1016/j.ijpddr.2017.10.001

269. Jones LM, Flemming AJ, Urwin PE. NHR-176 regulates cyp-35d1 to control hydroxylation-dependent metabolism of thiabendazole in Caenorhabditis elegans. Biochem J. 2015;466: 37–44. doi:10.1042/BJ20141296

270. Doyle SR, Bourguinat C, Nana-Djeunga HC, Kengne-Ouafo JA, Pion SDS, Bopda J, 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. 2017;11: e0005816. doi:10.1371/journal.pntd.0005816

271. Holden-Dye L, Walker RJ. Anthelmintic drugs and nematicides: studies in Caenorhabditis elegans. WormBook. 2014; 1–29. doi:10.1895/wormbook.1.143.2

272. Hunt PW, Kotze AC, Knox MR, Anderson LJ, McNally J, LE Jambre LF. The use of DNA markers to map anthelmintic resistance loci in an intraspecific cross of Haemonchus contortus. Parasitology. 2010;137: 705–717. doi:10.1017/S0031182009991521

273. Stear MJ, 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. 2009;31: 274–282. doi:10.1111/j.1365-3024.2009.01105.x

274. Doyle SR, Illingworth CJR, Laing R, Bartley DJ, Redman E, Martinelli A, et al. A major locus for ivermectin resistance in a parasitic nematode [Internet]. bioRxiv. 2018. p. 298901. doi:10.1101/298901

275. Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J. WormBase: network access to the genome and biology of Caenorhabditis elegans. Nucleic Acids Res. 2001;29: 82–86. doi:10.1093/nar/29.1.82

276. C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. 1998;282: 2012–2018. Available: https://www.ncbi.nlm.nih.gov/pubmed/9851916

277. Strange K. Drug Discovery in Fish, Flies, and Worms. ILAR J. 2016;57: 133–143. doi:10.1093/ilar/ilw034

278. Zamanian M, Cook DE, Zdraljevic S, Brady SC, Lee D, Lee J, et al. Discovery of unique loci that underlie nematode responses to benzimidazoles [Internet]. 2017. doi:10.1101/116970

279. Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics. 2016;32: 1423–1426. doi:10.1093/bioinformatics/btw079

280. Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei L-J, et al. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010;86: 832–838. doi:10.1016/j.ajhg.2010.04.005

281. Browning BL, Browning SR. Genotype Imputation with Millions of Reference Samples. Am J Hum Genet. 2016;98: 116–126. doi:10.1016/j.ajhg.2015.11.020

282. Prior H, Jawad AK, MacConnachie L, Beg AA. Highly Efficient, Rapid and Co-CRISPR Independent Genome Editing in Caenorhabditis elegans. G3: Genes|Genomes|Genetics. 2017; g3.300216.2017. doi:10.1534/g3.117.300216

283. Zhao Y, Long L, Xu W, Campbell RF, Large EL, Greene JS, et al. Laboratory evolution from social to solitary behavior in the N2 reference strain is unnecessary for its fitness advantages [Internet]. bioRxiv. 2018. p. 309997. doi:10.1101/309997

284. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10: 845–858. doi:10.1038/nprot.2015.053

285. De Lano WL, California U. PyMOL: An Open-Source Molecular Graphics Tool. Available:

http://www.ccp4.ac.uk/newsletters/newsletter40/11_pymol.pdf

286. Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ. PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. Mol Biol Evol. Oxford University Press; 2014;31: 1929–1936. doi:10.1093/molbev/msu136

287. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2014. 2016.

288. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–2158. doi:10.1093/bioinformatics/btr330

289. McVean G, Awadalla P, Fearnhead P. A Coalescent-Based Method for Detecting and Estimating Recombination From Gene Sequences. Genetics. Genetics; 2002;160: 1231–1241. Available: http://www.genetics.org/content/160/3/1231

290. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32: 1792–1797. doi:10.1093/nar/gkh340

291. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25: 1189–1191. doi:10.1093/bioinformatics/btp033

292. Warnes G, Gorjanc G, Leisch F, Man M. Genetics: Population genetics. R package version 1.3. 6. 2012.

293. Kishore SP, Blank E, Heller DJ, Patel A, Peters A, Price M, et al. Modernizing the World Health Organization List of Essential Medicines for Preventing and Controlling Cardiovascular Diseases. J Am Coll Cardiol. 2018;71: 564–574. doi:10.1016/j.jacc.2017.11.056

294. Aguayo-Ortiz R, Méndez-Lucio O, Medina-Franco JL, Castillo R, Yépez-Mulia L, Hernández-Luis F, et al. Towards the identification of the binding site of benzimidazoles to β-tubulin of Trichinella spiralis: insights from computational and experimental data. J Mol Graph Model. 2013;41: 12–19. doi:10.1016/j.jmgm.2013.01.007

295. Aguayo-Ortiz R, Méndez-Lucio O, Romo-Mancillas A, Castillo R, Yépez-Mulia L, Medina-Franco JL, et al. Molecular basis for benzimidazole resistance from a novel β-tubulin binding site model. J Mol Graph Model. 2013;45: 26–37. doi:10.1016/j.jmgm.2013.07.008

296. Robinson MW, McFerran N, Trudgett A, Hoey L, Fairweather I. A possible model of benzimidazole binding to β-tubulin disclosed by invoking an inter-domain movement. J Mol Graph Model. 2004;23: 275–284. doi:10.1016/j.jmgm.2004.08.001

297. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. Nucleic Acids Res. 2016;44: D81–9. doi:10.1093/nar/gkv1272

298. Oosumi T, Garlick B, Belknap WR. Identification of putative nonautonomous transposable elements associated with several transposon families in Caenorhabditis elegans. J Mol Evol. 1996;43: 11–18. Available:

https://www.ncbi.nlm.nih.gov/pubmed/8660424

299. Li WH, Wu CI, Luo CC. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol. 1985;2: 150–174. doi:10.1093/oxfordjournals.molbev.a040343

300. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. 2000;155: 1405–1413. Available: https://www.ncbi.nlm.nih.gov/pubmed/10880498

301. Zeng K, Fu Y-X, Shi S, Wu C-I. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics. 2006;174: 1431–1439. doi:10.1534/genetics.106.061432

302. Furtado LFV, de Paiva Bello ACP, Rabelo ÉML. Benzimidazole resistance in helminths: From problem to diagnosis. Acta Trop. 2016;162: 95–102. doi:10.1016/j.actatropica.2016.06.021

303. Clarke NE, Doi SAR, Wangdi K, Chen Y, Clements ACA, Nery SV. Efficacy of anthelminthic drugs and drug combinations against soil-transmitted helminths: a systematic review and network meta-analysis. Clin Infect Dis. 2018; doi:10.1093/cid/ciy423

