High rate of mutation and efficient removal by selection of structural variants from natural populations of *Caenorhabditis elegans*

Ayush Shekhar Saxena[1,2] and Charles F. Baer[1,3]

1 – Department of Biology, University of Florida, Gainesville, FL, USA

2 – Present address – Regeneron Pharmaceuticals Inc., Tarrytown, NY, USA

3 – University of Florida Genetics Institute, Gainesville, FL, USA

Correspondence to:

Ayush Shekhar Saxena

ayushsaxena.iitg@gmail.com

Abstract

The importance of genomic structural variants (SVs) is well-appreciated, but much less is known about their mutational properties than of single nucleotide variants (SNVs) and short indels. The reason is simple: the longer the mutation, the less likely it will be covered by a single sequencing read, thus the harder it is to map unambiguously to a unique genomic location.

Here we report SV mutation rate estimates from six mutation accumulation (MA) lines from two strains of *C. elegans* (N2 and PB306) using long-read (PacBio) sequencing. The inferred SV mutation rate ~1/10 the SNV rate and ~1/4 the short indel rate. We identified 40 mutations, and removed 52 false positives (FP) by manual inspection of each SV call. Excluding one atypical line (5 mutations, 35 FPs), the signal (inferred mutant) to noise (FP) ratio is approximately 2:1. False negative rates were determined by simulating variants in the reference genome, and observing 'recall'. Recall rate ranges from >90% for short indels and declines as SV length increases. Small deletions have nearly the same recall rate as small insertions (~100bp), but deletions have higher recall rates than insertions as size increases. The reported SV mutation rate is likely an underestimate.

A quarter of identified SV mutations occur in SV hotspots that harbor pre-existing low complexity repeat variation. By comparison of the spectrum of spontaneous SVs to wild isolates, we infer that natural selection is not only efficient at removing SVs in exons, but also removes roughly half of SVs in intergenic regions.

Introduction

Based on the total number of bases affected, structural variants (SV) — large insertions and deletions, repeat expansions and contractions, inversions, and translocations — are the largest source of genetic diversity in the genomes of multicellular organisms, and they have the largest average phenotypic and fitness effects of any type of mutation (Collins et al. 2020; Beyter et al. 2021). SVs are associated with a host of human disorders, including autism (Marshall et al. 2008; Weiss et al. 2008; Pinto et al. 2010; Sanders et al. 2011), schizophrenia (Walsh et al. 2008; Karayiorgou et al. 2010), and several types of cancers (Friedman et al. 1994; Wooster et al. 1995; Meyerson and Pellman 2011; Stransky et al. 2014; Nattestad et al. 2018). The functional consequences of SVs are well-documented (Hurles et al. 2008; Conrad et al. 2010; Yi and Ju 2018).

In recent years, human population geneticists have converged on the idea that the genetic basis of complex disease can be largely accounted for by the combined influence of very rare, recently arisen variants of large effect (Wang et al. 2021) and variants of small effect at very many loci (the "omnigenome", Boyle et al. 2017). Both classes of variants imply a significant role for mutation, the former for obvious reasons and the latter because, even if the per-locus mutation rate is low, the cumulative genetic variance contributed by a large fraction of the genome at mutation-selection-drift balance may be non-trivial.

Given the outsized contribution of SVs to genetic diversity on the one hand, and the apparently important contribution of mutation to complex disease on the other, understanding the mutational properties of SVs is an important problem in biomedical genetics. More broadly, characterizing SV mutation is necessary for an unbiased estimate of the distribution of fitness effects (DFE) of new mutations, which is a longstanding goal of evolutionary genetics (Fisher 1922; Eyre-Walker and Keightley 2007). Nearly all estimates of the DFE are restricted to the effects of SNPs and small indels (and most are restricted to SNPs, e.g., Kousathanas and Keightley 2013; Kim et al. 2017; Tataru et al. 2017; Böndel et al. 2019; Gilbert et al. 2021; James et al. 2023; Crombie et al. 2024). Because SVs are

ignored, all the variation in fitness is ascribed to small indels and SNPs, artificially inflating their importance by an unknown amount that may be quite large.

The reason that the mutational properties of SVs are poorly characterized is not because their biological relevance is not appreciated; it obviously is. Rather, the problem is technical: SVs are challenging to unambiguously identify and characterize with short-read sequencing technology, which remains the dominant mode of whole-genome sequencing. Long-read sequencing is superior to short-read sequencing for SV calling for several reasons. Longer reads may span the entire region of interest in one contiguous molecule, including problematic regions such as low complexity repeats. Even when a single read does not span the entire region of interest, longer reads result in superior local *de novo* assemblies that can span regions much larger than the read length itself. The obvious disadvantages include cost and error rate with respect to base-substitution and small indel calls, although the accuracy of the latest generation of some long-read platforms approaches that of short-read sequencing (Harvey et al. 2023).

A complete *de novo* assembly (whole genome assembly, WGA) of an individual is the only method certain to capture the full array of structural variants in a genome (Chaisson et al. 2015). As the cost of long-read technology declines, genome assemblies from single individuals have become economically feasible. Audono et al. (2019) published fully assembled reference-quality genomes of 13 humans, from which they estimated genetic diversity resulting from SVs as a proof of concept. More recently, Nurk et al. (2022) employed long-read sequencing to finish a complete telomere-to-telomere sequence of a human genome. The application of long-read sequencing technology to characterize spontaneous mutations is still nascent, although it is evident that long-read sequencing picks up variants that are missed with short-read sequencing (Noyes et al. 2022; Lopez-Cortegano et al. 2023; Zhou et al. 2023; Hénault et al. 2024; López-Cortegano et al. 2025).

Here we present estimates of the spontaneous mutation rate and spectrum of SVs in a set of *C. elegans* mutation accumulation (MA) lines propagated under minimal selection for

~250 generations. A schematic diagram of the MA experiment is depicted in **Figure 1**. Four MA lines derived from the N2 strain and two MA lines derived from the PB306 strain, and their ancestral progenitors, were sequenced using Pacific Biosciences Sequel and Sequel 2 technology. In addition, we report estimates of the standing SV diversity derived from long-read sequences of the genomes of four wild isolates of *C. elegans*. Comparison of the relative proportions of SVs to other variants (SNPs and small indels) among spontaneous mutations to the same proportions in the standing variation provides an estimate of the relative strength of selection against SVs in nature.

Results

Workflow

We identify structural variants (SVs), defined as mutations larger than 30 base pairs (bp), using a combination of alignment and assembly, followed by conventional variant calling. Recognizing the inherent challenges in SV detection, we attempted to automate the process of discovering the optimal parameters for the variant calling pipeline using segregating SVs as 'ground truth', under the assumption that a variant called across many wild isolates is likely real. We discovered that the signal-to-noise ratio of spontaneous SVs is significantly lower than that of segregating SVs, complicating the development of an appropriate pipeline. These complications are detailed in section 2 of the **Supplementary Discussion**. We ultimately adopted a workflow loosely based on the methodology of Audano et al. (2019), with three key modifications:

1. Utilization of a superior aligner, **minimap2** (Li 2018), instead of **blasr** (https://github.com/jcombs1/blasr);
2. Variant calling with **PBSV (**https://github.com/PacificBiosciences/pbsv); and, critically,
3. Manual ("by eye") verification of each SV call by inspection in **IGV** (Robinson et al. 2017).

This workflow involves aligning noisy PacBio raw reads ('subreads') to the *C. elegans* reference genome, followed by *de novo* assembly of aligned reads in 60-kb tiles. The

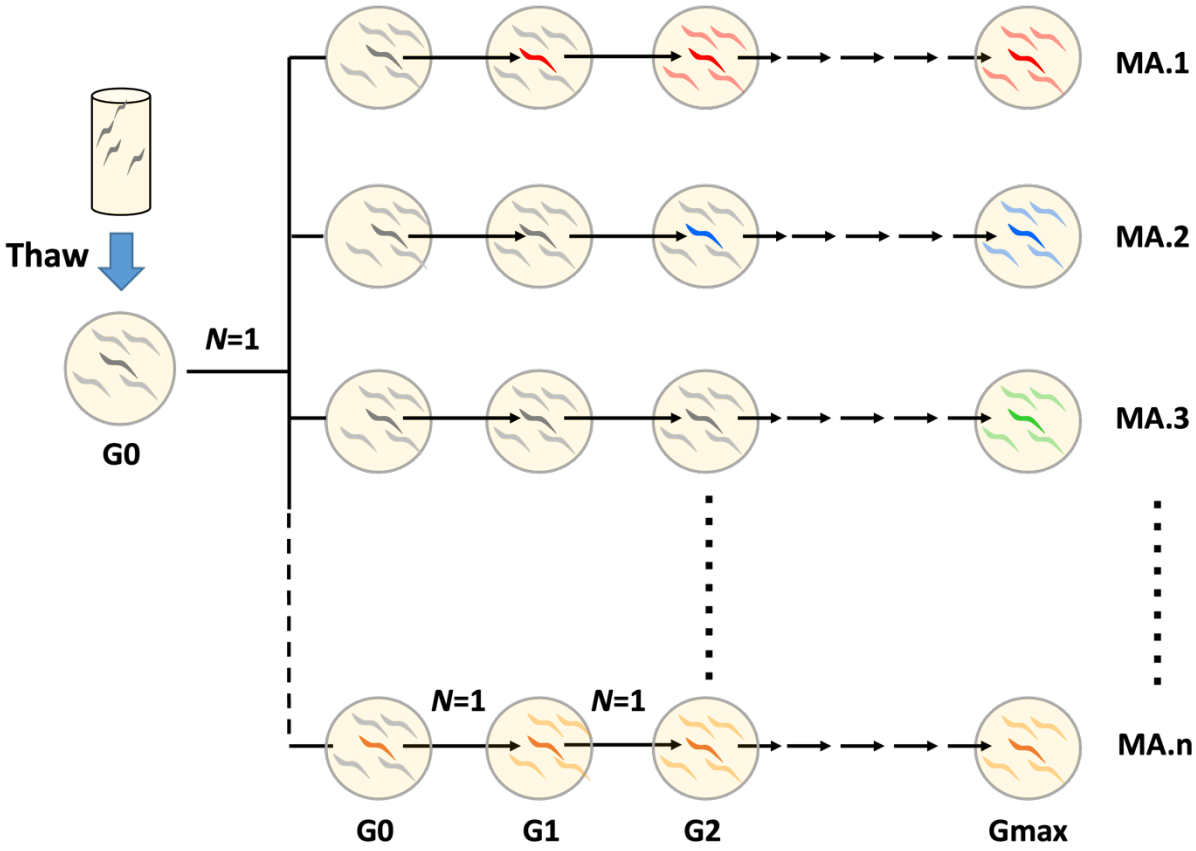


Figure 1. Schematic diagram of the mutation accumulation (MA) experiment. A cryopreserved sample of the ancestor (N2 or PB306) was thawed onto an agar plate (G0, left) and grown for a few days until the cryopreserved worms had reached the L4 larval stage. L4-stage worms were transferred singly to new NGM agar plates; those were the founders of the MA lines (MA.1, MA.2,...MA.n; n=100). At four-day intervals (one generation), a single immature worm was picked at random onto a new plate. This procedure was repeated until the experiment reached Gmax=250 generations. Individual MA lines experienced fewer than Gmax generations, G(ave)=238 generations. The different colored worms represent (the set of) unique mutations fixed in each MA line, which contribute to the among-line (mutational) variance.

assembled contigs —henceforth referred to as “pseudo-reads”— are then re-aligned to the reference genome, followed by conventional variant calling using PBSV. The pseudo-reads exhibit higher sequence quality than raw PacBio subreads, and aligned pseudo-reads are free from alignment errors that may occur with subread alignment. Accurate genotype calling in both the MA progenitor and the mutation accumulation (MA) line is required for

spontaneous SV calls, which increases the error rate of a spontaneous SV calling workflow as opposed to a conventional SV calling workflow. Consequently, each final mutation call is manually verified using IGV.

