Mutational Landscape of Spontaneous Base Substitutions and Small Indels in Experimental *Caenorhabditis elegans* Populations of Differing Size

Anke Konrad, Meghan J. Brady, Ulfar Bergthorsson, and Vaishali Katju¹

Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, Texas 77845

ORCID IDs: 0000-0003-3994-460X (A.K.); 0000-0003-1419-1349 (U.B.); 0000-0003-4720-9007 (V.K.)

ABSTRACT Experimental investigations into the rates and fitness effects of spontaneous mutations are fundamental to our understanding of the evolutionary process. To gain insights into the molecular and fitness consequences of spontaneous mutations, we conducted a mutation accumulation (MA) experiment at varying population sizes in the nematode *Caenorhabditis elegans*, evolving 35 lines in parallel for 409 generations at three population sizes ($N = 1, 10$, and 100 individuals). Here, we focus on nuclear SNPs and small insertion/deletions (indels) under minimal influence of selection, as well as their accrual rates in larger populations under greater selection efficacy. The spontaneous rates of base substitutions and small indels are 1.84 (95% C.I. ± 0.14) $\times 10^{-9}$ substitutions and 6.84 (95% C.I. ± 0.97) $\times 10^{-10}$ changes/site/generation, respectively. Small indels exhibit a deletion bias with deletions exceeding insertions by threefold. Notably, there was no correlation between the frequency of base substitutions, nonsynonymous substitutions, or small indels with population size. These results contrast with our previous analysis of mitochondrial DNA mutations and nuclear copy-number changes in these MA lines, and suggest that nuclear base substitutions and small indels are under less stringent purifying selection compared to the former mutational classes. A transition bias was observed in exons as was a near universal base substitution bias toward A/T. Strongly context-dependent base substitutions, where 5'-Ts and 3'-As increase the frequency of A/T $\rightarrow$ T/A transversions, especially at the boundaries of A or T homopolymeric runs, manifest as higher mutation rates in (i) introns and intergenic regions relative to exons, (ii) chromosomal cores vs. arms and tips, and (iii) germline-expressed genes.

KEYWORDS *Caenorhabditis elegans*; mutation accumulation line; base substitution; small indel; selection; genetic drift

SPONTANEOUS mutation is central to our understanding of the evolutionary process, given its role as the preeminent source of genetic variation. A detailed understanding of the rate and spectrum of spontaneous mutations is critical for the interpretation of genetic variation in natural populations, the evolutionary dynamics of mutations under the forces of natural selection and genetic drift, the limits to adaptation, the nature of complex human disease and cancer, and the genetic and phenotypic consequences of maintaining populations at small sizes, among others. Because natural variation is the result of an interplay between mutations, genetic drift, and

natural selection, having a realistic hypothesis for genetic variation in the absence of selection is essential. Furthermore, features of the genome such as base composition can be shaped by prevailing mutational biases and, in turn, the base composition itself can influence mutation rates (Smith *et al.* 2002; Krasovec *et al.* 2017). Moreover, mutation rates themselves are not uniformly distributed across genes in the genome. In addition to base composition, variables such as age, replication timing, chromatin organization, and gene expression have been suggested to influence the mutation rate (Hodgkinson and Eyre-Walker 2011).

Mutation accumulation (MA) experiments have had a rich history in evolutionary biology since the late 1960s, having provided us with a relatively unbiased view of the mutation process by enabling the study of newly originated mutations with minimal interference from the eradicated influence of purifying selection. Replicate lines descended from a single ancestral genotype are evolved independently under extreme

Copyright © 2019 by the Genetics Society of America

doi: <https://doi.org/10.1534/genetics.119.302054>

Manuscript received February 25, 2019; accepted for publication May 16, 2019; published Early Online May 20, 2019.

Supplemental material available at Figshare: <https://doi.org/10.25386/genetics.8120783>.

¹Corresponding author: Department of Veterinary Integrative Biosciences, Texas A&M University, 4458 TAMU, College Station, TX 77843-4458. E-mail: vkajtu@cvm.tamu.edu

bottlenecks each generation to diminish the efficacy of selection, thereby promoting evolutionary divergence due to the accumulation of mutations by random genetic drift. This experimental evolution design of MA experiments circumvents the challenges associated with studying newly arisen mutations in natural or wild populations where strong selection may purge the very mutational variants of interest [reviewed in Halligan and Keightley (2009) and Katju and Bergthorsson (2019)].

MA experiments typically maintain all replicate lines at the same minimal population size. A variation of this theme, comparing the rates and properties of mutations between MA lines maintained at different population sizes, enables one to manipulate the efficacy of selection as a function of population size. In our spontaneous *Caenorhabditis elegans* MA experiment, all MA lines descended from a single N2 hermaphrodite ancestor, were bottlenecked each generation at $N = 1$, 10, or 100 hermaphrodites for > 400 generations. This experimental design permits a simultaneous investigation of the effects of spontaneous mutation and selection on genetic variation, as well as indirect inferences of the fitness consequences of different classes of mutations. Here, we employ the same set of spontaneous *C. elegans* MA lines comprising three population-size treatments (Katju *et al.* 2015, 2018; Konrad *et al.* 2017, 2018), and leverage this experimental framework with high-throughput sequencing to identify *de novo* nuclear base substitutions and small insertion/deletions (indels) on a genome-wide scale since the divergence of the MA lines from their common ancestor. With the completion of this study, we are able to: (i) offer a comprehensive view of the spontaneous mutation process in *C. elegans*, across both the organellar and nuclear genomes, and all major classes of mutations (base substitutions, small indels, and copy-number variations); (ii) compare our spontaneous mutation rates for nuclear SNPs to previously generated rates that employed older sequencing technologies; (iii) provide one of the first direct, genome-wide estimates of the spontaneous small indel rate for a nematode; and (iv) investigate selective constraints that may impinge on nuclear base substitutions and small indels.

Materials and Methods

MA experiment

As a self-fertilizing nematode with a generation time of 3.5 days at 20°, and the ability to survive long-term cryogenic storage, *C. elegans* is an ideal organism for MA studies. The spontaneous MA experiment was initiated with a single wild-type Bristol (N2) hermaphrodite originally isolated as a virgin L4 larva. The F1 hermaphrodite descendants of this single worm were further inbred by self-fertilization, before establishing 35 MA lines and cryogenically preserving thousands of excess animals at -86° for use as ancestral controls. Twenty of these 35 lines were established with a single worm and propagated at $N = 1$ individual per generation. Ten lines were initiated with 10 randomly chosen L4 hermaphrodite

larvae and subsequently bottlenecked each generation at $N = 10$. Five lines were initiated and subsequently maintained each generation with 100 randomly chosen L4 hermaphrodite larvae ($N = 100$). A new generation was established every 4 days. The $N = 1$, 10, and 100 population-size treatments corresponded to effective population sizes (N_e) of 1, 5, and 50, respectively (Katju *et al.* 2015, 2018). The worms were cultured using standard techniques with maintenance at 20° on NGM agar in (i) 60 × 15 mm Petri dishes seeded with 250 µl suspension of *Escherichia coli* strain OP50 in YT media ($N = 1$ and $N = 10$ lines) or (ii) 90 × 15 mm Petri dishes seeded with 750 µl suspension of *E. coli* strain OP50 in YT media ($N = 100$ lines). Stocks of the MA lines were cryogenically preserved at -86° every 50 generations. The experiment was terminated following 409 MA generations because the $N = 1$ lines displayed a highly significant fitness decline. Three lines were already extinct due to the accumulation of a significant mutation load and five additional lines were on the verge of extinction (displaying great difficulty in generation-to-generation propagation) (Katju *et al.* 2015).

DNA preparation and sequencing

Following the completion of the MA phase, a total of 86 worms were prepared for DNA whole-genome sequencing: one worm from every population of size $N = 1$, four individuals from every population of size $N = 10$, five individuals from every population of size $N = 100$, and one individual from the ancestral strain. Each of the 86 individuals were allowed to go through several self-fertilization and reproductive cycles to generate enough offspring necessary for genomic DNA extraction. Genomic DNA extraction and library preparation for sequencing followed a previously described methodology (Konrad *et al.* 2017, 2018). The multiplexed DNA libraries were sequenced on Illumina HiSeq sequencers with default quality filters at the Northwest Genomics Center (University of Washington).

Sequence alignment and identification of putative variants

The demultiplexed raw reads stored as individual fastq files for each genome were aligned to the reference N2 genome (version WS247; www.wormbase.org; Harris *et al.* 2010) via the Burrows-Wheeler Aligner (BWA Version 0.5.9) (Li and Durbin 2009) and via Phaster (Green laboratory), and prepared for analysis as previously described (Konrad *et al.* 2018). The two alignment tools utilize different approaches to the alignment of sequences to homologous regions, as well as read-splitting and realignment. By employing both tools and requiring variants to be detected in both alignments, the false-positive rate of variants that arises due to potential alignment issues is minimized. Similarly, more than one variant caller was employed for the identification of putative mutations (see below) to reduce the false-positive rate.

Seventeen lines of size $N = 1$ were included in the final analysis (1A–1H, 1K, and 1M–1T). The alignment files were used to identify all putative base substitutions and indels within the 82 individual descendants relative to the ancestral

genome. Putative substitutions and indels were identified separately for the Phaster and BWA alignments using Platypus (Rimmer *et al.* 2014), Freebayes (Garrison and Marth 2012), and a pipeline consisting of mpileup (Li *et al.* 2009), bcftools (Li 2011), vcfutils (Danecek *et al.* 2011), and custom filters written in Perl. Indel calls were based primarily on Phaster alignments, but were verified in the BWA alignments. Indelminer (Ratan *et al.* 2015) was used as an additional approach to call indels with the ancestral line as a direct reference. A minimum root-mean-square mapping quality of 30 was required for SNPs to be retained, while a mapping quality of 40 was required for indels. SNPs were required to have a minimum support of three quality reads, while indels were required to be covered by a minimum of five quality reads. Variants that occurred even with low quality or coverage in the ancestral line were removed from the analysis. Only variants supported by $\geq 80\%$ of the high-quality reads at its position were retained in the data set. Each variant had to be confirmed by at least two of the variant callers to be considered for further analysis.