304. Gilleard JS. Understanding anthelmintic resistance: the need for genomics and genetics. Int J Parasitol. 2006;36: 1227–1239. doi:10.1016/j.ijpara.2006.06.010

305. Redman E, Sargison N, Whitelaw F, Jackson F, Morrison A, Bartley DJ, et al. Introgression of ivermectin resistance genes into a susceptible Haemonchus contortus strain by multiple backcrossing. PLoS Pathog. 2012;8: e1002534. doi:10.1371/journal.ppat.1002534

306. Valentim CLL, Cioli D, Chevalier FD, Cao X, Taylor AB, Holloway SP, et al. Genetic and Molecular Basis of Drug Resistance and Species-Specific Drug Action in Schistosome Parasites. Science. American Association for the Advancement of Science; 2013;342: 1385–1389. doi:10.1126/science.1243106

307. Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. WormBase ParaSite - a comprehensive resource for helminth genomics. Mol Biochem Parasitol. 2017;215: 2–10. doi:10.1016/j.molbiopara.2016.11.005

308. Crofts TS, Men Y, Alvarez-Cohen L, Taga ME. A bioassay for the detection of benzimidazoles reveals their presence in a range of environmental samples. Front Microbiol. 2014;5. doi:10.3389/fmicb.2014.00592

309. Oliver RP, Geoff Hewitt H. Fungicides in Crop Protection, 2nd Edition [Internet]. CABI; 2014. Available: https://market.android.com/details?id=book-4x0xBQAAQBAJ

310. Prchal L, Podlipná R, Lamka J, Dědková T, Skálová L, Vokřál I, et al. Albendazole in environment: faecal concentrations in lambs and impact on lower development stages of helminths and seed germination. Environ Sci Pollut Res Int. 2016;23: 13015–13022. doi:10.1007/s11356-016-6472-0

311. Perruchon C, Pantoleon A, Veroutis D, Gallego-Blanco S, Martin-Laurent F, Liadaki K, et al. Characterization of the biodegradation, bioremediation and detoxification capacity of a

bacterial consortium able to degrade the fungicide thiabendazole. Biodegradation. 2017;28: 383–394. doi:10.1007/s10532-017-9803-z

312. Cycoń M, Mrozik A, Piotrowska-Seget Z. Bioaugmentation as a strategy for the remediation of pesticide-polluted soil: A review. Chemosphere. 2017;172: 52–71. doi:10.1016/j.chemosphere.2016.12.129

313. Taube J, Vorkamp K, Förster M, Herrmann R. Pesticide residues in biological waste. Chemosphere. 2002;49: 1357–1365. doi:10.1016/S0045-6535(02)00503-9

314. de Oliveira Neto OF, Arenas AY, Fostier AH. Sorption of thiabendazole in sub-tropical Brazilian soils. Environ Sci Pollut Res Int. 2017;24: 16503–16512. doi:10.1007/s11356-017-9226-8

315. Hawkins NJ, Fraaije BA. Predicting Resistance by Mutagenesis: Lessons from 45 Years of MBC Resistance. Front Microbiol. 2016;7: 1814. doi:10.3389/fmicb.2016.01814

316. Oakley BR. Tubulins in Aspergillus nidulans. Fungal Genet Biol. 2004;41: 420–427. doi:10.1016/j.fgb.2003.11.013

317. Hesse J, Thierauf M, Ponstingl H. Tubulin sequence region beta 155-174 is involved in binding exchangeable guanosine triphosphate. J Biol Chem. 1987;262: 15472–15475. Available: https://www.ncbi.nlm.nih.gov/pubmed/3680207

318. McKean PG, Vaughan S, Gull K. The extended tubulin superfamily. J Cell Sci. 2001;114: 2723–2733. Available: https://www.ncbi.nlm.nih.gov/pubmed/11683407

319. Löwe J, Li H, Downing KH, Nogales E. Refined structure of alpha beta-tubulin at 3.5 A resolution. J Mol Biol. 2001;313: 1045–1057. doi:10.1006/jmbi.2001.5077

320. Lu C, Srayko M, Mains PE. The Caenorhabditis elegans microtubule-severing complex MEI-1/MEI-2 katanin interacts differently with two superficially redundant beta-tubulin isotypes. Mol Biol Cell. 2004;15: 142–150. doi:10.1091/mbc.E03-06-0418

321. Ellis GC, Phillips JB, O'Rourke S, Lyczak R, Bowerman B. Maternally expressed and partially redundant beta-tubulins in Caenorhabditis elegans are autoregulated. J Cell Sci. 2004;117: 457–464. doi:10.1242/jcs.00869

322. Hutter H, Suh J. GExplore 1.4: An expanded web interface for queries on Caenorhabditis elegans protein and gene function. Worm. 2016;5: e1234659. doi:10.1080/21624054.2016.1234659

323. Hurd DD. Tubulins in C. elegans. WormBook. 2018; 1–34. doi:10.1895/wormbook.1.182.1

324. Hao L, Thein M, Brust-Mascher I, Civelekoglu-Scholey G, Lu Y, Acar S, et al. Intraflagellar transport delivers tubulin isotypes to sensory cilium middle and distal segments. Nat Cell Biol. 2011;13: 790–798. doi:10.1038/ncb2268

325. Bounoutas A, O'Hagan R, Chalfie M. The multipurpose 15-protofilament microtubules in C. elegans have specific roles in mechanosensation. Curr Biol. 2009;19: 1362–1367. doi:10.1016/j.cub.2009.06.036

326. Savage C, Hamelin M, Culotti JG, Coulson A, Albertson DG, Chalfie M. mec-7 is a beta-tubulin gene required for the production of 15-protofilament microtubules in Caenorhabditis elegans. Genes Dev. 1989;3: 870–881. Available: https://www.ncbi.nlm.nih.gov/pubmed/2744465

327. Munkácsy E, Khan MH, Lane RK, Borror MB, Park JH, Bokov AF, et al. DLK-1, SEK-3 and PMK-3 Are Required for the Life Extension Induced by Mitochondrial Bioenergetic Disruption in C. elegans. PLoS Genet. 2016;12: e1006133. doi:10.1371/journal.pgen.1006133

328. Saunders GI, Wasmuth JD, Beech R, Laing R, Hunt M, Naghra H, et al. Characterization and comparative analysis of the complete Haemonchus contortus β-tubulin gene family and implications for benzimidazole resistance in strongylid nematodes. Int J Parasitol. 2013;43: 465–475. doi:10.1016/j.ijpara.2012.12.011

329. Kwa MS, Kooyman FN, Boersema JH, Roos MH. Effect of selection for benzimidazole resistance in Haemonchus contortus on beta-tubulin isotype 1 and isotype 2 genes. Biochem Biophys Res Commun. 1993;191: 413–419. doi:10.1006/bbrc.1993.1233

