GENETIC VARIATION IN CAENORHABDITID NEMATODE WORMS

By

MATTHEW PHILIP SALOMON

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2011

To my brother, Craig

ACKNOWLEDGMENTS

I would like to thank the members of my graduate committee for all their support over the course of the past six years. Each of my committee members helped me to mature as a scientist, and for that I am very grateful: I truly appreciate the patience shown by Dr. Luciano Brocchieri as he assisted me in developing my programing skills, the open door to Dr. Edward Braun's office (where I could always receive impromptu advice), the big picture perspective and encouragement that Dr. Jamie Gilooly always provided, the guidance in science and life that Dr. Michael Miyamoto offered, and the amazing opportunity provided to me by Dr. Charles Baer. I can't find the words to convey my appreciation for the friendship and support that Charlie has provided, nor can I begin to express how much it means to me that he believed in me through this whole process. I would like to thank my best friend and wife, Gigi Ostrow, for always being there to support me over the past six years. Also, I would also like to thank all my friends for their support throughout my education. I was very fortunate to have the help of my brother Keith Salomon always was there to help me with any computer programing challenges I came across. Finally, I would like to thank my mother and father for everything they have done for me throughout my entire life. Without their love and support none of this would have been possible.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

H^2	Broad-sense heritability.
I_G	Total genetic variance scaled by the square mean.
I_M	Per-generation change in the mean with mutation accumulation.
μ	Molecular mutation rate.
N_e	Effective population size.
t_P	Persistence time.
S	Average selection coefficient against a new mutation.
U	Genomic mutation rate for fitness.
V_E	Environmental component of variance.
V_G	Total genetic variance.

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

GENETIC VARIATION IN CAENORHABDITID NEMATODE WORMS

By

Matthew Philip Salomon

August 2011

Chair: Charles F. Baer
Major: Zoology

When we observe the natural world we cannot help but notice the amount of variation all around us. Where does this variation come from? What maintains it? Traditionally most of the attention has focused on the role of natural selection and/or genetic drift as main evolutionary forces responsible for differences in standing genetic variation among taxa. That is, differences among taxa are due to variation in the strength or efficiency of selection and/or in the differences in effective population size among groups. However, much less attention has been paid to the alternative view that differences among taxa are reflections of their underlying rate of mutation.

The following three studies examine how differential mutational properties among and within closely related species manifests in varying levels of standing genetic variation observed in nature.

CHAPTER 1
INTRODUCTION

When we observe the natural world we cannot help but notice the vast amount of genetic variation all around us. Whether we observe this variation at the level of the phenotype or at molecular level, it is remarkable just how variable organisms are within a population or between different species. Attempting to explain this variability has been a driving motivation for many biologists, especially in the field of evolutionary genetics. Because of this, much of the energies of evolutionary biologists have focused on answering two fundamental questions: Where does variation come from? And what evolutionary forces maintain it?

In general, most of the attention has focused on the role of natural selection and/or genetic drift as the main evolutionary forces responsible for differences in standing genetic variation among taxa. That is, differences among taxa are due to variation in the strength or efficiency of selection and/or in the differences in effective population size among groups. However, much less attention has been paid to an alternative view, that differences among taxa may reflect differences in the underlying mutational process. The focus of the three studies presented here, are an attempt to examine how different mutational properties among and within closely related species manifest as varying levels of standing genetic variation.

The majority of this work has relied on estimating mutational properties in several sets of long-term mutation accumulation (MA) lines of closely related species of Rhadiditid nematode worms and has explored how these estimates compare to levels of standing genetic variation in natural populations. Prior work in the Baer lab has

demonstrated that, between the two related nematode species *Caenorhabditis elegans* and *C. briggsae, C. briggsae* accumulates mutations at twice the rate of *C. elegans*.

The first study looks at how different mutational properties between two closely related species affect levels of standing genetic variation for two quantitative traits, comparing the mutational variation estimated from mutation accumulation lines to the levels of standing genetic variation in a worldwide collection of natural isolates of both species to estimate the persistence times of new mutations affecting quantitative traits (fitness and body size). This work is motivated by the theory that at mutation-(purifying) selection balance (MSB) in a large population, the standing genetic variance for a trait (V_G) is predicted to be proportional to the mutational variance for the trait (V_M); V_M is proportional to the mutation rate for the trait. Thus, the ratio V_M/V_G predicts the average strength of selection (S) against a new mutation. By comparing V_M and V_G for lifetime reproductive success (≈ fitness) and body size in two species of self-fertilizing rhabditid nematodes, *Caenorhabditis briggsae* and *C. elegans*, the role of mutation can be contrasted between these two species to look at its effect on levels of standing genetic variation.

The goal of the second study is to complement the quantitative genetic results and examine whether the difference in mutation rates between *C. elegans* and *C. briggsae*, which were estimated from phenotypic traits, are reflected at the molecular level. To test this, estimates were generated of the rate and spectrum of di-nucleotide microsatellite repeats in four sets of 250-generation mutation accumulation (MA) lines, two sets in the species *Caenorhabditis briggsae* and two sets in *C. elegans*. These estimates were then compared to the standing variation at the same set of di-nucleotide

microsatellite loci for each species. In addition, the mutational properties of microsatellites are also compared to the cumulative effects of mutations on fitness in the same MA lines.

The third study examines how mating system (selfing vs. outcrossing) and chromosomal context (autosome vs. sex chromosome) affect the rate and spectrum of di-nucleotide microsatellite mutations using a set of MA lines derived from the outcrossing species of nematode worm*, C. remanei*. Theory predicts that natural selection should lead to a reduced mutation rate in selfing taxa and this study provides an ideal experimental test, at the molecular level, of this theoretical prediction. In addition, this study examines if mutation rates at a given locus are reflected in the number of alleles at the same locus in natural populations, thus allowing for comparisons on how differences in mating systems contribute to levels of standing genetic variation in related taxa.

CHAPTER 2
COMPARING MUTATIONAL AND STANDING GENETIC VARIABILITY FOR FITNESS AND SIZE IN *CAENORHABDITIS BRIGGSAE* AND *C. ELEGANS*

The genetic variation present in a population or species is a composite function of mutation, population size, and natural selection. Historically, efforts to understand differences (or similarities) between groups in genetic variation have focused on the interplay between population size and natural selection (Ohta 1973; Houle 1989; Byers and Waller 1999; Gillespie 2000; Lynch 2007). Considerably less attention has been paid to the possibility that differences among groups are due to systematic differences in the underlying rate of mutation. Although it has been known for a long time that mutation rates differ among taxa (Sturtevant 1937; Drake et al. 1998; Baer et al. 2007), the extent to which variation in mutation rate underlies variation among taxa in the standing genetic variation is poorly understood, particularly for quantitative traits.

The relationship between the standing genetic variance (V_G) and the per-generation input of genetic variance by mutation (the mutational variance, V_M) has a straightforward interpretation under two evolutionary scenarios. Under a deterministic mutation-(purifying) selection balance (MSB) model, $V_G \approx \frac{V_M}{S}$, where S is the average selection coefficient against a new mutation (Barton 1990; Crow 1993; Houle et al. 1996). Put another way, the ratio $\frac{V_G}{V_M}$ can be interpreted as the "persistence time" (t_P) of a new mutation, i.e., the expected number of generations a mutant allele is present in the (infinite) population before it is eliminated by selection (Crow 1993; Houle et al. 1996). The more deleterious the mutant allele, the faster it is removed from the population by selection. At the opposite extreme, under a strict neutral model at

mutation-drift equilibrium (MDE), for self-fertilizing taxa, $V_G \approx 4N_eV_M$, where N_e is the genetic effective population size (Lynch and Hill 1986). For a quantitative trait, $V_M = UE(a^2)$ where U is the genomic mutation rate and a is the additive phenotypic effect of a new mutation (Lynch and Walsh 1998, p. 329).

The unifying factor in these different scenarios is the mutational variance, V_M. Under both the MSB and MDE scenarios, we expect V_G to be proportional to V_M and thus the persistence time $t_P = \frac{V_G}{V_M}$ to be constant if selection is uniform. Changes in the relationship between V_G and V_M among groups must be due to the action of natural selection. Thus, if the persistence time differs between groups, the difference must be due to historical differences in the strength or efficiency of natural selection. This principle has been demonstrated by Houle et al. (1996), who found that the average persistence time for life history traits was about half that for morphological traits in a variety of taxa, consistent with the expected stronger correlation of life history traits with fitness . Average persistence times differed significantly between species. However, the traits considered varied between species, the species were phylogenetically disparate and included taxa not likely to be at equilibrium and/or to have experienced recent strong artificial selection (e.g., crop plants and inbred lines of mice). To our knowledge, there has been no comparison of mutational variance and standing genetic variance for the same traits in natural populations of related taxa.

In this study we report the standing genetic variation for two quantitative traits - lifetime reproduction weighted by survivorship (≈ fitness, which we call "Total Fitness" and designate *W*) and body size - in worldwide collections of two species of rhabditid nematodes, *Caenorhabditis briggsae* and *C. elegans*. Several lines of evidence

suggest that the mutation rate in *C. briggsae* is greater than that of *C. elegans* for these traits (Baer et al. 2005, 2006; Ostrow et al. 2007) and for microsatellites (Phillips et al. 2009), and perhaps for single nucleotide substitutions (D. Denver, CFB, et al., unpublished data). Thus, we predict that the standing genetic variance (V_G) should be consistently greater in *C. briggsae* than in *C. elegans*. If it is not, it suggests that idiosyncratic natural selection (e.g., genetic draft; Gillespie 2000) is of primary importance in shaping the standing genetic variation.

