Genome structural variation in natural *Caenorhabditis elegans* populations

Kyle J. Lesack, James D. Wasmuth*

Faculty of Veterinary Medicine, University of Calgary, Calgary, Alberta, Canada

* Corresponding author: jwasmuth@ucalgary.ca

Abstract

Structural variation, involving large alterations in chromosome structure, drives genetic diversification and the emergence of new phenotypes. These changes are widespread in natural populations and play an important role in adaptation and speciation. For many species, research has been limited to laboratory adapted strains and experimental evolution, which may not reflect the diversity of structural variants in the wild. Furthermore, technological limitations have proved to be a major barrier to accurate and comprehensive variant calling. In this study, we used PacBio sequencing data from 14 wild *Caenorhabditis elegans* strains to characterize structural variants in a natural population. With long-reads, we overcame limitations associated with short-read approaches and leveraged population-level data to further refine the accuracy of variant calls. We found that large, rapidly evolving gene families, such as GPCR, F-Box, and C-Type lectin, were prominent among the variants predicted to have phenotypic consequences. These results shed light on the significant role of structural variants in the evolution of *Caenorhabditis elegans* in its wild habitats and the limitations of treating N2 as the reference wild-type.

Introduction

Structural variants (SVs) refer to large-scale alterations in a genome that can change the presence, order, or quantity of genetic elements. These changes can have profound phenotypic effects through a variety of mechanisms, including the disruption of genes, modulation of gene dosage, or rearrangement of regulatory elements. Unsurprisingly, SVs are often deleterious and have been linked to numerous human diseases[1–3]. However, SVs also play an important role in adaptive evolution. For instance, gene duplication can provide a selective advantage through an increase in gene dosage or neofunctionalization, where one copy acquires a novel function[4–6]. Additionally, chromosomal inversions, even of relatively short regions, can suppress recombination and have been implicated in speciation and the emergence of sex chromosomes[7,8].

Despite their importance, SVs remain poorly characterized in most species, especially at the population level. This is partly due to the focus on human studies, but technological limitations are also a factor. Most studies rely on short-read DNA sequencing to identify SVs, but, despite considerable work, short-read SV calling remains error-prone, with many commonly used tools often generating drastically different results from the same data[9]. The technical limitations of short-reads, such as difficulty in resolving larger or complex variants and poor mapping accuracy in repetitive regions, contribute to these challenges[10]. Long-read sequencing technologies have the potential to address some of these short-comings by providing sequence reads that may span entire variants and have improved mappability. Long-read approaches have already shown improvements over short-reads for certain SVs and benchmarks from recent tools have demonstrated promising results[11].

For non-human studies, much of the population-level work on SVs has focused on domesticated plants and animals. In crops, copy number variants (CNVs) are widespread and have been proposed as a mechanism for the rapid adaptation associated with agriculture[12]. Unsurprisingly, CNVs associated with desirable traits, such as flowering time and disease resistance, feature prominently[13,14]. SVs have also been described in a variety of domesticated animals, including livestock[15], fish[16,17], and dogs[18,19]. Valuable knowledge can be gained from model organisms that are amenable to standard genetic manipulation techniques. In particular, mutation accumulation experiments in *Caenorhabditis elegans* have revealed high rates of spontaneous duplications, making it a suitable organism for studying SVs[20–22]. For example, a recombinant inbred line was recently described as having higher fitness than its parental strains due to a partial duplication and inversion of five different regions of *rcan*-1, which regulates thermotactic behavior[23]. It is notable because it describes a beneficial complex allele that is believed to have arisen from a single mutational event.

However, while these studies provide valuable information, they provide an incomplete picture of SV evolution in natural populations. Artificial selection in domesticated and laboratory environments exerts strong pressures that may not be as relevant in the wild. Domesticated or long-established laboratory populations can also accumulate deleterious alleles at a higher rate than their wild counterparts[24]. Notably, the *C. elegans* reference strain, N2, has accumulated numerous mutations, including multiple laboratory-adapted alleles and pseudogenized genes[25,26].

Studies describing SVs in natural populations have complemented insights gained from experimental evolution. An array-based comparative genomic hybridization analysis revealed considerable CNV diversity across 12 natural *C. elegans* strains, affecting over 5% of the genes in the N2 reference genome. The number of deletions ranged from 1 to 172, with the CB4856

strain harboring the most. This strain, first isolated in Hawaii, has diverged considerably from N2 and has been the focus of other studies on SVs and presence-absence variation in *C. elegans*[27–30]. Alignments between N2 and a *de novo* assembly of CB4856 identified 79,529 insertion and deletion events, as well as 816 segments that failed to align[31]. This comparison also identified 61 highly diverged regions, which led to the discovery of additional "hyper-divergent" regions between N2 and other strains collected from the wild[32].

The *Caenorhabditis elegans* Natural Diversity Resource (CeNDR) presents an opportunity to expand our knowledge of SVs in naturally occurring populations[33]. CeNDR maintains a collection of whole-genome sequencing data for 1,524 strains and 550 isotypes (as of June 2023). This resource has facilitated discoveries about the genetic and phenotypic diversity in *C. elegans*. For example, variable anthelmintic responses have been observed in wild strains and facilitated the discovery of novel quantitative trait loci linked to anthelmintic drug resistance[34–37]. CeNDR strains have also been tested against a range of toxicants, providing valuable information on xenobiotic metabolism and detoxification[38–40]. Collectively, these studies highlight the importance of studying genetic variation in natural populations. In this study, we leveraged long-read PacBio data from 14 CeNDR isolates to obtain a high-quality, population-scale profile of SVs in C. elegans. We identified thousands of deletions, duplications, and inversions, many of which affected genes involved in environmental response, stress, and immune function. These SVs could be leveraged as a high-quality reference set in future hybrid sequencing approaches that aim to mitigate the high false positive rates of short-read SV calling[41].

Materials and Methods

Strains and Sequencing Data

PacBio sequencing data from the *Caenorhabditis elegans* Natural Diversity Resource (CeNDR) [33] were used to predict SVs in 14 *C. elegans* strains (DL238, ECA36, ECA396, EG4725, JU1400, JU2526, JU2600, JU310, MY2147, MY2693, NIC2, NIC526, QX1794, XZ1516) collected from the wild in Hawaii, New Zealand, and Western Europe (Figure 1). Size selection was performed using two different cut-offs: a 15-kb cut-off was used for samples that had 30- to 50-kb fragment sizes, and a 20-kb cut-off was used for all others. Further details of the library preparation and sequencing methods can be found in a previous study that characterized species-wide hyper-divergent regions in *C. elegans* [32].