The summary of raw PacBio data is presented in **Supplementary Table 1**, which includes the raw read error rate as well as the error rate of the assembly-generated pseudo-reads. The error rate of raw reads, calculated from all N2-derived lines, is 14%, while the pseudo-reads exhibit a significantly reduced error rate of 2.2%. The figures for other lines are also available in **Supplementary Table 1**. Error rates were calculated after alignment using the 'NM' tag in the BAM file, which represents the Levenshtein distance between the aligned segment of the read and the reference. Notably, pseudo-reads also exhibit fewer unaligned segments, measured as soft or hard clips after alignment. The percentage of clipped bases upon alignment is approximately 5.3% for raw reads, compared to 4.3% for pseudo-reads. To minimize the effect of reference bias on error rate estimates, we only report results from N2-derived lines, from which the reference genome is derived, thereby providing more reliable estimates. The gains to alignment quality are even higher for wild isolates and PB306-derived MA lines.

Raw sequence data generated in this study are archived in the NCBI Short Read Archive - PRJNA1233366. We also utilized publicly available long-read data of *C. elegans* wild isolates from the following SRA repositories (PRJNA523481, and PRJNA1025857).

SV mutation rate

The SV mutation rate among observed SVs (i.e., not accounting for false negatives), is estimated to be ~7 SVs per MA line, over approximately 238 generations per MA line, on average. In comparison, we identified an average of 62 single nucleotide variants (SNV) and 25 short indel mutations per line in the same set of lines. The inferred SV mutation rate is roughly 10% of the SNV rate and ~30% of the short indel rate, indicating that SVs comprise about 8% of new mutations, or roughly one new SV mutation per-genome every ~30 generations. Summary statistics of SV mutations are presented in **Table 1**. The full set

Table 1. Summary statistics. Column headings: *Line* (strain.MA line # or wild isolate ID); *Type* (MA or Wild isolate); *#Muts* (number of inferred structural variants > 30bp in size); *Man. Corr* (# of SVs that required manual correction; see Methods for explanation); *#False pos.* (# of false positives); *%Recall (100 bp, 1Kb, 10Kb)* (% of simulated SVs of that length recalled; see Methods for explanation); *Ave. Cov*. (average sequencing coverage); *Raw error rate* (Error rate of raw PacBio subreads calculated after alignment to reference); *Pseudo-reads Error rate* (Error rate of assembled contigs after alignment to reference)*; #gens MA* (number of MA generations; see Methods); *Rel. fit (%)* (Fitness of MA lines relative to ancestor N2; unpublished data from Saxena et al.); #SNVs (number of base-substitution mutations); #Indels (number of indel mutations ≤ 30 bp). "Mutations" in wild isolates are variant compared to the N2 reference genome. For a more detailed discussion of false positives in wild isolates, see section 2 of Supplementary Discussion.

Line	Type	#Muts	Man. Corr	# False pos.	Recall rate (100-bp)	Recall rate (1 kb)	Recall rate (10 kb)	Ave. Cov.	Raw Error Rate (%)	Pseudo-read Error Rate (%)	#gens MA	Rel. Fit (%)	# SNVs	# Indels
MA517	N2 - MA	5	3	35	49	53	29	17	14	1%	242	72.6	53	17
MA530	N2 - MA	7	2	1	65	77	47	56	13	3%	243	122.2	51	26
MA563	N2 - MA	5	1	0	63	73	43	33	15	3%	237	66.0	70	24
MA566	N2 - MA	8	4	6	62	73	45	37	13	2%	238	58.6	71	21
MA445	PB306 - MA	10	5	6	57	68	42	50	16	6%	232	126.8	76	30
MA459	PB306 - MA	5	2	4	55	62	42	79	16	6%	235	61.4	61	31
PB306	WI	7691	-	-	59	69	43	188	15	7%	-	-	189278	73876
CB4856	WI	11028	-	-	56	65	37	59	18	7%	-	-	285787	104591
QX1211	WI	16528	-	-	50	49	30	54	18	8%	-	-	505065	164826
JU1088	WI	6181	-	-	60	69	39	52	17	6%	-	-	151734	61013
N2	Baer - N2	-	-	-	62	76	54	83	15	2%	-	-	-	-

of SVs and their attributes are given in **Supplementary Tables 2 and 3** for PB306 derived lines and N2 derived lines respectively.

Estimation of False Discovery Rates

(*a*) False positives. The accuracy of SV mutation rate estimates is contingent upon the accuracy of the variant calling workflow. False positives can be identified only by either (*i*) "by eye" validation of putative variants using Integrative Genomics Viewer (IGV), or (*ii*) by using a different sequencing technique, ideally by PCR amplification of the putative mutant locus followed by Sanger sequencing of the resulting amplicon. However, those approaches become impractical as the number of samples and/or putative variants increases.

We scrutinized each putative variant in the MA lines by visualizing the aligned pseudo-reads, raw PacBio subreads, and Illumina reads from a separate study in IGV. We identified 52 false positives in the original set of called putative SVs, resulting in a signal (inferred mutant, n=40) to noise (false positive, n=52) ratio of 40/52 = 0.77. However, 2/3 of the false positives occurred in one MA line (N2 line MA517). MA517 had lower sequencing coverage than the other lines (17X vs 51X for other MA lines), which potentially introduced alignment and/or assembly artifacts. When line MA517 is excluded, the signal-to-noise ratio increases to ~2:1.

Among the true positive calls, minor corrections were made after visualizing the SVs in IGV. Corrections are required when an MA SV occurs at a locus that already carries an ancestral SV. Including other minor reasons for adjustments, corrections were necessary for 17 of the 40 true positive calls. The reasons for corrections are provided in **Supplementary Tables 2 and 3**.

(*b*) False negatives. In contrast to identification of false positives, which must be done by brute force, false negatives can be inferred by means of bioinformatics. We estimated the false negative rate by (*i*) simulating SVs ("pseudo-variants") in the reference genome and (*ii*) attempting to identify ("recall") the pseudo-variant in the sequenced sample. For example, a pseudo-deletion introduced in the reference genome should appear as a variant pseudo-

insertion in the sequenced sample. We measured recall in both the MA lines and wild isolates, with the latter providing information about the extent to which reference bias impacts false negative rates.

Recall rates exceed 90% for short (~100-bp) indels and decline as SV length increases, with deletions exhibiting higher recall rates than insertions (**Figure 2** and **Supplementary Table S4**). Notably, recall rates were lower in wild isolates compared to the N2 strain, presumably because the *C. elegans* reference genome is N2. As a result, slightly

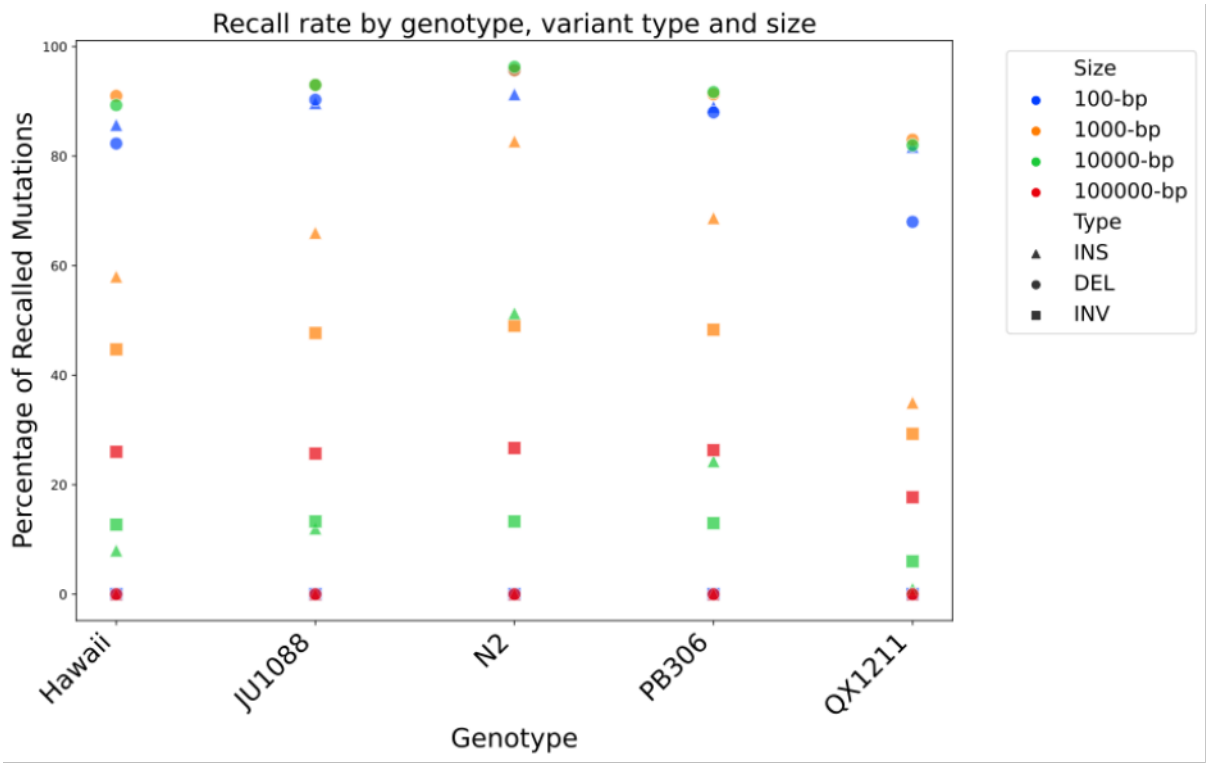


Figure 2. Recall rates by genotype, SV type, and SV size. N2 has the highest recall rate across SV types and sizes. Small deletions (100 bp) have the highest recall rates, comparable to 1-kb and 10-kb deletions. Recall rates of insertions drop sharply beyond 100 bp. Recall rate decreases with increasing SV size and genetic divergence from N2, a trend also observed in PB306-derived MA lines (not shown). Inversions have the lowest recall rates. QX1211, the most divergent genotype from N2, is the most challenging for SV detection.

more SVs were missed in the PB306 MA lines than in the N2 MA lines (~4%). For reasons that are unclear, we miss almost all small inversions (100-bp), whereas we are able to recall roughly a quarter of 100-kb inversions.