Methods and Materials

Wild Isolates

We initially obtained all the (at the time) publicly available strains of wild-caught nematodes in the species *Caenorhabditis briggsae* (6 strains) and *C. elegans* (40 strains) from the *Caenorhabditis* Genetics Center at the University of Minnesota in the spring of 2005. Subsequent to our initial assay in 2005, additional *C. briggsae* strains became publicly available. We obtained 50 additional strains of *C. briggsae* from various sources in the spring of 2007. Each strain is descended from a single wild-caught individual that was allowed to reproduce to large population size and then cryopreserved. Upon receipt, each strain was allowed to expand to large population size (~ 2 generations) under standard conditions of worm husbandry (Wood 1988), at which time three replicate sub-lines were initiated from a single individual. Each sub-line was inbred for six generations by randomly picking a single L4 stage (juvenile) worm and allowing it to self-fertilize. This inbreeding protocol is identical to the initial stage of our mutation accumulation (MA) protocol (Baer et al. 2005) and was done both to assure similar starting conditions of wild strains and MA stocks and to account for genetic variation resulting from within-strain heterozygosity.

Mutation Accumulation Lines and Fitness Assays

Details of the mutation accumulation and fitness assay protocols have been reported elsewhere (Vassilieva et al. 1999; Baer et al. 2005, 2006). All 40 *C. elegans* strains and six of the *C. briggsae* strains were assayed in the summer of 2005; the remaining 50 *C. briggsae* strains were assayed in the summer of 2007. Worms were assayed under standard conditions (20º C, fed on the OP50 strain of *E. coli*) except that density of agar plates was 30% greater to prevent worms from burying. Worms were counted using the same protocol as in Baer et al. (2005). At the beginning of the assay, each sub-line was replicated 10 times and taken through an additional three generations of single-worm transfer to account for parental and grand-parental effects. If a worm did not reproduce at the first generation the replicate was replaced; if the worm did not reproduce subsequent to the first generation the replicate was not included in the assay. Sample sizes therefore differ between strains and sub-lines. We define "Total Fitness" (W) of a worm as the lifetime reproductive output of an individual; $W = 0$ for worms that died prior to maturation or failed to reproduce.

Some plates became contaminated with mold prior to counting. Mold contamination was scored as None/Some/Heavy = 0/1/2 in 2005 and None/Present = 0/1 in 2007. On average, plates with mold contamination had fewer worms than uncontaminated plates. There are two potential non-exclusive explanations: we know that worms are more difficult to see on plates with mold present, leading to predictable undercounts. We suspect, but cannot prove, that worms eat mold spores and newly germinated molds, such that mold grows more quickly on plates with fewer worms. Based on extensive observation, we believe, but again cannot prove, that this mold does not have direct deleterious effects on the worms. To statistically account for the

effects of mold, we constructed a "mold index" (MI) by weighting the density of mold (0/1/2) by the average fraction of total reproduction contributed by that day's reproduction. For example, for *C. elegans* in 2005, the first day's reproduction (R1) accounted for 53.0% of total reproduction, R2 accounted for 40.7% and R3 accounted for 6.3%. Thus, if the three plates representing an individual worm's reproductive output were scored 2/0/1 representing heavy mold, no mold, and some mold on days 1, 2, and 3, respectively, the MI would be (2)(0.53)+(0)(0.407)+(1)(0.063) = 1.123. We included MI as a covariate in subsequent analyses of fitness; the effect was essentially zero for *C. briggsae* in 07; it was marginally significant ($0.10 < P < 0.11$) for *C. briggsae* in 2005 and highly significant for *C. elegans* ($P < 0.0001$) in 2005.

Body Size

Body volume was assayed using a slight modification of the protocol outlined in Ostrow *et al.* (2007). For each inbred line in each strain, we randomly chose half of the replicates in the fitness assay for inclusion in the body size analysis. Approximately 10 age-synchronized young adult worms (72-74 hrs post-laying) were randomly picked into a 1.5 ml microcentrifuge tube containing a fixative solution of 4% gluteraldehyde buffered with PBS. Fixed worms were then pipetted onto a slide micrometer and digitally photographed at 20X magnification using a Leica MZ75 dissecting microscope fitted with a Leica DFC280 camera controlled by the Leica IM50 software package (Leica Microsystems Imaging Solutions Ltd). Digital images were first manually procesed by removing the background surrounding each worm using Adobe Photoshop (version 6), leaving only the worm/s in the image. Edited images were then batch processed using a custom macro written in the ImageJ64 software package (http://rsb.info.nih.gov/ij/) (see supplemental information for ImageJ64 macro details).

After processing, the measured outline of each worm was visually inspected and all worms that were damaged (e.g., broken during pipetting) were removed from the final analysis. A subset of the images were analyzed by manually adjusting the threshold of the image and then outlining the worm using the "Analyze Particles" option in ImageJ; no significant difference between manually analyzed and batch processed images were found (data not shown). The area (*A*) and perimeter (*P*) of each worm were calculated and used to estimate body volume (*S*) under the assumption that the worm is cylindrical using the equation from Azevedo *et al.* 2002 (note that a typographical error in the original publication omitted the exponent in the numerator):

$$S = \frac{\pi\left(P + \sqrt{P^2 - 16A}\right)\left(P - \sqrt{P^2 - 16A}\right)^2}{256}$$

Data Analysis

Mutational Variance

For traits under directional selection (direct or indirect), the most meaningful measure of genetic variation is the genetic variance divided by the square of the trait mean. This quantity establishes the upper bound on the rate of response to selection and is commonly referred to as the "opportunity for selection", I_G, where $I_G = \frac{V_G}{\bar{z}^2}$ and $\bar{z}$ represents the trait mean (Crow 1958; Houle 1992; Wade 2006). The mutational variance V_M is half the among-line component of variance divided by the number of generations of mutation accumulation (Lynch and Walsh 1998, p. 331). Thus, $I_M = \frac{V_M}{\bar{z}^2} = \frac{1}{2t}\left(\frac{V_{L,t}}{\bar{z}_t^2} - \frac{V_{L,0}}{\bar{z}_0^2}\right)$ where t is the number of generations of MA, $V_{L,i}$ is the among-line

component of variance after *i* generations of MA and $\bar{z}_i$ is the trait mean at generation *i*. Values of I_M presented here are averages of the two strains of each species at t_{MA} = 200 and are from reanalysis of data reported elsewhere (Baer et al. 2005, 2006; Ostrow et al. 07). Empirical 95% confidence intervals for I_M were calculated using a bootstrap resampling method outlined in Baer et al. (2005). Briefly, a pseudo-data set was initially constructed by sampling the data with replacement at the level of line (i.e., all replicates within a line were included), maintaining the block structure of the original data. Means and variance components of the pseudo-data set were estimated separately for MA and control lines within each of the four strains by REML as implemented in the MIXED procedure of SAS v. 9.2. For the trait *W*, we used the model *y* = *Block* + *Line* + *Error*; for body volume we used the model *y* = *Block* + *Line* + *Replicate(Line)* + *Error*. The calculation was repeated 1000 times; the upper and lower 25 pseudo-estimates establish the approximate 95% confidence interval. The average of the two strains within each species for each pseudo-replicate is taken as the species average.

Standing Genetic Variance

Variance components were estimated for each species using restricted maximum likelihood (REML) as implemented in the MIXED procedure of SAS v. 9.2. Since different strains of *C. briggsae* were assayed in different years (2005, 2007), for the trait *W* we initially analyzed the model *y* = *MI* + *Year* + *Strain*(*Year*) + *Inbred_Line*(*Strain*(*Year*)) + *error*, with mold index (*MI*) as a covariate, year considered a fixed effect and the other effects random. Means did not differ significantly among years, so subsequent analyses were done without regard to assay year. For the trait body volume we initially considered the model *y* = *Year* + *Strain*(*Year*) +

Inbred_Line(*Strain*(*Year*)) + *Replicate*(*Inbred_Line*(*Strain*(*Year*))) + *error*. *C. briggsae* measured in 2005 were marginally smaller than those measured in 2007, so we retained year in the model for subsequent analyses.

Population genetic structure introduces a potential complication into the analysis. Strictly speaking, the approximation $V_G \approx \frac{V_M}{S}$ is a statement about within-population variation. However, the model tacitly assumes that deleterious mutations never reach appreciable frequency in the population, so it is reasonable to assume that sub-populations do not diverge substantially due to genetic drift, and positive selection is ignored in any case. Both *C. elegans* and *C. briggsae* show a species-wide clade structure, with two deep clades in *C. elegans* (Denver et al. 2003) and three in *C. briggsae* (Dolgin et al. 2008). The clades in *C. briggsae* largely reflect geography, with the three lineages representing Temperate, Tropical, and Equatorial clades (Dolgin et al. 2008), whereas the deep clades in *C. elegans* are not geographically structured. In addition, some collections contain several individuals from the same collecting location, but many locations include only a single individual. Almost half (19/40) of the *C. elegans* strains were collected in a single location (Roxel; see supplementary Table 3). Moreover, because of the nature of *Caenorhabditis* life history (self-fertile, very short generation time) it is possible that some or all of the strains within a collecting location are descended from the same ancestral worm only a few generations back. To accommodate the variable collecting structure, we further subdivide the data into "locations", some of which contain multiple strains and some of which do not. For the trait *W*, the full model is: *y = MI + Clade + Location(Clade) + Strain(Location(Clade)) + Inbred_Line(Strain(Location(Clade))) + error*, with MI as a covariate, clade as a fixed

effect and the others random. Clade identity is unknown for 6/40 *C. elegans* lines from two locations. The effect of clade was not significant for either trait in either species (*C. briggsae*, P>0.13 or greater in all cases) so we omitted clade from further analyses.