Assessing the false-negative rate for variant detection

The false-negative rate for variant detection was quantified by randomly introducing 1000 synthetic mutations into the reference genome for each of five different comparisons between an altered reference genome and an MA genome (1A–1D, and 1O), yielding a synthetic mutation data set of 5000 mutations. By altering the reference genome, we maintained the inherent heterogeneity of the sequencing data, thus allowing us to evaluate each of these sites in light of potential sequencing and alignment issues. The complete raw sequencing data of each of these five lines was realigned to their new reference genomes and the variant calling pipeline was applied with filters identical to those used previously for putative mutation calling. The percentage of callable sites and recovered synthetic mutations among those 5000 sites was determined as before.

Binomial probability verification

When sequencing multiple genomes, some sites with a higher than average error rate during sequencing or due to alignment error may pass the frequency threshold due to chance in a single line. Sequencing reads containing these variants may then be present at low frequency in multiple sequenced genomes. Every variant was independently verified by calculating its binomial probability, given the number of variant calls at the same location in the genome across all other genomes sequenced. The average frequency of the variant across lines was used as the probability of any given read calling the variant by chance (P), assuming that the site does not contain a mutation, and all variants at the site are due to sequencing or alignment error. For each site containing a putative mutation, we calculated the binomial probability of the variant within a given line as $\frac{N!}{K!(N-K)!} \times P^K (1-P)^{N-K}$, where K is the number of reads containing the variant within a line and N is the total number of reads spanning the variant

site within the line. The probabilities across all lines were sorted from most significant to least significant, and a Holm–Bonferroni correction was applied to determine if the previously identified putative mutations met their critical P -value thresholds. If the P -value did not reach significance, the variant was removed from further analysis.

Independent validation of SNP and small indel variants

All substitutions and indels identified in the exons of the $N = 1$ lines were checked against the RNA-sequencing (RNA-Seq) data set previously described in Konrad *et al.* (2018). The RNA-Seq reads were realigned using STAR to allow for indel-aware alignment of these reads (Dobin *et al.* 2013). Verification of all variants was done via computational analysis of the CIGAR (Concise Idiosyncratic Gapped Alignment Report) scores in the BAM files and finalized manually using the Integrative Genomics Viewer (Thorvaldsdóttir *et al.* 2013). Of the 199 substitutions detected in the exons, 195 were verified by RNA-Seq data. The four variants that could not be validated by RNA-Seq were associated with line 1T, which went extinct at MA generation 309 (Katju *et al.* 2015, 2018). RNA for line 1T was extracted from an earlier stock cryopreserved at MA generation 305. In total, 35 indels were detected in exons, all of which were verified in an independent RNA-Seq data set.

In addition, 46 SNP and small indel variants, identified by whole-genome sequencing in the introns and intergenic regions of the 17 $N = 1$ MA lines, were randomly selected for independent confirmation via PCR and Sanger sequencing. Primers were designed to amplify regions containing candidate mutations. The locus of interest was sequenced in the candidate MA line as well as the ancestral control. PCR products were purified using a silica membrane protocol and Sanger sequenced by Eton Biosciences. Sequences were mapped to the reference genome using the Basic Local Alignment Search Tool (Altschul *et al.* 1990), and alignments were inspected to verify either the ancestral sequence or the new variant. Chromatograms were examined to ensure sequence quality. Forty-four of the 46 variants were independently validated using this approach. Two mutations in MA line 1T could not be verified. Both these mutations were initially detected within segmental duplications. This line demonstrated evidence of chromothripsis and went extinct prior to the termination of the MA experiment, which may have been a complicating factor (Konrad *et al.* 2018).

Annotation, characterization, and mutation-rate calculations for SNPs and indels

All variants were annotated based on the GFF file available for the N2 reference genome of *C. elegans* (version WS247; www.wormbase.org; Harris *et al.* 2010) using a custom script. Mutations were assigned to exons, introns, and intergenic regions (if the mutation occurred outside a protein-coding gene), and to chromosomal arms, cores, and tips based on boundaries predicted by Rockman and Kruglyak (2009). The expected distribution of variants across these

regions was estimated based on the proportion of the genome falling within each category. The mutation rate, μ_i , was estimated individually for each population as variants (or sum of variant frequencies) per base per generation ($\mu_i = \frac{M_i}{G_i \cdot B_i}$), where M_i refers to the number (or sum of frequencies) of SNPs or indels within the line, G_i refers to the number of generations through which the line was propagated, and B_i refers to the total number of bases in the genome that meet the same thresholds (base and mapping qualities, and quality read depth) required for variant identification relative to the N2 reference genome (version WS247). For populations of size $N > 1$, the sum of frequencies of variants was calculated from the proportion of individuals sequenced for each population that carried each of the variants of interest. B_i in populations of size $N > 1$ was averaged across the genomes of the individuals sequenced for that population. Mutation rates for each of the population sizes were calculated by averaging the population-specific mutation rates within each population-size treatment: $\mu_N = \frac{\sum_{i=1}^n \mu_i}{n}$, where μ_i refers to the population-specific mutation rate and n refers to the total number of populations of a given population size (N) (17, 10, and 5 for populations of size $N = 1, 10$, and 100, respectively). Every protein-coding gene was categorized as either a germline- or nongermline-expressed gene based on the data of Wang *et al.* (2009). Mutation rates for germline- and nongermline-expressed genes were calculated as above, based on the number of mutations within these genes and the total number of high-quality bases within both gene categories.

We calculated the mean amino acid radicality for the pool of amino acid replacement substitutions by first calculating a radicality score for each amino acid change. For this, we used the six biochemical classification schemes described in Sharbrough *et al.* (2018) to determine how radical any given amino acid change is. For instance, if a pair of amino acids is assigned into the same class for all six schemes, the amino acid substitution is assigned a score of 0. If only three out of the six schemes assign the amino acids into the same category, the substitution will have a score of 0.5, and if no scheme classifies the amino acids the same, the substitution will have a radicality of 1. Before the mean of the radicality scores for each substitution within a line was calculated, we normalized each score by the frequency of the variant within its population.

Normalization of mutation spectra and category-specific mutation rates (arms, cores, tips, exons, and introns, *etc.*) were calculated by dividing the raw variant counts or frequencies for each category by the number of bases in the genome belonging to each category (which met the same quality thresholds as those required for variant calling), and by generation time. Sequence complexity was calculated as previously described (Morgulis *et al.* 2006). Briefly, given a sequence (a) of length n and 64 possible triplets of {A, C, G, T}, the occurrence of each possible triplet (t) was counted across the sequence and yields $c_t(a)$. The total number of

overlapping triplets occurring in any sequence (l) equals $n - 2$. Sequence complexity [$S(a)$] was then calculated as $S(a) = \frac{\sum_{t \in R} c_t(a)(c_t(a) - 1)/2}{(l - 1)}$. All statistical tests were performed in R (R Core Development Team 2014).

Analysis of genome-wide nucleotide mutability

To test the contribution of different genomic features to the mutability of a given site, a regularized logistic regression approach was used as previously described in Ness *et al.* (2015). The training set comprised 2355 SNPs and 1,000,000 random unmutated sites across the *C. elegans* genome. Annotations (chromosomal location, functional properties, and germline expression) and genomic properties [recombination rate, GC content, sequence complexity (s), repeat sequence, and trinucleotide sequence context (upstream and downstream bases, as well as the focal base)] for each of the 1,002,355 sites ($\sim 1\%$ of the genome) were used as potential predictors for mutability in the logistic regression. GC content and sequence complexity (s) were calculated across 41-bp windows (20 bp upstream and downstream of the focal site) (Morgulis *et al.* 2006). Recombination rates were assigned as per Rockman and Kruglyak (2009). All categorical predictors (chromosome, functional category, and trinucleotide context) were converted to a series of binary predictors referring to each category level. Recombination rate, GC content, and sequence complexity were treated as numeric predictors, while germline expression (Wang *et al.* 2009) and repeat sequence were binary predictors.

A generalized linear model fit was performed in R using the GLMnet package (v1.9-8), which implements penalized maximum likelihood via ridge and/or lasso regression, thus yielding more precise fits for models containing intercorrelated predictor variables (Friedman *et al.* 2010). A binary response variable (SNP = 1; random site = 0) was used. The regularization parameter λ , which determines the penalty against large correlations between coefficients, was set to a value of 5.73×10^{-5} . This value for λ ("lambda.min") minimizes cross-validated error via the built-in cross-validation function, and is provided by the model-fitting step. The coefficients used were retrieved using an α of 0.01. Changes in α did not change the fit of the model. Odds ratios (ORs) for each coefficient (c) were calculated as $OR = e^c$.

The mutability of each site in the genome was estimated as the relative probability of bearing a mutation using the *predict* function of GLMnet, and given the genomic properties at each site as well as the estimated model coefficients. Mutability at a site is affected by the proportion of mutated sites over all sites used for the training data set. Hence, the relative mutability values are of interest here. Given the 2355 SNP sites and 1,000,000 unmutated sites, the mean predicted mutability is expected to equal ~ 0.002 . Predicted mutability across the genome ranged from 1.7×10^{-4} to 0.57, but 100% of SNPs were covered between mutabilities of 0 and 0.15. Mutabilities from 0 to 0.15 were combined into bins of size

Table 1 Summary of the spontaneous rates of base substitutions and small indels under three population-size treatments

	<i>N</i> = 1	<i>N</i> = 10	<i>N</i> = 100
μ_{bs} (per site per generation) ^a	$1.84 (\pm 0.14) \times 10^{-9}$	$1.95 (\pm 0.13) \times 10^{-9}$	$1.83 (\pm 0.14) \times 10^{-9}$
μ_{indel} (per site per generation) ^b	$6.84 (\pm 0.97) \times 10^{-10}$	$9.46 (\pm 1.50) \times 10^{-10}$	$6.95 (\pm 0.71) \times 10^{-10}$
μ_{ins} (per site per generation) ^c	$1.79 (\pm 0.40) \times 10^{-10}$	$2.28 (\pm 0.72) \times 10^{-10}$	$1.90 (\pm 0.60) \times 10^{-10}$
μ_{del} (per site per generation) ^d	$5.06 (\pm 0.83) \times 10^{-10}$	$7.18 (\pm 1.93) \times 10^{-10}$	$5.05 (\pm 1.01) \times 10^{-10}$

Rate estimates for the *N* = 1 mutation accumulation lines represent the spontaneous rates of origin of the various classes of mutations with minimal influence of selection. 95% C.I.s are provided in parentheses.