SV mutation spectrum

Frequencies of SVs are listed by type and average size in **Table 2**. The total number of bases affected is approximately 16 kb in the four N2-derived lines, with a net loss of 1.5 kb across four lines (7 insertions and 18 deletions). For the two PB306-derived lines, approximately 15 kb of bases were affected in total, with a net gain of 7 kb (5 insertions and 10 deletions). Deletions outnumber insertions 2:1, the same bias observed among smaller indels in a larger set of N2 and PB306 MA lines (Rajaei et al. 2021). Among wild isolates, the number of insertions and deletions are nearly equal, with large insertions outnumbering large deletions by roughly 10% among SVs, and insertions and deletions being nearly equal among smaller indels. This suggests stronger purifying selection against deletions, on average, as opposed to insertions in the wild.

Table 2. SV spectrum by variant type and number of bases affected.

Line	Type	Insertions	Deletions	Inversions	Net bases affected	Net base gain (loss)
N2.517	MA	2	3	0	5,342	184
N2.530	MA	3	4	0	1,509	843
N2.563	MA	2	3	0	7,606	(309)
N2.566	MA	0	8	0	2,166	(2,166)
PB.445	MA	3	7	0	3,501	(1,397)
PB.459	MA	2	3	0	12,184	8,536
PB306	WI	4347	3280	64	4,027,284	1,791,544
CB4856	WI	5887	5056	85	5,205,431	1,562,847
QX1211	WI	9148	7279	101	7,949,707	2,849,281
JU1088	WI	3597	2525	59	3,331,096	1,465,676

Spontaneous SV mutations are substantially larger in size compared to SVs identified in wild isolates (mean SV size: MA=1039 bp, wild=602 bp; median SV size: MA=404 bp, wild=125 bp; Kolmogorov-Smirnov test, $p<0.006$). This difference is primarily driven by the size of insertions (mean insertion size: MA=1574 bp, wild=799 bp; median insertion size: MA=840 bp, wild=162; K-S test, $p<0.004$), whereas deletions show no

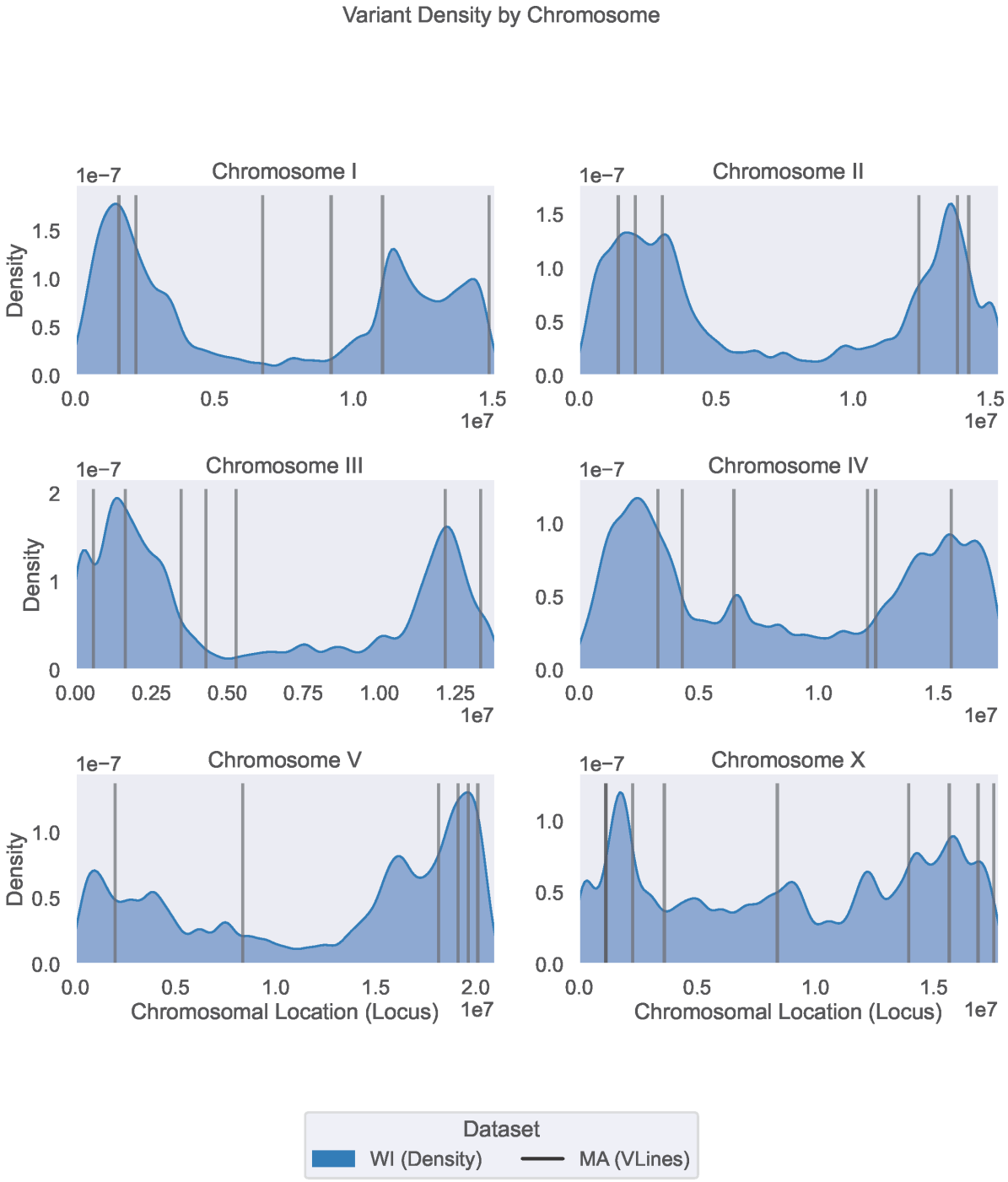
significant difference in size (mean deletion size: MA=538 bp, wild=415 bp; median deletion size: MA=131 bp, wild=105 bp; K-S test, p>0.21). Although the difference in deletion sizes is not statistically significant, we observe a heavier tail in the size distribution of spontaneous deletions, similar to spontaneous insertions. This distribution exhibits a multimodal pattern, indicative of larger deletions in the MA lines (**Supplementary Figures S1-S3**).

A slight majority of spontaneous SVs (22/40) result from simple repeat expansions or deletions, while four of the 40 SVs are associated with transposon-related events. The average size of the repeat family across 22 independent events is 93 bp. In the PB306-derived MA line 445, a CER1 transposon insertion was identified on chromosome 1, resulting in an 8.8-kb insertion within a short 80-bp intron of the *smg-1* gene. The inserted sequence contains 3' end of the *plg-1* gene and the full CER1 transposon sequence. Similarly, in the N2-derived MA line 566, an RTE transposon insertion was observed, leading to a 3.2-kb insertion at the 5' end of the *alg-1* gene. Additionally, seven of the 40 SVs are located within hyperdivergent genomic regions, consistent with their frequency in the *C. elegans* genome (~20%; Lee et al. 2021).

Distribution of SVs across the genome

The distribution of spontaneous SV mutations among exonic, intronic, and intergenic regions does not deviate significantly from expectations based on genomic composition (Fisher Exact test; chi squared statistic=1.89, df=2, sample size=140, p=0.39). Thus, the distribution of spontaneous SVs reflects the relative proportion of the genome associated with those regions.

In the wild isolates, the distribution of segregating SVs across the genome mirrors the base-substitution distribution – higher in arms and lower in the centers (**Figure 3**); see Figure 3 of Andersen et al. (2012) for the distribution of base-substitutions along *C. elegans* chromosomes. The recombination rate is higher in chromosomal arms (of autosomes) than centers in *C. elegans* (Rockman and Kruglyak 2009). The greater genetic diversity in chromosomal arms is consistent with reduced Hill-Robertson interference in the arms



265

266 **Figure 3.** Distribution of SVs within and among chromosomes in wild isolates (density
267 shown in blue) and MA lines (vertical black lines). The overall distributions between these
268 two groups are not significantly different (Kolmogorov-Smirnov test; See text for details).

269

compared to the chromosome centers. We do not see this pattern on the X chromosome, which is also consistent with expectation, since the recombination rate is nearly uniform on the X.

It is possible, however, that the chromosomal distribution of genetic diversity is not solely due to variation in the efficacy of purifying selection. It is possible that the SV mutation rate may be higher in the arms. This view is supported by the fact that low complexity DNA is overrepresented in the arms of *C. elegans* chromosomes (see Supplementary Figures S4 and S5). Even with the small sample size of this study, we observe more spontaneous SVs in the arms of the chromosomes than in the centers. A K-S test revealed no significant deviation between the genomic distribution of spontaneous SVs and those in the wild (p-values range from 0.29 to 0.92 across the six chromosomes). It may be that the distribution of SVs across the genome can be explained by a combination of mutation (more in the arms than the centers) and selection (stronger Hill-Robertson interference in the centers).

Fitness effects of SVs

To infer the fitness effects of SVs, we compared the spectrum of spontaneous SVs to that of wild isolates. Multiple aspects of the spectrum can be characterized – size, type, location, etc. For example, spontaneous SVs are, on average, larger in size compared to segregating SVs, aligning with the expectation that larger SVs are more deleterious and are thus removed by natural selection. As noted, the deletion-to-insertion ratio is higher among spontaneous SVs as opposed to segregating SVs, implying that deletions are more deleterious than insertions, on average. Functionally, the most important aspect of SV annotation is based on their genomic location (e.g., exon, intron, intergenic). We analyzed nucleotide diversity across different SV categories in MA lines and wild isolates, as well as among shorter

296 variants such as missense mutations and small indels, as shown in **Table 3.** SNVs in
297 introns, intergenic regions, and synonymous SNVs in exons were classified as
298 "neutral" base-substitution for this analysis. After normalizing for mutation rate, we
299 find that SV diversity in exons is approximately 14% of that of (putatively) neutral
300

301 **Table 3. Distribution of variant types by genomic context**. See text for details of
302 calculations of diversity measures.