330. Beech RN, Prichard RK, Scott ME. Genetic variability of the beta-tubulin genes in benzimidazole-susceptible and -resistant strains of Haemonchus contortus. Genetics. 1994;138: 103–110. Available: https://www.ncbi.nlm.nih.gov/pubmed/8001777

331. Prichard R. Genetic variability following selection of Haemonchus contortus with anthelmintics. Trends Parasitol. 2001;17: 445–453. Available: https://www.ncbi.nlm.nih.gov/pubmed/11530357

332. Demeler J, Krüger N, Krücken J, von der Heyden VC, Ramünke S, Küttler U, et al. Phylogenetic characterization of β-tubulins and development of pyrosequencing assays for benzimidazole resistance in cattle nematodes. PLoS One. 2013;8: e70212. doi:10.1371/journal.pone.0070212

333. Fontaine P, Choe K. The transcription factor SKN-1 and detoxification gene ugt-22 alter albendazole efficacy in Caenorhabditis elegans. Int J Parasitol Drugs Drug Resist. 2018;8: 312–319. doi:10.1016/j.ijpddr.2018.04.006

334. Kim C, Kim J, Kim S, Cook DE, Evans KS, Andersen EC, et al. Long-read sequencing reveals intra-species tolerance of substantial structural variations and new subtelomere formation in C. elegans. Genome Res. 2019;29: 1023–1035. doi:10.1101/gr.246082.118

335. Yoshimura J, Ichikawa K, Shoura MJ, Artiles KL, Gabdank I, Wahba L, et al. Recompleting the Caenorhabditis elegans genome. Genome Res. 2019;29: 1009–1022. doi:10.1101/gr.244830.118

336. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018;36: 875–879. doi:10.1038/nbt.4227

337. Bloom JS, Boocock J, Treusch S, Sadhu MJ, Day L, Oates-Barker H, et al. Rare variants contribute disproportionately to quantitative trait variation in yeast [Internet]. bioRxiv. 2019. p. 607291. doi:10.1101/607291

338. Ben-David E, Burga A, Kruglyak L. A maternal-effect selfish genetic element in Caenorhabditis elegans. Science. 2017;356: 1051–1055. doi:10.1126/science.aan0621

339. Jiang D, McPeek MS. Robust rare variant association testing for quantitative traits in samples with related individuals. Genet Epidemiol. 2014;38: 10–20. doi:10.1002/gepi.21775

340. Zhang Q, Guldbrandtsen B, Calus MPL, Lund MS, Sahana G. Comparison of gene-based rare variant association mapping methods for quantitative traits in a bovine population with complex familial relationships. Genet Sel Evol. 2016;48: 60. doi:10.1186/s12711-016-0238-5

341. Privé F, Aschard H, Ziyatdinov A, Blum MGB. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. Bioinformatics. 2018;34: 2781–2787. doi:10.1093/bioinformatics/bty185

342. Klasen JR, Barbez E, Meier L, Meinshausen N, Bühlmann P, Koornneef M, et al. A multi-marker association method for genome-wide association studies without the need for population structure correction. Nat Commun. 2016;7: 13299. doi:10.1038/ncomms13299

343. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet. 2018;19: 491–504. doi:10.1038/s41576-018-0016-z

344. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. Genetics. 2014;198: 497–508. doi:10.1534/genetics.114.167908

345. Kichaev G, Yang W-Y, Lindstrom S, Hormozdiari F, Eskin E, Price AL, et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. 2014;10: e1004722. doi:10.1371/journal.pgen.1004722

346. Duveau F, Félix M-A. Role of pleiotropy in the evolution of a cryptic developmental variation in Caenorhabditis elegans. PLoS Biol. 2012;10: e1001230. doi:10.1371/journal.pbio.1001230

347. Keele GR, Quach BC, Israel JW, Zhou Y, Chappell GA, Lewis L, et al. Integrative QTL analysis of gene expression and chromatin accessibility identifies multi-tissue patterns of genetic regulation [Internet]. bioRxiv. 2019. p. 588723. doi:10.1101/588723

348. Pierce BL, Tong L, Chen LS, Rahaman R, Argos M, Jasmine F, et al. Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians. PLoS Genet. 2014;10: e1004818. doi:10.1371/journal.pgen.1004818

349. O'Donnell MP, Chao P-H, Kammenga JE, Sengupta P. Rictor/TORC2 mediates gut-to-brain signaling in the regulation of phenotypic plasticity in C. elegans. PLoS Genet. 2018;14: e1007213. doi:10.1371/journal.pgen.1007213

350. Greene JS, Dobosiewicz M, Butcher RA, McGrath PT, Bargmann CI. Regulatory changes in two chemoreceptor genes contribute to a Caenorhabditis elegans QTL for foraging behavior. Elife. elifesciences.org; 2016;5. doi:10.7554/eLife.21454

351. McGrath PT, Rockman MV, Zimmer M, Jang H, Macosko EZ, Kruglyak L, et al. Quantitative mapping of a digenic behavioral trait implicates globin variation in *C. elegans* sensory behaviors. Neuron. 2009;61: 692–699. doi:10.1016/j.neuron.2009.02.012

352. Quarrie SA, Lazić-Jančić V, Kovačević D, Steed A, Pekić S. Bulk segregant analysis with molecular markers and its use for improving drought resistance in maize. J Exp Bot. Narnia; 1999;50: 1299–1306. doi:10.1093/jxb/50.337.1299

353. Hill WG. Selection with recurrent backcrossing to develop congenic lines for quantitative trait loci analysis. Genetics. 1998;148: 1341–1352. Available: https://www.ncbi.nlm.nih.gov/pubmed/9539447

354. Lee D, Yang H, Kim J, Brady S, Zdraljevic S, Zamanian M, et al. The genetic basis of natural variation in a phoretic behavior. Nat Commun. 2017;8: 273. doi:10.1038/s41467-017-00386-x

355. Zamanian M, Cook DE, Zdraljevic S, Brady SC, Lee D, Lee J, et al. Discovery of genomic intervals that underlie nematode responses to benzimidazoles. PLoS Negl Trop Dis. 2018;12: e0006368. doi:10.1371/journal.pntd.0006368

356. Pedersen B. smoove [Internet]. Github; Available: https://github.com/brentp/smoove

357. Pedersen BS, Layer RM, Quinlan AR. Vcfanno: fast, flexible annotation of genetic variants. Genome Biol. 2016;17: 118. doi:10.1186/s13059-016-0973-5

358. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics. 1993;133: 693–709. Available: https://www.ncbi.nlm.nih.gov/pubmed/8454210

359. Wall JD. Recombination and the power of statistical tests of neutrality. Genet Res . Cambridge University Press; 1999;74: 65–79. doi:10.1017/S0016672399003870

360. Rozas J, Gullaud M, Blandin G, Aguadé M. DNA variation at the rp49 gene region of Drosophila simulans: evolutionary inferences from an unusual haplotype structure. Genetics. 2001;158: 1147–1155. Available: https://www.ncbi.nlm.nih.gov/pubmed/11454763