^a Rate of base substitution.

^b Rate of small insertions and deletions.

^c Rate of small insertions.

^d Rate of small deletions.

0.015, and mutation rates for each bin were calculated:

$\mu = \frac{\#SNPs \text{ in bin}}{\#Sites \text{ in bin} \times \text{Average Generations}}$. Correlation coefficients and R^2 values were calculated for a linear regression of mutation rate over mutability in R (R Core Development Team 2014).

Data availability

Whole-genome sequence data from this MA experiment has been deposited under National Center for Biotechnology Information BioProject PRJNA448413. Supplemental material, comprising additional figures and data sets of all SNP and indel mutations identified from whole-genome sequencing data, is available at Figshare: <https://doi.org/10.25386/genetics.8120783>.

Results

We sequenced the genomes of 86 *C. elegans* MA lines and their N2 ancestor from a long-term MA experiment with differing population sizes (Katju *et al.* 2015, 2018; Konrad *et al.* 2017, 2018). The MA phase of the experiment lasted for 409 generations and comprised three population-size treatments, wherein a new worm generation was established with *N* = 1, 10, or 100 hermaphrodite worms (Supplemental Material, Figure S1A). For the 20 MA lines (1A–1T) maintained at population size *N* = 1 and the ancestral pre-MA N2 control, the genome of a population of worms derived from a single hermaphrodite per line was sequenced (Figure S1B). In MA lines comprising larger population sizes, the genomes of four and five individuals were sequenced per *N* = 10 (10 lines; 10A–10J) and *N* = 100 (five lines; 100A–100E) line, respectively. This sequencing design yielded 40 and 25 genomes for the *N* = 10 and *N* = 100 MA lines, respectively (Figure S1B). The average read depth was 27.3×, 15.5×, and 16.8× per individual genome within the *N* = 1, 10, and 100 population-size treatments, respectively. A total of 2355 SNPs (Table S1) and 699 small indels (1–100 bp) (Table S2) were called across all sequenced MA lines (Figure S2). Because differing efficacies of selection vs. drift were hypothesized for the three different population sizes, we analyzed the mutation rates and spectra separately for each population-size treatment.

The false-negative rate of the variant analysis pipeline was estimated by altering the reference genome with 1000 independent mutations separately for five MA lines (1A–D and 1O). Of the 5000 mutations thus introduced, 4790 (95.8%)

mutations were recovered, and 4932 (98.6%) of these positions passed quality filters to be included into the denominator of mutation rates. In total, 142 mutations were not recovered despite their positions passing raw quality filters. This amounts to an unaccounted false-negative rate of ~2.8%. Thus, the mutation rates presented below are conservatively low estimates.

Genome-wide estimate of the spontaneous base substitution rate in *C. elegans*

Single-nucleotide substitutions accounted for 1112 mutations across the *N* = 1 lines, yielding a spontaneous base substitution rate of 1.84 (95% C.I. ± 0.14) $\times 10^{-9}$ /site/generation (Figure S2 and Table 1). The per base substitution rates between the individual *N* = 1 lines ranged from 1.43×10^{-9} to 2.54×10^{-9} per generation. The variation among lines was not greater than expected by chance ($\chi^2 = 7.8 \times 10^{-10}$, d.f. = 16, and $P = 1$), and there was no correlation between mutation rate and the relative fitness of individual *N* = 1 MA lines ($r = -0.26$ and $P = 0.31$) (Katju *et al.* 2018). Our estimate of the spontaneous base substitution rate falls within the range previously reported for *C. elegans*, other nematodes, and multicellular eukaryotes (Figure 1). However, it is 4.5-fold lower than the earliest direct estimate for *C. elegans*, which was based on Sanger sequencing of ≤ 30 kb of the nuclear genome (Denver *et al.* 2004). Furthermore, our estimate of the nuclear base substitution rate is lower than that reported by Denver *et al.* (2009) (Student's *t*-test = 3.76 and $P = 0.004$) but higher than that of Denver *et al.* (2012) (Student's *t*-test = 3.15 and $P = 0.004$) (Figure 1). However, there is no significant difference when the average rate in the N2 strain from these two previous studies (Denver *et al.* 2009, 2012) is compared to our estimate (Student's *t*-test = 2.03 and $P = 0.058$). Our estimate of the base substitution rate is significantly lower than that reported recently by Saxena *et al.* (2019) (Student's *t*-test = 2.7 and $P = 0.01$) for an independent set of spontaneous MA lines of *C. elegans*.

Estimate of the genome-wide spontaneous indel mutation rate in a nematode and a pronounced deletion bias

We characterized small indel events as comprising the addition or removal of ≤ 100 -bp sequences. We detected 357 small indel events in the *N* = 1 lines, resulting in a genome-wide spontaneous indel rate of 6.84 (95% C.I.

Figure 1 Estimated genome-wide spontaneous base substitution and indel (insertion/deletion) rates for various multicellular eukaryotes. Substitution rates are shown in gray, blue, purple, rust orange, and green for nematodes, crustaceans, insects, mammals, and plant species, respectively. Where available, the yellow bar indicates the indel rate for the corresponding species/study. Data from: the current study¹, Denver *et al.* (2012)², Denver *et al.* (2009)³, Weller *et al.* (2014)⁴, Flynn *et al.* (2017)⁵, Keith *et al.* (2016)⁶, Assaf *et al.* (2017)⁷, Sharp and Agrawal (2016)⁸, Huang *et al.* (2016)⁹, Schrider *et al.* (2013)¹⁰, Keightley *et al.* (2009)¹¹, Uchimura *et al.* (2015)¹², and Ossowski *et al.* (2010)¹³.

$\pm 0.97) \times 10^{-10}$ /site/generation (Figure 1, Table 1, and Figure S2). Spontaneous indel rates have been reported for *Drosophila melanogaster* (Keightley *et al.* 2009; Schrider *et al.* 2013; Huang *et al.* 2016; Sharp and Agrawal 2016) and *Arabidopsis thaliana* (Ossowski *et al.* 2010), ranging from 3.38×10^{-10} to 1.37×10^{-9} /site/generation (Figure 1). Our estimate of the indel rate for *C. elegans* falls within this reported range and is similar to a recently reported indel rate for a different set of *C. elegans* MA lines (Saxena *et al.* 2019) (Student's *t*-test = 1.21 and *P* = 0.23).

In the *N* = 1 MA lines reflecting the spontaneous mutation spectrum, we observed small deletion and insertion rates of 5.06 (95% C.I. ± 0.83) $\times 10^{-10}$ /site/generation and 1.79 (95% C.I. ± 0.40) $\times 10^{-10}$ /site/generation, respectively (Table 1). This results in a significant deletion bias of 2.83 deletions per insertion. This finding is in stark contrast to the study of Denver *et al.* (2004), which reported a predominance of insertion mutations based on a partial genome analysis (14–29 kb) of a different set of *C. elegans* *N* = 1 MA lines. If all MA lines across our three population-size treatments are considered, we observed 519 deletions and 180 insertions, resulting in a deletion bias of 2.88 deletions per insertion. Hence, the deletion bias is consistent across population sizes (Figure S3, A and B) and deletion rates among all MA lines are significantly higher than insertion rates (Figure 2A; paired Student's *t*-test = 10.22 and *P* < 0.0001; Wilcoxon signed-rank: *W* = 264 and *P* < 0.0001). The vast majority of indels in our study (67% in *N* = 1 lines) are single-nucleotide insertions or deletions, and 76% of the indels comprise three or fewer nucleotides. The size distribution is also different between insertions and deletions as a greater proportion of deletions relative to insertions exceed two nucleotides (Figure 2B; Student's *t*-test = -6.57 and *P* < 0.0001; Wilcoxon rank-sum test: *Z* = -5.0 and *P* < 0.0001). This strong deletion bias, as well as the differences in length distributions between insertions and deletions, resulted in a spontaneous net loss of 1495 bp from the genomes of the *N* = 1 MA lines, an average of 88 bp per

genome over the entire experiment, or 0.24 bp per genome per generation.

No significant difference in the base substitution or indel rates between populations of different sizes

Our analysis identified 788 and 455 independent base substitutions in the *N* = 10 and *N* = 100 lines, respectively. The average base substitution rate in the *N* = 10 and *N* = 100 MA lines was 1.95 (95% C.I. ± 0.13) $\times 10^{-9}$ and 1.83 (95% C.I. ± 0.14) $\times 10^{-9}$ /site/generation (Table 1), respectively. There was no correlation between population size and the base substitution rate (ANOVA *F* = 0.073 and *P* = 0.79; Kendall's τ = 0.06 and *P* = 0.65) (Figure 3A). We identified 226 and 116 independent indel events in the *N* = 10 and *N* = 100 lines, respectively. This yielded average indel rates of 9.46 (95% C.I. ± 1.50) $\times 10^{-10}$ and 6.95 (95% C.I. ± 0.71) $\times 10^{-10}$ /site/generation for the *N* = 10 and *N* = 100 lines, respectively (Table 1). As was the case for base substitutions, we found no correlation between population size and the indel rate (ANOVA *F* = 1.17 and *P* = 0.29; Kendall's τ = 0.22 and *P* = 0.13) (Figure 3B).