Type	Context	Wild Isolates	PB306 MA	N2 MA	All MA	dN / dS*
SV	Exon	2866	4	6	10	0.14
	Intron	11818	5	14	19	0.31
	Intergenic	11292	6	5	11	0.52
SNV	Splice Variant	2840	2	1	3	0.48
	Disruptive Start/Stop Codon	1830	2	2	4	0.23
	Missense Variant	78191	11	30	41	0.96
	Intergenic Region	243014	53	88	141	Neutral
	Intron Variant	274542	52	94	146	Neutral
	Synonymous Variant	82326	6	8	14	Neutral
Indel	Frameshift Variant	8446	4	5	9	0.47
	In-frame Indel	2905	0	4	4	0.36
	Intergenic Region	74279	20	26	46	0.81
	Intron Variant	100183	32	45	77	0.65

303

304 base-substitutions, consistent with strong purifying selection (see Methods for details of the
305 calculation). This analysis is analogous to the dN/dS ratio as a test for selective constraint,
306 with the denominator (dS*) expanded to increase the sample size of MA variants. In
307 contrast to SVs in exons, the SNV diversity of missense mutations is ~96% of the neutral
308 expectation. This unexpectedly high frequency could be due to the sampling variance
309 associated with the small number of spontaneous mutations. Interestingly, SV diversity in
310 intergenic regions is ~52% of the neutral expectation, suggesting moderately strong
311 selection against SVs in these regions. This observation is, to our knowledge, the first of its
312 kind and warrants further investigation into the functional impacts of intergenic SVs, which

may have significant implications for human genetics research. We discuss the implications of this result to the concept of 'junk DNA' later.

Discussion

Limitations in Estimating the Contribution of Structural Variants to Fitness Decline

Over the course of an average of 238 generations of mutation accumulation (MA), the six MA lines in this study exhibited an average decline in absolute fitness of ~15% (Baer et al. 2006). Structural variants (SVs) constitute only ~7% of all detected variants. However, accurately quantifying their contribution to fitness decline is challenging without detailed knowledge of the distribution of fitness effects (DFE), which is likely influenced by factors such as variant size and genomic context.

For instance, SVs occurring in exons represent a mutation class with particularly low diversity in natural populations. After adjusting for differences in mutation rates across variant classes, SV diversity in exons is approximately 14% of the neutral expectation, which implies that selection removes 86 out of every 100 SV mutations that arise in an exon. Even that is an underestimate, however, given that the standing variants potentially include relatively new deleterious mutations that have persisted long enough to be observed but are on their way to being selected out of the population (i.e. mutation-selection balance).

In contrast, small indels in introns, the most abundant non-neutral mutation class in our dataset, exhibit diversity at ~65% of the neutral expectation, implying that selection removes 35 out of every 100 introduced mutations in this category. Despite this, the MA lines contain approximately eight times as many small indels in introns than SVs in exons. While it is difficult to estimate relative selection coefficients (s) for these two classes without a well-characterized DFE, their relative abundance suggests that small indels in introns could contribute substantially to fitness decline, even if their average effect is weaker than that of SVs in exons. Ignoring variant frequency in favor of a focus on average effects risks underestimating the evolutionary consequences of more frequent but potentially less deleterious mutations, such as small indels.

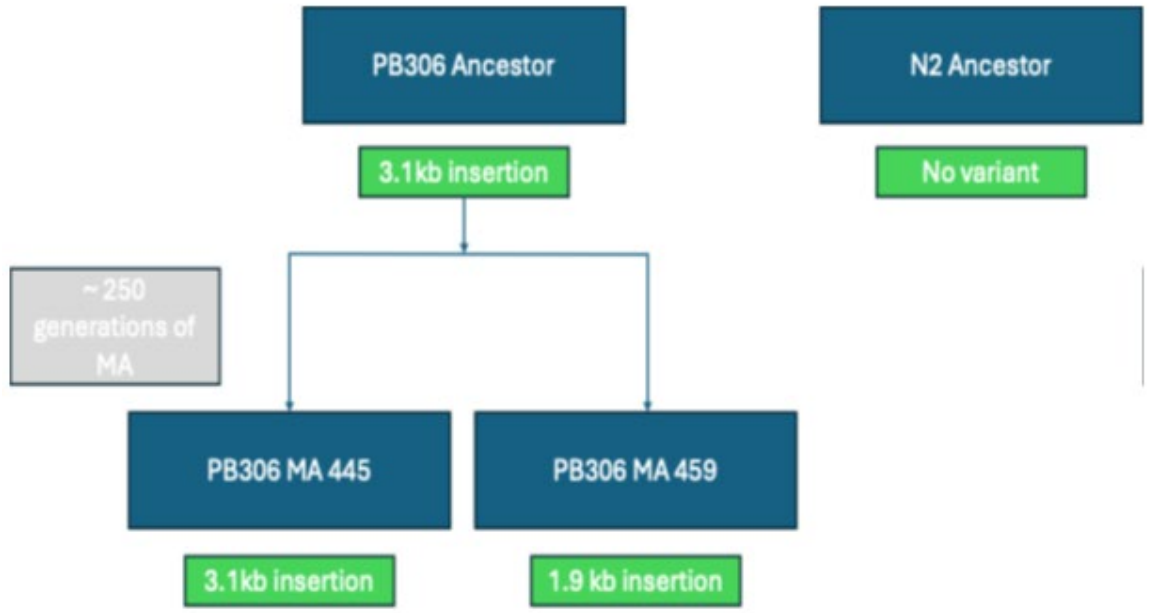
Structural Variants in Low-Complexity Mutational Hotspots

A significant proportion (~22.5%; 9 out of 40) of structural variants (SVs) identified in mutation accumulation (MA) lines occur in low-complexity mutational hotspots. Previously, Konrad et al. (2018) reported a similar estimate (~31%) for duplications and a much higher estimate (~81%) for deletions occurring in loci with pre-existing CNVs in the ancestor, based on short read sequencing of N2-dervied MA lines in *C. elegans*. Among the SV calls in MA lines, the majority (22 out of 40) are simple repeat expansions or deletions. These annotations were determined through manual inspection of variants using the JBrowse genome browser as well as IGV rather than automated annotation, which, while scalable, may introduce noise. The repeating units range in size from as small as dinucleotide repeats (e.g., AT, CA, CT) to larger structural events, such as the deletion of a single copy within a four-copy 360-mer repeat.

An unexpected observation is that 9 of the 22 SVs identified in repeat-rich regions required a 'manual correction' due to a common misclassification in the workflow – where the MA variant call occurs on a locus with a pre-existing variant in the ancestor. This suggests that these loci represent mutational hotspots. Notably, the variant caller PBSV did not correctly handle these cases, even in joint genotyping mode. **Supplementary Tables S2 and S3** provide annotations for SV calls in MA lines, detailing the necessary corrections following manual verification using IGV and genome browser inspection.

For example, a variant call on chromosome I of PB306-derived MA line MA445 (I: 1567493), illustrated in **Figure 4a** with a pedigree and **4b** with an IGV snapshot highlights this issue. In the PB306 ancestor, this region contains a 3.1 kb insertion composed of perfectly repeating copies of a 173-mer interspersed with additional sequence. Some of this additional sequence contains similar repeating motifs, but of smaller units, and precise annotation is difficult due to sequencing errors. The sequence of this insertion, along with its repeating unit, is provided in **Supplementary Sequence File S1** in FASTA format. We show the alignment of raw PacBio reads in **Supplementary Figure S6** and show the

(a)



(b)

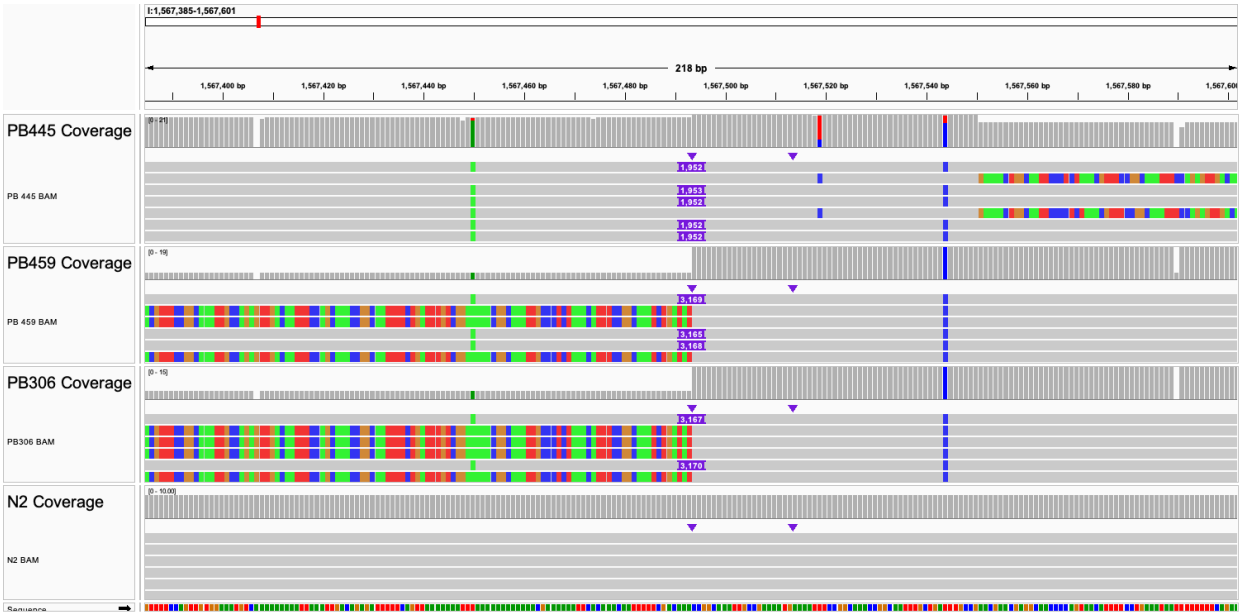


Figure 4. (**a**) A schematic representation of the variants identified at the locus (I: 1567493) in the pedigree of MA lines. PBSV initially classified the 1.9 kb insertion and the 3.1 kb insertion as two independent mutations. However, a deeper analysis—accounting for sequencing errors—reveals that a 1.2 kb segment was deleted within the inserted sequence in the ancestor, leading to the observed 1.9 kb insertion. (**b**) IGV snapshot of the 1.2 kb deletion in PB306-derived MA 445. The IGV snapshot shows pseudo-read alignments, which are assembled contigs representing longer sequences. Some pseudo-reads fail to fully capture the

inserted segment and appear as soft-clipped sequences. If the inserted segment is not entirely covered within the pseudo-read, it is clipped from the alignment, appearing as base calls (colored bars) instead of clean (gray bars), fully aligned sequences. Some insertion sequences are misaligned because of the repetitive nature of the locus; however, all pseudo-reads carry the insertion. Supplementary Figure S6 presents the same analysis using error-prone PacBio raw reads.

repeating copies of 173-mer highlighted in text in **Supplementary Figure S7**. The same 3.1 kb insertion is also present in the PB306-derived MA line MA459, consistent with the expectation that a variant present in the ancestor is inherited by its descendant. This insertion is absent in the N2 ancestor, indicating that it is polymorphic in *C. elegans*. Interestingly, in MA445, this region harbors a 1.9 kb insertion, effectively representing a 1.2 kb deletion relative to the PB306 ancestor. However, variant calling software failed to annotate this as a deletion event, necessitating extensive manual validation of both raw and assembled reads. Because some repeat copies are perfectly conserved while others exhibit minor variations, the identification of this type of event remains beyond the resolution of current variant calling algorithms.