361. Kelly JK. A test of neutrality based on interlocus associations. Genetics. 1997;146: 1197–1206. Available: https://www.ncbi.nlm.nih.gov/pubmed/9215920

362. Browning BL, Browning SR. Detecting Identity by Descent and Estimating Genotype Error Rates in Sequence Data. Am J Hum Genet. 2013;93: 840–851. doi:10.1016/j.ajhg.2013.09.014

363. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019; doi:10.1093/bioinformatics/btz305

9. Appendix 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 publications I co-authored.

Strategies to regulate transcription factor–mediated gene positioning and interchromosomal clustering at the nuclear periphery

Carlo Randise-Hinchliff, Robert Coukos, Varun Sood, Michael Chas Sumner, Stefan Zdraljevic, Lauren Meldi Sholl, Donna Garvey Brickner, Sara Ahmed, Lauren Watchmaker, and Jason H. Brickner

Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208, USA

Abstract

In budding yeast, targeting of active genes to the nuclear pore complex (NPC) and interchromosomal clustering is mediated by transcription factor (TF) binding sites in the gene promoters. For example, the binding sites for the TFs Put3, Ste12, and Gcn4 are necessary and sufficient to promote positioning at the nuclear periphery and interchromosomal clustering. However, in all three cases, gene positioning and interchromosomal clustering are regulated. Under uninducing conditions, local recruitment of the Rpd3(L) histone deacetylase by transcriptional repressors blocks Put3 DNA binding. This is a general function of yeast repressors: 16 of 21 repressors blocked Put3-mediated subnuclear positioning; 11 of these required Rpd3. In contrast, Ste12-mediated gene positioning is regulated independently of DNA binding by mitogen-activated protein kinase phosphorylation of the Dig2 inhibitor, and Gcn4-dependent targeting is up-regulated by increasing Gcn4 protein levels. These different regulatory strategies provide either qualitative switch-like control or quantitative control of gene positioning over different time scales.

Contributions

I constructed the Ste12 plasmids to visualize the localization of the Ste12 locus during pheromone induction. I also constructed yeast strains with this plasmid integrated and generated images of yeast cells after pheromone induction and measured the distance to locus was to the nuclear periphery.

The Genetic Basis of Natural Variation in Caenorhabditis elegans Telomere Length

Daniel E. Cook,*,† Stefan Zdraljevic,*,† Robyn E. Tanny,* Beomseok Seo,‡ David D. Riccardi,§, ** Luke M. Noble,§, ** Matthew V. Rockman,§, ** Mark J. Alkema,†† Christian Braendle,‡‡ Jan E. Kammenga,§§ John Wang,*** Leonid Kruglyak,†††,‡‡‡ Marie-Anne Félix,§§§ Junho Lee,‡, **** and Erik C. Andersen*,††††,‡‡‡‡,§§§§,1

*Department of Molecular Biosciences, †Interdisciplinary Biological Science Program††††Robert H. Lurie Comprehensive Cancer Center ‡‡‡‡Chemistry of Life Processes Institute §§§§Northwestern Institute on Complex Systems, Northwestern University, Evanston, Illinois 60208
‡Department of Biological Sciences, Institute of Molecular Biology and Genetics, Seoul National University, 08826, Korea
§Department of Biology, and **Center for Genomics and Systems Biology, New York University, New York 10003
††Department of Neurobiology, University of Massachusetts Medical School, Worcester, Massachusetts 01605
‡‡Centre National de la Recherche Scientifique, Institut National de la Santé et de la Recherche Médicale, Institut de Biologie Valrose, Université Nice Sophia Antipolis, 06100 Nice, France
§§Laboratory of Nematology, Wageningen University, 6708 PB, Netherlands
***Biodiversity Research Center, Academia Sinica, Taipei 115, Taiwan
†††Departments of Human Genetics and Biological Chemistry, University of California, Los Angeles, California 90095
‡‡‡Howard Hughes Medical Institute, Chevy Chase, Maryland 20815
§§§Institut de Biologie de l'École Normale Supérieure, Centre National de la Recherche Scientifique, Institut National de la Santé et de la Recherche Médicale, 75005 Paris, France ****Department of Biophysics and Chemical Biology, Seoul National University, 08826, Korea

This manuscript was published in *Genetics* in July 2016 [44]

Abstract

Telomeres are involved in the maintenance of chromosomes and the prevention of genome instability. Despite this central importance, significant variation in telomere length has been observed in a variety of organisms. The genetic determinants of telomerelength variation and their effects on organismal fitness are largely unexplored. Here, we describe natural variation in telomere length across the Caenorhabditis elegans species. We identify a large-effect variant that contributes to differences in telomere length. The variant alters the conserved oligonucleotide/oligosaccharide-binding fold of protection of telomeres 2 (POT-2), a homolog of a human telomere-capping shelterin complex subunit. Mutations within this domain likely reduce the ability of POT-2 to bind telomeric DNA, thereby increasing telomere length. We find that telomere-length variation does not correlate with offspring production or longevity in C. elegans wild isolates, suggesting that naturally long telomeres play a limited role in modifying fitness phenotypes in C. elegans.

Contributions

I developed the fine-mapping approach that led to the discovery of *pot-2* variation within the QTL described in this manuscript.

CeNDR, the *Caenorhabditis elegans* natural diversity resource

Daniel E. Cook[1,2], Stefan Zdraljevic[1,2], Joshua P. Roberts[2] and Erik C. Andersen[2],*

[1]Interdisciplinary Biological Sciences Program, Northwestern University, Evanston, IL 60208, USA
[2]Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208, USA
*Corresponding author, erik.andersen@northwestern.edu

This manuscript was published in *Nucleic Acids Research* in October 2016 [27]

Abstract

Studies in model organisms have yielded considerable insights into the etiology of disease and our understanding of evolutionary processes. Caenorhabditis elegans is among the most powerful model organisms used to understand biology. However, C. elegans is not used as extensively as other model organisms to investigate how natural variation shapes traits, especially through the use of genome-wide association (GWA) analyses. Here, we introduce a new platform, the C. elegans Natural Diversity Resource (CeNDR) to enable statistical genetics and genomics studies of C. elegans and to connect the results to human disease. CeNDR provides the research community with wild strains, genome-wide sequence and variant data for every strain, and a GWA mapping portal for studying natural variation in C. elegans. Additionally, researchers outside of the C. elegans community can benefit from public mappings and integrated tools for comparative analyses. CeNDR uses several databases that are continually updated through the addition of new strains, sequencing data, and association mapping results. The CeNDR data are accessible through a freely available web portal located at http://www.elegansvariation.org or through an application programming interface.

Contributions

I developed the genome-wide association mapping framework that is used on the CeNDR website.