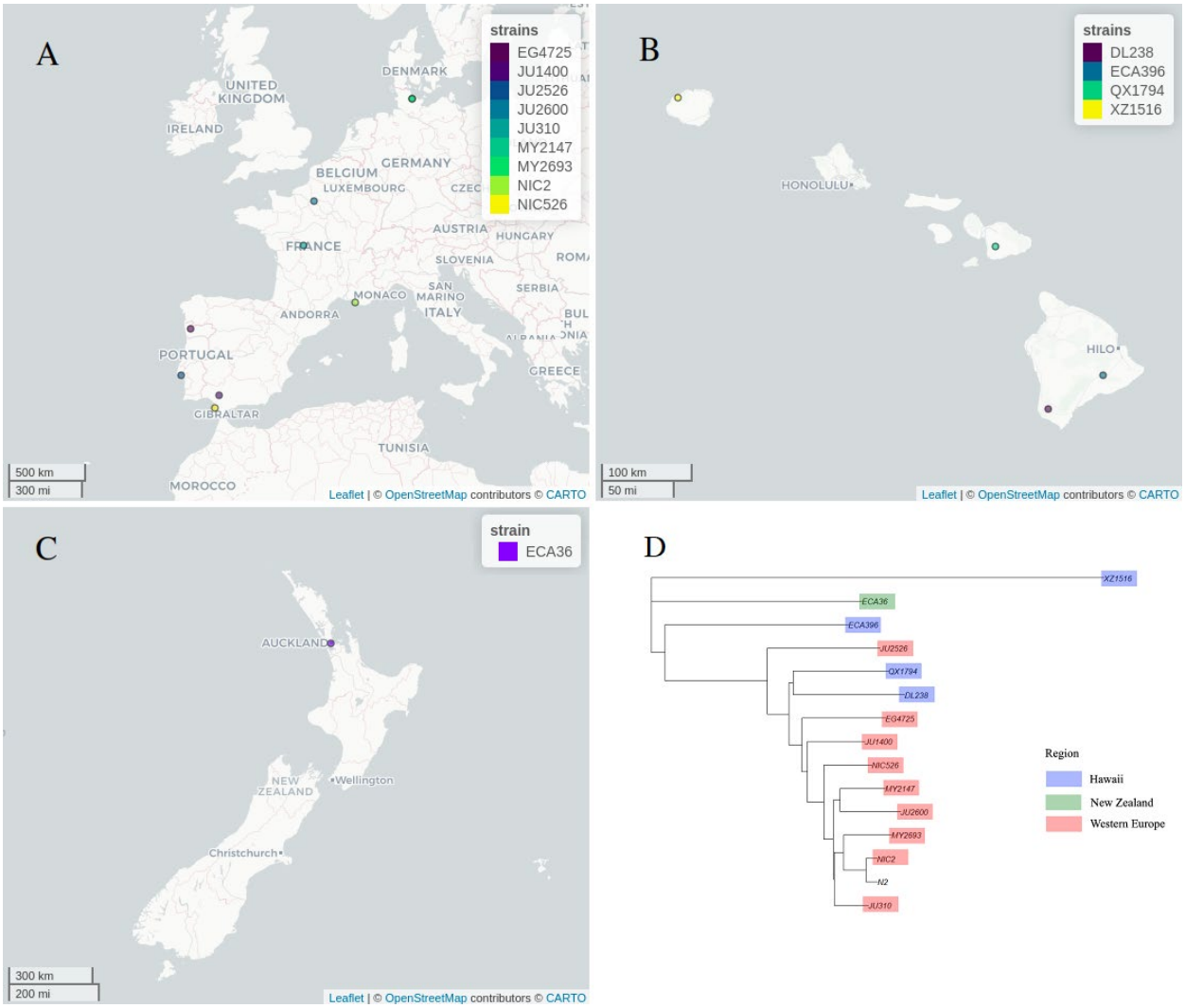

Figure 1 CeNDR sampling. (A) Nine strains were collected from Western Europe. (B) Four strains were collected from Hawaii. (C) One strain was collected from New Zealand. (D) Phylogenetic tree for the 14 strains (adapted from the CeNDR species tree available at https://caendr.org/data/data-release/c-elegans/latest).

Sequence Alignment and Structural Variant Calling

The CeNDR FASTQ files were aligned using NGMLR (v0.2.7; median depth: 142X)[42] and sorted using Picard (v2.27.5)[43]. Structural variant calling was performed using Sniffles2

(v2.0.7)[11]. Variant calls below 100bp were filtered out and only deletions, duplications, and inversions were included for further study. Further quality control was performed using Jasmine (v1.1.5) and Iris (v1.0.4)[44]. Iris was used to refine the breakpoints and sequences of predicted duplications, while Jasmine was used to remove redundant SV calls and to create a merged call set containing SVs from all strains.

Structural Variant Annotation

Variant Effect Predictor (v.109.3) was used to predict the functional consequences of each variant[45]. Genes affected by putative high impact SVs were selected for functional profiling using the easyGSEA tool from eVITTA (v1.3.1)[46]. Overrepresented genes were annotated using the annotated gene sets from WormCat (V2), a curated database of functional annotations for *C. elegans*[47]. Each gene was assigned to a statistically significant term in each category level (FDR ≤ 0.05), unless otherwise specified. WormCat provides three category levels, with the C1 categories providing broad descriptions of large sets of genes and the C3 categories providing more detailed descriptions of smaller sets of genes. Category level choice is difficult as no agreed upon standard has been adopted and choosing the more detailed C2 or C3 categories may exclude genes of interest that are recategorized in a smaller category containing fewer members than the minimum threshold. Therefore, no single category was chosen for this analysis.

Data Presentation

The results in the text are reported using either counts or as the mean (SD). The densities of SVs along chromosomes were plotted using karyoploteR (v1.26.0)[48]. UpSetPlot (v0.8.0)[49] was used to plot the SV allele frequencies among the strains. The 25 largest set intersections were

plotted for duplications. However, only the 20 largest intersections were plotted for deletions and inversions because it was not possible to select the 25 largest intersections for these variants without excluding intersections of equal size to those included in the plot. The remaining plots were created using ggplot2 (v.3.4.2)[50].

Results

Structural Variants in Wild *C. elegans* Strains

PacBio sequencing data from the CeNDR project was used to identify structural variants in 14 wild *C. elegans* strains. The analysis pipeline included Sniffles2, Jasmine and Iris. A total of 14,473 deletions, 3,149 duplications, and 1,748 inversions were identified (Table 1). Deletions were, on average, smaller compared to duplications and inversions. More deletions and duplications were found in the autosome arms, whereas inversions were most equally distributed along the length of the chromosomes (Figure 2).

Table 1. Total Structural Variants.

		All Variants		
Variant Type	Count	Mean Size ±SD (bp)	Minimum Size (bp)	Maximum Size (bp)
Deletion	14,373	705 ±2,459	100	97,753
Duplication	3,149	1,523±4,163	103	75,314
Inversion	1,748	3,822 ±8,684	100	96,813

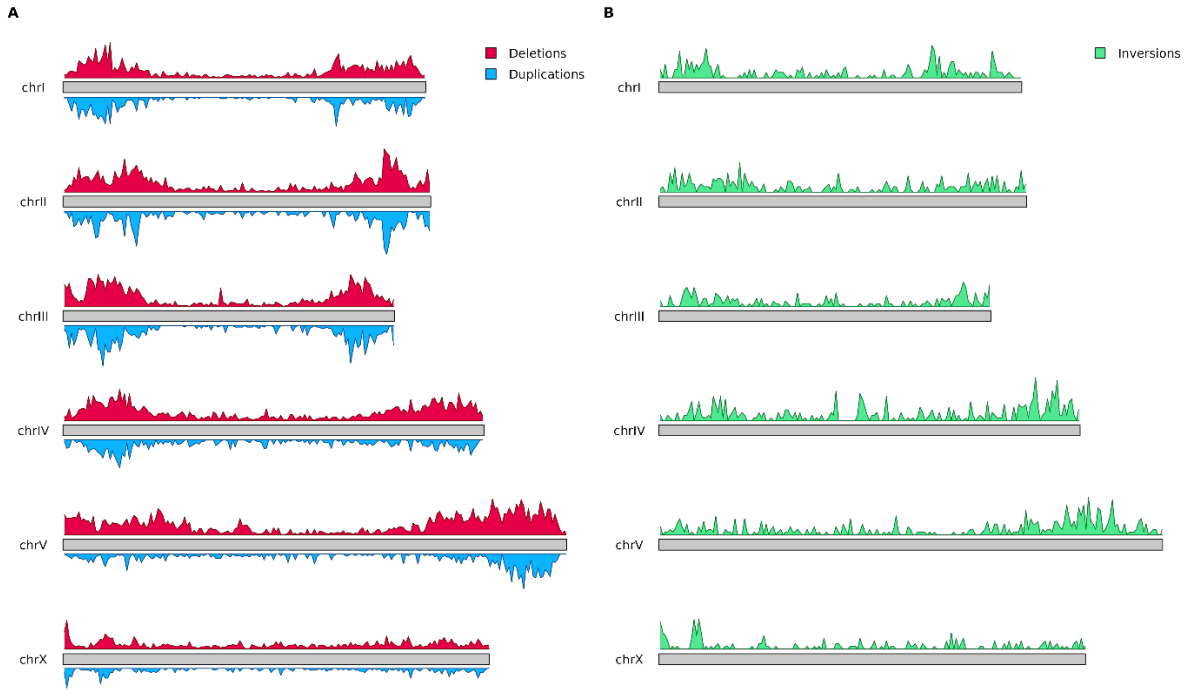

Figure 2 Distribution of SVs along chromosomes. (A) Deletions and duplications are shown in blue and red, respectively. (B) Inversions are shown in green.

Across individual strains, deletions were the most common variant type (mean = 2066, SD = 1000). Duplications were typically more frequent than inversions (mean = 427, SD = 79; mean = 293, SD = 96). The exception was strain XZ1516, which harboured 519 inversions and 443 duplications (Figure 3).