The need for manual correction in 9 out of 22 repeat-associated SVs raises an important question: Are certain repeat regions disproportionately prone to structural variation? The *C. elegans* genome contains a vast number of repeat-associated sequences, with 116,921 entries for repeat regions in the WS270 GFF3 annotations. Our findings suggest that loci already carrying an SV in the ancestor are prone to recurrent SVs, making them significant hotspots for genetic variation. Furthermore, these regions are likely underrepresented or misannotated by automated variant calling pipelines. With only 9 such identified events, it is beyond the scope of this study to understand the factors that cause some repeat sequences to be hotspots. Notably, 2 of the 9 manually corrected repeat-associated SVs occur in hyperdivergent regions—an observation worth noting, though we refrain from drawing broader conclusions.

SV Mutation Rate and Genome Size

Genome size increased in five of the six MA lines assayed, by an average of ~1000 bp over ~238 generations (~4bp/generation). Comparison of the estimate of the increase in genome size by mutation with variation in genome size estimated from the wild isolates allows us to estimate the strength of selection acting on genome size. If genome size is considered to be a quantitative trait under selection (directional or stabilizing), at mutation-selection balance (MSB), the ratio of the mutational variance (V_M) to the standing genetic variance (V_G) is approximately equal to the strength of selection acting on a mutation that affects the trait (Barton 1990; Kondrashov and Turelli 1992). Alternatively, if the trait of interest is a neutral trait at mutation drift equilibrium (MDE), in a predominantly selfing population (such as *C. elegans*) $V_G \approx 4N_eV_M$, where N_e is the effective population size (Lynch and Hill 1986; Denver et al. 2005). The ratio of V_G to V_M for genome size is ~27000, which implies a selection coefficient $s \approx 3.7 \times 10^{-5}$. If we assume the demarcation of effective neutrality is $4N_es<1$ (Kimura 1962; Kondrashov et al. 2006) and that N_e of *C. elegans* is on the order of 10^5 (Gilbert et al. 2021), it is plausible that genome size is under weak, but effective selection, with the obvious caveat that any measure of variance estimated from a sample of fewer than ten genomes is not robust.

Estimating the Fraction of the Genome under Evolutionary Constraint

We observe that selection efficiently removes most SVs, regardless of their genomic locations (genic versus intergenic). This finding challenges the conventional view of "junk DNA," which is often assumed to evolve free from the constraints of purifying selection. Our results suggest that a larger fraction of the genome may be functionally relevant than previously inferred from studies focusing on base substitutions. Unconstrained genomic regions are typically defined by the absence of purifying selection. However, it is possible that some genomic regions can tolerate base substitutions but not SVs. For instance, consider a hypothetical low-complexity region between an enhancer and a promoter (Bateman and Johnson 2022). Base substitutions in such regions may not disrupt gene

function, whereas SVs that alter the distance between the enhancer and promoter might, which could in turn negatively impact fitness. Similarly, regulatory sequences flanking genes may be more tolerant of base substitutions than of small indels or larger SVs. In a similar vein, Payer et al. (2017) found that polymorphic *Alu* elements in the human genome were significantly associated with Trait Associated SNPs (TASs) in human GWAS. Larger-scale studies, capable of identifying more SV mutations, could reveal many such regions that appear to evolve freely based on SNP variation but are subject to purifying selection against SVs.

Implications for GWAS

Our findings have significant implications for genome-wide association studies (GWAS), which typically rely on exome sequencing data and the analysis of short variants such as single-nucleotide variants (SNVs) and small indels. While this approach has been highly successful in identifying genetic variants associated with numerous traits and diseases, our results highlight a potential blind spot in the detection and interpretation of structural variants (SVs), particularly those in noncoding regions such as intergenic areas. Specifically, we observed that SVs in exons are under strong purifying selection, with diversity reduced to 14% of the expected value after accounting for mutation rates. This underscores the functional importance of exonic SVs and aligns with the focus on exome sequencing in GWAS.

However, our finding that intergenic SV diversity is reduced to 52% of the expected value suggests that many of these variants are also functionally significant, likely due to their impact on regulatory elements or chromatin structure. This is in contrast to the relatively weaker selection observed for missense SNVs, which retain 96% of their expected diversity. The data reveal a striking contrast in the selective pressures acting on different classes of mutations. Variants that strongly disrupt gene function—such as splice site mutations, frameshifts, and structural variants in coding regions—exhibit significantly reduced diversity in the wild, suggesting that purifying selection effectively removes them from the population.

In contrast, missense mutations, which alter protein structure but do not necessarily abolish function, persist at much higher frequencies. The accumulation of mildly deleterious mutations could also contribute, in the long run, to compensatory adaptation via the introduction of variants of small positive effects.

These observations raise critical questions about the extent to which current GWAS approaches may overlook the contributions of intergenic SVs to phenotypic variation and disease. Expanding the scope of GWAS to include comprehensive SV profiling, particularly in noncoding regions, may uncover a substantial reservoir of previously unrecognized genetic contributors to human health and disease.

Conclusion

The spontaneous SV mutation rate is inferred to be approximately 8% of the SNV/small indel rate, but that is likely an underestimate. SVs are efficiently removed from natural populations, indirectly implying a function (biochemical or otherwise) to a larger fraction of the genome than is typically assumed.

Material and Methods

MA Protocol

The MA protocol has been described in detail previously (Baer et al. 2005) and is depicted in **Figure 1**. The method by which we estimated the number of generations of MA for a line (below Gmax) is reported in Saxena et al. (2019).

DNA extraction, library preparation, and sequencing

Cryopreserved samples of the MA lines were thawed onto 100 mm NGMA agar plates seeded with a lawn of *E. coli* OP50 and grown for 2-3 days, until there were many gravid worms, at which time worms were harvested, "bleached" to remove bacterial and fungal contamination (Stiernagle 2006), and embryos transferred to a new, seeded 100 mm plate. L1 larvae were collected the next day and transferred to seven seeded 100 mm plates and

grown until the plates were nearly starved. Worms were washed from the plates, double-rinsed in M9 buffer, and pelleted for DNA extraction. DNA was extracted from each sample with the Quiagen Gentra Puregene kit, following the manufacturer's instructions. Extracted DNA samples were sent to the University of Maryland genomics core facility for library preparation and sequencing on the Pacific Biosciences Sequel platform.

Variant Calling Pipeline

We evaluated multiple variant calling strategies, including direct alignment-based approaches and de novo assembly followed by genome-to-genome alignment using Assemblytics (Nattestad and Schatz 2016). Ultimately, we adopted a modified version of the SMRT-SV2 (Audano et al. 2019) pipeline, as it provided a balance between sensitivity and specificity while minimizing false positives. A detailed discussion of our rationale for selecting this approach over others is provided in section 1 of the **Supplementary Discussion**.

The SMRT-SV2 pipeline processes noisy PacBio subreads by aligning them to the reference genome using blasr (Chaisson and Tesler 2012). These alignments are then segmented into 60-kb tiles, which are assembled using Canu (Koren et al. 2017). This localized assembly approach reduces interference from reads originating from other genomic regions. Importantly, the assembled contigs effectively "clean" the PacBio subreads, leading to improved downstream alignments. While SMRT-SV2 automatically re-aligns these contigs using blasr, we opted to replace blasr with minimap2 (Li 2018), which provided superior alignment performance for our dataset.

SMRT-SV2 includes its own variant caller. However, it led to an unacceptably high false positive rate for de novo mutations. To improve accuracy, we benchmarked several alternative structural variant (SV) callers, including Sniffles (Sedlazeck et al. 2018) and PBSV (Wenger et al. 2019), ultimately selecting PBSV due to its optimized performance on high-quality PacBio reads. Further details on the performance of SMRT-SV2, Sniffles, and SMRT-SV2 variant calling are provided in section 3 of the **Supplementary Discussion**.

In cases where a potential variant was detected, SMRT-SV2 produced multiple assemblies—referred to as "pseudo-reads"—for the candidate locus. These pseudo-reads were then aligned with minimap2 and variants were called using PBSV. Despite this refinement, we still observed many false positives, albeit a manageable number, further necessitating manual curation. Each candidate variant was inspected in IGV (Robinson et al. 2017) and JBrowse genome browser (Diesh et al. 2023), comparing error-prone raw PacBio subread alignments, pseudo-read alignments, and independent Illumina sequencing data generated for a separate study. Only variants present in the mutation accumulation (MA) line but absent in both the MA ancestor and other MA lines were considered true de novo mutations. Once validated, variants were functionally annotated using genome browser annotations to classify their genomic context (e.g., exon, intron, UTR, repetitive element, hyperdivergent region, potential regulatory elements).

For identifying segregating SVs, we followed the same pipeline with one key difference: instead of an ancestor-descendant comparison, all wild isolates were jointly called using PBSV.

Variant Annotation

Annotation of structural variants (SVs) identified in the mutation accumulation (MA) lines was performed manually. Basic genomic classifications—such as exon, intron, and intergenic regions—were straightforward to assign. Similarly, regulatory annotations, including DNase hypersensitive sites, transcription factor binding sites, and hyperdivergent regions, could be systematically determined using existing datasets. However, annotating repetitive elements within SVs was significantly more time-consuming. Many SVs overlapped repetitive sequences that were not explicitly annotated by RepeatMasker, requiring manual inspection of the variant and read sequences to deduce repeat structure.