Natural Variation in the Distribution and Abundance of Transposable Elements Across the *Caenorhabditis elegans* Species

K.M. Laricchia,[1] S. Zdraljevic,[1,2] D.E. Cook,[1,2] and E.C. Andersen*,[1,3,4,5]

[1] Department of Molecular Biosciences, Northwestern University, Evanston, IL
[2] Interdisciplinary Biological Sciences Graduate Program, Northwestern University, Evanston, IL
[3] Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL
[4] Chemistry of Life Processes Institute, Northwestern University, Evanston, IL
[5] Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL

*Corresponding authors, erik.andersen@northwestern.edu and elegans@snu.ac.kr

This manuscript was published in *Molecular Biology and Evolution in 2017*

Abstract

Transposons are mobile DNA elements that generate both adaptive and deleterious phenotypic variation thereby driving genome evolution. For these reasons, genomes have mechanisms to regulate transposable element (TE) activity. Approximately 12–16% of the Caenorhabditis elegans genome is composed of TEs, of which the majority are likely inactive. However, most studies of TE activity have been conducted in the laboratory strain N2, which limits our knowledge of the effects of these mobile elements across natural populations. We analyzed the distribution and abundance of TEs in 208 wild C. elegans strains to better understand how transposons contribute to variation in natural populations. We identified 3,397 TEs as compared with the reference strain, of which 2,771 are novel insertions and 241 are TEs that have been excised in at least one wild strain. Likely because of their hypothesized deleterious effects, we find that TEs are found at low allele frequencies throughout the population, and we predict functional effects of TE insertions. The abundances of TEs reflect their activities, and these data allowed us to perform both genome-wide association mappings and rare variant correlations to reveal several candidate genes that impact TE regulation, including small regulatory piwi-interacting RNAs and chromatin factors. Because TE variation in natural populations could underlie phenotypic variation for organismal and behavioral traits, the transposons that we

identified and their regulatory mechanisms can be used in future studies to explore the genomics of complex traits and evolutionary changes.

Contributions

I performed and interpreted the genome-wide association mapping described in this study.

The genetic basis of natural variation in a phoretic behavior

Daehan Lee[1,2], Heeseung Yang[1], Jun Kim[1], Shannon C. Brady[2], Stefan Zdraljevic[2], Mostafa Zamanian[2,3], Heekyeong Kim[4], Young-ki Paik[4,5,6], Leonid Kruglyak[7,8], Erik C. Andersen[2]*, and Junho Lee[1]*

[1]Department of Biological Sciences, Institute of Molecular Biology and Genetics, Seoul National University, Seoul 08826, Korea
[2]Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208, USA
[3]Department of Pathobiological Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA
[4]Yonsei Proteome Research Center, Yonsei University, Seoul 03722, Korea
[5]Department of Integrated OMICS for Biomedical Science, Yonsei University, Seoul 03722, Korea
[6]Department of Biochemistry, Yonsei University, Seoul 03722, Korea
[7]Department of Human Genetics and Biological Chemistry, University of California, Los Angeles, CA 90095, USA
[8]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
*Corresponding authors, erik.andersen@northwestern.edu and elegans@snu.ac.kr

This manuscript was published in *Nature Communications* in August 2017 [354]

Abstract

Phoresy is a widespread form of commensalism that facilitates dispersal of one species through an association with a more mobile second species. Dauer larvae of the nematode *Caenorhabditis elegans* exhibit a phoretic behavior called nictation, which could enable interactions with animals such as isopods or snails. Here, we show that natural *C. elegans* isolates differ in nictation. We use quantitative behavioral assays and linkage mapping to identify a genetic locus (*nict-1*) that mediates the phoretic interaction with terrestrial isopods. The *nict-1* locus contains a Piwi-interacting small RNA (piRNA) cluster; we observe that the Piwi Argonaute PRG-1 is involved in the regulation of nictation. Additionally, this locus underlies a trade-off between offspring production and dispersal. Variation in the *nict-1* locus contributes directly to differences in association between nematodes and terrestrial isopods in a laboratory assay. In summary, the piRNA-rich *nict-1* locus could define a novel mechanism underlying phoretic interactions.

Contributions

I assisted in the generation of the CRISPR/Cas9-mediated deletion alleles of *prg-1* in the N2 and CB4856 backgrounds.

Discovery of genomic intervals that underlie nematode responses to benzimidazoles

Mostafa Zamanian[1,*], Daniel E. Cook[2,3], Stefan Zdraljevic[2,3], Shannon C. Brady[2,3], Daehan Lee[2,4], Junho Lee[4], Erik C. Andersen[2,5,6*]

[1]Department of Pathobiological Sciences, University of Wisconsin-Madison, Madison, Wisconsin, USA
[2]Department of Molecular Biosciences, Northwestern University, Evanston, Illinois, USA
[3]Interdisciplinary Biological Science Program, Northwestern University, Evanston, Illinois, USA
[4]Institute of Molecular Biology and Genetics, Department of Biological Sciences, Seoul National University, Seoul, Korea
[5]Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, Illinois, USA
[6]Northwestern Institute on Complex Systems, Northwestern University, Evanston, Illinois, USA
*Corresponding authors, mzamanian@wisc.edu and erik.andersen@northwestern.edu

This manuscript was published in *PLoS Neglected Tropical Diseases* in March 2018 [355]

Abstract

Parasitic nematodes impose a debilitating health and economic burden across much of the world. Nematode resistance to anthelmintic drugs threatens parasite control efforts in both human and veterinary medicine. Despite this threat, the genetic landscape of potential resistance mechanisms to these critical drugs remains largely unexplored. Here, we exploit natural variation in the model nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* to discover quantitative trait loci (QTL) that control sensitivity to benzimidazoles widely used in human and animal medicine. High-throughput phenotyping of albendazole, fenbendazole, mebendazole, and thiabendazole responses in panels of recombinant lines led to the discovery of over 15 QTL in *C. elegans* and four QTL in *C. briggsae* associated with divergent responses to these anthelmintics. Many of these QTL are conserved across benzimidazole derivatives, but others show drug and dose specificity. We used near-isogenic lines to recapitulate and narrow the *C. elegans* albendazole QTL of largest effect and identified candidate variants correlated with the resistance phenotype. These QTL do not overlap with known benzimidazole target resistance genes from parasitic nematodes and present specific new leads for the discovery of

novel mechanisms of nematode benzimidazole resistance. Analyses of orthologous genes reveal conservation of candidate benzimidazole resistance genes in medically important parasitic nematodes. These data provide a basis for extending these approaches to other anthelmintic drug classes and a pathway towards validating new markers for anthelmintic resistance that can be deployed to improve parasite disease control.

Contributions

I assisted in the generation of the CRISPR/Cas9-mediated deletion alleles of *prg-1* in the N2 and CB4856 backgrounds.