No significant difference in the accumulation of nonsynonymous and frameshift mutations with differing intensity of selection

Natural selection is expected to have greater consequences for the accumulation of nonsynonymous substitutions and frameshift mutations, relative to synonymous mutations or mutations in noncoding DNA. Synonymous mutations should be predominantly neutral and we did not expect their rates to vary between different population-size treatments. Indeed, there was no significant difference between the synonymous substitution rates at different population sizes (Figure 4A: ANOVA *F* = 0.77 and *P* = 0.39; Kendall's τ = 0.13 and *P* = 0.37). In contrast, many nonsynonymous and frameshift mutations were expected to be deleterious and subject to purifying selection in larger populations. However, we did not find significant differences in the nonsynonymous

Figure 2 Rates and size distributions of small insertion/deletion (indel) events. (A) The deletion rates among all mutation accumulation lines are significantly higher than insertion rates (paired Student's *t*-test: $t = 10.22$, $P < 0.0001$; Wilcoxon signed-rank: $W = 264$, $P < 0.0001$). (B) The size distribution of indels reveals that deletions tend to be larger than insertions (Student's *t*-test = -6.57 , $P < 0.0001$; Wilcoxon rank-sum: $Z = -5.0$, $P < 0.0001$).

substitution rates (Figure 4B: ANOVA $F = 0.08$ and $P = 0.77$; Kendall's $\tau = 0.03$ and $P = 0.81$), or the combined nonsynonymous substitution and frameshift mutation rates (Figure 4C: ANOVA $F = 0.05$ and $P = 0.82$; Kendall's $\tau = -0.01$ and $P = 0.93$) across different population-size treatments. In addition, we did not find a significant correlation between the non-synonymous/synonymous substitution ratio (K_a/K_s) and population size, although there appeared to be a negative trend in the predicted direction (Figure 4D: ANOVA $F = 0.19$ and $P = 0.67$; Kendall's $\tau = -0.16$ and $P = 0.27$). Similarly, the mean radicality of amino acid changes did not correlate significantly with population size (Figure S4; ANOVA $F = 1.93$ and $P = 0.16$; Kendall's $\tau = -0.214$ and $P = 0.13$), despite the appearance of a negative trend.

Base substitution spectrum exhibits a strong A/T bias

The pattern of base substitutions in the $N = 1$ lines that are under minimal influence of selection should reflect the spontaneous mutation spectrum. The base substitution rate exhibits a strong G/C $\rightarrow$ A/T mutation bias, primarily driven by G/C $\rightarrow$ A/T transitions (Figure 5). The mutation rate from a G/C pair to an A/T pair was 2.1 (95% C.I. ± 0.21), 2.3 (95% C.I. ± 0.23), and 2.1 (95% C.I. ± 0.23) $\times 10^{-9}$, for the $N = 1$, 10, and 100 lines, respectively. Conversely, the mutation rate from an A/T pair to a G/C pair was 0.56 (95% C.I. ± 0.07), 0.57 (95% C.I. ± 0.08), and 0.51 (95% C.I. ± 0.09) $\times 10^{-9}$, respectively, for the corresponding population sizes as listed above. Taking $N = 1$ as the best estimate of the mutation rate in the absence of selection, the A/T mutation bias was 3.75. The expected equilibrium G+C content (GC_{eq}), where the number of G/C $\rightarrow$ A/T mutations equals A/T $\rightarrow$ G/C mutations, was calculated as 26% for the *C. elegans* nuclear genome. The *C. elegans* nuclear genome has a G+C content of 36%.

Base substitutions in the $N = 1$ lines exhibit a slight but nonsignificant transition bias, leading to a transition:transversion ratio (Ts:Tv) of 0.64 ($N = 1$ line-specific values range from 0.36 to 1.04). If all mutations between the four nucleotides are equally likely, the expected transition bias is 0.5. The relative overrepresentation of transitions compared to transversions is therefore 0.64/0.5, or 1.28. The Ts:Tv ratio

did not vary with population size ($F = 0.016$ and $P = 0.73$; Kendall's $\tau = 0.31$ and $P = 0.76$), and the relative overrepresentation of transitions in the $N = 10$ and 100 lines was 1.41 and 1.28, respectively. The lack of a strong transition bias was partly due to high rates of A/T $\rightarrow$ T/A transversions in introns and intergenic regions. If we analyze the transition bias in coding and noncoding sequences separately, the relative overrepresentation of transitions was 1.93 in exons and 1.14 in introns in the $N = 1$ lines.

Strong context dependence of A/T $\rightarrow$ T/A transversions in noncoding DNA

Compared to previous studies, our data indicate a greater frequency of A/T $\rightarrow$ T/A transversions. The majority of these mutations are flanked by A and T base pairs on each side, and occur more frequently in introns and intergenic regions compared to exons (Figure 6A). A/T $\rightarrow$ T/A transversions are particularly common in introns and intergenic regions, when the focal nucleotide is flanked by a 5'–T and a 3'–A. A flanking 5'–A and 3'–T also appears to elevate the rate of A/T $\rightarrow$ T/A transversion in introns (Figure 6A). Additionally, these substitutions primarily occur on the boundaries of homopolymeric runs of 7–11 bases of either As or Ts (Figure 6B).

Elevated base substitution rate in chromosomal arms relative to cores

There was no significant effect of population size on the base substitution rate either at the interchromosomal or intrachromosomal level. Hence, much of the subsequent analysis of the distribution of base substitutions across the *C. elegans* genome will be based on the pooled results from all of the MA lines ($N = 1$, 10, and 100 populations). The nucleotide substitution rates were analyzed in a three-way ANOVA for chromosomes (five autosomes and one sex chromosome), functional regions (exons, introns, and intergenic regions), and recombination domains (arms, cores, and tips). The nucleotide substitution rates did not vary significantly between chromosomes (Figure 7A: $F = 0.86$ and $P = 0.51$). There is a significant difference between the nucleotide substitution rates in exons, introns, and intergenic regions (Figure 7B: $F = 6.51$

Figure 3 The base substitution and insertion/deletion (indel) rates do not vary with population size. (A) The base substitution rates do not differ significantly between population sizes of $N = 1, 10$, and 100 individuals (ANOVA $F = 0.073$, $P = 0.79$; Kendall's $\tau = 0.06$, $P = 0.65$). (B) The three population sizes do not differ significantly with respect to the indel rates (ANOVA $F = 1.17$, $P = 0.29$; Kendall's $\tau = 0.22$, $P = 0.13$).

and $P = 0.0015$). The substitution rate in introns is significantly higher than that in exons [2.25 (95% C.I. ± 0.14) $\times 10^{-9}$ /site/generation and 1.51 (95% C.I. ± 0.15) $\times 10^{-9}$ /site/generation, respectively; Tukey's multiple comparisons of means, $P = 0.001$], whereas the nucleotide substitution rate in intergenic regions [1.82 (95% C.I. ± 0.14) $\times 10^{-9}$ substitutions/site/generation] falls between that of introns and exons, and is not statistically different from either one. The per nucleotide substitution rates differ significantly between chromosomal arms, cores, and tips (Figure 7C: $F = 6.62$ and $P = 0.0014$), and it is higher in arms than cores [2.18 (95% C.I. ± 0.13) $\times 10^{-9}$ /site/generation and 1.58 (95% C.I. ± 0.10) $\times 10^{-9}$ /site/generation, respectively; Tukey's multiple comparisons of means, $P = 0.0019$], while the arms and tips [2.18 (95% C.I. ± 0.13) $\times 10^{-9}$ /site/generation and 1.96 (95% C.I. ± 0.28) $\times 10^{-9}$ /site/generation, respectively) do not differ significantly in their substitution rates (Tukey's multiple comparisons of means, $P = 0.82$). The difference in base substitution rates between the arms and the cores is evident for coding and noncoding sequences alike (Figure 7D).

A/T and G/C homopolymeric runs differ in their mutational properties

The number of single-nucleotide A or T indels was as expected in the absence of strand bias (Figure 8A). Similarly, G or C single-nucleotide indels did not show any evidence of strand bias and occurred at roughly equal frequency (Figure 8A). Furthermore, there was no significant difference in the spectrum of indels between different population sizes (Figure 8B; Fisher's exact test: $P = 0.51$). While A/T indels were more common across the genome, the G/C indel rates were higher than A/T indel rates after standardizing the rates by mutational opportunity (Figure 8C). The rates of indels in runs of As and Ts increased with the length of the run (Figure 8C). Deletion rates tend to be higher than insertion rates in long A/T homopolymeric runs. Similarly, longer runs of G/Cs have higher deletion rates than short G/C runs. In contrast, shorter G+C runs have higher insertion rates relative to deletion rates (Figure 8C). The mean complexity of the sequence that incurred indels was significantly lower than both (i) random

sites in the genome (Student's t -test = 17.03 , $P < 2.2 \times 10^{-16}$) and (ii) sequences that incurred nucleotide substitution (Student's t -test = 10.28 , $P < 2.2 \times 10^{-16}$). This is likely due to the propensity of indels to occur mainly in A+T-rich regions, which are by nature of low complexity (Figure 8D).

Intrachromosomal location significantly affects the indel rate

The effect of chromosomal location on indel rates mirrors that of base substitutions. Due to the paucity of indels in chromosomal tips (only 1.8 indels/line), we performed a three-way ANOVA including data for cores and arms, while excluding that for tips. There were no significant interactions between the effects of the chromosome and the recombination domain ($F = 1.84$ and $P = 0.10$). However, there were significant interactions between (i) chromosomes and coding content (exons, introns, and intergenic regions) ($F = 2.06$ and $P = 0.025$), as well as (ii) recombination domain and coding content ($F = 3.24$ and $P = 0.04$). The indel rates were not significantly different between individual chromosomes (Figure 9A: $F = 1.95$ and $P = 0.08$). As was the case for base substitutions, the indel rates differed significantly between exons, introns, and intergenic regions (Figure 9B: $F = 46.47$ and $P < 2.0 \times 10^{-16}$). Indel rates were observed to be the lowest for exonic regions [3.15 (95% C.I. ± 0.74) $\times 10^{-10}$ /site/generation], whereas intronic and intergenic regions had higher indel rates [10.15 (95% C.I. ± 1.14) $\times 10^{-10}$ /site/generation and 8.64 (95% C.I. ± 1.10) $\times 10^{-10}$ /site/generation, respectively]. These differences are likely attributable to different amounts of low-complexity sequence within these regions. Furthermore, the indel rates differed between chromosomal arms and cores (Figure 9C: $F = 23.14$ and $P = 1.7 \times 10^{-6}$). The indel rate in cores [5.86 (95% C.I. ± 0.74) $\times 10^{-10}$ /site/generation] was significantly lower than that in chromosomal arms [9.62 (95% C.I. ± 0.99) $\times 10^{-10}$ /site/generation]. The low indel rates in the cores compared to the arms were detected for all functional regions (exons, introns, and intergenic regions) (Figure 9D). The distribution of indels across the three recombination domains did not differ significantly between population-size treatments (Figure S5; Fisher's exact test: $P = 0.74$).