In contrast, annotation of SVs in wild isolates was automated using GFF3-based genome annotations. Variant calls were intersected with genomic features using *bedtools* (Quinlan and Hall 2010), allowing for systematic classification. Automated annotation tools

549 such as *snpEff* (Cingolani et al. 2012) were not suitable for SV annotation, as they do not
550 properly account for the full extent of large structural variants. For example, an 8-kb deletion
551 may only be annotated at its breakpoints rather than across its entire span. To ensure
552 accurate functional classification, we assigned the highest-impact annotation observed
553 across the deleted region. If a deletion started upstream of a gene but extended into coding
554 sequences, it was annotated based on the most functionally significant feature it disrupted
555 (e.g., an exon rather than an upstream region).
556 Short variants (SNPs and small indels) in both MA and wild isolates were annotated
557 using *snpEff* with the following flags: -no-downstream -no-upstream -no-utr. As a result,
558 upstream, downstream, and UTR annotations were instead categorized as intergenic,
559 maintaining consistency with our SV annotation approach. *snpEff* produces a diverse set of
560 annotations, many of which are redundant or overly specific for our purposes. To
561 standardize annotation categories, we applied a systematic simplification approach using a
562 predefined mapping (**Supplementary Table 5**). This mapping condensed detailed
563 annotation labels into a manageable number of biologically meaningful categories, ensuring
564 consistency across variant types.
565
566 *Estimating the relative strength of selection for different categories of mutations*
567 To estimate the relative strength of selection, analogous to the dN/dS analysis, we first
568 defined a neutral class of variants. This neutral class comprises intergenic SNPs, intronic
569 SNPs, and synonymous substitutions within exons. We included intronic and intergenic
570 variants alongside synonymous variants to increase the sample size of MA variants in the
571 neutral class, given the limited number of synonymous MA variants in our dataset (only 14
572 observed). To reflect the expanded definition of the neutral class — which includes
573 intergenic and intronic variants alongside synonymous mutations — we refer to our
574 approach as the dN/dS* analysis.
575 To calculate the relative strength of selection for a particular class of mutation (e.g.,
576 structural variants in exons), we used the following formula:

$$dN\ /\ dS = \frac{\left(\frac{\#\ segregating\ variants\ in\ the\ focal\ class}{\#\ segregating\ variants\ in\ the\ neutral\ class}\right)}{\left(\frac{\#\ spontaneous\ mutations\ in\ the\ focal\ class}{\#\ spontaneous\ mutations\ in\ the\ neutral\ class}\right)}$$

This approach quantifies the relative diversity within a particular class of mutation against the diversity in neutral classes after accounting for the mutation rates in both categories (data obtained from spontaneous variants). Mutation counts for each category in the wild isolate and MA populations are provided in **Table 3**.

Simulation of Reference Genomes and Estimation of False Negative Rates

To evaluate the sensitivity of our variant calling pipeline, we simulated structural variants in the *C. elegans* reference genome (WS270) using the R package *RSVSim* (Bartenhagen and Dugas 2013). For each variant type—insertions, deletions, and inversions—we generated three independent simulated genomes. Within each variant type, we introduced 100 variants of lengths 100 bp, 1,000 bp, and 10,000 bp, as well as 50 variants of 100,000 bp.

Following simulation, we applied our variant calling pipeline to assess recall rates. Introducing structural variants into the reference genome shifts the positions of subsequent variants, making direct locus-based comparisons dependent on tracking the precise order of variant introduction. While this could be addressed if the software explicitly recorded variant placement order, doing so would introduce additional dependencies into the analysis. To avoid this complexity, we instead used a size-based evaluation approach, considering a variant correctly detected if its length fell within predefined tolerance ranges: (98–102) bp, (995–1,005) bp, (9,995–10,005) bp, and (99,995–100,005) bp for their respective size categories. This approach simplified the evaluation while providing a systematic estimate of false negative rates.

Data Analysis

All statistical analyses (e.g., Kolmogorov-Smirnov test, Fisher's exact test) were performed in Python using SciPy (Virtanen et al. 2020). Data visualization was conducted using Matplotlib (Hunter 2007) and Seaborn (Waskom 2021).

Acknowledgments

We thank Tim Crombie, Joanna Dembek, Lindsay Johnson, and Mike Snyder for assistance in the lab, and Erik Andersen and Annalise Paaby for providing the wild isolate data. This study was supported by the NIH award GM10722.

Declaration of interests

The authors declare no competing interests. A.S. contributed to this article in his personal capacity as a graduate student at the University of Florida, and the views expressed do not necessarily represent the views of Regeneron Pharmaceuticals Inc.

Data availability

Raw sequence data generated in this study are archived in the NCBI Short Read Archive - PRJNA1233366. We also utilized publicly available data from C. elegans wild isolates from the following SRA repositories (PRJNA523481, and PRJNA1025857). All code used for structural variant (SV) calling, variant annotation, and manuscript figure generation is available at https://github.com/Baer-Lab/c_elegans_spontaneous_sv.

Figure Legends

Figure 1. Schematic diagram of the mutation accumulation (MA) experiment. A cryopreserved sample of the ancestor (N2 or PB306) was thawed onto an agar plate (G0, left) and grown for a few days until the cryopreserved worms had reached the L4 larval stage. L4-stage worms were transferred singly to new NGM agar plates; those were the founders of the MA lines (MA.1, MA.2,...MA.n; n=100). At four-day intervals (one generation), a single immature worm was picked at random onto a new plate. This

631 procedure was repeated until the experiment reached Gmax=250 generations. Individual
632 MA lines experienced fewer than Gmax generations, G(ave)=238 generations. The different
633 colored worms represent (the set of) unique mutations fixed in each MA line, which
634 contribute to the among-line (mutational) variance.
635
636 **Figure 2.** Recall rates by genotype, SV type, and SV size. N2 has the highest recall rate
637 across SV types and sizes. Small deletions (100 bp) have the highest recall rates,
638 comparable to 1-kb and 10-kb deletions. Recall rates of insertions drop sharply beyond 100
639 bp. Recall rate decreases with increasing SV size and genetic divergence from N2, a trend
640 also observed in PB306-derived MA lines (not shown). Inversions have the lowest recall
641 rates. QX1211, the most divergent genotype from N2, is the most challenging for SV
642 detection.
643
644 **Figure 3.** Distribution of SVs within and among chromosomes in wild isolates (density shown
645 in blue) and MA lines (vertical black lines). The overall distributions between these two
646 groups are not significantly different (Kolmogorov-Smirnov test; See text for details).
647
648 **Figure 4.** (Top) A schematic representation of the variants identified at the locus (I:
649 1567493) in the pedigree of MA lines. PBSV initially classified the 1.9 kb insertion and the
650 3.1 kb insertion as two independent mutations. However, a deeper analysis—accounting for
651 sequencing errors—reveals that a 1.2 kb segment was deleted within the inserted sequence
652 in the ancestor, leading to the observed 1.9 kb insertion.
653 (Bottom) IGV snapshot of the 1.2 kb deletion in PB306-derived MA 445. The IGV snapshot
654 shows pseudo-read alignments, which are assembled contigs representing longer
655 sequences. Some pseudo-reads fail to fully capture the inserted segment and appear as
656 soft-clipped sequences. If the inserted segment is not entirely covered within the pseudo-
657 read, it is clipped from the alignment, appearing as base calls (colored bars) instead of clean
658 (gray bars), fully aligned sequences. Some insertion sequences are misaligned because of

659 the repetitive nature of the locus; however, all pseudo-reads carry the
660 insertion. **Supplementary Figure S6** presents the same analysis using error-prone PacBio
661 raw reads.
662

References

Andersen EC, Gerke JP, Shapiro JA, Crissman JR, Ghosh R, Bloom JS, Felix MA, Kruglyak L. 2012. Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. Nature Genetics 44:285-U283.

Audano PA, Sulovari A, Graves-Lindsay TA, et al. 2019. Characterizing the major structural variant alleles of the human genome. Cell 176:663-+.

Baer CF, Phillips N, Ostrow D, Avalos A, Blanton D, Boggs A, Keller T, Levy L, Mezerhane E. 2006. Cumulative effects of spontaneous mutations for fitness in Caenorhabditis: Role of genotype, environment and stress. Genetics 174:1387-1395.

Baer CF, Shaw F, Steding C, et al. 2005. Comparative evolutionary genetics of spontaneous mutations affecting fitness in rhabditid nematodes. Proceedings of the National Academy of Sciences of the United States of America 102:5785-5790.

Bartenhagen C, Dugas M. 2013. RSVSim: an R/Bioconductor package for the simulation of structural variations. Bioinformatics 29:1679-1681.

Barton NH. 1990. Pleiotropic models of quantitative variation. Genetics 124:773-782.

Bateman JR, Johnson JE. 2022. Altering enhancer-promoter linear distance impacts promoter competition in cis and in trans. Genetics 222.

Beyter D, Ingimundardottir H, Oddsson A, et al. 2021. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nature Genetics 53:779-+.

Böndel KB, Kraemer SA, Samuels T, McClean D, Lachapelle J, Ness RW, Colegrave N, Keightley PD. 2019. Inferring the distribution of fitness effects of spontaneous mutations in *Chlamydomonas reinhardtii*. PLoS Biology 17:e3000192.

Boyle EA, Li YI, Pritchard JK. 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169:1177-1186.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. Bmc Bioinformatics 13:238.

Chaisson MJP, Wilson RK, Eichler EE. 2015. APPLICATIONS OF NEXT-GENERATION SEQUENCING Genetic variation and the de novo assembly of human genomes. Nature Reviews Genetics 16:627-640.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu XY, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. Fly 6:80-92.

Collins RL, Brand H, Karczewski KJ, et al. 2020. A structural variation reference for medical and population genetics. Nature 581:444-+.

Conrad DF, Pinto D, Redon R, et al. 2010. Origins and functional impact of copy number variation in the human genome. Nature 464:704-712.

Crombie TA, Rajaei M, Saxena AS, Johnson LM, Saber S, Tanny RE, Ponciano JM, Andersen EC, Zhou J, Baer CF. 2024. Direct inference of the distribution of fitness effects of spontaneous mutations from recombinant inbred *Caenorhabditis elegans* mutation accumulation lines. Genetics 228.

Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, Thomas WK. 2005. The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. Nature Genetics 37:544-548.

Diesh C, Stevens GJ, Xie P, et al. 2023. JBrowse 2: a modular genome browser with views of synteny and structural variation. Genome Biology 24:74.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nature Reviews Genetics 8:610-618.

Fisher RA. 1922. On the Dominance Ratio. Proceedings of the Royal Society of Edinburgh 42:321-341.