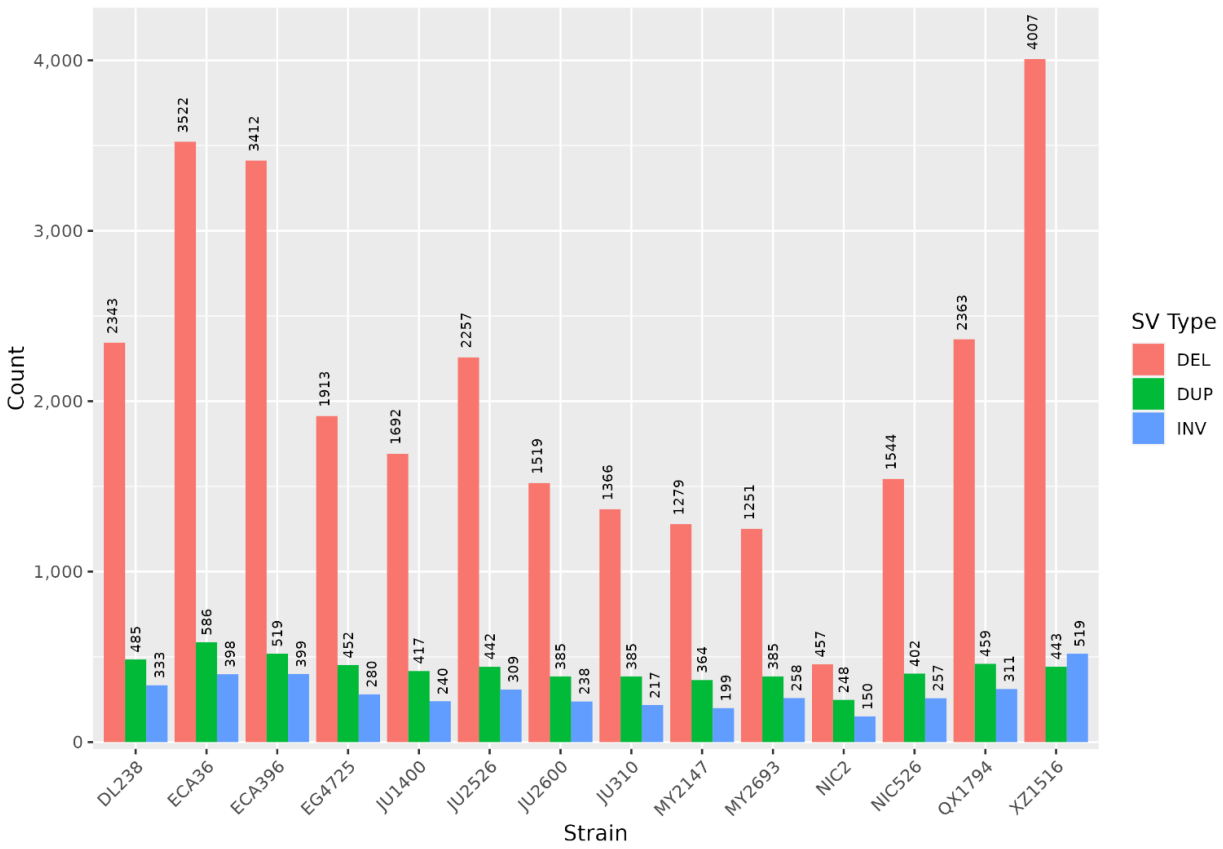

Figure 3 Predicted Structural Variants per Strain.

SV genotypes were assigned to each strain using Jasmine [44]. Among all identified SVs, 64% were singletons, unique to a single strain (Table 2). Conversely, 25 variants were found in all 14 strains: 15 deletions, two duplications, and eight inversions. The proportion of singletons was slightly higher for deletions and duplications (0.65) compared to inversions (0.60). However, at the individual strain level, the average proportion of singleton deletions was 0.25 (SD = 0.16) and ranged from 0.08 to 0.61 (Table 3). Duplications had the highest proportion of singleton variants (mean = 0.33, SD = 0.15; range = 0.19 to 0.60), while inversions had the lowest (mean = 0.21, SD = 0.14; range = 0.07 to 0.60). Strains isolated from Western Europe had the fewest singleton SVs (0.08 to 0.20 for deletions; 0.19 to 0.31 for duplications; 0.07 to 0.18

for inversions). UpSet plots depicting the allele frequencies of SVs among multiple strains can be found in Figures 4, 5, 6.

Table 2. Distribution of structural variant allele frequencies among wild *C. elegans* strains.

Strains Containing Variant[1]	**Deletions**	**Duplications**	**Inversions**
1	9314	2062	1041
2	1897	456	245
3	1095	239	153
4	673	147	78
5	415	82	57
6	323	51	39
7	194	35	32
8	154	25	23
9	97	17	17
10	82	18	21
11	52	10	13
12	44	4	12
13	18	1	9
14	15	2	8

1. Jasmine was used to assign genotypes to each strain for each SV call. Strains with the same genotype for a given SV call were considered to contain the same variant.

Table 3. Singleton variants among individual *C. elegans* strains.

	Region	**Deletions**		**Duplications**		**Inversions**	
Strain		**Total**	**Singleton Count (proportion)**	**Total**	**Singleton Count (proportion)**	**Total**	**Singleton Count (proportion)**
DL238	Hawaii	2343	804 (0.34)	485	209 (0.43)	333	95 (0.29)
ECA396	Hawaii	3412	1643 (0.48)	519	300 (0.58)	399	140 (0.35)
QX1794	Hawaii	2363	756 (0.32)	459	186 (0.41)	311	74 (0.24)
XZ1516	Hawaii	4007	2440 (0.61)	443	266 (0.60)	519	301 (0.58)
ECA36	New Zealand	3522	1671 (0.47)	586	308 (0.53)	398	147 (0.37)
EG4725	Western Europe	1913	290 (0.15)	452	94 (0.21)	280	42 (0.15)
JU1400	Western Europe	1692	225 (0.13)	417	120 (0.29)	240	34 (0.14)
JU2526	Western Europe	2257	452 (0.20)	442	96 (0.22)	309	55 (0.18)
JU2600	Western Europe	1519	290 (0.19)	385	87 (0.23)	238	39 (0.16)
JU310	Western Europe	1366	170 (0.12)	385	72 (0.19)	217	16 (0.07)

MY2147	Western Europe	1279	217 (0.17)	364	85 (0.23)	199	29 (0.15)
MY2693	Western Europe	1251	97 (0.08)	385	84 (0.22)	258	35 (0.14)
NIC2	Western Europe	457	80 (0.18)	248	77 (0.31)	150	12 (0.08)
NIC526	Western Europe	1544	179 (0.12)	402	78 (0.19)	257	22 (0.09)

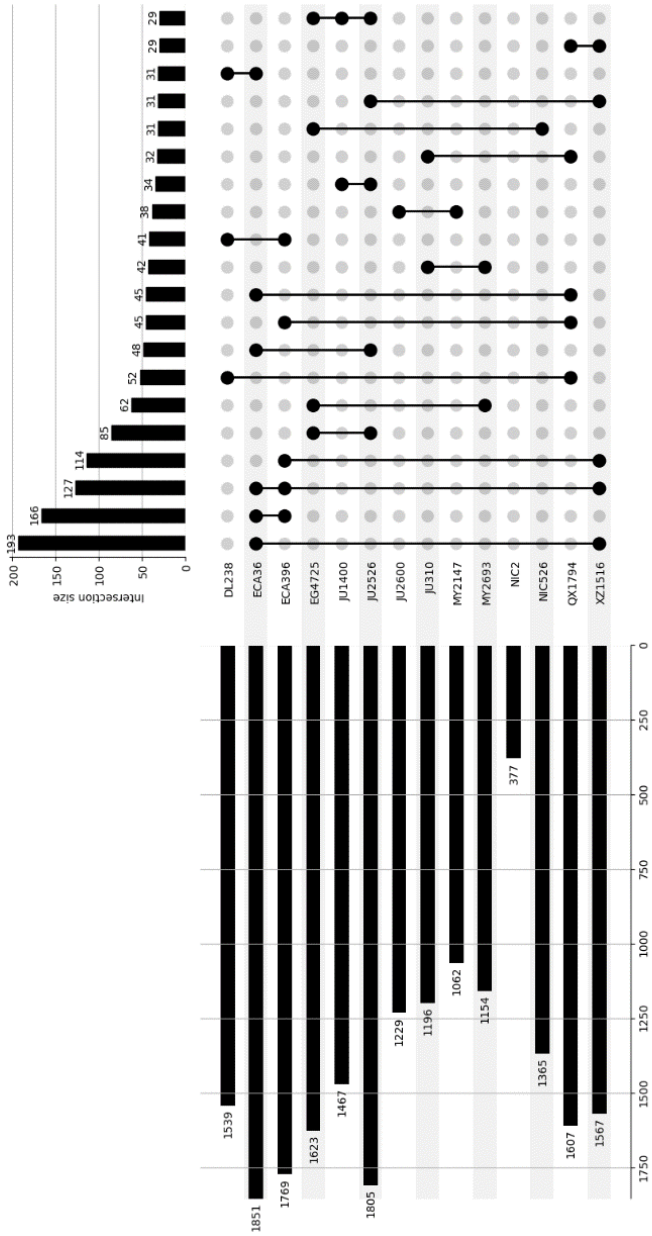

Figure 4 Distribution of deletion allele frequencies in wild *C. elegans* strains. The count totals only include variants present in at least two strains. The plot is limited to the 20 largest sets.