Figure 4 The rates of synonymous and nonsynonymous mutations did not vary with population size. (A) No significant effect of population size is detected in synonymous substitution rates (ANOVA $F = 0.77$, $P = 0.39$; Kendall's $\tau = 0.13$, $P = 0.37$). (B) Nonsynonymous substitution rates do not vary significantly with population size (ANOVA $F = 0.08$, $P = 0.77$; Kendall's $\tau = 0.03$, $P = 0.81$). (C) Pooled nonsynonymous and frameshift mutations rates do not vary significantly with population size (ANOVA $F = 0.05$, $P = 0.82$, Kendall's $\tau = -0.01$, $P = 0.93$). (D) The K_a/K_s ratio does not vary with population size (ANOVA $F = 0.19$, $P = 0.67$; Kendall's $\tau = -0.16$, $P = 0.27$).

Germline-expressed genes have higher mutation rates than nongermline-expressed genes

The transcription of a gene has the potential to influence its mutation rate, and some studies have found a positive association between transcription and mutation rate (Hudson *et al.* 2003; Kim and Jinks-Robertson 2012; Alexander *et al.* 2013). To determine whether germline expression of *C. elegans* genes is correlated with the mutation rate, we classified the protein-coding genes into germline- and nongermline-expressed genes using published results (Wang *et al.* 2009). The substitution rate across all MA lines was significantly higher in germline-expressed genes than in nongermline-expressed genes (Figure S6A; two-way ANOVA: $F = 12.05$ and $P = 0.0007$). Moreover, there was a significant between germline expression and the recombination domain (Figure S6B; two-way ANOVA: $F = 12.8$ and $P = 0.0007$). With respect to the core regions, there was no significant difference in the mutation rates of germline- and nongermline-expressed genes. In contrast, germline-expressed genes had higher mutation rates than nongermline-expressed genes when residing in the chromosomal arms and tips.

Context-dependent A/T $\rightarrow$ T/A transversions contribute to intrachromosomal variation in substitution rates

There are significant differences in the frequencies of homopolymeric runs between coding and noncoding DNA. Because strongly context-dependent A/T $\rightarrow$ T/A transversions occur frequently at the boundaries of A/T homopolymers, we

tested if any of the positional or transcription-related differences in mutation rate could be accounted for by these A/T $\rightarrow$ T/A transversions. If all A/T $\rightarrow$ T/A transversions are excluded from the analysis, we no longer observe significant differences in mutation rates between (i) exons and noncoding DNA (Figure S7A; ANOVA: $F = 0.91$ and $P = 0.41$), nor (ii) between germline- and nongermline-transcribed genes (Figure S7B; Welch's paired t -test = 1.60, $P = 0.12$). In contrast, a significant mutation rate variation still exists among chromosomal cores, arms, and tips, despite the exclusion of A/T $\rightarrow$ T/A transversions (Figure S7C; ANOVA $F = 3.90$, $P = 0.024$). This variation is primarily due to a significant difference in mutation rates between chromosomal cores and arms (Tukey's multiple comparisons of means, $P = 0.02$). In summary, the nonrandom distribution of mutable motifs can account for the differences between coding and noncoding DNA, as well as transcription-related differences in mutation rates, and they contribute to the differences in mutation rates between cores, arms, and tips. However, the differences in mutation rates between cores, arms, and tips are not fully explained by context-dependent A/T $\rightarrow$ T/A transversions. Thus, the higher rates of mutations in arms compared with cores could also be due to higher recombination frequency.

Genomic properties and their effects on site mutability

The genomic properties of substitution sites in conjunction with 1,000,000 random nonmutated sites were used to predict the mutability of each site in the genome using a logistic regression approach. The strongest positive predictors of

Figure 5 The mutational spectrum at different population sizes. The transition bias is not significantly different from random. The mutational spectrum and the transition:transversion ratio do not vary with population size ($F = 0.016$, $P = 0.73$; Kendall's $\tau = 0.31$, $P = 0.76$).

mutability are repeat sequences (OR = 5.34) and the immediate nucleotides flanking the mutated site (Figure S8A and Table S3). A and T sites with a flanking 5'–T and a 3'–A (TAA or TTA) had a strong positive effect on mutability (OR = 4.89), while other A+T nucleotide triplets were negatively related to mutability (Figure S8A). The GC contents of 20-bp tracts, both upstream and downstream of a site, negatively affects its mutability (OR = 0.35) (Figure S8A), whereas G+C triplets at the focal site have positive effects on mutability (ORs: CGC = 2.99, CCG = 2.65, and GCC = 1.94). Other genomic properties—including 5' and 3'–UTRs, chromosomal location, functional annotation (exons, introns, and intergenic regions), recombination rate, arm and tip recombination domains, germline expression, and sequence complexity—had negligible effects on a site's mutability (Figure S8A and Table S3). Of the total genome, 20% contains sites with mutability > 0.003, which contribute to 50% of the observed SNPs. An additional 25% of the SNPs are found at sites with predicted mutability of > 0.0075, representing < 4% of the genome (Figure S8B). The predicted mutability of sites explains 82% of the observed variation in mutability (Figure S8C; linear regression $F = 348.5$, $R^2 = 0.82$, and $P < 2.2 \times 10^{-16}$). These results suggest that the mutability of a site is primarily influenced by the focal base pair and its immediate flanking nucleotides, and that the association between mutational patterns and other genomic properties are due to the nonrandom distribution of mutable motifs.

Discussion

MA experiments typically consist of passaging experimental replicate lines through a minimum population bottleneck in each generation of the experiment. In contrast, our *C. elegans* MA experiment comprised three population-size treatments aimed at assessing the rates of origin of diverse classes of

mutations and their differential accumulation under varying regimes of natural selection. We have previously assessed the phenotypic consequences of mutation and selection under benign laboratory (Katju *et al.* 2015) and osmotic stress conditions (Katju *et al.* 2018). In addition, we have employed modern genomic approaches to investigate the interplay of mutation and selection on mtDNA (mitochondrial DNA) SNPs and small indels (Konrad *et al.* 2017), and nuclear copy-number variants (Konrad *et al.* 2018). In both preceding genomic analyses, there was evidence of selection in the larger population-size treatments. With regard to the mitochondrial genome, there was no difference in the accumulation of synonymous mutations across different population-size treatments, whereas nonsynonymous mutations, frame-shifts, and deletions accumulated at a higher rate in MA lines maintained at the most extreme population bottleneck of $N = 1$ (Konrad *et al.* 2017). The accumulation of copy-number variants in the nuclear genome also showed a significant relationship with population size (Konrad *et al.* 2018). Gene deletions accumulated at a higher rate in the smallest $N = 1$ populations and the frequency of gene duplications in the larger populations ($N = 10$ or 100 individuals) was significantly influenced by gene expression, which suggested that high ancestral transcription levels of genes, as well as the degree of increase in transcript abundance of duplicated genes, contribute to the fitness cost of gene duplications. In this study, we investigated two additional major classes of mutational variants in the nuclear genome, namely SNPs and small indels, to provide a comprehensive picture of the spontaneous mutation process in *C. elegans* through the lens of experimental evolution.

The $N = 1$ lines provide the baseline for the spontaneous rate of origin of different classes of mutations and the expected rate of neutral evolution. In this study, the spontaneous rate of origin of nuclear base substitutions (μ_{bs}) and small indels of < 100-bp length (μ_{indel}) in *C. elegans* were determined to be 1.84×10^{-9} substitutions/site/generation and 0.68×10^{-9} indels/site/generation, respectively. Hence, the rate of accumulation of nuclear SNPs exceeds that of small nuclear indels by approximately threefold. Based on this study and our preceding mtDNA genome analysis on the same set of MA lines (Konrad *et al.* 2017), we find that the spontaneous rates of different classes of mutations per nucleotide in *C. elegans* range from 10^{-10} to 10^{-8} per base per generation, representing a ~90-fold difference. This relationship can be expressed as follows: $\mu_{indel} < \mu_{bs} < \text{mtDNA } \mu_{bs} < \text{mtDNA } \mu_{indel}$. While the small indel rate is lower than the base substitution rate in the nuclear genome, the inverse is true for the mitochondrial genome. A higher indel rate in the mtDNA is largely due to a higher incidence of homopolymeric runs and a greater AT skew in this genome. In addition, nuclear copy-number changes (gene duplications and deletions) represent a major component of the genetic variation arising due to spontaneous mutation, with rates of origin in the order of 10^{-5} per gene per generation (Konrad *et al.* 2018).

Figure 6 Context dependence of base substitutions. (A) The vast majority of mutations in intron and intergenic regions are 5′–TTA–3′ ↔ 5′–TAA–3′ transversions. (B) Substitutions occurring at boundaries of A or T homopolymeric runs are responsible for the disproportionate contribution of A/T → T/A transversions. The A → T and T → A transversions are equally frequent in homopolymeric runs, consistent with the absence of a strand bias.

Our empirical estimate of the spontaneous nuclear base substitution rate for *C. elegans* is similar to three previous estimates for the species using high-throughput sequencing of MA lines (Denver *et al.* 2009, 2012; Saxena *et al.* 2019), but substantially lower than the first estimate, which was based on Sanger sequencing (9.1×10^{-9} ; Denver *et al.* 2004). Additionally, our spontaneous base substitution rate is similar to estimates for the congeneric species *C. briggsae* (average 1.33×10^{-9} ; Denver *et al.* 2012) and another nematode species, *Pristionchus pacificus* (2.0×10^{-9} ; Weller *et al.* 2014). The divergence times for *C. elegans*–*C. briggsae* and *Pristionchus*–*Caenorhabditis* are estimated at 80–120 MYA (Hillier *et al.* 2007) and 280–430 MYA (Dieterich *et al.* 2008), respectively. Despite the uncertainty in divergence times based on the molecular clock, the mutation rates of these nematodes under experimental conditions are remarkably similar given the considerable evolutionary time since their divergence, and suggests that the mutation rates are under stabilizing selection. The base substitution rate in these nematodes is lower relative to other invertebrates for which similar information exists. For example, the base substitution rate in the cladoceran *Daphnia pulex* (Flynn *et al.* 2017) is roughly twice as high as in nematodes, whereas *D. melanogaster* has an approximately threefold-higher rate than *Caenorhabditis* (Huang *et al.* 2016; Sharp and Agrawal 2016; Assaf *et al.* 2017). Furthermore, the spontaneous mitochondrial base substitution rate for the very same *C. elegans*

MA lines (Konrad *et al.* 2017) is 24-fold higher than the nuclear base substitution rate generated from this study.