In the above analysis, the environmental component of phenotypic variance (V_E) is represented by the error (i.e., within inbred line) variance. Under an additive model, the total genetic variance (V_G) is represented by half the sum of the remaining components of variance, and can be partitioned into within- and among-group components. Note that we assume that each inbred line is homozygous at all loci, which cannot be strictly true. The among-inbred line component of variance is twice the genetic variance within each strain resulting from polymorphism in the ancestral worm and the among-strain component of variance represents twice the genetic variation among individuals within a sampling location.

Standard errors and 95% confidence limits for I_G were calculated using a delete-one jackknife protocol in which each strain was sequentially deleted from the data and the mean and variance components of the redacted data set re-calculated as described above (Knapp et al. 1989). Confidence limits were calculated using the jackknife standard error and the usual Student's *t*-distribution formulation (Sokal and Rohlf 1981, p. 145).

Standard errors of persistence time (I_G/I_M), average selection coefficient (I_M/I_G), and the ratios of the mutational and standing variances between species ($I_{M,Cbr}/I_{M,Cel}$ and $I_{G,Cbr}/I_{G,Cel}$) were determined using the Delta method for the variance of a ratio where $SE\left(\frac{x}{y}\right) = \left(\frac{x}{y}\right)\{\lfloor CV(x)\rfloor^2 + \lfloor CV(y)\rfloor^2\}^{1/2}$ and *CV(x)* is the coefficient of sampling variation (the

ratio of the standard error to the estimated value) of x (Lynch and Walsh 1998, equation A1.19b; Vassilieva et al. 1999). 95% confidence limits were calculated as above.

We also report the broad-sense heritability, H^2, defined as the total genetic variance scaled as a fraction of the environmental variance, i.e., $H^2 = V_G/V_E$.

Results

Fitness

Summary statistics are presented in Table 1. First, we can safely rule out the mutation-drift equilibrium (MDE) scenario. At MDE, $V_G \approx 4N_eV_M$; given the observed V_G and V_M, N_e would have to be on the order of a few dozen individuals to explain the results, an implausibly small number (Cutter et al. 2006). This is unsurprising, given that we expect lifetime reproductive output to be closely related to fitness; however, ruling out MDE is a necessary first condition.

Under the null hypothesis that the standing genetic variance is entirely explained by the mutational variance, we predict the ratio of the standing variances in the two species, $I_{G,Cbr}/I_{G,Cel}$ should equal the ratio of mutational variances, $I_{M,Cbr}/I_{M,Cel}$, or put another way, the persistence times should be the same in the two species. The ratio of the mutational variances of *C. briggsae* to *C. elegans*, $I_{M,Cbr}/I_{M,Cel}$ is 3.99, with the 95% bootstrap confidence interval between 1.78 and 8.01. The observed species-wide value of $I_{G,Cbr}/I_{G,Cel}$ is 1.85, with the asymptotic 95% confidence interval between 0 and 5.07. Thus, although the ratio of the standing variances in the two species is somewhat less than predicted from the respective mutational variances (i.e., too little variance in *C. briggsae* and/or too much in *C. elegans*), the null hypothesis of $I_{G,Cbr}/I_{G,Cel} = 3.99$ cannot be rejected.

Looking across species and populations, with one exception (the Viosne population of *C. briggsae*, to which we will return), estimated (homozygous) selection coefficients (I_M/I_G) are quite consistent and fall within a relatively narrow range of a few percent (Table 1, last row). Broad-sense heritabilities are on the order of 5-15%. These results are similar to analogous results from *Drosophila melanogaster* and *Daphnia pulex*, which suggest heterozygous selection coefficients of mutations affecting components of fitness (e.g., viability) on the order of a few percent, and comparable heritabilities (Houle et al. 1996). Unfortunately, our results cannot be directly compared to results from other taxa, for two reasons. First, because there are no similar estimates of lifetime reproductive success (≈ absolute fitness) for other taxa (we note that absolute fitness is much easier to measure in worms than in most other organisms because reproduction is both rapid and confined to a few days), and second, because these Caenorhabditis species are (believed to be) predominantly selfing, the relevant selection coefficient against new mutations is presumably the homozygous effect, whereas in obligately outcrossing taxa such as Drosophila it is the heterozygous effect that is estimated by V_M/V_G. To our knowledge, there are no analogous data from any other predominantly selfing organism.

Selection coefficients inferred from the ratio of mutational to standing variance (= V_M/V_G) can be compared to those calculated directly from the MA data (typically referred to as E(a) in the MA literature). Selection coefficients calculated from MA data are widely believed to be overestimates, perhaps gross overestimates, because mutations of very small effect are difficult to detect with the available methods (Keightley and Eyre-Walker 1999; Davies et al. 1999). However, the standing genetic variance has

presumably accumulated over a long time, such that the accumulation of many alleles of small effect would contribute to the total genetic variance. If so, we expect V_G to be much larger than V_M and the inferred selection coefficient to be very small. In fact, the selection coefficients inferred from this study (~0.02) are somewhat smaller than those estimated from the MA data (~0.05-0.2; Vassilieva et al. 2000; Baer et al. 2005; Begin and Schoen 2006), perhaps by as much as an order of magnitude. This implies that estimates of genomic mutation rates (U) calculated from MA data are underestimates of similar order. This conclusion is reinforced by indirect estimates of U inferred from direct sequencing of MA lines using the method of Kondrashov and Crow (1993), in which (diploid) U for *C. elegans* was estimated to be on the order of 1 per generation (Denver et al. 2004).

The relationship $V_G \approx \frac{V_M}{S}$ (and thus $t_P = \frac{V_G}{V_M}$) is valid only when the population is (1) large, and (2) at mutation-selection balance (Keightley and Hill 1988). Together, these two assumptions imply that the standing genetic variation observed within a single population will be similar to the genetic variation present in the entire species, because deleterious alleles will never achieve appreciable frequency in any population, precluding divergence due to fixation of slightly deleterious alleles. If positive selection and/or genetic drift had caused populations to diverge, there should be substantially more variation present in the species as a whole than in any individual population. Again with one exception, the results appear to lend credence to this assumption. The point estimate for the standing genetic variance for fitness (I_G) in the worldwide collection of *C. elegans* not including the Roxel population (N=21 lines) is 0.011; it is twice that in the Roxel population alone (I_G = 0.026, N=19 lines). The species-wide

standing genetic variance I_G for fitness in *C. briggsae* (N=50 lines) is 0.027; within the Merlet population (N=12 lines) it is 0.058. The non-significant effect of clade in either species is consistent with the deterministic MSB scenario.

The inconvenient exception is the Viosne population of *C. briggsae* (N=12 lines), which is genetically depauperate (I_G = 0.0015). Taken at face value, the point estimate of the selection coefficient ($S = I_M/I_G$) is 0.59, implying very strong purifying selection acting solely within this population. A more plausible scenario is that the worms in the Viosne sample are not at any kind of equilibrium, having descended from a single hermaphroditic founder in the very recent past. In fact, given the reproductive biology of Caenorhabditis, we find it somewhat surprising that the other populations do not exhibit the same pattern.

The fact that the average lifetime reproductive success ($\overline{W}$) we report here for *C. elegans* (~135) is substantially less than is typically reported (>200, e.g., Hodgkin and Doniach 1997) deserves comment. The low fitness cannot be primarily attributed to undercounts caused by moldy plates, because the total fecundity of individuals that reproduced entirely on plates with no mold was not appreciably larger than on plates that did have mold (an average difference of ~7 worms, or about 5%). It seems likely that some combination of systematic undercounts for reasons unrelated to mold and/or some factor in our worm husbandry contributes to low fecundity. However, many previous fitness assays with mutation accumulation lines using essentially identical techniques have not resulted in similarly low counts. The fact that some lines had fecundity that was within the normal range and some lines (and inbred lines within lines) had very low fitness (e.g., < 20 offspring) suggests that there might be inbreeding

depression, although a previous study with a subset of these wild isolates of *C. elegans* found no evidence for inbreeding depression (Dolgin et al. 2007). An alternative explanation is that different (sub)lines have differences in the extent of heteroplasmy of deletions in the mitochondrial genome, which are known to be correlated with fitness and are much more prevalent in *C. briggsae* than in *C. elegans* (Howe and Denver 2008).

Body Volume

Summary statistics are presented in Table 1. The ratio of the mutational variances for body size in the two species, $I_{M,Cbr}/I_{M,Cel}$, is 1.69, with the 95% bootstrap confidence interval between 0.43 and 3.61. The observed species-wide value of $I_{G,Cbr}/I_{G,Cel}$ is 1.92, with the asymptotic 95% confidence interval between 0 and 4.57, very close to the value predicted under the null hypothesis of equal ratios of standing genetic and mutational variances.