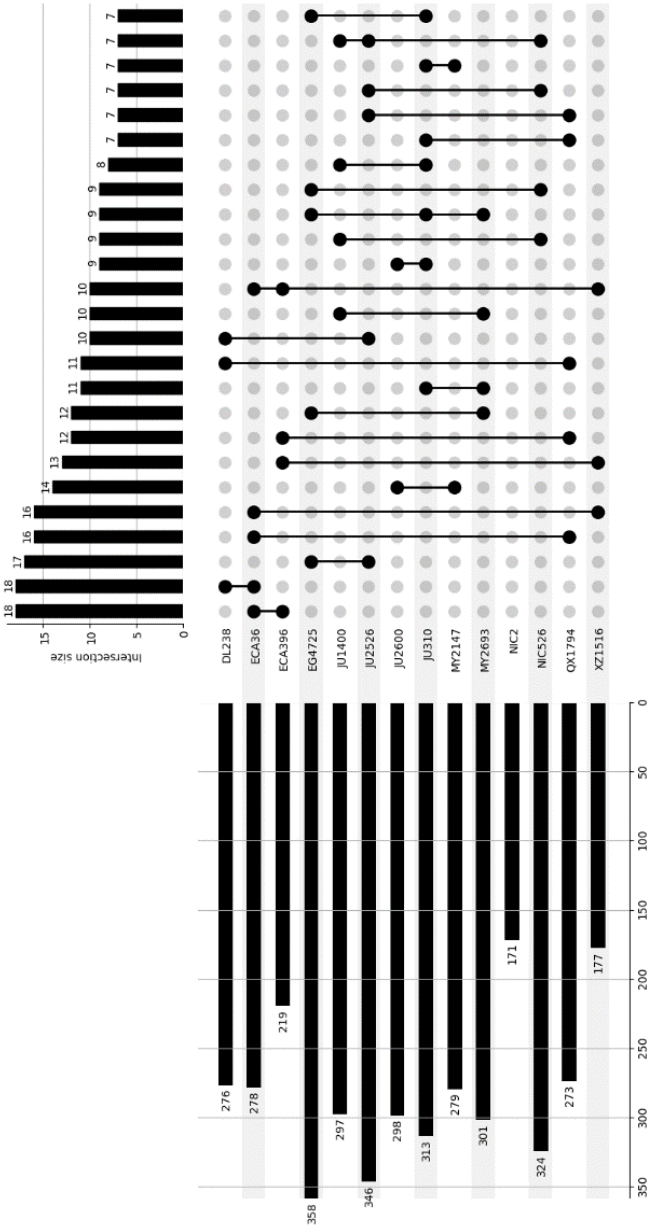

Figure 5 Distribution of duplications allele frequencies in wild *C. elegans* strains. The count totals only include variants present in at least two strains. The plot is limited to the 25 largest sets.

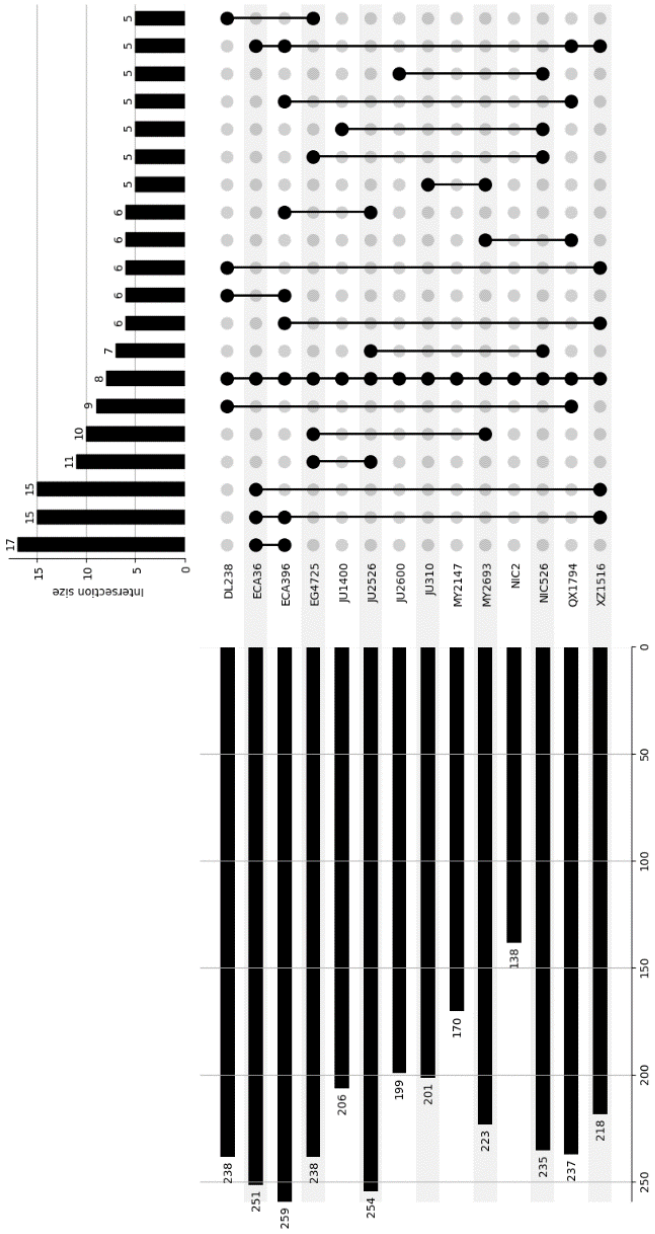

Figure 6 Distribution of inversion allele frequencies in wild *C. elegans* strains. The count totals only include variants present in at least two strains. The plot is limited to the 20 largest sets.

Variant Annotation

The Variant Effect Predictor [45] tool was used to predict the consequences of each SV (Table S1-S3). To identify the variants most likely to cause meaningful phenotypic outcomes in the wild, only variants predicted to have high impacts were selected for downstream analyses. Deletions accounted for the majority of high-impact variants among all strains (mean = 205, SD = 88), followed by duplications (mean = 29, SD = 8). Only three high-impact inversions were identified, each in a different strain. In total, 1,485 deletions, 241 duplications, and three inversions were classified as high-impact SVs (Figure 7). Duplications were the largest variants on average (mean = 10,009, SD =10,898), followed by deletions (mean = 2,923, SD = 6,811), and inversions (mean = 543, SD = 418).