Tightly-linked antagonistic-effect loci underlie polygenic demographic variation in *C. elegans*

Max R. Bernstein[1] , Stefan Zdraljevic[2] , Erik C. Andersen[2] , Matthew V. Rockman[1]

1. Department of Biology and Center for Genomics & Systems Biology, New York University, New York, NY 10003
2. Molecular Biosciences and Interdisciplinary Biological Sciences Program, Northwestern University, Evanston, IL 60208

This manuscript was published in *eLife* in April 2019 [32].

Abstract

Recent work has provided strong empirical support for the classic polygenic model for trait variation. Population-based findings suggest that most regions of genome harbor variation affecting most traits. This view is hard to reconcile with the experience of researchers who define gene functions using mutagenesis, comparing mutants one at a time to the wild type. Here, we use the approach of experimental genetics to show that indeed, most genomic regions carry variants with detectable effects on complex traits. We used high-throughput phenotyping to characterize demography as a multivariate trait in growing populations of Caenorhabditis elegans sensitized by nickel stress. We show that demography under these conditions is genetically complex in a panel of recombinant inbred lines. We then focused on a 1.4-Mb region of the X chromosome. When we compared two near isogenic lines (NILs) that differ only at this region, they were phenotypically indistinguishable. When we used additional NILs to subdivide the region into fifteen intervals, each encompassing ~0.001 of the genome, we found that eleven of intervals have significant effects. These effects are often similar in magnitude to those of genome-wide significant QTLs mapped in the recombinant inbred lines but are antagonized by the effects of variants in adjacent intervals. Contrary to the expectation of small additive effects, our findings point to large-effect variants whose effects are masked by epistasis or linkage disequilibrium between alleles of opposing effect.

Contributions

I performed the high-throughput assay to generate the phenotype data set that was used in this study.

Evolution of sperm competition: Natural variation and genetic determinants of *Caenorhabditis elegans* sperm size

Clotilde Gimond[1]*, Anne Vielle1*, Nuno SilvaSoares[1,2], Stefan Zdraljevic[3], Patrick T. McGrath[4], Erik C. Andersen[3], and Christian Braendle[1#]

1. Université Côte d'Azur, CNRS, Inserm, IBV, Nice, France
2. Instituto Gulbenkian de Ciencia, Oeiras, Portugal
3. Department of Molecular Biosciences, Northwestern University, Evanston, IL USA
4. School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA USA

* these authors contributed equally
# corresponding author

Abstract

Sperm morphology is critical for sperm competition and thus for reproductive fitness. In the male-hermaphrodite nematode *Caenorhabditis elegans*, sperm size is a key feature of sperm competitive ability. Yet despite extensive research, the molecular mechanisms regulating *C. elegans* sperm size and the genetic basis underlying its natural variation remain unknown. Examining 97 genetically distinct *C. elegans* strains, we observe significant heritable variation in male sperm size but genome-wide association mapping did not yield any QTL (Quantitative Trait Loci). While we confirm larger male sperm to consistently outcompete smaller hermaphrodite sperm, we find natural variation in male sperm size to poorly predict male fertility and competitive ability. In addition, although hermaphrodite sperm size also shows significant natural variation, male and hermaphrodite sperm size do not correlate, implying a sex-specific genetic regulation of sperm size. To elucidate the molecular basis of intraspecific sperm size variation, we focused on recently diverged laboratory strains, which evolved extreme sperm size differences. Using mutants and quantitative complementation tests, we demonstrate that variation in the gene *nurf-1* – previously shown to underlie the evolution of improved hermaphrodite reproduction – also explains the evolution of reduced male sperm size. This result illustrates how adaptive changes in *C. elegans* hermaphrodite function can cause the deterioration of a male-specific fitness trait due to a sexually antagonistic variant, representing

an example of intralocus sexual conflict with resolution at the molecular level. Our results further provide first insights into the genetic determinants of *C. elegans* sperm size, pointing at an involvement of the NURF chromatin remodelling complex.

Contributions

I performed and interpreted the genome-wide association mapping described in this study.

Selection and gene flow shape niche-associated copy-number variation of pheromone receptor genes

Daehan Lee[1], Stefan Zdraljevic[1,2], Daniel E. Cook[1,2,3], Lise Frézal[4], Jung-Chen Hsu[5], Mark G. Sterken[6], Joost A.G. Riksen[6], John Wang[5], Jan E. Kammenga[6], Christian Braendle[7], Marie-Anne Félix[4], Frank C. Schroeder[8] , Erik C. Andersen[1] *

1. Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208, USA
2. Interdisciplinary Biological Sciences Program, Northwestern University, Evanston, IL 60208, USA
3. Present address: The Francis Crick Institute, London NW1 1ST, UK
4. Institut de Biologie de l'Ecole Normale Supérieure, Centre National de la Recherche Scientifique, INSERM, École Normale Supérieure, Paris Sciences et Lettres, Paris, France
5. Biodiversity Research Center, Academia Sinica, Taipei, 11529, Taiwan
6. Laboratory of Nematology, Wageningen University and Research, 6708PB, The Netherlands
7. Université Côte d'Azur, CNRS, Inserm, IBV, France, 06100 Nice, France
8. Boyce Thompson Institute and Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853, USA

* Corresponding author

Abstract

From quorum sensing in bacteria to pheromone signaling in social insects, chemical communication mediates interactions among individuals in a local population. In *Caenorhabditis elegans*, ascaroside pheromones can dictate local population density, in which high levels of pheromones inhibit the reproductive maturation of individuals. Little is known about how natural genetic diversity affects the pheromone responses of individuals from diverse habitats. Here, we show that a niche-associated copy-number variation (CNV) of pheromone receptor genes contributes to natural differences in pheromone responses. We found putative loss-of-function deletions that reduce copy number of duplicated pheromone receptor genes (*srg-36* and *srg-37*), which were shown previously to be selected in population-dense laboratory cultures. A common natural deletion in the less functional copy (*srg-37*) arose from a single ancestral population that spread throughout the world and underlies reduced pheromone sensitivity across the global *C. elegans* population. This deletion is enriched in wild strains that were isolated from a rotting fruit niche, where proliferating populations are often found. Taken together, these results demonstrate that selection and gene flow together shape the copy

number of pheromone receptor genes in natural *C. elegans* populations to facilitate local adaptation to diverse niches.

Contributions

I established the genome-wide association mapping pipeline Dr. Daehan Lee used in this manuscript to identify the QTL on chromosome X. I also assisted Daehan with the population genomic analysis and the characterization of structural variation described in this manuscript.