Two patterns emerge from the data on body volume. First, estimated (homozygous) selection coefficients (I_M/I_G) are very similar in all cases and are nearly equal in the two species. The Viosne population of *C. briggsae* is again less genetically variable than the others, although the difference is less than for fitness. Just as for fitness, the species-wide genetic variance for body size is not substantially larger than the genetic variance present within individual populations, lending credence to the MSB interpretation of the relationship between standing genetic and mutational variance. Second, and interestingly, the inferred selection coefficients for body volume are of the same order as those for fitness, and if anything are slightly larger. At first glance this seems counterintuitive; we expect "Total Fitness" (a function of fecundity and survivorship) to be under strong direct selection. A possible resolution is that the

genetic architectures of fitness and body volume differ in such a way that many alleles have small effects on fitness, whereas only alleles with larger effects cause body size to vary.

These species of Caenorhabditis are believed to be primarily self-fertilizing in nature, and are expected to be largely homozygous. However, in *C. briggsae*, for both fitness and body volume, approximately 25% of the total genetic variance is among inbred lines within a line; the among-inbred line component is significantly different from 0 for body size but not for fitness. Presumably, the among-inbred line variance results from residual heterozygosity in the ancestor, although new mutations arising during the six generations of inbreeding could also contribute. Conversely, for the worldwide collection of *C. elegans*, the fraction of the genetic variance attributable to variation among inbred lines within a line is about 2% for fitness and 0 for body volume. In the Roxel population of *C. elegans*, however, about half the genetic variance for fitness is among inbred lines, although the REML estimate of the among-inbred line component of variance is 0 for body volume. A prosaic explanation is that some of the *C. elegans* lines were kept in lab culture before cryopreservation, thus more of the residual genetic variation has been purged in *C. elegans* than in *C. briggsae*. The qualitative difference of the Roxel population from the species-wide collection of *C. elegans* reinforces that interpretation. If in fact some of the genetic variation originally present in *C. elegans* has been purged subsequent to collection, it implies that the estimates of persistence times are underestimates.

If differences in mutational variance between species are due to differences in mutation rate and not the distribution of effects and/or the mutational target size, the

ratio $I_{M,Cbr}/I_{M,Cel}$ is expected to be equal for all traits. Species-wide $I_{M,Cbr}/I_{M,Cel}$ is 3.99 for total fitness and 1.76 for body volume; the difference is not significant. The ratio of the standing genetic variances, $I_{G,Cbr}/I_{G,Cel}$, does not differ between traits (Species-wide $I_{G,Cbr}/I_{G,Cel} \approx 1.8$ for Total Fitness and 1.9 for body volume). The close concordance of $I_{G,Cbr}/I_{G,Cel}$ between the two traits is at least suggestive that similar factors (e.g., differences in mutation rate) shape the standing genetic variance for both traits.

An important caveat, however, is that the values of the mutational properties (i.e., I_M) assigned to each species are derived from estimates from only two starting genotypes. We can put legitimate confidence limits on the estimates, but they represent a fixed effect, i.e., the sampling universe is the two strains represented in this experiment. In reality the mutational variance is a random effect and we have little understanding of the scope of the variation in mutational properties within and between species; it is entirely possible that, on average, the mutational properties of *C. elegans* and *C. briggsae* are identical in every way and the variation we observe is simply due to sampling a very small number (two) of starting genotypes. This consideration is of course not unique to Caenorhabditis.

Finally, some perspective is in order with respect to power. The way we analyzed the data, by nesting lines within locations, is very conservative, providing a sample size of only 15 or 16, respectively. If we are willing to accept that genetic variation among locations is not much greater than within locations (the Viosne population of *C. briggsae* notwithstanding), we can legitimately increase power by resampling over lines rather than locations. However, the results for the basic statistics - I_G and H^2 - do not change substantially. For example, the upper 95% confidence limit on I_G in *C. elegans*

decreases from 1.9 times the point estimate to 1.6 times the point estimate with an increase in sample size from 15 to 40 (the point estimates are within a few percent). This exercise is obviously not the same as actually sampling an additional 25 locations, but it does suggest that doing so might not change the results much.

Discussion

Two primary results emerge from this study. First, the strength of selection acting on mutations affecting both fitness and body size is consistently inferred to be on the order of a few percent. This result is broadly compatible with results from other taxa. Second, the total genetic variance in a worldwide collection does not appear to be very much greater than the genetic variance within sampling locations (with one exception). Taken together, these results suggest that genetic variation for these traits is maintained by mutation-(purifying) selection balance. However, there is an alternative possibility. The relative consistency of genetic variance within and between populations and species, and especially between traits, suggests that recurrent hitchhiking events - genetic draft (Gillespie 2000) - may play a predominant role in structuring the genetic variation in these species. Genetic draft is likely to be particularly important in predominantly self-fertilizing taxa. Moreover, the relative consistency of persistence time (V_G/V_M) across disparate taxa and traits (~ 50 for life history traits; Houle *et al.* 1996; this study) is intriguing. Draft theory predicts relative independence of standing molecular genetic variation from population size (Gillespie 2000). Equivalent theory for quantitative traits does not exist, but the parallels are obvious.

Comparisons of standing and mutational variation between taxa and traits provide the best way to tease apart the relative contributions of neutral and non-neutral processes in shaping quantitative genetic variation. We envision two classes of studies

that would be particularly informative, especially in tandem. First, the best-characterized metazoan species (*Drosophila melanogaster, Daphnia pulex, Caenorhabditis*) all have large effective population sizes, such that the deterministic MSB formulation is plausible. It would be extremely interesting to perform the kind of study we report here in groups with known large differences in population size. Second, it would be useful to consider multiple traits with different expected levels of selection, in the spirit of Houle et al. (1996). Discerning among the alternative possibilities (e.g., draft, uniform mutational properties) will require that studies include very many genotypes from multiple populations.

Table 2-1. Total Fitness (W). See text for details of calculations. 95% confidence interval follows point estimate in parentheses. Abbreviations are: U : % per-generation change in the mean with mutation accumulation; IM: per-generation increase in the genetic (mutational) variance, scaled by the square of the mean; VLOC: component of genetic variance among locations (N = # of locations); VLINE: component of genetic variance among lines within locations (N = # of lines): VINLINE: component of genetic variance among inbred lines within lines; VE: environmental (= error) component of variance; VG: total genetic variance; : mean lifetime reproductive output; IG: total genetic variance scaled by the square of the mean; H2: broad-sense heritability; tP: persistence time; S: average selection coefficient against new mutations.

Estimate	$U\bar{a}$ $\times 10^3$	I_M $\times 10^4$	V_{LOC} (N)	V_{LINE} (N)	V_{INLINE}	V_E	V_G	$\overline{W}$	I_G	H^2	t_P ($=I_G/I_M$)	S
C. briggsae (all lines)	-2.89 (-3.17, -2.60)	8.96 (4.22, 13.91)	252.8 (16)	68.3 (51)	129.4 (0, 287.6)	2034.2	225.3 (20.9, 429.6)	90.8 (80.8, 100.8)	0.027 (0.0008, 0.054)	0.11 (0.007, 0.21)	30.5 (0, 65.0)	0.033 (0, 0.070)
Cbr (Merlet)	-	-	-	392.8 (12)	187.9 (0, 562.7)	1634.3	290.4 (0, 779.0)	71.0 (55.1, 86.8)	0.058 (0, 0.163)	0.17 (0, 0.49)	65.2 (0, 188.3)	0.015 (0, 0.044)
Cbr (Viosne)	-	-	-	0.13 (12)	23.0 (0, 119.1)	1853.4	11.6 (0, 59.1)	87.4 (82.4, 92.4)	0.0015 (0, 0.008)	0.006 (0, 0.033)	1.7 (0, 8.8)	0.59 (0,1)
C. elegans (all lines)	-1.07 (-1.38, -0.75)	2.24 (1.34, 3.30)	185.1 (13)	280.8 (34)	73.5 (0, 316.6)	4113.1	269.7 (0,548.5)	140.5 (123.8, 157.1)	0.0137 (0, 0.0290)	0.07 (0.0013, 0.130)	61.0 (0, 134.9)	0.016 (0, 0.036)
Cel (Roxel only)	-	-	-	399.5 (19)	306.9 (0,1171.2)	4433.6	353.2 (0, 779.1)	116.3 (97.1, 135.6)	0.026 (0, 0.063)	0.08 (0, 0.19)	116.8 (0, 291.0)	0.009 (0, 0.021)
Cel (Roxel omitted)	-	-	131 (12)	192.4 (15)	28.9 (0, 197.8)	3864	176.4 (0, 418.4)	145.1 (129.5, 160.7)	0.0084 (0, 0.020)	0.05 (0, 0.11)	37.4 (0, 92.0)	0.027 (0, 0.066)

Table 2-2. Body volume. See text for details of calculations. 95% confidence interval follows point estimate in parentheses. $\bar{z}$: mean body volume (mm3), all other abbreviations are the same as in Table 1.