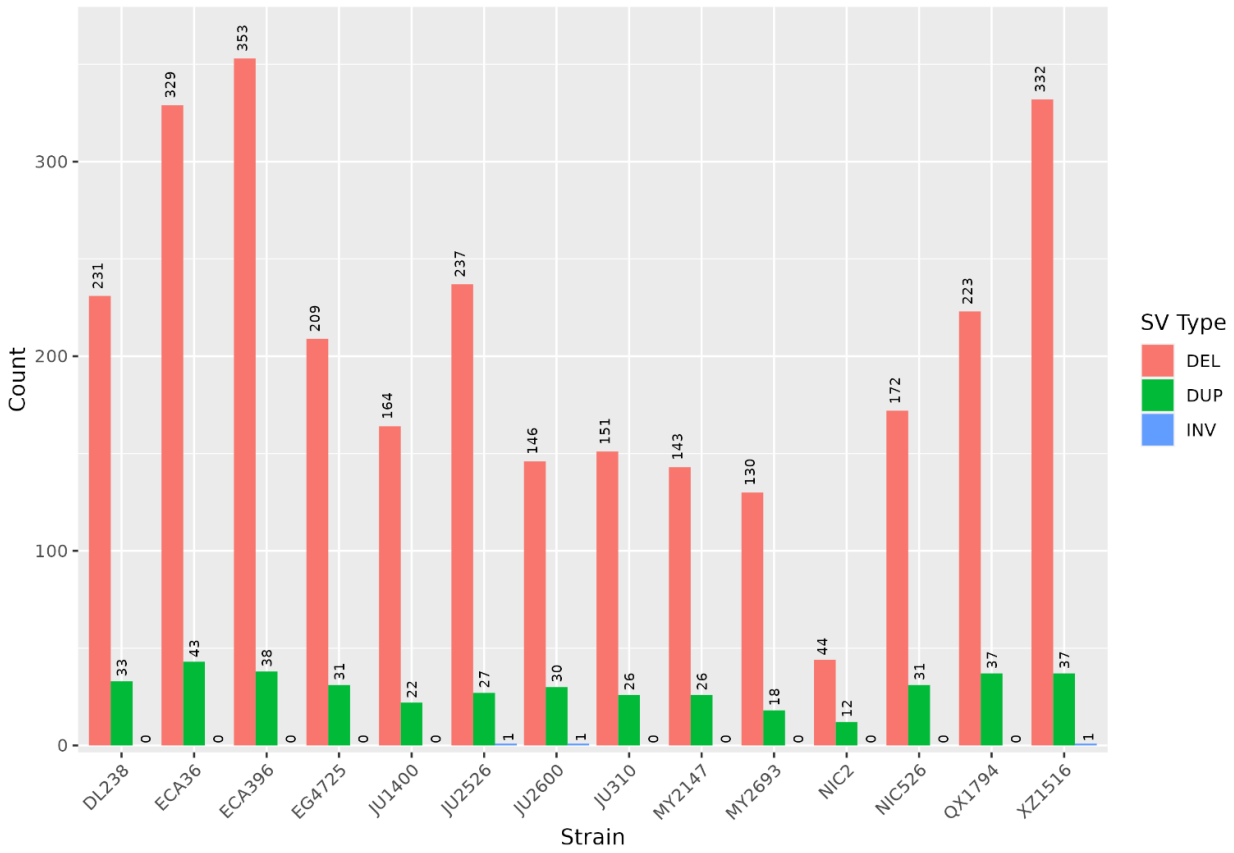

Figure 1 High impact structural variants.

Several types of molecular consequences were identified for the SVs predicted to have high impact. For deletions, a total of 2,233 consequences were predicted from seven categories with splicing variants being the most common types (Table 4). The *transcript_amplification* category accounted for 240 duplications, along with a single *stop_lost* variant. All three inversions were classified as *stop_lost* variants.

Table 4. Predicted Consequences of High Severity Structural Variants.

SV Type	**Term**	**Description[1]**	**Count**
Deletion (n = 1,485)	splice_donor_variant	A splice variant that changes the 2 base region at the 5' end of an intron	733
	splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	724
	transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	411

	frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	223
	stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	120
	start_lost	A codon variant that changes at least one base of the canonical start codon	20
	stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a	2

		premature stop codon, leading to a shortened transcript	
Duplication (n = 241)	transcript_amplification	A feature amplification of a region containing a transcript	240
	start_lost	A codon variant that changes at least one base of the canonical start codon	1
Inversion (n=3)	start_lost	A codon variant that changes at least one base of the canonical start codon	3

1- Consequence descriptions downloaded on May 22, 2023 from: https://useast.ensembl.org/info/genome/variation/prediction/predicted_data.html

Functional Enrichment Analysis

We performed an overrepresentation analysis of the genes affected by high impact SVs using the easyGSEA tool and WormCat annotations [47] (Figure 8). Most genes were categorized as "Unassigned" (C1 = 444; C2 = 444; C3 = 295). After excluding the "Unassigned" categories, four

categories remained at the C1 level, which were subdivided into five and fifteen categories at the C2 and C3 levels, respectively. Many of these genes were linked to responses to external stimuli, such as xenobiotics, pathogens, and other stressors (Table S4).

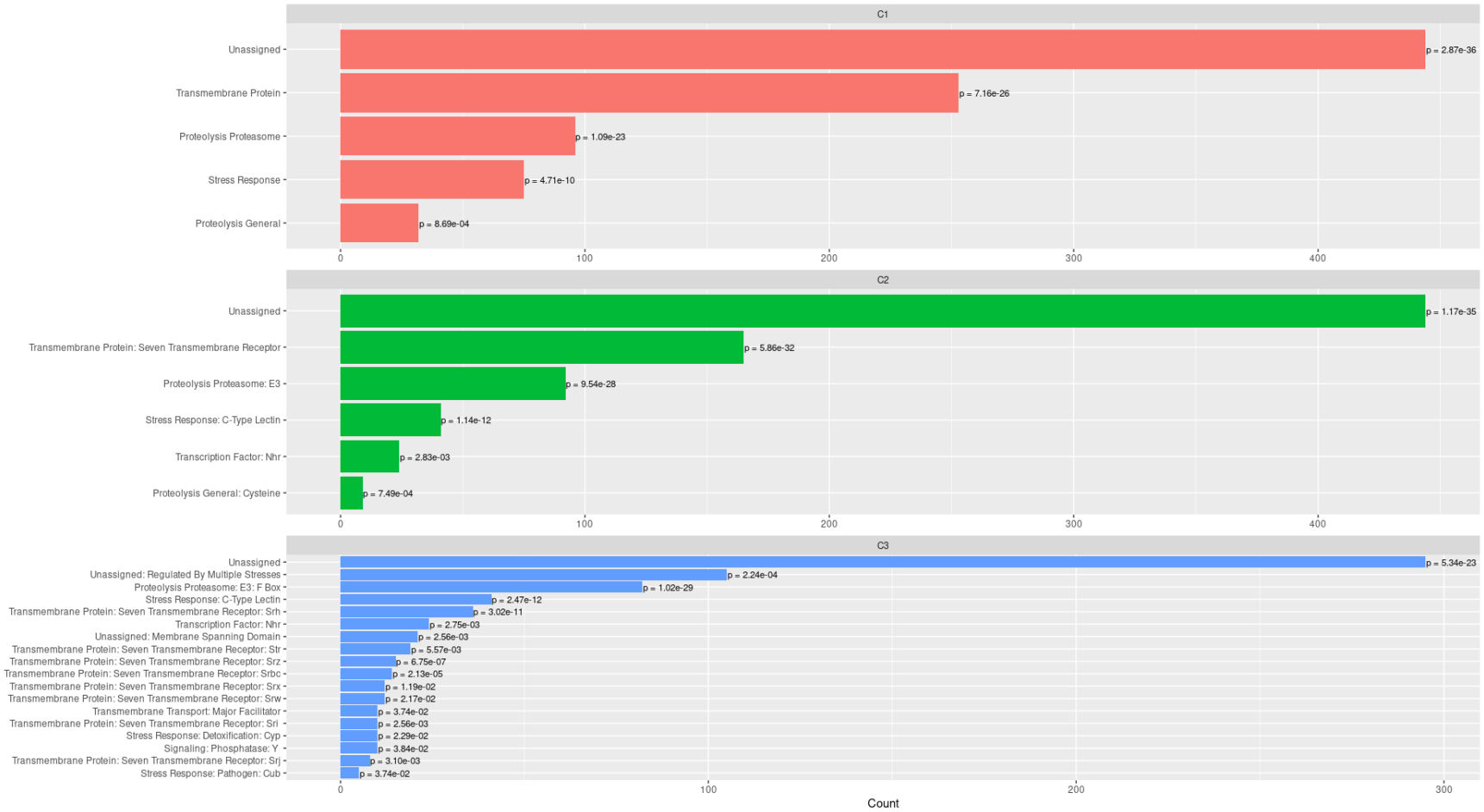

Figure 8 Statistically significant overrepresented WormCat terms for putative high impact SVs. The red, green, and blue colours represent the categories for the C1, C2, and C3 levels, respectively.

GPCRs

The "Transmembrane protein" category included 253 genes, of which 152 were seven-transmembrane G-protein-coupled receptors (GPCRs). There are approximately 1,300 GPCRs encoded in the *C. elegans* genome and they are organised into 19 different families based on sequence similarity [51,52]. The SRH family had the largest representation, with 36 genes affected by SVs.

A duplication of the *srh-234* gene was identified in two strains: DL238, and QX1794. Previous work revealed that *srh-234* expression is regulated by sensory input related to the presence or absence of food, suggesting that this gene plays a role in the chemosensory basis of feeding behaviour in *C. elegans* [53].

Two deletions were identified, which may affect chemosensory behaviour. First, a deletion spanning the entire length of the *sri-14* gene was present in the ECA36 and ECA396 strains. Loss of function experiments linked *sri-14* to impaired avoidance to high diacetyl concentrations [54]. At low concentrations diacetyl is an indicator of potential food, but at high concentrations it is toxic. Second, a deletion was present within the *str-208* gene of the QX1794 strain. Previously, a missense substitution mutation (Cys226Ser/Tyr) in STR-208 was identified in three Bergerac strains—CB4851, RW6999, and RW7000—and was implicated in the erratic movement of these strains, possibly due to protein misfolding [55]. Unfortunately, Bergerac strains were not included in this study due to the lack of long-read data.