A nematode-specific gene underlies bleomycin-response variation in *Caenorhabditis elegans*

Shannon C. Brady[*,†], Stefan Zdraljevic[*,†], Karol W. Bisaga[‡], Robyn E. Tanny[*], Daniel E. Cook[§], Daehan Lee[*], Ye Wang[*], Erik C. Andersen[*,†,**,1]

[*] Molecular Biosciences, Northwestern University, Evanston, IL 60208
[†] Interdisciplinary Biological Sciences Program, Northwestern University, Evanston, IL 60208
[‡] Weinberg College of Arts and Sciences, Northwestern University, Evanston, IL 60208
[§] The Francis Crick Institute, London NW1 1ST, UK
[**] Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL 60611
[1] Corresponding author

Abstract

Bleomycin is a powerful chemotherapeutic drug used to treat a variety of cancers. However, individual patients vary in their responses to bleomycin. The identification of genetic differences that underlie this response variation could improve treatment outcomes by tailoring bleomycin dosages to each patient. We used the model organism *Caenorhabditis elegans* to identify genetic determinants of bleomycin-response differences by performing linkage mapping on recombinants derived from a cross between the laboratory strain (N2) and a wild strain (CB4856). This approach identified a small genomic region on chromosome V that underlies bleomycin-response variation. Using near-isogenic lines and strains with CRISPR-Cas9 mediated deletions and allele replacements, we discovered that a novel nematode-specific gene (*scb-1*) is required for bleomycin resistance. Although the mechanism by which this gene causes variation in bleomycin responses is unknown, we suggest that a rare variant present in the CB4856 strain might cause differences in the potential stress-response function of *scb-1* between the N2 and CB4856 strains, thereby leading to differences in bleomycin resistance.

Contributions

I assisted Dr. Shannon Brady with the construction of near-isogenic lines used to validate the QTL discussed in this manuscript. I also assisted in performing the genome-wide association mapping.

Deep sampling of Hawaiian *Caenorhabditis elegans* reveals high genetic diversity and admixture with global populations

Timothy A. Crombie[1], Stefan Zdraljevic[1,2], Daniel E. Cook[1,2], 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 Zwaag[1], Karin C. Kiontke[3], and Erik C. Andersen[1,*]

1. Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208, USA
2. Interdisciplinary Biological Sciences Program, Northwestern University, Evanston, IL 60208, USA
3. Department of Biology, New York University, New York, NY 10003, USA
* Corresponding author

Abstract
The genetic diversity of the world-wide *Caenorhabditis elegans* population is low, partly because of recent chromosome-scale selective sweeps, which are theorized to contain loci that increase fitness in human-associated habitats. However, strains from Hawaii are highly divergent from non-Hawaiian strains and do not harbor the swept haplotypes. This result suggests that Hawaiian strains might contain ancestral genetic diversity purged from most non-Hawaiian strains by selective sweeps. To characterize the genetic diversity of Hawaiian *C. elegans*, we sampled across the Hawaiian Islands and isolated 100 new *C. elegans* strains. Admixture analysis identified seven global populations—three Hawaiian and four non-Hawaiian. One of the three Hawaiian populations shows admixture with non-Hawaiian populations, including portions of globally swept haplotypes. This discovery provides the first evidence of gene flow between Hawaiian and non-Hawaiian populations. Most importantly, the high levels of diversity observed in Hawaiian strains might represent the complex patterns of ancestral genetic diversity in the *C. elegans* species before human influence.

Contributions

I was one of the members of the Andersen lab that went to Hawaii in 2017 to collect *C. elegans* strains. When we returned, I assisted with preparing samples for Sanger sequencing to establish the species identity. I helped analyze the isolation data with the rest of the lab over the

course of many group coding sessions. I also performed a majority of the population genomic analysis described in the manuscript. Dr. Tim Crombie and I went back to Hawaii in 2018 for another round of field collection. Though the results of our second trip are not discussed in this manuscript, we will apply many of the same analyses to that dataset once it is ready.

10. Appendix B: *cegwas2-nf*

Since I started in the Andersen lab I have been constantly updating and improving the lab's standard genome-wide association (GWA) mapping workflow. *cegwas2-nf* represents the culmination of my efforts to standardize this process for members in the lab. That being said, new methods for GWA mapping are constantly being developed and it is difficult to keep up. I validated the mapping methods in *cegwas2-nf* using simulations during my rotation and have not updated the core EMMA algorithm since [56]. I developed c*egwas2-nf* using Nextflow (v0.32) and custom scripts [73], which will enable future users to update the pipeline to their needs. The *cegwas2-nf* workflow can be found on https://github.com/AndersenLab/cegwas2-nf

Explanation of functionality

At the core of *cegwas2-nf* is the EMMA algorithm for single-marker GWA mapping. By default, genotype data are downloaded from the latest VCF release (Release 20180527) on CeNDR that was imputed as described previously [27] or the user can provide their own VCF using the --vcf flag. BCFtools [87] is used to filter variants that had any missing genotype calls and variants that were below 5% minor allele frequency. We used PLINK v1.9 [94,95] to LD-prune the genotypes at a threshold of $r^2 < 0.8$, using *--indep-pairwise 50 10 0.8.* The resulting genotype matrix is used to generate the realized additive kinship matrix using the *A.mat* function in the *rrBLUP* R package [149]. These markers were also used for genome-wide mapping. However, because these markers still have substantial LD within this genotype set, we performed eigen decomposition of the correlation matrix of the genotype matrix using *eigs_sym* function in Rspectra package [221]. The correlation matrix was generated using the *cor* function in the correlateR R package [222]. We set any eigenvalue greater than one from this analysis to one and summed all of the resulting eigenvalues [223]. The sum of the eigenvalues corresponds to

the number of independent tests within the genotype matrix. We used the *GWAS* function in the *rrBLUP* package to perform genome-wide mapping. Using the --p3d flag, user can define the mapping algorithm that will be used, where --p3d=TRUE corresponds to the EMMAx [58] method and --p3d=FALSE corresponds to the EMMA method [56]. For each mapping, a manhattan plot is generated and if significant QTL are identified, a phenotype by genotype plot is generated for the most significant marker within each QTL.

To perform fine-mapping, we confidence intervals from the genome-wide mapping as +/- 100 SNVs from the rightmost and leftmost markers above the user-defined significance threshold (--sthresh=BF for Bonferroni significance threshold or --sthresh=EIGEN for the eigen method described above). For each QTL, a QTL region of interest genotype matrix is filtered as described above, except no LD pruning is performed. We used PLINK v1.9 to extract the LD between the markers used for fine mapping and the peak QTL marker identified from the genome-wide scan. We used the same command as above to perform fine mapping, but with the reduced variant set.

In parallel to single-marker mappings, *cegwas2-nf* performs burden-based mapping using SKAT and the VT methods [280,339], which are both implemented in the RVtests software [279]. The --freqUpper flag determines the maximum allele frequency to be used for the burden mapping and --minburden determines the minimum number of strains that share a variant to be considered. Manhattan plots are generated for each method at the end of the pipeline.