Estimate	$U\bar{a} \times 10^3$	$I_M \times 10^4$	V_{LOC} (N)	V_{LINE} (N)	V_{INLINE}	V_{REP}	V_E	V_G $(x\ 10^9)$	$\bar{z}$ $(x\ 10^3)$	I_G $(x\ 10^3)$	H^2	t_P $(=I_G/I_M)$	S
C. briggsae (all lines)	-1.43 (-1.77, -1.10)	2.48 (0.30, 4.66)	7.85 E-9 (16)	7.03 E-9 (51)	4.53 E-9 (1.49, 7.57 E-9)	4.04 E-8	2.48 E-8	9.71 (4.14, 15.3)	1.23 (1.15, 1.30)	6.45 (2.46, 10.44)	0.15 (0.05, 0.25)	26.0 (0, 55.1)	0.038 (0, 0.081)
Cbr (Merlet)	-	-	-	1.44 E-8 (11)	5.83 E-9 (0, 1.47 E-8)	3.02 E-8	2.14 E-8	10.10 (0, 25.2)	1.06 (0.96, 1.15)	9.02 (0, 21.93)	0.20 (0, 0.50)	36.4 (0, 99.2)	0.027 (0, 0.075)
Cbr (Viosne)	-	-	-	3.83 E-9 (12)	4.18 E-9 (0, 1.71 E-8)	4.28 E-8	2.54 E-8	4.00 (0, 11.41)	1.20 (1.14, 1.27)	2.76 (0, 7.82)	0.06 (0, 0.17)	11.2 (0, 34.1)	0.090 (0, 0.275)
C. elegans (all lines)	-0.68 (-0.96, -0.40)	1.54 (0.89, 2.20)	6.71 E-10 (13)	2.08 E-8 (34)	5.06 E-11 (0, 1.36 E-9)	1.12 E-7	8.86 E-8	10.74 (4.26, 17.23)	2.15 (2.06, 2.23)	2.33 (0.86, 3.87)	0.053 (0.018, 0.089)	15.2 (3.3, 26.9)	0.066 (0.014, 0.118)
Cel (Roxel only)	-	-	-	2.08 E-8 (19)	0	1.42 E-7	1.04 E-7	10.41 (0, 24.83)	2.15 (20.56, 22.50)	2.25 (0, 5.35)	0.04 (0, 0.10)	14.6 (0, 35.7)	0.069 (0, 0.169)
Cel (Roxel ommitted)	-	-	2.12 E-9 (12)	1.86 E-8	8.51 E-11	1.01 E-7	7.55 E-8	10.42 (8.3 E-10, 20.02)	2.12 (2.03, 2.22)	2.32 (0, 7.55)	0.10 (0.01, 0.20)	10.31 (0, 21.3)	0.097 (0, 0.201)

CHAPTER 3
SPONTANEOUS MUTATIONAL AND STANDING GENETIC (CO)VARIATION AT DINUCLEOTIDE MICROSATELLITES IN *CAENORHABDITIS BRIGGSAE* AND *C. ELEGANS*

For many reasons, understanding evolution requires understanding mutation - the rate at which mutations occur, the molecular spectrum, and their effects on fitness. First, the standing genetic variation within a population (H) is a composite function of the effective population size (N_e) and the mutation rate (μ): at equilibrium, $\hat{H} \approx 4N_e\mu$ (Hartl and Clark 2007). Differences in the standing genetic variation between populations or species, or between regions of the genome in the same species, may be due to differences in N_e, in mutation rate, or both; N_e itself depends on many factors, including natural selection (Hill and Robertson 1966). Similarly, the rate of neutral divergence (k) between taxa is equal to the neutral mutation rate (μ_0) (Kimura 1968). Differences between lineages in the rate of molecular evolution may be due either to differences in the absolute mutation rate (μ) or to differences in the fraction of mutations that are effectively neutral, which is a function of N_e.

Second, the evolution of genome size and/or composition may depend both on natural selection (e.g., small genomes may be favored due to the increased speed of replication) and on the mutational process (e.g., a deletion bias will lead to a reduction in genome size). Differences among taxa in genome size and/or composition, or in the properties of particular genomic components (such as introns) may result from the effects of selection, mutation, or both.

Third, the cumulative effect on fitness of deleterious mutations depends on both the rate (U) and the distribution of fitness effects ($g[a]$) (Lynch and Walsh 1998), and those parameters are of utmost importance in evolutionary theory. However, the rate

and distribution of effects are confounded both statistically (Begin and Schoen 2006) and conceptually (Baer, Miyamoto, and Denver 2007), and taxa that differ in the degree to which they suffer the effects of deleterious mutation may have different mutation rates, different distributions of mutational effects, or both.

The common theme of these examples is that the manifestation of mutational processes is almost always confounded with natural selection and/or population size, from which it follows that an unambiguous characterization of mutational processes can greatly facilitate understanding of evolution. The most effective way to dissociate mutational processes from selection is to allow mutations to accumulate at very small N_e, thereby minimizing the efficiency of selection (Kimura 1962; Kondrashov, Ogurtsov, and Kondrashov 2006). Here we report the results of a comparative study of the mutational properties of dinucleotide short tandem repeat - STR, or microsatellite - loci in two species of Rhabditid nematode, *Caenorhabditis briggsae* and *C. elegans*. Mutations were allowed to accumulate at $N_e \approx 1$ for 250 generations; selection has thus been "turned off" as much as is possible.

We target STRs for scrutiny for three reasons. First, we are interested in the relationship between the mutation rate at a well-defined class of molecular loci and the overall impact of new mutations on fitness. We have previously documented significant variation in the cumulative effects of new mutations on fitness and body size between and potentially within *C. briggsae* and *C. elegans* (Baer et al. 2005; Baer et al. 2006; Ostrow et al. 2007; Baer 2008). The weight of the evidence suggests a scenario in which *C. briggsae* declines in fitness at about twice the rate as *C. elegans*. However, the mutational decay of fitness may differ between groups due to a difference in the

distribution of mutational effects rather than in mutation rate. To date there is very limited information on the relationship between the rate and spectrum of new mutations at the molecular level (μ) and the genomic mutation rate for fitness (U). It would be of considerable interest to identify an easily-screened class of marker loci whose mutational properties vary with U in a consistent way.

Second, we are interested in the relationship between standing variation and mutation rate. Almost all studies that report variation in "mutation rate" infer mutation rate indirectly, from standing genetic variation and/or divergence among taxa at a class of loci *assumed* to be evolving neutrally. Any inference derived from such a study is only as robust as the underlying assumptions. There have been remarkably few studies (we know of only one in a eukaryote) in which the standing variance at a set of loci has been directly compared to the *demonstrated* mutation rate at the same set of loci. A strong positive correlation between the standing variance and the demonstrated mutation rate provides the best possible justification for inferring mutation rate from standing variation. STR loci are ideal for this purpose because their high mutation rate provides much greater power than could be obtained from other classes of loci (e.g., base substitutions at single nucleotides).

Third, an influential model of genome evolution invokes a general mutational bias in favor of small deletions relative to insertions as a driving force behind the evolution of genome size (Petrov 2001; Petrov 2002). A survey of random nuclear mutations in the *C. elegans* genome found a significant excess of short insertions relative to deletions (Denver et al. 2004), contrary to the phylogenetic pattern observed in nuclear pseudogenes (Witherspoon and Robertson 2003). Similarly, a survey of STR mutations

in the N2 strain of *C. elegans* showed a significant insertion bias (Seyfert et al. 2008), although repeat motifs were not uniformly represented, nor was an effect of repeat motif tested in that study.

Finally, our most fundamental motivation is "genomic natural history". The study is part of an ongoing effort to understand the factors underlying taxonomic variation in the genomic mutation rate. Evolutionary theory provides clear guidance regarding the evolution of mutation rate with respect to mating system and chromosomal context (reviewed in Drake et al. 1998; Sniegowski et al. 2000; Baer, Miyamoto, and Denver 2007), but there have been very few empirical studies conducted in a systematic comparative context. The two species studied here have very similar life histories, thus there is no *a priori* reason to expect systematic differences in the strength of selection on mutation rate.

Materials and Methods

Mutation Accumulation Lines

The MA protocol has been described in detail elsewhere (Vassilieva and Lynch 1999; Baer et al. 2005). Briefly, highly inbred stocks of each strain were replicated 100 times and perpetuated by single-hermaphrodite transfer for 250 generations. This protocol results in a genetic effective population size of $N_e \approx 1$ (the approximation is the result of occasionally having to use backup stocks of worms when the original worm did not survive), thereby minimizing the efficiency of natural selection and ensuring that all but the most deleterious mutations behave according to neutral dynamics.

Choice of Loci and Primer Design

We initially identified all dinucleotide repeat loci of ≥ 5 perfect repeats in the draft *C. briggsae* genome (strain AF16) by BLAST search against the NCBI nr database (ca.

October 2002) for the oligonucleotide sequence $XX_{(5)}$, where XX is the dinucleotide sequence AC, AG, AT, CG, using the "short, nearly exact match" algorithm. The dinucleotide repeat and 200 bp of upstream and downstream flanking sequence were saved and screened for duplicates by pairwise BLAST search of the flanking sequence. From the list of unique sequences we chose at random 96 loci of at least five repeats, allowing at most one nucleotide indel over the entire repeat sequence. Primers were designed using Primer3 software (Rozen and Skaletsky 2000) using the default parameters with an optimum fragment length of 250, a minimum(maximum) allowable fragment length of 150(350) bases, and a minimum distance of 10 bases between the repeat and the primer termini. Loci for *C. elegans* were chosen from those previously published by (Sivasundar and Hey 2003) and (Frisse 1999) and supplemented with loci chosen randomly from the *C. elegans* genome (WS137).