F-box Genes and Pathogen Response

Two categories related to proteolysis were assigned to 128 genes: "Proteolysis proteasome" (n = 96) and "Proteolysis general" (n = 32). Among these were 74 F-box genes,

which are known to play critical roles in protein degradation and have undergone significant expansion in *C. elegans* [56,57]. One of the genes considered to be under positive selection, *fbxa-158*, promotes thermotolerance associated intracellular pathogen response in *C. elegans* [56,58], but we found that it was lost in four strains: ECA396, EG4725, JU1400, and MY2693.

Other genes associated with response to pathogen infection were included in the overrepresented terms. The strains ECA396 and XZ1516 harboured a deletion in *hecw-1*, a E3 ubiquitin-protein ligase that has been shown to inhibit *Pseudomonas aeruginosa* avoidance behaviour [59]. Thirty-nine members of the C-Type Lectin family were also affected by SVs. This superfamily is characterized by a C-type lectin-like domain containing approximately 110 to 130 amino acids. C-Type lectins (CLEC) are abundant in *C. elegans* (approximately 283 genes) and include many critical immune-related genes. We identified deletions involving seven CLECs: *clec-2* (induced by *P. aeruginosa*), *clec-10* (suppressed by *P. aeruginosa*), *clec-13* (induced by *Microbacterium nematophilum*), *clec-67* (induced by *P. aeruginosa* and *Stenotrophomonas maltophilia*), *clec-122* (induced by *P. aeruginosa*), *clec-143* (induced by *P. aeruginosa*), and *clec-174* (induced by *P. aeruginosa*, *Photorhabdus luminescens*, and *Vibrio cholerae*) [60–63]. Three CLECs were duplicated, one of which, *clec-70* (present in DL238), is induced by *M. nematophilum* and *Staphylococcus aureus* [60]. No functional role has been determined for the other two duplicated CLECs: *clec-189* (present in EG4725 and JU2526) and *clec-258* (present in QX1794).

Detoxification and Xenobiotic Response

The category "stress response", related to detoxification, was overrepresented and included the cytochrome P450s (CYPs), glutathione S-transferases (GSTs), UDP-glucuronosyl

transferases (UGTs), and nuclear hormone receptors (NHRs). CYP enzymes, involved in the oxidative metabolism of numerous endogenous and exogenous substrates, are particularly noteworthy. The *cyp-35a4* was deleted in strains DL238 and XZ1516, and entirely duplicated in strain QX1794. This gene is involved in the regulation of fatty acid accumulation, is inducible using caffeine, which suggests a role in mitigating stress or toxin exposure [64,65]. A deletion overlapping another caffeine inducible CYP, *cyp-33e2*, was found in strains ECA396, JU2526, and XZ1516. Deletions were identified in *cyp-33c5* and *cyp-13a5*; both genes are induced by cadmium and associated with its detoxification[66]. The strain ECA396 contained a deletion overlapping *cyp-33d3*, a gene inducible by ethidium bromide and low pH [67,68]. Finally, three strains— EG4725, JU1400, JU2526—contained a deletion in *cyp-33c5* and two other strains—JU310, QX1794—contained a deletion overlapping *cyp-13a5*. A deletion overlapping *cyp-33d3* was also identified in ECA396.

Several GSTs and UGTs were affected by deletions, including *gst-17* (XZ1516) and *gst-18* (EG4725), and *ugt-14* and *ugt-32* (both in ECA36). Notably, *ugt-14* was upregulated following exposure to albendazole, mebendazole, thiabendazole, and oxfendazole [69,70]. A total of 26 NHRs were affected by SVs. In strain XZ1516, *nhr-196* was partially deleted. This gene was differentially expressed in *C. elegans* following exposure to the mycotoxin aflatoxin B1, and may have a role in its detoxification [71].

Discussion

Structural Variants in Wild *C. elegans* Strains

SVs are a major source of genetic diversity, associated with both disease and key evolutionary processes. Despite considerable progress in developing computational methods for variant calling, accurate, reproducible and comprehensive identification of SVs remains challenging. Technical limitations and a bias towards human studies and laboratory-adapted strains continue to hamper our understanding of SVs, especially in natural populations. In this study, we profiled SVs in 14 *C. elegans* strains which had PacBio long-reads available. These long-reads can mitigate many of the issues common in short-read-based variant calling.

Our final merged SV call set contained 14,473 deletions, 3,149 duplications, and 1,748 inversions. Consistent with previous studies, most deletions and duplications were concentrated in the chromosome arms of the five autosomes[28,32,72,73]. In *C. elegans*, the chromosome arms exhibit lower gene densities and higher recombination rates in comparison to the gene rich centres that contain most core genes[74,75]. Therefore, the chromosome arms provide favourable conditions in which to generate new SVs through recombination mediated mechanisms[76–78], while minimizing the risk of disrupting essential core genes. In fact, most of the genetic diversity in *C. elegans* is maintained in the chromosome arms, particularly in hyper-divergent regions that span approximately 20% of the reference genome[32]. Notably, the hyper-divergent regions are maintained through balancing selection and are enriched for genes associated with responses to environmental stimuli and stresses. This suggests that the hyper-divergent regions may have enabled *C. elegans* to adapt to a range of environments by maintaining genetic variation, which would typically be lost in a selfing species[32].

Genome-wide SV calling often generates hundreds or thousands of calls that precludes a manual review of the entire call set. To prioritise functionally relevant SVs, we used the Variant Effect Predictor tool to identify SVs presumed to have significant phenotypic consequences. We acknowledge that by taking this strategy, we ignored SVs limited to intergenic regions and likely overlooked some functionally important variants[79,80]. Altogether, 1,253 unique genes were selected from a set of SVs that included 1,485 deletions, 241 duplications, and three inversions. An overrepresentation analysis, using functional annotations from the WormCat database, classified 444 genes as "Unassigned", a term used to classify poorly annotated genes in WormCat[81]. Targeted investigation of these genes would be highly valuable, particularly with the remaining categories containing an abundance of genes associated with responses to environmental stimuli and stressors.

Genes belonging to the GPCR superfamily and F-Box family were significantly overrepresented among SVs, accounting for 152 and 74 genes, respectively. Comparisons with other *Caenorhabditis* species have revealed large expansions of the F-Box family within members of the genus, as well as large expansions of several GPCR families (*sra*, *srab*, *srh*, *sri*, *srz*, and *str*) families in *C. elegans* [51,82–84]. For both *C. briggsae* and *C. elegans,* higher non-synonymous to synonymous substitution ratios have been found in F-Box and srz genes, suggesting that these families are under positive selection [51,83]. Recently, non-synonymous substitutions in 330 CeNDR strains were used to quantify the genetic diversity and population structure of *C. elegans*[56]. Further signatures of positive selection in F-Box and GPCR genes were identified, which were strongest for the F-Box and *srw* families. The highest genetic diversity was observed in strains isolated from Hawaii, which also retained many of the ancestral genotypes. These results are consistent with prior studies that reported selective sweeps in non-

Hawaiian strains and support the hypothesis that *C. elegans* originated from Hawaii or a similar niche[85,86]. Currently, the functions of most GPCR and F-Box genes are unknown, but chemosensation and immune defence are believed to play prominent roles. As *C. elegans* inhabits microbiota-rich habitats and has a limited ability to detect light or sound, these large gene families may reflect a need to respond to a range of environmental stimuli and pathogens[87–89].

The immune system of *C. elegans* also appears to be significantly affected by SVs; 39 of the 283 *C. elegans* C-type lectins (CLECs) were present in SVs. CLECs are involved in pathogen recognition and immune defense, and the variation of this family within the genus *Caenorhabditis*—from 111 copies in *C. japonica* to 319 in *C. brenneri*— along with differential expression in response to pathogens, suggests an evolutionary adaptation to diverse microbial environments in the wild[60,90].

Several genes associated with xenobiotic detoxification—CYPs, GSTs, UGTs, and NHRs—were overrepresented. These gene families have undergone multiple lineage specific expansions in nematodes and are associated with responses to environmental toxins and anthelmintics[91–93]. As with the CLECs, the specific function of most detoxification genes remains unknown.

Our study provides a foundation for future research. Those genes affected by SVs should be prioritised for targeted investigations into their specific function, particularly around *C. elegans* response to its environment and exposure to pathogens. Hybrid approaches that combine long- and short sequence reads[41] would likely allow us to accurately expand our understanding of SVs in natural populations of *C. elegans* and, arguably, in animals. Additionally, mapping SVs through quantitative trait loci (eQTL) could offer new insights into the genetic basis of traits that

are difficult to resolve using single nucleotide variants alone[2]. The recent availability of transcriptomic data for wild *C. elegans* strains now offers an unique opportunity to explore these relationships in greater detail[94].