Spontaneous nuclear small indel rates are observed to be considerably lower than base substitution rates for a wide range of surveyed genomes [reviewed in Katju and Bergthorsson (2019)]. Our spontaneous small indel rate of 6.84×10^{-10} changes/site/generation is approximately one-third of the base substitution rate in the *C. elegans* nuclear genome. However, comparing the indel rates with other taxa can be problematic because of the great variation in estimates of indel rates within taxa. For example, indel rate estimates within *D. melanogaster* differ by fourfold whereas the base substitution rates vary less than twofold [reviewed in Katju and Bergthorsson (2019)]. Furthermore, many whole-genome sequencing studies of MA lines do not report indel rates. However, the small indel rate for *C. elegans* from this study falls within the range reported from MA studies in a few metazoans (0.31×10^{-9} to 1.37×10^{-9} ; Katju and Bergthorsson 2019). Our genome-wide estimate of the small indel rate is considerably lower, namely $< 6\%$, of the originally reported rate for *C. elegans* (Denver *et al.* 2004), but similar to that recently reported by Saxena *et al.* (2019). In another notable departure from previous results, which found that insertions outnumbered deletions in the *C. elegans* genome (Denver *et al.* 2004), we find a strong deletion bias wherein deletions exceed insertions by threefold. This is consistent with an almost universal deletion bias observed in MA experiments

Figure 7 Variation in base substitution rates across different genomic regions. (A) There was no significant difference in the base substitution rate between chromosomes ($F = 0.86$, $P = 0.51$). (B) The base substitution rates differ significantly between exons, introns, and intergenic regions ($F = 6.51$, $P = 0.0015$). (C) Base substitution rates are significantly different between chromosomal arms, cores, and tips ($F = 6.62$, $P = 0.0014$). (D) A lower base substitution rate in cores relative to arms and tips applies to exons, introns, and intergenic regions.

[reviewed in Katju and Bergthorsson (2019)] as well as in comparative analyses of sequenced genomes (Kuo and Ochman 2009). The vast majority of indels occur in homopolymeric runs, and their frequency increases as a function of the length of the run. However, in contrast to A/T runs, short G/C runs appear to have an insertion bias, although long G/C runs have a deletion bias. Moreover, the indel rates are higher in G/C runs relative to their A/T counterparts. The differences in the mutational properties of low-complexity repeats such as homopolymeric runs is likely to play a role in the evolution of their frequency and length distribution in the genome.

The varying population-size design of our spontaneous MA experiment allowed us to investigate the influence of increasing selection efficacy on the evolutionary dynamics, and persistence, of newly occurring nuclear SNPs and small indel mutations. Notably, there was no correlation between the frequency of base substitutions, nonsynonymous substitutions, or small indels with population size. Previous phenotypic analyses of these MA lines for two fitness-related traits indicated that: (i) the $N = 10$ and $N = 100$ populations did

not suffer significant decline in fitness due to deleterious mutations, and (ii) most of the decline in fitness in the $N = 1$ populations was due to mutations of large effects (Katju *et al.* 2015, 2018). Alternatively, the observed decline in fitness traits could be due to a large number of mutations with small fitness effects. Our combined phenotypic and genomic results are consistent with the former hypothesis that a fitness decline of $N = 1$ is primarily due to a few mutations of large effect. In this study, there were, on average, 90 substitutions and small indels per genome, whereas our preceding analysis of fitness-related traits suggested that only two to three mutations per genome (Katju *et al.* 2015) were responsible for the decline in the $N = 1$ lines. The scale of this experiment lacks the power to detect such small differences in the number of mutations between replicate lines across different population sizes. Our genomic results are also concordant with a growing suite of MA studies empirically implicating a few mutations of large effect as the primary contributors to fitness decline (Keightley and Caballero 1997; Davies *et al.* 1999; Ávila and Garcio-Dorado 2002; Halligan *et al.* 2003; Estes *et al.* 2004; Sanjuán *et al.* 2004;

Figure 8 Different rates and patterns of A/T and G/C insertion/deletions (indels) in homopolymeric runs. (A) The numbers of single-nucleotide A or T indels are almost identical, and G or C indels are also equally frequent, as expected in the absence of strand bias in the indel calls. (B) There is no difference in the frequency of different kinds of single-nucleotide indels between different population sizes. (C) G/C homopolymeric runs have higher indel rates than A/T homopolymeric runs. The frequency of A/T indels rises with increasing length of a homopolymeric run but then tapers off. The deletion bias is more pronounced for A/T indels in longer runs as the deletion rates tend to be higher than the insertion rates in long A/T homopolymeric runs. Longer G/C runs have higher deletion rates than short G/C runs, whereas shorter G/C runs have increased insertion rates relative to long runs. (D) The mean sequence complexity surrounding indels is significantly lower than for both random sites in the genome (Student's t -test = 17.03, $P < 2.2 \times 10^{-16}$) and sequences surrounding base substitutions (Student's t -test = 10.28, $P < 2.2 \times 10^{-16}$).

Heilbron *et al.* 2014; Luijckx *et al.* 2018). Although the K_a/K_s ratio for different population sizes trends in the expected direction, this negative correlation was not significant. Furthermore, the lack of significant decline in fitness in the $N = 10$ and 100 populations suggests that the vast majority of nuclear base substitutions and small indels have small fitness effects, and that the vast majority of these escape selection even at the larger population sizes ($N = 10$ or 100) within this study. The results from this study are interesting in light of the significant negative correlations observed in this very set of MA lines between population size and (i) nonsynonymous mitochondrial mutations (Konrad *et al.* 2017), and (ii) many aspects of gene copy-number changes (Konrad *et al.* 2018). For example, gene deletions accumulated at a higher rate in the $N = 1$ populations than in the larger populations (Konrad *et al.* 2018). Similarly, duplications of highly expressed genes, and those that strongly increased the transcript levels of duplicated genes, also accumulated more rapidly in the $N = 1$ than in the $N = 10$ or $N = 100$ populations (Konrad *et al.* 2018). This suggests that both mitochondrial mutations and gene copy-number changes are under more stringent purifying selection than nuclear base substitutions or small indels.

The predominance of transitions over transversions is commonly observed in molecular evolution studies (Vogel and Röhrborn 1966; Fitch 1967; Wakeley 1996). The key mechanisms contributing to this transition bias are held to be (i) selection against transversions, which are more likely to cause missense mutations than transitions, and (ii) mutational bias due to the structural similarities among purines and pyrimidines (Stoltzfus and Norris 2016). We did not observe a genome-wide mutational bias toward transitions in our *C. elegans* MA lines, a pattern that has been noted by others (Denver *et al.* 2009, 2012). Without any base substitutional bias, transversions are expected to be twice as frequent as transitions, and the frequency of transitions and transversions in our study was not significantly different from this expectation. However, in exons where a transition/transversion bias is most likely to have consequences for fitness, we did in fact observe a transition bias. The numbers of transitions and transversions were roughly equal in exons, which means that transitions are twice as frequent as expected if there was no bias. The near-universal base substitution bias toward A/T nucleotides was also observed in our results as G/C $\rightarrow$ A/T substitutions were 3.75-fold more likely than mutations in the opposite direction. This base substitution bias predicts an equilibrium base composition of 26% G/C,

Figure 9 Variation in small insertion/deletion (indel) rates across different genomic regions. (A) There was no significant difference in the small indel rate between chromosomes ($F = 1.95$, $P = 0.08$). (B) The indel rate differs significantly between exons, introns, and intergenic regions ($F = 46.47$, $P < 2.0 \times 10^{-16}$). (C) The indel rates are significantly different between chromosomal arms and cores ($F = 23.14$, $P = 1.7 \times 10^{-6}$). (D) A lower indel rate in cores compared to arms and tips applies to exons, introns, and intergenic regions.

which is lower than either the total G/C content of the *C. elegans* genome (36%), or the G/C content of intergenic DNA and introns (33%). Assuming that the mutational biases under experimental conditions are the same as the prevailing mutational biases in the wild, the departure of the observed G+C content from that expected suggests that other mechanisms than the biases of spontaneous mutations are influencing the base composition of the *C. elegans* nuclear genome. Higher G+C content than is expected by mutation pressure alone seems to be the rule in genome evolution, and it is usually presumed that natural selection for higher G+C content and/or biased gene conversion is responsible. However, this departure from equilibrium G+C content also has the effect of increasing the mutation rate (Krasovec *et al.* 2017).

Furthermore, there are interesting context-dependent patterns in the frequency of substitutions. In particular, a 5'-T and a 3'-A have a strong positive effect on the A/T $\rightarrow$ T/A substitution rate, especially at the boundaries of A or T homopolymeric runs. Similar observations have been made in

mismatch repair-deficient lines of *C. elegans* (Meier *et al.* 2018). The combination of this strong context-dependence of base substitutions and the genomic distribution of A and T homopolymeric runs explains three other observations about the base substitution patterns in our MA lines. Introns and intergenic regions have significantly higher mutation rates than exons in our study. It is usually assumed that differences in substitution rates between introns and exons are due to selection rather than intrinsic differences in mutation rates. However, lower mutation rates in coding sequences relative to noncoding ones have been observed in other MA experiments and were ascribed to transcription-coupled repair, and differential efficiency of mismatch repair between coding and noncoding DNA (Krasovec *et al.* 2017). Additionally, a recent study of somatic mutation rates in humans concluded that introns have higher mutation rates than exons, due in part to the greater efficiency of mismatch repair in exons (Frigola *et al.* 2017). The data presented here suggest that the difference in mutation rates between introns and exons in

C. elegans is caused by strongly context-dependent A/T → T/A substitution mutations. These mutations, which are particularly frequent at the boundaries of A and T homopolymeric runs, are in turn more common in introns and intergenic regions, and less prevalent in exons. Indeed, if we exclude A/T → T/A mutations from our analysis, the difference in mutation rates between exons and introns disappears. Hence, the higher mutation rates in introns and intergenic regions, compared to exons in *C. elegans*, are due to a higher prevalence of mutagenic motifs in introns and intergenic regions.