Presence and (approximate) length of microsatellite loci were confirmed by direct sequencing of cloned PCR products for most loci in both species (Appendix A). PCR products were cloned into TOPO TA cloning vectors (Invitrogen, Carlsbad CA) and sequenced on an ABI 3730 automated capillary sequencer at the University of Florida's Interdisciplinary Center for Biotechnology Research (ICBR). Sequences were aligned in Sequence Viewer 4 (CLC Bio, Cambridge MA) using default parameters. Microsatellite repeat number was determined by taking the average of at least two sequencing reads per locus and searching for dinucleotide repeat motifs with the PHOBOS algorithm (Christoph Mayer, Ruhr-Universität Bochum).

Genotyping

Genomic DNA was extracted from two replicate cultures of each MA line and the ancestral control stock using a modified protocol of (Williams et al. 1992). We

employed a nested PCR strategy with fluorescently tagged primers, via a modification of the "three-primer" method of (Schuelke 2000). PCR reactions of 15 µl were done in 96-well plates, using 1 ul of DNA sample, 60 pmol of selective primer, 6 pmol of M13-tail primer, 60 pmol of labeled M13 primer, 1.5 mM MgCl2, 10 mM of dNTPs, and 0.375 units of Eppendorf Taq polymerase. Reactions were initially run for 10 cycles of 40 seconds denaturing at 94°C, 40 seconds annealing at 60°C, and 40 seconds extension at 72°C; the annealing temperature was then decreased to 48°C and the reaction was continued for an additional 30 cycles.

Three different fluorescent labels were used (fam, hex and ned), with only a single label used for a given reaction (= locus). Products labeled with different fluorescent tags were pooled and analyzed using an Applied Biosystems 3730XL DNA analyzer (The Center for Applied Genomics, Hospital for Sick Children, Toronto, ON, Canada). Data were analyzed using Genemapper 3.0 software (Applied Biosystems) and GeneMarker (SoftGenetics). Fragment length was established relative to a known size-standard ladder (GeneScan 500, Applied Biosystems). We employed an iterative binning procedure to identify putative mutants. In the first iteration we calculated the mean fragment length for all replicates. Fragments that deviated by > 1.5 bases from the mean were removed from the dataset and the mean recalculated. Alleles that differed by > 1.5 bases from the re-calculated mean were scored as putative mutants. Putative mutants were then re-amplified from an independent DNA extraction and re-genotyped as above. Putative mutants that had the same fragment length in both reactions were scored as mutant alleles; putative mutants that did not have the same fragment length in both reactions were considered false positives and scored as non-

mutant. Averaged over all loci, the initial false-positive rate was approximately 40%, with a final false-positive rate equal to (p/L)(0.40), where p is the probability of observing a mutant allele at a locus, L is the number of MA lines genotyped at a locus and assuming a complete mutational bias to the observed allele. Averaged over all loci, the final false-positive rate is on the order of 0.5%.

A conservative estimate of the false-negative rate is to inflate the observed fraction of mutant alleles by the initial false-positive rate (≈ 40%). Thus, it is possible that we have underestimated the true mutation rate by 20% if there is no mutational bias and as much as 40% if there is a complete mutational bias to the wild-type allele. The only solution to the false-negative problem is to genotype each line twice at each locus, which would have been prohibitively expensive. Reported values of mutation rates are uncorrected values.

The possibility of contamination is an inherent problem in MA experiments; the most likely cause is mis-labeling of plates or tubes, e.g., line 421 is inadvertently labeled 412. Although our MA protocol was explicitly designed to make mis-labeling immediately detectable, no system is perfect. We found no lines that had high frequencies of mutations (≥3 in *C. briggsae*, >2 in *C. elegans*) that shared the identical set of mutations. As an additional test, we fit the distribution of the number of mutations among lines to a Poisson distribution; non-independence of mutations in lines would increase the variance above the Poisson expectation. The number of lines differed among loci, so we used the weighted mean number of lines for each repeat type as the sample size. In three of the four strains the distribution of numbers of mutations was an excellent fit to the Poisson (goodness-of-fit chi-square, $P > 0.18$ in all cases). In PB306

there was a marginally significant excess of lines with one mutation and a deficit of lines with two or more mutations (obs/exp. 0 = 45/49; obs/exp 1 = 21/29; obs/exp ≥2 = 2/5; goodness-of-fit χ^2_1 = 4.59, P > 0.03). Undetected contamination would have the effect of reducing the actual number of lines surveyed, thus potentially biasing the estimated mutation rate downward.

Data Analysis

Mutation Rate

Allelic state was initially modeled as a binomially-distributed random variable X with state 0 = wild-type and state 1 = variant (length differences among variant alleles are not considered); each locus is assumed to constitute an independent manifestation of the mutational process. The null hypothesis is no difference among species or strains in the binomial parameter p = Pr(X=1). The mutation rate μ can be approximated by p/t, where t is the number of generations of MA. The full data set can be represented by an $n \times m$ matrix with the rows representing the n MA lines and the columns representing the m loci. Since different loci were examined in the two species, loci are nested within species. Differences among groups were assessed via generalized linear mixed model as implemented in SAS v. 9.1 PROC GLIMMIX, using a logit link function (http://support.sas.com/rnd/app/papers/glimmix.pdf). Significance of approximate F-tests for fixed effects was determined by the residual pseudo-likelihood method (Wolfinger and O'Connell 1993); degrees of freedom were calculated by the Kenward-Rogers method. Length-specific mutation rates were calculated by estimating the least-squares mean of the binomial parameter for each strain/repeat type combination and dividing by the number of generations of MA.

We first tested for variation between strains within each species. In principle, strain is a random effect, but two strains cannot provide a meaningful estimate of the within-species variance, so strain is treated as a fixed effect. The initial model was *p* = *rep_num* + *rep_type* + *strain* + *locus*(*rep_type*(*strain*)) + all interactions, where *p* is the binomial parameter, *rep_num* is the number of dinucleotide repeats and *rep_type* is the dinucleotide motif (AC, AG, AT). *Strain*, *rep_type* and their interactions are categorical fixed effects, *locus* is a categorical random effect, and *rep_num* is a continuous covariate. To account for overdispersion of the data, the among-locus (residual) component of variance was estimated separately for each *rep_type*/*strain* combination. Among-locus variance in mutation rate was assessed by likelihood-ratio test (LRT), comparing the (pseudo)likelihoods of the models with and without the residual variance term. Twice the difference in the negative log-likelihood of the two models is chi-square distributed with degrees of freedom equal to the difference in the number of parameters between the two models.

We next tested for differences among species using the model *p* = *rep_num* + *rep_type* + *species* + *strain(species)* + *locus*(*rep_type*(*species*)) + all interaction terms; the among-locus component of variance is estimated separately for each *rep_type*/*strain* combination as described previously. A complication is that the within-species analysis revealed significant interactions with strain in *C. elegans* but not in *C. briggsae* (see Results) and the fixed-effect model fits a single effect of strain nested within species. To account for the variation in the effects of strain between the two species, we pooled the two strains of *C. briggsae* and compared the pooled data to *C. elegans*, using the same model as above.

Mutational Spectrum

Mutations were characterized as insertions or deletions (without respect to length), and the proportion of deletions *q* was calculated for each locus in each strain, where *q* = #deletions/total # of mutations. Allelic state (insertion or deletion) is modeled as a binomially-distributed random variable, with the null hypothesis of no difference among groups in the binomial parameter *q*. Differences among groups were analyzed using PROC GLIMMIX with the logit link function. We initially tested for differences between strains within species, using the model *q* = *rep_num* + *rep_type* + *strain* + all interactions + *locus*(*rep_type*). The analysis failed to converge, so we removed repeat number from the model, subsuming repeat number in the among-locus variance. That model also failed to converge, so we considered differences among strains for each repeat type separately, using the model *q* = *strain* + *locus*, employing a Bonferroni corrected global α=0.05 with three tests.

We next tested for differences between species; the full model is *q* = *species* + *rep_type* + *species*×*rep_type* + *locus*(*rep_type*(*species*)). The full model did not converge, so we considered differences among species for each repeat type separately, using the model *q* = *species* + *strain*(*species*) + *locus*(*species*), employing a Bonferroni-corrected global α=0.05 with three tests. Species and strain are fixed effects, locus is a random effect.