Acknowledgments

We thank Dr. Erik Andersen and members of the Andersen lab (Johns Hopkins University) for discussions on using CeNDR datasets. We also acknowledge the high-performance computing resources by the Faculty of Veterinary Medicine and Research Computing at the University of Calgary.

Funding

This work was funded through grants to JDW from the Natural Sciences and Engineering Research Council of Canada (NSERC; #04589-2020).

References

1. Hurles, M. E., Dermitzakis, E. T. & Tyler-Smith, C. The functional impact of structural variation in humans. *Trends Genet.* **24**, 238–245 (2008).

2. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).

3. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61**, 437–455 (2010).

4. Hurles, M. Gene duplication: The genomic trade in spare parts. *PLoS Biol.* **2**, (2004).

5. Katju, V. & Bergthorsson, U. Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Front. Genet.* **4**, 1–12 (2013).

6. Zhang, J. Evolution by gene duplication: An update. *Trends Ecol. Evol.* **18**, 292–298 (2003).

7. Kirkpatrick, M. How and why chromosome inversions evolve. *PLoS Biol.* **8**, (2010).

8. Faria, R., Johannesson, K., Butlin, R. K. & Westram, A. M. Evolving inversions. *Trends Ecol. Evol.* **34**, 239–248 (2019).

9. Lesack, K., Mariene, G. M., Andersen, E. C. & Wasmuth, J. D. Different structural variant prediction tools yield considerably different results in *Caenorhabditis elegans*. *PLoS One* **17**, (2022).

10. Koboldt, D. C. Best practices for variant calling in clinical sequencing. *Genome Med.* **12**, 1–13 (2020).

11. Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* **42**, 1571–1580 (2024).

12. Lye, Z. N. & Purugganan, M. D. Copy number variation in domestication. *Trends Plant Sci.* **24**, 352–365 (2019).

13. Żmieńko, A., Samelak, A., Kozłowski, P. & Figlerowicz, M. Copy number polymorphism in plant genomes. *Theor. Appl. Genet.* **127**, 1–18 (2014).

14. Dolatabadian, A., Patel, D. A., Edwards, D. & Batley, J. Copy number variation and disease resistance in plants. *Theor. Appl. Genet.* **130**, 2479–2490 (2017).

15. Clop, A., Vidal, O. & Amills, M. Copy number variation in the genomes of domestic animals. *Anim. Genet.* **43**, 503–517 (2012).

16. Brown, K. H. *et al.* Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 529–534 (2012).

17. Liu, S. *et al.* Identification of high-confidence structural variants in domesticated rainbow trout using whole-genome sequencing. *Front. Genet.* **12**, 1–9 (2021).

18. Serres-Armero, A. *et al.* Copy number variation underlies complex phenotypes in domestic dog breeds and other canids. *Genome Res.* **31**, 762–774 (2021).

19. Serres-Armero, A. *et al.* Similar genomic proportions of copy number variation within gray wolves and modern dog breeds inferred from whole genome sequencing. *BMC Genomics* **18**, 1–15 (2017).

20. Farslow, J. C. *et al.* Rapid increase in frequency of gene copy-number variants during experimental evolution in *Caenorhabditis elegans*. *BMC Genomics* **16**, 1–18 (2015).

21. Konrad, A. *et al.* Mutational and transcriptional landscape of spontaneous gene duplications and deletions in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 7386–7391 (2018).

22. Lipinski, K. J. *et al.* High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Curr. Biol.* **21**, 306–310 (2011).

23. Zhao, Y. *et al.* A spontaneous complex structural variant in *rcan-1* increases exploratory behavior and laboratory fitness of *Caenorhabditis elegans*. *PLoS Genet.* **16**, 1–29 (2020).

24. Lu, J. *et al.* The accumulation of deleterious mutations in rice genomes: A hypothesis on

the cost of domestication. *Trends Genet.* **22**, 126–131 (2006).

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

26. Andersen, E. C. & Rockman, M. V. Natural genetic variation as a tool for discovery in *Caenorhabditis* nematodes. *Genetics* **220**, (2022).

27. Maydan, J. S. *et al.* Efficient high-resolution deletion discovery in *Caenorhabditis elegans* by array comparative genomic hybridization. *Genome Res.* **17**, 337–347 (2007).

28. Lee, B. Y., Kim, J. & Lee, J. Intraspecific de novo gene birth revealed by presence-absence variant genes in *Caenorhabditis elegans*. *NAR Genomics Bioinforma.* **4**, 1–13 (2022).

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

30. Vergara, I. A. *et al.* Genome-wide variations in a natural isolate of the nematode *Caenorhabditis elegans*. *BMC Genomics* **15**, (2014).

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

32. Lee, D. *et al.* Balancing selection maintains hyper-divergent haplotypes in *Caenorhabditis elegans*. *Nat. Ecol. Evol.* **5**, 794–807 (2021).

33. Cook, D. E., Zdraljevic, S., Roberts, J. P. & Andersen, E. C. CeNDR, the *Caenorhabditis elegans* natural diversity resource. *Nucleic Acids Res.* **45**, D650–D657 (2017).

34. Evansid, K. S. *et al.* Two novel loci underlie natural differences in *Caenorhabditis elegans* abamectin responses. *PLoS Pathog.* **17**, 1–26 (2021).

35. Hahnel, S. R. *et al.* Extreme allelic heterogeneity at a *Caenorhabditis elegans* beta-tubulin locus explains natural resistance to benzimidazoles. *PLoS Pathog.* **14**, 1–26 (2018).

36. Shaver Amanda *et al.* Variation in anthelmintic responses are driven by genetic

differences among diverse *C. elegans wild* strains. *PLOS Pathog.* 1–28 (2022) doi:10.1371/journal.ppat.1011285.

37. Wit, J., Rodriguez, B. C. & Andersen, E. C. Natural variation in *Caenorhabditis elegans* responses to the anthelmintic emodepside. *Int. J. Parasitol. Drugs Drug Resist.* **16**, 1–8 (2021).

38. Evans, K. S. *et al.* Shared genomic regions underlie natural variation in diverse toxin responses. *Genetics* **210**, 1509–1525 (2018).

39. Widmayer, S. J., Crombie, T. A., Nyaanga, J. N., Evans, K. S. & Andersen, E. C. *C. elegans* toxicant responses vary among genetically diverse individuals. *Toxicology* **479**, 153292 (2022).

40. Zdraljevic, S. *et al.* Natural variation in *C. elegans* arsenic toxicity is explained by differences in branched chain amino acid metabolism. *Elife* **8**, 1–28 (2019).

41. Quan, C., Lu, H., Lu, Y. & Zhou, G. Population-scale genotyping of structural variation in the era of long-read sequencing. *Comput. Struct. Biotechnol. J.* **20**, 2639–2647 (2022).

42. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).

43. Broad Institute. Picard. (2022).

44. Kirsche, M. *et al.* Jasmine and Iris: Population-scale structural variant comparison and analysis. *Nat. Methods* **20**, 408–417 (2023).

45. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 1–14 (2016).

46. Cheng, X., Yan, J., Liu, Y., Wang, J. & Taubert, S. EVITTA: A web-based visualization and inference toolbox for transcriptome analysis. *Nucleic Acids Res.* **49**, W207–W215 (2021).

47. Holdorf, A. D. *et al.* WormCat: An online tool for annotation and visualization of *Caenorhabditis elegans* genome-scale data. *Genetics* **214**, 279–294 (2020).

48. Gel, B. & Serra, E. KaryoploteR: An R/Bioconductor package to plot customizable

genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).

49. Nothman, J. UpSetPlot.

50. Wickham, H. *ggplot2: Elegant graphics for data analysis*. (2016).

51. Thomas, J. H. & Robertson, H. M. The Caenorhabditis chemoreceptor gene families. *BMC Biol.* **6**, (2008).

52. Vidal, B. *et al.* An atlas of *Caenorhabditis elegans* chemoreceptor expression. *PLOS Biol.* **16**, e2004218 (2018).

53. Gruner, M. *et al.* Feeding state, insulin and NPR-1 modulate chemoreceptor gene expression via integration of sensory and circuit inputs. *PLoS Genet.* **10**, (2014).

54. Taniguchi, G., Uozumi, T., Kiriyama, K., Kamizaki, T. & Hirotsu, T. Screening of odor-receptor pairs in *Caenorhabditis elegans* reveals different receptors for high and low odor concentrations. *Sci. Signal.* **7**, 1–10 (2014).