Nucleotide polymorphisms in natural populations are correlated with recombination rates (Begun and Aquadro 1992; Cutter and Choi 2010; McGaugh *et al.* 2012). These correlations are usually attributed to the combination of natural selection and genetic linkage, where genetic hitchhiking or background selection on linked sites depresses genetic variation in regions of low recombination. However, mutation rates are also positively correlated with recombination rates in several well-studied systems such as humans, *Arabidopsis*, and honey bees (Arbeithuber *et al.* 2015; Francioli *et al.* 2015; Yang *et al.* 2015; Smith *et al.* 2018). The *C. elegans* chromosomes can be divided into three regions with respect to recombination frequency (Rockman and Kruglyak 2009). The most central regions of the chromosomes, the cores, comprise 47% of the genome and have higher gene densities, lower repetitive element content, and lower recombination rates. The chromosomal arms comprise 46% of the genome and are marked by a higher incidence of repetitive elements, lower gene densities, and increased recombination. Chromosomal tips are much shorter sections at the ends of chromosomes (7% of the genome), which are not thought to experience recombination (Barnes *et al.* 1995; Rockman and Kruglyak 2009). Our previous study of spontaneous gene copy-number changes in these *C. elegans* MA lines found that duplication and deletion breakpoints were more frequent in arms and tips than in the cores (Konrad *et al.* 2018). In this study, the distributions of nuclear base substitutions and indels followed the same pattern, with significantly lower mutation rates in the cores relative to the arms and tips. Our comparison of the base substitution spectrum in cores vs. arms and tips revealed that A/T → T/A substitutions are disproportionately more common in the arms and tips than in the cores. Even when A/T → T/A mutations are excluded from the analysis, there is still a difference in substitution rates between recombination domains. However, just as with the difference in mutation rates between exons, introns, and intergenic regions, the difference in mutation rates between cores vs. arms and tips is also a function of the frequency of A/T homopolymeric runs. The mutation rate in the tips (negligible recombination) was similar to that in the arms (high recombination frequency). Furthermore, the logistic regression analysis did not detect an increase in mutations due to recombination. It appears that the differences in the mutation rates between arms and cores are best explained by the differential abundance of mutagenic motifs, and not by recombination rate *per se*.

Experiments in several organisms have suggested that frequent transcription can render the transcribed DNA more vulnerable to mutations (Klapacz and Bhagwat 2002; Hudson *et al.* 2003; Kim and Jinks-Robertson 2012). For such an effect to influence the mutation rates in multicellular animals, germline-transcribed genes could hypothetically have higher mutation rates than genes that are only expressed in somatic tissues. Our results initially suggested that germline-expressed genes may have higher substitution rates than nongermline-expressed genes. However, this effect was only detected in germline-transcribed genes located in the chromosomal arms, and not in the cores. Upon further analysis, we found that the association between germline transcription and the base substitution was due to context-dependent A/T → T/A substitutions in the introns of germline-transcribed genes. Hence, the higher mutation rates of germline-expressed genes in our MA lines were not due to a general increase in the substitution rate and it did not extend to exons of these genes.

This study contains the largest set of mutations for a spontaneous MA experiment employing the *C. elegans* N2 wild-type strain. The analysis of base substitutions in our MA lines confirmed some previous results regarding mutation rates and mutational biases. Other results add context to previous observations. For example, the lack of transition bias is primarily due to high transversion rates, specifically A/T → T/A, in introns and intergenic regions, and does not extend to exons. The analysis also illustrates that correlations between genomic location and transcription with mutation rate can arise from the nonrandom distribution of mutagenic motifs. The efficacy of natural selection vs. genetic drift depends on the effective population size. These MA experiments utilized different population sizes to reveal the effects of differing efficacy of selection on the accumulation of mutations. The lack of a correlation between nuclear base substitution rates and population sizes suggests that the vast majority of these mutations are either neutral or have extremely small fitness effects. In direct contrast, a negative correlation was indeed found between population size and the accumulation of mitochondrial mutations, gene deletion rates, and transcript abundance of duplicated genes in the same set of experimental lines. The differences between the results for mitochondrial mutations and gene copy-number changes on the one hand, and nuclear base substitutions and small indels on the other, are consistent with the view that the former have, on average, more detrimental effects on fitness.

Acknowledgments

The authors thank Associate Editor Michael Nachman and two anonymous reviewers for their immensely valuable suggestions that helped improve the manuscript; Lucille Packard for assistance in the creation of the MA lines; Robert Waterston (University of Washington) and Donald Moerman (University of British Columbia) for their help with genome sequencing; and Philip Green from the University of

Washington for providing the program Phaster. This research was supported by National Science Foundation grant MCB-1330245 to V.K. U.B. and V.K. were additionally supported by start-up funds from the Department of Veterinary Integrative Biosciences, College of Veterinary Medicine and Biomedical Sciences at Texas A&M University.

Literature Cited

- Alexander, M. P., K. J. Begins, W. C. Crall, M. P. Holmes, and M. J. Lippert, 2013 High levels of transcription stimulate transversions at GC base pairs in yeast. *Environ. Mol. Mutagen.* 54: 44–53. <https://doi.org/10.1002/em.21740>
- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990. Basic local alignment search tool. *J. Mol. Biol.* 215: 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Arbeithuber, B., A. J. Betancourt, T. Ebner, and I. Tiemann-Boege, 2015 Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc. Natl. Acad. Sci. USA* 112: 2109–2114. <https://doi.org/10.1073/pnas.1416622112>
- Assaf, Z. J., S. Tilk, J. Park, M. L. Siegal, and D. A. Petrov, 2017 Deep sequencing of natural and experimental populations of *Drosophila melanogaster* reveals biases in the spectrum of new mutations. *Genome Res.* 27: 1988–2000. <https://doi.org/10.1101/gr.219956.116>
- Ávila, V., and A. Garcio-Dorado, 2002 The effects of spontaneous mutation on competitive fitness in *Drosophila melanogaster*. *J. Evol. Biol.* 15: 561–566. <https://doi.org/10.1046/j.1420-9101.2002.00421.x>
- Barnes, T. M., Y. Kohara, A. Coulson, and S. Hekimi, 1995 Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. *Genetics* 141: 159–179.
- Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356: 519–520. <https://doi.org/10.1038/356519a0>
- Cutter, A. D., and J. Y. Choi, 2010 Natural selection shapes nucleotide polymorphism across the genome of the nematode *Caenorhabditis briggsae*. *Genome Res.* 20: 1103–1111. <https://doi.org/10.1101/gr.104331.109>
- Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. *Bioinformatics* 27: 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- Davies, E. K., A. D. Peters, and P. D. Keightley, 1999 High frequency of cryptic deleterious mutations in *Caenorhabditis elegans*. *Science* 285: 1748–1751. <https://doi.org/10.1126/science.285.5434.1748>
- Denver, D. R., K. Morris, M. Lynch, and W. K. Thomas, 2004 High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature* 430: 679–682. <https://doi.org/10.1038/nature02697>
- Denver, D. R., P. C. Dolan, L. J. Wilhelm, W. Sung, J. I. Lucas-Lledó *et al.*, 2009 A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proc. Natl. Acad. Sci. USA* 106: 16310–16314. <https://doi.org/10.1073/pnas.0904895106>
- Denver, D. R., L. J. Wilhelm, D. K. Howe, K. Gafner, P. C. Dolan *et al.*, 2012 Variation in base-substitution mutation in experimental and natural lineages of *Caenorhabditis* nematodes. *Genome Biol. Evol.* 4: 513–522. <https://doi.org/10.1093/gbe/evs028>
- Dieterich, C., S. W. Clifton, L. N. Schuster, A. Chinwalla, K. Delehaunty *et al.*, 2008 The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nat. Genet.* 40: 1193–1198. <https://doi.org/10.1038/ng.227>
- Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2013 STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15–21. <https://doi.org/10.1093/bioinformatics/bts635>
- Estes, S., P. C. Phillips, D. R. Denver, W. K. Thomas, and M. Lynch, 2004 Mutation accumulation in populations of varying size: the distribution of fitness effects for fitness correlates in *Caenorhabditis elegans*. *Genetics* 166: 1269–1279. <https://doi.org/10.1534/genetics.166.3.1269>
- Fitch, W. M., 1967 Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. *J. Mol. Biol.* 26: 499–507. [https://doi.org/10.1016/0022-2836\(67\)90317-8](https://doi.org/10.1016/0022-2836(67)90317-8)
- Flynn, J. M., F. J. Chain, D. J. Schoen, and M. E. Cristescu, 2017 Spontaneous mutation accumulation in *Daphnia pulex* in selection-free vs. competitive environments. *Mol. Biol. Evol.* 34: 160–173. <https://doi.org/10.1093/molbev/msw234>
- Francioli, L. C., P. P. Polak, A. Koren, A. Menelaou, S. Chun *et al.*, 2015 Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* 47: 822–826. <https://doi.org/10.1038/ng.3292>
- Friedman, J., T. Hastie, and R. Tibshirani, 2010 Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33: 1–22. <https://doi.org/10.18637/jss.v033.i01>
- Frigola, J., R. Sabarinathan, L. Mularoni, F. Muñios, S. Gonzalez-Perez *et al.*, 2017 Reduced mutation rate in exons due to differential mismatch repair. *Nat. Genet.* 49: 1684–1692 [corrigenda: *Nat. Genet.* 50: 1196 (2018)]. <https://doi.org/10.1038/ng.3991>
- Garrison, E., and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. arXiv: 1207.3907v2 [q-bio-GN].
- Halligan, D. L., and P. D. Keightley, 2009 Spontaneous mutation accumulation studies in evolutionary genetics. *Annu. Rev. Ecol. Evol. Syst.* 40: 151–172. <https://doi.org/10.1146/annurev.ecolsys.39.110707.173437>
- Halligan, D. L., A. D. Peters, and P. D. Keightley, 2003 Estimating numbers of EMS-induced mutations affecting life-history traits in *Caenorhabditis elegans* in crosses between inbred sublines. *Genet. Res.* 82: 191–205. <https://doi.org/10.1017/S0016672303006499>
- Harris, T. W., I. Antoshechkin, T. Bieri, D. Blasiar, J. Chan *et al.*, 2010 WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.* 38: D463–D467. <https://doi.org/10.1093/nar/gkp952>
- Heilbron, K., M. Toll-Riera, M. Kojadinovic, and R. C. MacLean, 2014 Fitness is strongly influenced by rare mutations of large effect in a microbial mutation accumulation experiment. *Genetics* 197: 981–990. <https://doi.org/10.1534/genetics.114.163147>
- Hillier, L. W., R. D. Miller, S. E. Baird, A. Chinwalla, L. A. Fulton *et al.*, 2007 Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny. *PLoS Biol.* 5: e167. <https://doi.org/10.1371/journal.pbio.0050167>
- Hodgkinson, A., and A. Eyre-Walker, 2011 Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* 12: 756–766. <https://doi.org/10.1038/nrg3098>
- Huang, W., R. F. Lyman, R. A. Lyman, M. A. Carbone, S. T. Harbison *et al.*, 2016 Spontaneous mutations and the origin and maintenance of quantitative genetic variation. *Elife* 5: e14625 (erratum: *Elife* 5: e22300). <https://doi.org/10.7554/eLife.14625>
- Hudson, R. E., U. Bergthorsson, and H. Ochman, 2003 Transcription increases multiple spontaneous point mutations in *Salmonella enterica*. *Nucleic Acids Res.* 31: 4517–4522. <https://doi.org/10.1093/nar/gkg651>