Friedman LS, Ostermeyer EA, Szabo CI, Dowd P, Lynch ED, Rowell SE, King MC. 1994. CONFIRMATION OF BRCA1 LAY ANALYSIS OF GERMLINE MUTATIONS LINKED TO BREAST AND OVARIAN-CANCER IN 10 FAMILIES. Nature Genetics 8:399-404.

Gilbert KJ, Zdraljevic S, Cook DE, Cutter AD, Andersen EC, Baer CF. 2021. The distribution of mutational effects on fitness in *Caenorhabditis elegans* inferred from standing genetic variation. Genetics 220.

Harvey WT, Ebert P, Ebler J, et al. 2023. Whole-genome long-read sequencing downsampling and its effect on variant-calling precision and recall. Genome Research 33:2029-2040.

Hénault M, Marsit S, Charron G, Landry CR. 2024. The genomic landscape of transposable elements in yeast hybrids is shaped by structural variation and genotype-specific modulation of transposition rate. Elife 12.

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering 9:90-95.

Hurles ME, Dermitzakis ET, Tyler-Smith C. 2008. The functional impact of structural variation in humans. Trends in Genetics 24:238-245.

James J, Kastally C, Budde KB, González-Martínez SC, Milesi P, Pyhäjärvi T, Lascoux M. 2023. Between but not within species variation in the Distribution of Fitness Effects. Mol Biol Evol 13.

Karayiorgou M, Simon TJ, Gogos JA. 2010. 22q11.2 microdeletions: linking DNA structural variation to brain dysfunction and schizophrenia. Nature Reviews Neuroscience 11:402-416.

Kim BY, Huber CD, Lohmueller KE. 2017. Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples. Genetics 206:345-361.

Kimura M. 1962. On probability of fixation of mutant genes in a population. Genetics 47:713-719.

Kondrashov AS, Turelli M. 1992. Deleterious mutations, apparent stabilizing selection and the maintenance of quantitative variation. Genetics 132:603-618.

Kondrashov FA, Ogurtsov AY, Kondrashov AS. 2006. Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. Journal of Theoretical Biology 240:616-626.

Konrad A, Flibotte S, Taylor J, Waterston RH, Moerman DG, Bergthorsson U, Katju V. 2018. Mutational and transcriptional landscape of spontaneous gene duplications and deletions in *Caenorhabditis elegans*. Proceedings of the National Academy of Sciences 115:7386-7391.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research 27:722-736.

Kousathanas A, Keightley PD. 2013. A comparison of models to Infer the distribution of fitness effects of new mutations. Genetics 193:1197-1208.

Lee D, Zdraljevic S, Stevens L, et al. 2021. Balancing selection maintains hyper-divergent haplotypes in Caenorhabditis elegans. Nat Ecol Evol 5:794-807.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094-3100.

López-Cortegano E, Chebib J, Jonas A, Vock A, Künzel S, Keightley PD, Tautz D. 2025. The rate and spectrum of new mutations in mice inferred by long-read sequencing. Genome Res 35:43-54.

Lopez-Cortegano E, Craig RJ, Chebib J, Balogun EJ, Keightley PD. 2023. Rates and spectra of de novo structural mutations in Chlamydomonas reinhardtii. Genome Research 33:45-60.

Lynch M, Hill WG. 1986. Phenotypic evolution by neutral mutation. Evolution 40:915-935.

Marshall CR, Noor A, Vincent JB, et al. 2008. Structural variation of chromosomes in autism spectrum disorder. American Journal of Human Genetics 82:477-488.

Meyerson M, Pellman D. 2011. Cancer Genomes Evolve by Pulverizing Single Chromosomes. Cell 144:9-10.

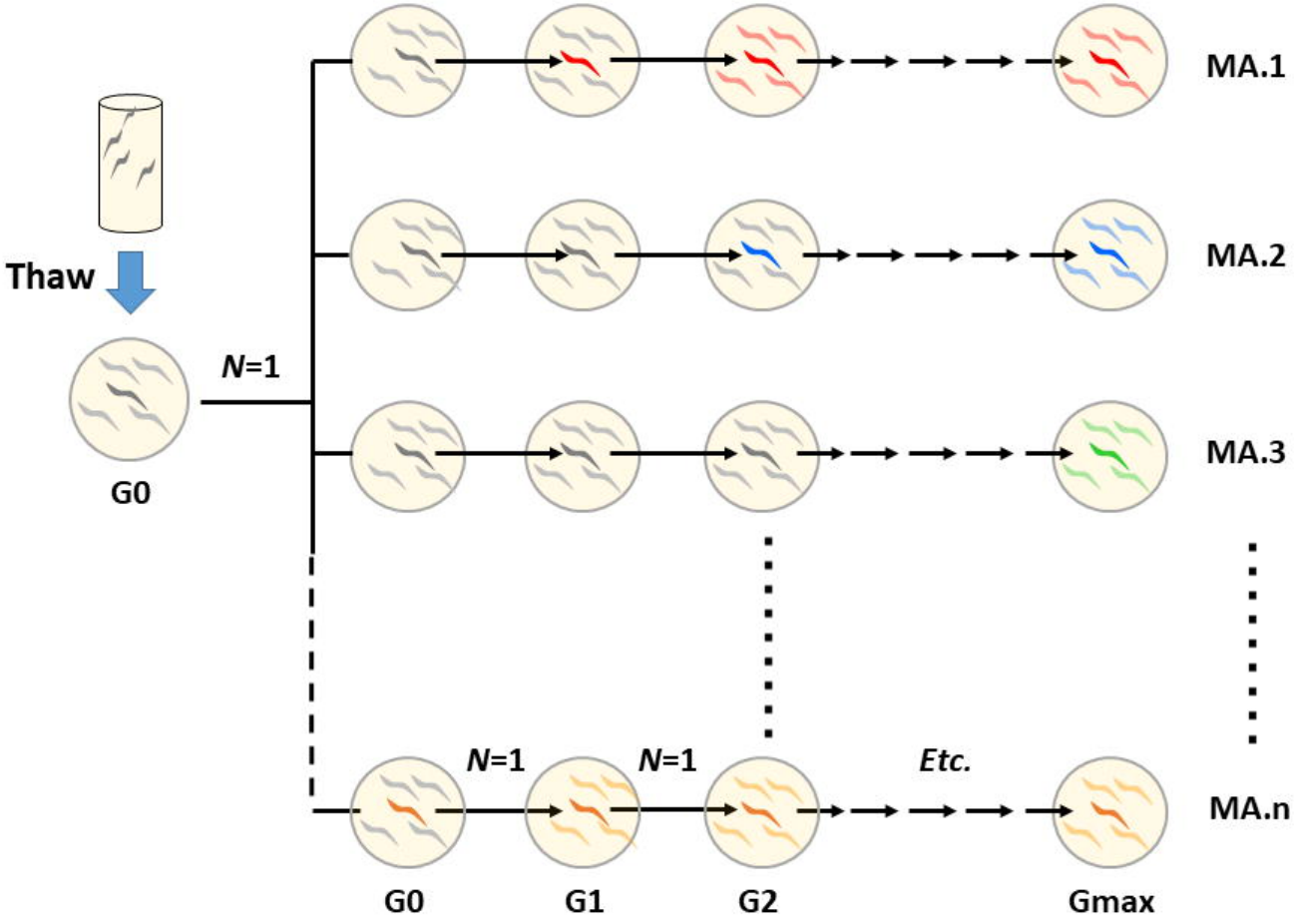
Nattestad M, Goodwin S, Ng K, et al. 2018. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. Genome Research 28:1126-1135.