Whole-Genome Distributions of STR Loci

To examine the genome wide abundance and distribution of dinucleotide microsatellite repeats we performed *in silica* searches of whole-genome sequences for both *C. elegans* (Wormbase version WS189) and *C. briggsae* (Wormbase version WS191). Because the current build of the *C. briggsae* genome (WS191) contains

"random" reads that share similarity but cannot be placed exactly on a particular chromosome assembly we omitted these sequences from the analysis. The inclusion of these random sequences increases the overall number of STR loci, but does not change the relative abundances of each repeat type (data not shown). This strategy provides a conservative estimate of the genome-wide distribution of STR loci in *C. briggsae*. All dinucleotide STRs of ≥ 5 perfect repeat units were identified using the PHOBOS algorithm (Christoph Mayer, Ruhr-Universität Bochum, http://www.ruhr-uni-bochum.de/spezzoo/cm/cm_phobos.htm). The PHOBOS parameters for all searches were: search method = imperfect, minimum unit length = 2, maximum unit length = 4, indel score = -2, recursion depth = 7, minimum score = 8. Here we present only perfect repeats of ≥ 5 repeat units. The expected genome-wide mutation rate μ_G was calculated as: $\mu_G = \mu_{AC}p_{AC} + \mu_{AG}p_{AG} + \mu_{AT}p_{AT}$, where μ_i is the expected mutation rate of repeat type *i* and p_i is the proportion of repeat type *i* in the genome. To estimate μ_i, we determined the average repeat number *n* for repeat type *i* and determined the expected mutation rate for a repeat of length *n* from the linear regression of the per-locus mutation rate on repeat number, averaged over the two strains of each species. Regressions were done using the MIXED procedure in SAS v. 9.1, including locus as a random effect.

Comparison of Mutational and Standing Variation

Six natural isolates ("strains") of *C. briggsae* were genotyped at 32 loci; these strains were the only wild strains of *C. briggsae* that were publicly available at the time. Strains were obtained from the Caenorhabditis Genetic Center stock collection at the University of Minnesota and cryopreserved upon receipt. All genotyping was performed

as described above for the MA lines. We calculated the locus-specific effective number of alleles $n_{e,i}$ at locus *i* under both the step-wise mutation model (SMM) (Kimura and Ohta 1975) and the infinite alleles model (IAM) (Kimura and Crow 1964) using the Microsatellite Analysis (MSA) software package (Dieringer and Schlotterer 2003). We used published values of n_e for 19 loci in 23 strains of *C. elegans* (Sivasundar and Hey 2003).

To assess the relationship between standing genetic variation and mutation rate, we first calculated the correlation between the per-locus mutation rate in the two strains within each species using SAS v.9.1 PROC MIXED with the TYPE=UNR covariance structure. The per-locus mutation rate was significantly positively correlated between the two strains in each species (*C. briggsae r* = 0.41, P < 0.002; *C. elegans r* = 0.65, P < 0.0001), so we used the average of the two strains. We then calculated the Spearman's correlation between $n_{e,i}$ in the wild isolates and μ_i, using SAS PROC CORR.

Results

Mutation Rate

Locus-by-locus statistics of the mutational properties (rate and spectrum) are presented in Appendix B and among-locus averages are presented in Table 1; the distribution of mutation rates for each repeat-type/strain combination are shown in Figure 1.

Variation within Species

Overall mutation rate does not differ between strains in either species. The large residual among-locus component of variance in both species (*C. briggsae*, LRT chi-square = 146.3, df = 6, P<0.0001; *C. elegans* LRT chi-square = 26.1, df = 6, P <

0.0003) is biologically relevant, because it means that there are locus-specific effects on mutation rate beyond the simple ones of repeat number and repeat type. Further, there is significant variation between strains in the among-locus variance in *C. elegans* but not in *C. briggsae* (*C. elegans* LRT chi-square = 5.4, df = 1, P < 0.03; *C. briggsae* LRT chi-square = 1.8, df = 1, P > 0.17). Mutation rate increases with repeat number in both species, although the effect is more pronounced in *C. elegans* than in *C. briggsae* (*C. elegans*, average slope of the regression of μ on repeat number = 3.84×10^{-6}, P < 0.0001; *C. briggsae*, average slope = 3.43×10^{-6}, P < 0.03). In addition, in *C. elegans* there is a marginally significant interaction between repeat number and repeat type (P < 0.04), indicating that the quantitative effect of repeat number on mutation rate varies depending on the particular repeat type (depicted in Figure 1).

There are marginally significant main effects of repeat type in both species (0.03 < P < 0.05), but the rank order differs between species (Table 1). In *C. briggsae*, the rank order averaged over strains is AG > AC > AT. In *C. elegans*, the rank order averaged over strains is AC > AG > AT, but there is a significant interaction between strain and repeat type (P < 0.02). The interaction between strain and repeat type results from the difference between strains in the average length-corrected AG mutation rate (4.3×10^{-5} in N2 vs. 0.82×10^{-5} in PB306). The qualitative rank order for N2 is AG > AC > AT; for PB306 it is AC > AT ≈ AG. Moreover, in *C. elegans* there is a marginally significant (P > 0.02) three-way interaction between repeat number, repeat type and strain.

Variation between Species

There is a highly significant difference in overall mutation rate between the two species (P < 0.002), with *C. briggsae* having an average length-corrected mutation rate almost three-fold higher than that of *C. elegans* (5.64×10^{-5} /generation vs. 1.98×10^{-5}

/generation). Further, there is a significant (P < 0.01) interaction of repeat number with species. However, the slope of the regression of mutation rate on repeat number differs by only about 10% between the two species, which suggests that the effect of repeat length is probably not qualitatively different in the two species. There is a marginally significant (P<0.03) interaction between repeat type and species. The source of the interaction can be primarily attributed to the much higher AG mutation rate in *C. briggsae* than in *C. elegans*. Finally, there is a marginally significant (P<0.03) three-way interaction between repeat number, repeat type, and species.

The preceding between-species analysis in which strain is considered a fixed effect fits an average effect of strain nested within species, but in reality some of the effects of strain differ between the two species, there being several significant interactions with strain in *C. elegans* but not in *C. briggsae*. When the two strains of *C. briggsae* are pooled and compared to (unpooled) *C. elegans*, the results are nearly identical; in particular, the main effect of species remains highly significant (P < 0.001).

Mutational Spectrum

The indel spectrum differs between species for two of the three repeat types (AC, P < 0.015; AG, P < 0.001; experiment-wide $\alpha = 0.05$ with three tests = 0.05/3 = 0.0167), with the bias being toward deletions in *C. briggsae* and towards insertions in *C. elegans* (Table 2, last column). In neither case is there a significant difference between strains within either species (P > 0.11 in all cases). The pattern differs for AT dinucleotides; there is a consistent insertion bias in *C. briggsae* (HK104, $q = 0.2$, PB800, $q = 0$), whereas the two strains of *C. elegans* have opposite indel biases (N2, $q = 0$; PB306, $q = 1$). However, there are fewer AT loci than AC or AG in the data set for *C. elegans* and

we were unable to assess the significance of the differences among groups at AT repeat loci.

Overall, the data provide a poor fit to the strict stepwise mutation model (Table 2). Averaged over strains and repeat types, the fraction of mutations that are insertions or deletions of a single repeat is very similar in the two species (73% in *C. briggsae*, 71% in *C. elegans*). This result is quite consistent with those reported by Seyfert et al. (2008) from the N2 strain of *C. elegans*. For loci which are comparable between the two studies (AC dinucleotides of < 100 repeats) the fraction of single step mutations is 56% in our study and 65% in theirs.

Genomic Mutational Properties

We can extrapolate from our experimental results to make inferences about the genome-wide mutational properties of the two species (Table 3). The complete distributions of perfect dinucleotide STR loci are presented in Figure 2; the estimated total for *C. briggsae* is probably an underestimate due to our use of the more conservative *C. briggsae* chromosome assemblies with the “random” sequences omitted from the analysis (see Methods). In the *C. elegans* genome, there are (approximately) 5586 STR loci of ≥ 5 perfect repeats of which (approximately) 35% are AC with an average length of ~7 repeats, 35% are AG (~ 8 repeats), 26% are AT(~8 repeats), and 5% are CG. In *C. briggsae* there are (approximately) 4408 STR loci, of which (approximately) 26% are AC with an average length of ~7 repeats, 58% are AG (~10 repeats), 14% are AT(8 repeats) and 3% are CG. Using the species-average point estimates of the linear regression parameters calculated from the mutation data, the expected number of mutations at dinucleotide STRs per generation in *C. briggsae* is about twice that of *C. elegans*.

The Relationship between Mutational and Standing Genetic Variation

Standing genetic variation in the six wild strains of *C. briggsae* is summarized in Supplementary Table S4. The correlation between the locus-specific mutation rate and n_e, the effective number of alleles at that locus (i.e., the standing genetic variance) is significantly positive in both species, but almost twice as great in *C. elegans* (Spearman's $r = 0.88$, $n = 19$, $P < 0.0001$) as in *C. briggsae* (Spearman's $r = 0.51$, $P < 0.003$). If the mutation rate of *C. briggsae* is in fact twice that of *C. elegans*, the expectation is that there will be twice as much standing genetic variance in *C. briggsae* as in *C. elegans*. However, because of the geographic disparity between our sample of wild *C. briggsae* strains and the sample of wild *C. elegans* of Sivasundar and Hey (2003) and because the global population genetic structure differs between the two species (Cutter et al. 2006; Dolgin, Felix, and Cutter 2008), comparison of standing genetic variation between the two data sets is not meaningful.