55. Daigle, A. T., Deiss, T. C., Melde, R. H., Bergthorsson, U. & Katju, V. Bergerac strains of *Caenorhabditis elegans* revisited: Expansion of Tc1 elements imposes a significant genomic and fitness cost. *G3 Genes, Genomes, Genet.* **12**, (2022).

56. Ma, F., Lau, C. Y. & Zheng, C. Large genetic diversity and strong positive selection in F-Box and GPCR genes among the wild isolates of *Caenorhabditis elegans*. *Genome Biol. Evol.* **13**, 1–20 (2021).

57. Ho, M. S., Tsai, P. I. & Chien, C. T. F-box proteins: The key to protein degradation. *J. Biomed. Sci.* **13**, 181–191 (2006).

58. Panek, J. *et al.* A cullin-RING ubiquitin ligase promotes thermotolerance as part of the intracellular pathogen response in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 7950–7960 (2020).

59. Chang, H. C., Paek, J. & Kim, D. H. Natural polymorphisms in *C. elegans* HECW-1 E3 ligase affect pathogen avoidance behaviour. *Nature* **480**, 525–529 (2011).

60. Pan, W., Huang, X., Guo, Z., Nagarajan, R. & Mylonakis, E. Identification and Functional

Analysis of Cytokine-Like Protein CLEC-47 in Caenorhabditis elegans. *MBio* **12**, 1–18 (2021).

61. Schulenburg, H., Hoeppner, M. P., Weiner, J. & Bornberg-Bauer, E. Specificity of the innate immune system and diversity of C-type lectin domain (CTLD) proteins in the nematode *Caenorhabditis elegans*. *Immunobiology* **213**, 237–250 (2008).

62. Sahu, S. N. *et al.* Genomic analysis of immune response against *vibrio cholerae* hemolysin in *Caenorhabditis elegans*. *PLoS One* **7**, 1–12 (2012).

63. White, C. V. & Herman, M. A. Transcriptomic, functional, and network analyses reveal novel genes involved in the interaction between *Caenorhabditis elegans* and *Stenotrophomonas maltophilia*. *Front. Cell. Infect. Microbiol.* **8**, 1–14 (2018).

64. Aarnio, V. *et al. Caenorhabditis elegans* mutants predict regulation of fatty acids and endocannabinoids by the CYP-35A gene family. *Front. Pharmacol.* **MAR**, 1–10 (2011).

65. Min, H., Kawasaki, I., Gong, J. & Shim, Y. H. Caffeine induces high expression of cyp-35A family genes and inhibits the early larval development in *Caenorhabditis elegans*. *Mol. Cells* **38**, 236–242 (2015).

66. Cui, Y., McBride, S. J., Boyd, W. A., Alper, S. & Freedman, J. H. Toxicogenomic analysis of *Caenorhabditis elegans* reveals novel genes and pathways involved in the resistance to cadmium toxicity. *Genome Biol.* **8**, 1–15 (2007).

67. Cong, Y. *et al.* Transcriptome analysis of the nematode *Caenorhabditis elegans* in acidic stress environments. *Front. Physiol.* **11**, 1–12 (2020).

68. Leung, M. C. K. *et al.* Effects of early life exposure to ultraviolet C radiation on mitochondrial DNA content, transcription, ATP production, and oxygen consumption in developing *Caenorhabditis elegans*. *BMC Pharmacol. Toxicol.* **14**, (2013).

69. Matoušková, P. *et al.* UDP-glycosyltransferase family in *Haemonchus contortus*: Phylogenetic analysis, constitutive expression, sex-differences and resistance-related differences. *Int. J. Parasitol. Drugs Drug Resist.* **8**, 420–429 (2018).

70. Stasiuk, S. J. *et al.* Similarities and differences in the biotransformation and transcriptomic

responses of *Caenorhabditis elegans* and *Haemonchus contortus* to five different benzimidazole drugs. *Int. J. Parasitol. Drugs Drug Resist.* **11**, 13–29 (2019).

71. Karengera, A. *et al.* Differential expression of genes in *C. elegans* reveals transcriptional responses to indirect-acting xenobiotic compounds and insensitivity to 2,3,7,8-tetrachlorodibenzodioxin. *Ecotoxicol. Environ. Saf.* **233**, 113344 (2022).

72. Cutter, A. D., Dey, A. & Murray, R. L. Evolution of the *Caenorhabditis elegans* genome. *Mol. Biol. Evol.* **26**, 1199–1234 (2009).

73. Laricchia, K. M., Zdraljevic, S., Cook, D. E. & Andersen, E. C. Natural variation in the distribution and abundance of transposable elements across the *Caenorhabditis elegans* species. *Mol. Biol. Evol.* **34**, 2187–2202 (2017).

74. C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science (80-.).* **282**, 2012–2018 (1998).

75. Rockman, M. V. & Kruglyak, L. Recombinational landscape and population genomics of *Caenorhabditis elegans*. *PLoS Genet.* **5**, (2009).

76. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).

77. Currall, B. B., Chiangmai, C., Talkowski, M. E. & Morton, C. C. Mechanisms for structural variation in the human genome. *Curr. Genet. Med. Rep.* **1**, 81–90 (2013).

78. Escaramís, G., Docampo, E. & Rabionet, R. A decade of structural variants: Description, history and methods to detect structural variation. *Brief. Funct. Genomics* **14**, 305–314 (2015).

79. Boyling, A., Perez-Siles, G. & Kennerson, M. L. Structural variation at a disease mutation hotspot: Strategies to investigate gene regulation and the 3D genome. *Front. Genet.* **13**, 1–16 (2022).

80. Mielczarek, M. *et al.* An effect of large-scale deletions and duplications on transcript expression. *Funct. Integr. Genomics* **23**, 1–9 (2023).

81. Higgins, D. P., Weisman, C. M., Lui, D. S., D'Agostino, F. A. & Walker, A. K. Defining characteristics and conservation of poorly annotated genes in *Caenorhabditis elegans* using WormCat 2.0. *Genetics* **221**, (2022).

82. Chen, N. *et al.* Identification of a nematode chemosensory gene family. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 146–151 (2005).

83. Thomas, J. H., Kelly, J. L., Robertson, H. M., Ly, K. & Swanson, W. J. Adaptive evolution in the SRZ chemoreceptor families of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 4476–4481 (2005).

84. Wang, A., Chen, W. & Tao, S. Genome-wide characterization, evolution, structure, and expression analysis of the F-box genes in *Caenorhabditis*. *BMC Genomics* **22**, 1–16 (2021).

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

86. Crombie, T. A. *et al.* Deep sampling of Hawaiian *Caenorhabditis elegans* reveals high genetic diversity and admixture with global populations. *Elife* **8**, 1–27 (2019).

87. Burr, A. H. The photomovement of *Caenorhabditis elegans*, a nematode which lacks ocelli. Proof that the response is to light not radiant heating. *Photochem. Photobiol.* **41**, 577–582 (1985).

88. Iliff, A. J. *et al.* The nematode *C. elegans* senses airborne sound. *Neuron* **109**, 3633-3646.e7 (2021).

89. Ward, A., Liu, J., Feng, Z. & Xu, X. Z. S. Light-sensitive neurons and channels mediate phototaxis in *C. elegans*. *Nat. Neurosci.* **11**, 916–922 (2008).

90. Pees, B., Yang, W., Zárate-Potes, A., Schulenburg, H. & Dierking, K. High innate immune specificity through diversified C-type lectin-like domain proteins in invertebrates. *J. Innate Immun.* **8**, 129–142 (2016).

91. Laing, S. T. *et al.* Characterization of the xenobiotic response of *Caenorhabditis elegans* to the anthelmintic drug albendazole and the identification of novel drug glucoside

metabolites. *Biochem. J.* **432**, 505–514 (2010).

92. Markov, G. V., Baskaran, P. & Sommer, R. J. The same or not the same: Lineage-specific gene expansions and homology relationships in multigene families in nematodes. *J. Mol. Evol.* **80**, 18–36 (2015).

93. Matoušková, P., Vokřál, I., Lamka, J. & Skálová, L. The role of xenobiotic-metabolizing enzymes in anthelmintic deactivation and resistance in helminths. *Trends Parasitol.* **32**, 481–491 (2016).

94. Zhang, G., Roberto, N. M., Lee, D., Hahnel, S. R. & Andersen, E. C. The impact of species-wide gene expression variation on *Caenorhabditis elegans* complex traits. *Nat. Commun.* **13**, 1–13 (2022).