As I described in the Discussion section of my dissertation, many new tools exist for performing GWA mapping. In its current state, the *cegwas2-nf* pipeline is extensible to these new approaches and I encourage people to make improvements to the pipeline as needed. One improvement that I was not able to accomplish is to incorporate the structural variation data I

generated into the mapping pipeline. Including this class of variation to the pipeline might reveal significant associations that might otherwise be missed. The incorporation of structural variation should be straightforward because one of the outputs described in Appendix B is a genotype structural variant VCF.

11. Appendix C: *joint-sv-nf*

I wanted to learn more genomics analyses during the last year and a half of my Ph.D.. As part of this learning effort I was tasked with calling structural variants in the *C. elegans* population. Structural variation is an important class of variation that affect phenotypic variation within a population, and at the time it was largely unexplored in the *C. elegans* species.

Explanation of functionality

Similar to the *cegwas2-nf* pipeline described in Appendix B, I developed the *joint-sv-nf* pipeline using Nextflow [73]. Prior to developing this pipeline, I performed simulations to identify the best variant caller for large insertions and deletions. I found Lumpy [83] and Delly2 [82] to have accuracy for large deletions and Manta [84] to be superior at identifying insertion variants. Lumpy was later optimized and incorporated in the smoove software package [356], which is what I use in the *joint-sv-nf* pipeline. The pipeline uses all three of these structural variant callers to identify variation in individual strains from alignment files. Once variants are identified in individual strains, the resulting files are merged to generate a population-wide VCF.

The pipeline used the BCFtools merge command to merge the individual VCFs generated by Manta. Because Manta does not have the capability to generate a square cohort VCF, this is the final step prior to annotating the predicted effects of the variants with SnpeEff [81]. Manta is unique relative to smoove and Delly2 because it performs local reassembly around detected breakpoints. Part of the Manta-generated VCF files includes a contig reassembly of the breakpoint, which is output by the pipeline in the BED file format [89]. Both Delly2 and smoove have similar workflows for merging and recalling variants, and I have incorporated the standard workflow described on the associated GitHub pages for these software packages. After calling

variants for individual strains, the generated files are merged step with the merge command built into Delly2 and smoove. This step generates a site list of all variants identified in the population. The resulting site list is then used to recall variants in individual strains using the Delly2 call and smoove genotype commands. This built-in recall functionality of Delly2 and smoove ensures that all strains are genotyped at all sites identified in the population. The pipeline makes use of the built-in germline filter command of Delly2 and uses the default smoove filters. Once individual strain variation is recalled, the Delly2 merge and smoove paste commands are used to generate a square cohort VCF and SnpEff is used to predict the effect of each identified variant.

After the three variant calling workflows are completed, the SURVIVOR software package is used identify overlapping variants called by the independent variant callers [88]. SURVIVOR is run for each strain and I have set the distance variants identified by different methods need to be to 1000 bp in order to be considered the same variant. The parameters I defined for SURVIVOR are (1000 1 0 0 1 30), which correspond to 1000 bp distance among callers, 1 caller needs to have found a variant, 0 means they do not have to be the same variant type, 0 means they do not need to be on the same strand, 0 means to not estimate the SV distance based on the SV size, and 30 means that SURRVIVOR will only consider variants at least 30 bp in size. SURVIVOR outputs a VCF for each strain that has the genotype data from each caller and the parameters of the merge. Next, the pipeline generates a BED file from the SURVIVOR merged VCF for each strain. Becauses C. elegans has a predominantly hermaphroditic life cycle, the pipeline also outputs a “High Quality” BED file for each strain that only contains sites with at least one homozygous ALT call. A custom R script processes each strain BED file and classifies “Complex” structural variants as variant sites that were identified to be different variant classes by the different callers. This script also generates a plot of the genomic location of each structural variant faceted by the predicted SnpEff annotation. Finally, the pipeline outputs a

structural variant genotype matrix that can be used for genome-wide association mapping or further population genomic analyses.

12. Appendix D: CePopulationGenetics-nf

I developed the *CePopulationGenetics-nf* pipeline using Nextflow [73] to perform standardized population genetic analyses. This pipeline is a work in progress and additional modules can be readily added to expand its functionality.

Explanation of functionality

The primary input for the pipeline is a small variant VCF file that can be defined with the *--snv_vcf* parameter. The user also defines an ancestral strain using the --anc parameter. The first step of the pipeline extracts the ancestor strain genotypes from the population VCF. Next, the INFO field of the population VCF is annotated with the ancestral allele (AA), transcription factor binding sites (TFid, TFname, TF_papersource), repetitive genomic regions (MASKED), histone binding sites (Histone_Binding), exons (exon), miRNA binding sites (miRNA_binding), splice sites (Splice_Sites), and the centimorgan position (cM) using vcfanno [357]. Using this annotated VCF, the *CePopulationGenetics-nf* calculates commonly used population genetics statistics, including Tajima's *D* [109], Fay and Wu's *H* [300], Zeng's E [301], Fu and Li's F and D [358], the corresponding thetas for each of these neutrality statistics, nucleotide diversity, Wall's B [359], Roza's ZZ and ZA [360], and Kelly's ZnS [361]. These statistics are all calculated using the PopGenome package in R [286]. Each of these statistics is calculated using sliding windows that can be defined with the *--popgenome_window* (window size in bp) and *--popgenome_slide* (step size in bp) parameters. The pipeline also performs admixture analysis using the ADMIXTURE software package [93], where users can define the number of populations using the *--admix_k* parameter and define the level of LD pruning to be done for admixture analysis using the *--admix_ld* parameter. Each population size (K) is run ten independent times and the top 5 best K sizes are chosen for additional cross-validation analysis using ADMIXTURE. Once the admixture analysis is completed, strains are assigned to the corresponding ancestral population and each of the population genetics statistics described above are performed on the subpopulations. The pipeline also calculates identity-by-descent regions using IBDseq [362] and generates a genome-wide haplotype plot for all analyzed strains. Additionally, the pipeline performs phylogenetic analysis using the RaXML-ng

software package [363] with the GTR substitution model. Finally, population structure is assessed using the *smartpca* command of the EIGENSTRAT software package [96] and 50 principal components are returned. *smartpca* is run in two modes, with and without outlier strain removal.

I have recently set up a variety of profiles for this pipeline so that only certain analyses can be performed. These profiles include *-profile full* to run the full analysis, *-profile admixture_cv* to perform admixture replicat analysis, *-profile admixture_full* to run admixture with user defined Ks (also requires a *--best_Ks* parameter which is a text file with a K size on each line), *-profile popgenome* that just calculates population genetic statistics on the input VCF. Because VCF annotation takes a long time, if the user does not provide the *--anc parameter*, the pipeline will assume that an annotated VCF was provided. This is useful when the pipeline was first run in the *admixture_cv* profile to identify the optimal K sizes before running the *admixture_full* profile.

As stated above, this pipeline is modular and users can include other analyses that use a VCF or PED file format as inputs, as they see fit.