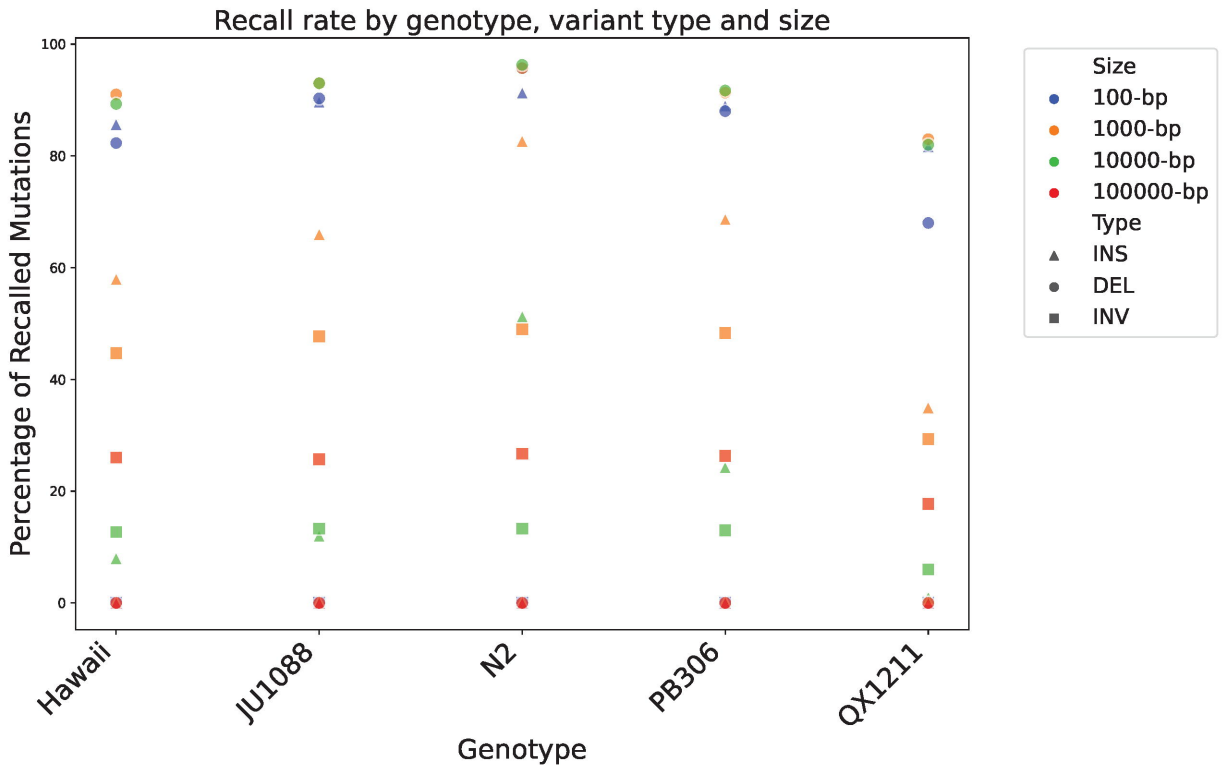
764 Nattestad M, Schatz MC. 2016. Assemblytics: a web analytics tool for the detection of variants from
765 an assembly. Bioinformatics 32:3021-3023.
766 Noyes MD, Harvey WT, Porubsky D, et al. 2022. Familial long-read sequencing increases yield of
767 de novo mutations. American Journal of Human Genetics 109:631-646.
768 Nurk S, Koren S, Rhie A, et al. 2022. The complete sequence of a human genome. Science 376:44-+.
769 Payer LM, Steranka JP, Yang WR, Kryatova M, Medabalimi S, Ardeljan D, Liu CH, Boeke JD,
770 Avramopoulos D, Burns KH. 2017. Structural variants caused by Alu insertions are associated
771 with risks for many human diseases. Proceedings of the National Academy of Sciences of the
772 United States of America 114:E3984-E3992.
773 Pinto DPagnamenta ATKlei L, et al. 2010. Functional impact of global rare copy number variation in
774 autism spectrum disorders. Nature 466:368-372.
775 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
776 Bioinformatics 26:841-842.
777 Rajaei M, Saxena AS, Johnson LM, Snyder MC, Crombie TA, Tanny RE, Andersen EC, Joyner-Matos J,
778 Baer CF. 2021. Mutability of mononucleotide repeats, not oxidative stress, explains the
779 discrepancy between laboratory-accumulated mutations and the natural allele-frequency
780 spectrum in *C. elegans*. Genome Research 31:1602-1613.
781 Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP. 2017. Variant Review with the
782 Integrative Genomics Viewer. Cancer Research 77:e31-e34.
783 Rockman MV, Kruglyak L. 2009. Recombinational Landscape and Population Genomics of
784 Caenorhabditis elegans. Plos Genetics 5.
785 Sanders SJ, Ercan-Sencicek AG, Hus V, et al. 2011. Multiple Recurrent De Novo CNVs, Including
786 Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism.
787 Neuron 70:863-885.
788 Saxena AS, Salomon MP, Matsuba C, Yeh SD, Baer CF. 2019. Evolution of the mutational process
789 under relaxed selection in *Caenorhabditis elegans*. Molecular Biology and Evolution 36:239-
790 251.
791 Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. 2018.
792 Accurate detection of complex structural variations using single-molecule sequencing.
793 Nature Methods 15:461-468.
794 Stiernagle T. 2006. Maintenance of *C. elegans*. In: T*Ce*R Community, editor. WormBook.
795 Stransky N, Cerami E, Schalm S, Kim JL, Lengauer C. 2014. The landscape of kinase fusions in cancer.
796 Nature Communications 5.
797 Tataru P, Mollion M, Glemin S, Bataillon T. 2017. Inference of Distribution of Fitness Effects and
798 Proportion of Adaptive Substitutions from Polymorphism Data. Genetics 207:1103-1119.
799 Virtanen PGommers ROliphant TE, et al. 2020. SciPy 1.0: fundamental algorithms for scientific
800 computing in Python. Nature Methods 17:261-272.
801 Walsh T, McClellan JM, McCarthy SE, et al. 2008. Rare structural variants disrupt multiple genes in
802 neurodevelopmental pathways in schizophrenia. Science 320:539-543.
803 Wang QL, Dhindsa RS, Carss K, et al. 2021. Rare variant contribution to human disease in 281,104 UK
804 Biobank exomes. Nature 597:527-+.
805 Waskom ML. 2021. seaborn: statistical data visualization. Journal of Open Source Software 6:3021.
806 Weiss LA, Shen YP, Korn JM, et al. 2008. Association between microdeletion and microduplication at
807 16p11.2 and autism. New England Journal of Medicine 358:667-675.
808 Wenger AM, Peluso P, Rowell WJ, et al. 2019. Accurate circular consensus long-read sequencing
809 improves variant detection and assembly of a human genome. Nature Biotechnology
810 37:1155-1162.
811 Wooster R, Bignell G, Lancaster J, et al. 1995. IDENTIFICATION OF THE BREAST-CANCER
812 SUSCEPTIBILITY GENE BRCA2. Nature 378:789-792.
813 Yi K, Ju YS. 2018. Patterns and mechanisms of structural variations in human cancer. Experimental
814 and Molecular Medicine 50.

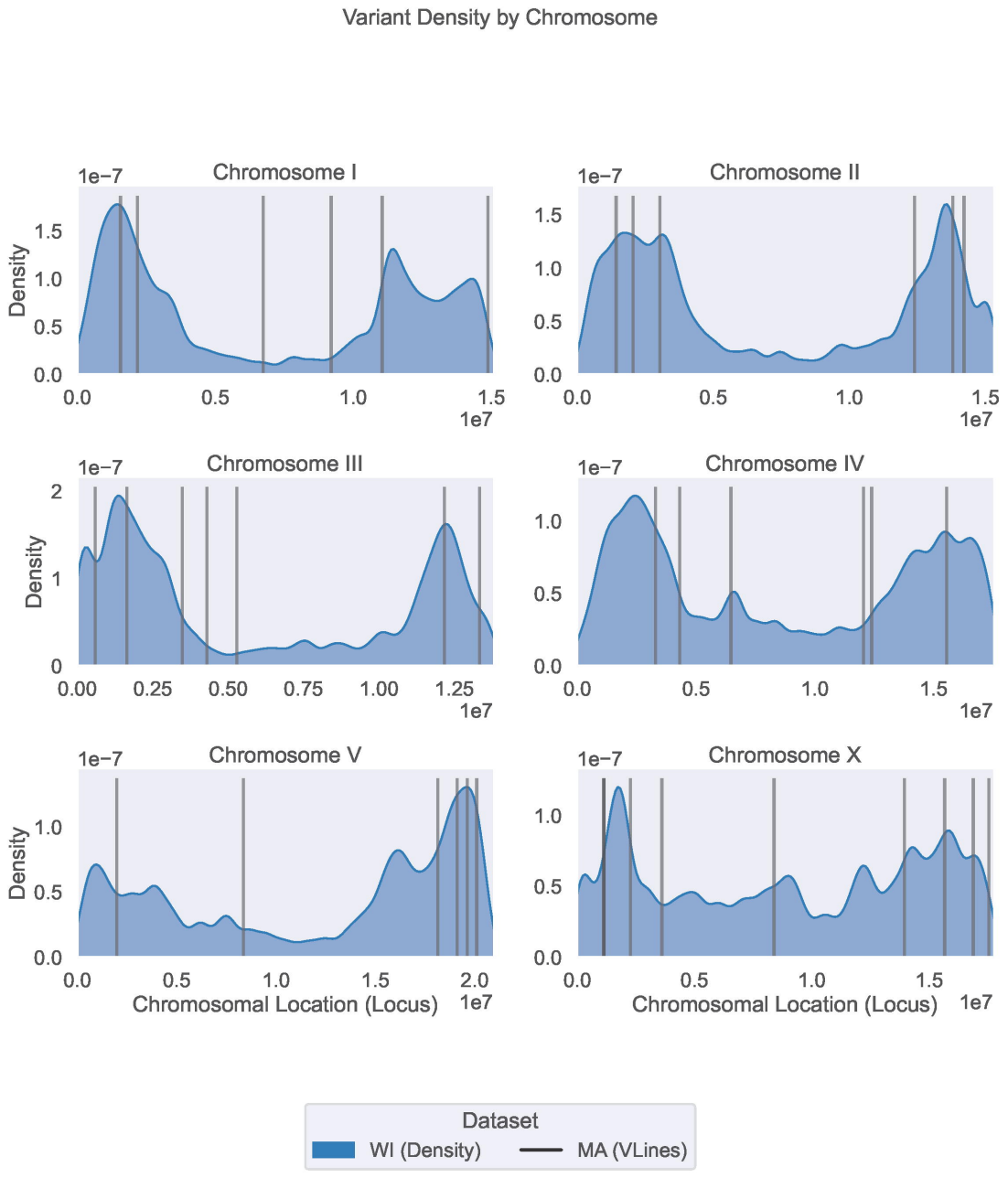
815 Zhou X, Pan J, Wang YH, Lynch M, Long HA, Zhang Y. 2023. De Novo structural variations of
816 *Escherichia coli* detected by Nanopore long-read sequencing. Genome Biology and Evolution
817 15.

818

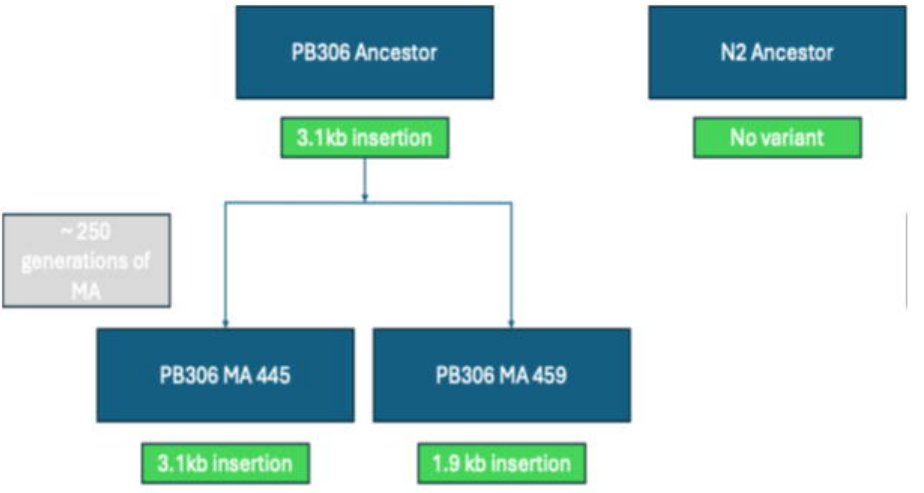
819


Thaw
N=1
G0
MA.1
MA.2
MA.3
MA.n
N=1
N=1
Etc.
G0
G1
G2
Gmax


Recall rate by genotype, variant type and size
Percentage of Recalled Mutations
100
80
60
40
20
0
Hawaii
JU1088
N2
PB306
QX1211
Genotype
Size
100-bp
1000-bp
10000-bp
100000-bp
Type
INS
DEL
INV


Variant Density by Chromosome
Chromosome I
Chromosome II
Chromosome III
Chromosome IV
Chromosome V
Chromosome X
Density
1e−7
1e7
Chromosomal Location (Locus)
Dataset
WI (Density)
MA (VLines)

(a)


PB306 Ancestor
N2 Ancestor
3.1kb insertion
No variant
~250 generations of MA
PB306 MA 445
PB306 MA 459
3.1kb insertion
1.9 kb insertion

(b)


PB445 Coverage
PB459 Coverage
PB306 Coverage
N2 Coverage