- Katju, V., and U. Bergthorsson, 2019 Old trade, new tricks: insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches. *Genome Biol. Evol.* 11: 136–165. <https://doi.org/10.1093/gbe/evy252>
- Katju, V., L. B. Packard, L. Bu, P. D. Keightley, and U. Bergthorsson, 2015 Fitness decline in spontaneous mutation accumulation lines of *Caenorhabditis elegans* with varying effective population sizes. *Evolution* 69: 104–116. <https://doi.org/10.1111/evo.12554>
- Katju, V., L. B. Packard, and P. D. Keightley, 2018 Fitness decline under osmotic stress in *Caenorhabditis elegans* populations subjected to spontaneous mutation accumulation at varying population sizes. *Evolution* 72: 1000–1008. <https://doi.org/10.1111/evo.13463>
- Keightley, P. D., and A. Caballero, 1997 Genomic mutation rates for lifetime reproductive output and lifespan in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* 94: 3823–3827. <https://doi.org/10.1073/pnas.94.8.3823>
- Keightley, P. D., U. Trivedi, M. Thomson, F. Oliver, S. Kumar *et al.*, 2009 Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res.* 19: 1195–1201. <https://doi.org/10.1101/gr.091231.109>
- Keith, N., A. E. Tucker, C. E. Jackson, W. Sung, J. I. Lucas Lledó *et al.*, 2016 High mutational rates of large-scale duplication and deletion in *Daphnia pulex*. *Genome Res.* 26: 60–69. <https://doi.org/10.1101/gr.191338.115>
- Kim, N., and S. Jinks-Robertson, 2012 Transcription as a source of genome instability. *Nat. Rev. Genet.* 13: 204–214. <https://doi.org/10.1038/nrg3152>
- Klapacz, J., and A. S. Bhagwat, 2002 Transcription-dependent increase in multiple classes of base substitution mutations in *Escherichia coli*. *J. Bacteriol.* 184: 6866–6872. <https://doi.org/10.1128/JB.184.24.6866-6872.2002>
- Konrad, A., O. Thompson, R. H. Waterston, D. G. Moerman, P. D. Keightley *et al.*, 2017 Mitochondrial mutation rate, spectrum and heteroplasmy in *Caenorhabditis elegans* spontaneous mutation accumulation lines of differing size. *Mol. Biol. Evol.* 34: 1319–1334.
- Konrad, A., S. Flibotte, J. Taylor, R. H. Waterston, D. G. Moerman *et al.*, 2018 Mutational and transcriptional landscape of spontaneous gene duplications and deletions in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* 115: 7386–7391. <https://doi.org/10.1073/pnas.1801930115>
- Krasovec, M., A. Eyre-Walker, S. Sanchez-Ferandin, and G. Piganeau, 2017 Spontaneous mutation rate in the smallest photosynthetic eukaryotes. *Mol. Biol. Evol.* 34: 1770–1779. <https://doi.org/10.1093/molbev/msx119>
- Kuo, C. H., and H. Ochman, 2009 Deletional bias across the three domains of life. *Genome Biol. Evol.* 1: 145–152. <https://doi.org/10.1093/gbe/evp016>
- Li, H., 2011 Improving SNP discovery by base alignment quality. *Bioinformatics* 27: 1157–1158. <https://doi.org/10.1093/bioinformatics/btr076>
- Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Luijckx, P., E. K. H. Ho, A. Stanić, and A. F. Agrawal, 2018 Mutation accumulation in populations of varying size: large effect mutations cause most mutational decline in the rotifer *Brachionus calyciflorus* under UV-C radiation. *J. Evol. Biol.* 31: 924–932. <https://doi.org/10.1111/jeb.13282>
- McGaugh, S. E., C. S. Heil, B. Manzano-Winkler, L. Loewe, S. Goldstein *et al.*, 2012 Recombination modulates how selection affects linked sites in *Drosophila*. *PLoS Biol.* 10: e1001422. <https://doi.org/10.1371/journal.pbio.1001422>
- Meier, B., N. V. Volkova, Y. Hong, P. Schofield, P. J. Campbell *et al.*, 2018 Mutational signatures of DNA mismatch repair deficiency in *C. elegans* and human cancers. *Genome Res.* 28: 666–675. <https://doi.org/10.1101/gr.226845.117>
- Morgulis, A., E. M. Gertz, A. A. Schäffer, and R. Agarwala, 2006 A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.* 13: 1028–1040. <https://doi.org/10.1089/cmb.2006.13.1028>
- Ness, R. W., A. D. Morgan, R. B. Vasanthakrishnan, N. Colegrave, and P. D. Keightley, 2015 Extensive de novo mutation rate variation between individuals and across the genome of *Chlamydomonas reinhardtii*. *Genome Res.* 25: 1739–1749. <https://doi.org/10.1101/gr.191494.115>
- Ossowski, S., K. Schneeberger, J. I. Lucas-Lledó, N. Warthmann, R. M. Clark *et al.*, 2010 The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327: 92–94. <https://doi.org/10.1126/science.1180677>
- Ratan, A., T. L. Olson, T. P. Loughran, and W. Miller, 2015 Identification of indels in next-generation sequencing data. *BMC Bioinformatics* 16: 42. <https://doi.org/10.1186/s12859-015-0483-6>
- R Core Development Team, 2014 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. <http://www.R-project.org/>.
- Rimmer, A., H. Phan, I. Mathieson, Z. Iqbal, S. R. F. Twigg *et al.*, 2014 Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* 46: 912–918. <https://doi.org/10.1038/ng.3036>
- Rockman, M. V., and L. Kruglyak, 2009 Recombinational landscape and population genomics of *Caenorhabditis elegans*. *PLoS Genet.* 5: e1000419. <https://doi.org/10.1371/journal.pgen.1000419>
- Sanjuán, R., A. Moya, and S. F. Elena, 2004 The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc. Natl. Acad. Sci. USA* 101: 8396–8401. <https://doi.org/10.1073/pnas.0400146101>
- Saxena, A. S., M. P. Salomon, C. Matsuba, S. D. Yeh, and C. F. Baer, 2019 Evolution of the mutational process under relaxed selection in *Caenorhabditis elegans*. *Mol. Biol. Evol.* 36: 239–251. <https://doi.org/10.1093/molbev/msy213>
- Schrider, D. R., D. Houle, M. Lynch, and M. W. Hahn, 2013 Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. *Genetics* 194: 937–954. <https://doi.org/10.1534/genetics.113.151670>
- Sharbrough, J., M. Luse, J. L. Boore, J. M. Logsdon, Jr., and M. Neiman, 2018 Radical amino acid mutations persist longer in the absence of sex. *Evolution* 72: 808–824. <https://doi.org/10.1111/evo.13465>
- Sharp, N. P., and A. F. Agrawal, 2016 Low genetic quality alters key dimensions of the mutational spectrum. *PLoS Biol.* 14: e1002419. <https://doi.org/10.1371/journal.pbio.1002419>
- Smith, N. G., M. T. Webster, and H. Ellegren, 2002 Deterministic mutation rate variation in the human genome. *Genome Res.* 12: 1350–1356. <https://doi.org/10.1101/gr.220502>
- Smith, T. C. A., P. F. Arndt, and A. Eyre-Walker, 2018 Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. *PLoS Genet.* 14: e1007254. <https://doi.org/10.1371/journal.pgen.1007254>
- Stoltzfus, A., and R. W. Norris, 2016 On the causes of evolutionary transition: transversion bias. *Mol. Biol. Evol.* 33: 595–602. <https://doi.org/10.1093/molbev/msv274>

- Thorvaldsdóttir, H., J. T. Robinson, and J. P. Mesirov, 2013 Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14: 178–192. <https://doi.org/10.1093/bib/bbs017>
- Uchimura, A., M. Higuchi, Y. Minakuchi, M. Ohno, A. Toyoda *et al.*, 2015 Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. *Genome Res.* 25: 1125–1134. <https://doi.org/10.1101/gr.186148.114>
- Vogel, F., and G. Röhrborn, 1966 Amino-acid substitutions in haemoglobins and the mutation process. *Nature* 210: 116–117. <https://doi.org/10.1038/210116a0>
- Wakeley, J., 1996 The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends Ecol. Evol.* 11: 158–162. [https://doi.org/10.1016/0169-5347\(96\)10009-4](https://doi.org/10.1016/0169-5347(96)10009-4)
- Wang, X., Y. Zhao, K. Wong, P. Ehlers, Y. Kohara *et al.*, 2009 Identification of genes expressed in the hermaphrodite germ line of *C. elegans* using SAGE. *BMC Genomics* 10: 213. <https://doi.org/10.1186/1471-2164-10-213>
- Weller, A. M., C. Röderlsperger, G. Eberhardt, R. I. Molnar, and R. J. Sommer, 2014 Opposing forces of A/T-biased mutations and G/C-biased gene conversions shape the genome of the nematode *Pristionchus pacificus*. *Genetics* 196: 1145–1452. <https://doi.org/10.1534/genetics.113.159863>
- Yang, S., L. Wang, J. Huang, X. Zhang, Y. Yuan *et al.*, 2015 Parent-progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* 523: 463–467. <https://doi.org/10.1038/nature14649>

Communicating editor: M. Nachman