Discussion

Comparison of Mutation Rate among Species/Strains

The primary motivation of this study is to compare the mutational properties (rate and spectrum) between strains and species, toward the end of understanding the factors underlying variation in those properties. Results from several experiments lead to the conclusion that, on average, the cumulative effects of new mutations on fitness (and body size) accrue about twice as fast in these two strains of *C. briggsae* as in these two strains of *C. elegans* (Baer et al. 2006; Ostrow et al. 2007). "Cumulative effects" in this context refers to both the change in mean phenotype over time (ΔM in the MA parlance) and in the increase in genetic variance due to the input of new mutations, V_M. Both ΔM and V_M are functions of both the rate and distribution of

phenotypic effects of new mutations (Lynch and Walsh 1998, pp. 328-335), and statistically separating the effects of the rate and distribution of effects is notoriously difficult because the sampling variances of the two are negatively correlated (Begin and Schoen 2006). The ~ two-fold difference between the two species in mutation rate supports the conclusion that it is a difference in mutation rate *per se* that underlies the greater cumulative mutational effects in these strains of *C. briggsae* (although of course the distribution of effects may differ as well). More generally, our results support the intuitive conclusion that there is a close correspondence between the genomic mutation rate for fitness, U, and the molecular mutation rate, μ.

There is a caveat to the conclusion that the mutation rate is higher in *C. briggsae*. In a MA experiment, mutational variance (V_M and n_e) is proportional to the effective population size, N_e. Our experiment was designed to have $N_e = 1$, by transferring a single hermaphrodite worm every generation. However, when the focal individual failed to reproduce, we "went to backup", and picked a worm from the previous generation (or, occasionally, from two generations previous), thereby the increasing the census size above 1. When census size fluctuates over time, N_e is a function of harmonic mean population size (Hartl and Clark 2007, p. 121). In the case where some generations have census size of 1 and other generations are large, the harmonic mean is insensitive to differences of many orders of magnitude of the large-size generations and depends only on the ratio $t(1):t(\text{large})$, where t represents number of generations of given size. From this calculation, it turns out that the HK104 strain of *C. briggsae* had a larger N_e (≈1.5) than did the other three strains in our experiment (≈1.1 in all cases), because we had to go to backup more often in that strain than in the other three strains. A caveat to

this caveat, however, is that selection acts to reduce N_e (Hill and Robertson 1966) and if selection were stronger in HK104 than in the other strains (which seems likely), then N_e would actually be closer to one than we infer from census size. The fact that the mutational properties of the HK104 and PB800 strains of *C. briggsae* are very similar, both in this study and in our previous studies, suggest that the potential difference in N_e between HK104 and the other three strains is not a major factor.

Comparison of Indel Spectrum among Species/Strains

Many, but by no means all, studies of the STR mutational spectrum report a bias toward insertions (summarized in Ellegren 2004; Paun and Horandl 2006). For two of the three repeat types (AC, AG), the direction of the indel spectrum differed significantly between the two species: the bias was toward deletions in *C. briggsae* ($\overline{q}_{AC}$ = 0.80, $\overline{q}_{AG}$ = 0.77) and toward insertions in *C. elegans* ($\overline{q}_{AC}$ = 0.42, $\overline{q}_{AG}$ = 0.21). The pattern at AT repeats was more variable (Table 2, last column), but the combination of fewer loci sampled in *C. elegans* and a substantially lower observed mutation rate for AT vs. AC and AG in *C. briggsae* limits the strength of inference about AT repeats. The overall insertion bias observed in *C. elegans*, and particularly N2, is consistent with previous findings of a mutational insertion bias in the N2 strain, both at STR loci (Frisse 1999; Seyfert et al. 2008) and for random nuclear sequence (Denver et al. 2004). Denver et al. speculated that the qualitative difference between the indel spectrum of mutations accumulated in an MA experiment and those observed over evolutionary time (Witherspoon and Robertson 2003) may be due to the effects of selection. A possible alternative explanation, given the results of this study, is that the ancestral mutational bias was toward deletion and that the *C. elegans* lineage evolved a bias toward

insertion in recent time. If so, the apparent discrepancy between the MA and evolutionary patterns can be resolved without invoking natural selection. However, the median size of insertion substitutions observed by Witherspoon and Robertson (+2 nucleotides) was smaller than the median deletion (-7 nucleotides) and the means were even more disparate, so we caution that conclusions drawn from these STR data may not be relevant to the genome at large.

Relationship between Mutational and Standing Genetic Variation

At mutation-drift equilibrium, the standing genetic variance at a locus is proportional to the mutation rate under both the IAM ($\hat{n}_e \approx 4N_e\mu$) and the SMM ($\hat{n}_e \gg \sqrt{1+8N_e}\,r$). In the absence of perturbing forces (i.e., natural selection), N_e is the same at all loci, so differences in standing genetic variance among loci must be due to either to differences in mutation rate or to natural selection (sampling variance notwithstanding). The mutation rate is usually known only imprecisely and is usually assumed to be uniform across loci, and differences among loci or classes of loci in the standing genetic variance is typically attributed to differences in the strength or efficiency of selection (e.g., Sabeti et al. 2006). Of the large body of studies of the mutational properties of STR loci (reviewed in Ellegren 2004), few of those that measure mutations directly (i.e., from the distribution over a pedigree or from reporter constructs) have attempted formal comparisons of the relative mutation rates of different repeat types, and those that draw inferences indirectly from the standing genetic variance or from comparisons of substitutions between species cannot unambiguously partition the effects of selection from those of mutation.

Our data show a (variably) strong positive correlation between the per-locus mutation rate and the standing variance at the locus, but the relationship is much stronger in *C. elegans* (Spearman's $r = 0.88$, $P < 0.0001$) than in *C. briggsae* (Spearman's $r = 0.51$, $P < 0.003$). A possible source of the discrepancy between the two species is that we only included six strains of *C. briggsae*, whereas Sivasundar and Hey (2003) had 23 strains of *C. elegans*. Alternatively, it is known that the population structure differs between *C. elegans* and *C. briggsae*, and our sample of strains from *C. briggsae* incorporates samples from two major clades (Cutter et al. 2006). If the migration rate ($4N_em$) differed among regions of the genome (there is no reason to expect it would), it could potentially reduce the correlation between standing genetic variance and mutation rate. Nevertheless, this is an important result because it demonstrates that, on average, the standing genetic variation at a locus does substantially reflect the underlying mutation rate. We know of one other study in which direct estimates of mutation rate were compared to standing variation at the same set of loci. Vigoroux et al. (2002) reported a Spearman's *r* of 0.42 of heterozyosity with mutation rate at a set of 98 STR loci in a collection of 193 maize plants (Matsuoka et al. 2002). Thus, studies that employ indirect methods of inference of mutation rate are justified. This was not a foregone conclusion - for example, recall the discrepancy between the indel spectra inferred from direct and indirect studies of *C. elegans* - and it remains to be verified for classes of loci other than STRs.

Taxon-Specific Differences in Mutability of Different Repeat Types

The most comprehensive study of the mutational properties of STRs is that of Kelkar et al. (2008), who indirectly inferred the mutational properties of a very large number(> 950,000 loci of four or more repeats) of orthologous dinucleotide STR loci in

the human and chimpanzee genomes. Comparison of our study with their vastly larger one is instructive in several respects. First, the fact that we (and Vigouroux et al. 2002) observe a positive relationship between mutation rate and standing variation validates their inferences, i.e., there is no evidence that their data are positively misleading because of hidden biases. Second, they observe substantial (order of magnitude) residual variation among loci in the mutation rate, after accounting for the effects of repeat number and type (= motif). Thus, the among-locus variance in mutation rate that we observe experimentally appears to be an honest manifestation of biologically-relevant properties that operate over long evolutionary time scales. Numerous studies have found both direct (i.e., position effects; (Lichtenauer-Kaligis et al. 1993) and (more often) indirect evidence that the mutation rate varies with local genomic context (e.g., Hardison et al. 2003; Lercher, Chamary, and Hurst 2004; Arndt, Hwa, and Petrov 2005). Various explanations have been offered; one that appears particularly convincing is that mutation rates are higher in regions of closed chromatin (Prendergast et al. 2007), although the specific mechanism is not known.

An intriguing disparity between our results and the findings of Kelkar et al. (2008) is the difference in the relative magnitudes of mutation rates of the different repeat types. In the human/chimp genome, AT dinucleotides mutate significantly faster than AC or AG repeats. Our data from *C. briggsae* show that AT repeats have a significantly lower mutation rate than AC and AG repeats, and the standing variance at AT repeats in *C. briggsae* is lower than for AG repeats, which have the highest observed mutation rate, although nothing like the predicted three-fold difference. Kelkar et al. posit that the higher mutation rate at AT repeats results from the smaller number of hydrogen bonds

in double-stranded AT repeats relative to AC or AG repeats. If we provisionally accept that the difference between our data and the human/chimp is real and not Type-I error on our part resulting from our vastly smaller sample size, it leads to the conclusion that "DNA is not just DNA", i.e., that DNA with the same sequence mutates in different ways in different taxa.

Extrapolation to Genome-wide Mutation Rate

Two properties of the genomic distribution of dinucleotide STRs differ qualitatively between (the AF16 strain of) *C. briggsae* and (the N2 strain of) *C. elegans* (Figure 2). First, the fraction of AG repeats is much greater in *C. briggsae* than in *C. elegans* (58% vs. 35%) and there are only half as many AT repeats in *C. briggsae* as in *C. elegans* (14% vs. 26%), and second, the average AG locus is over two repeats longer in *C. briggsae* than in *C. elegans* (10.1 vs. 7.6 repeats; Table 3). Taken together and extrapolating over the genomic distribution from the linear regression parameters inferred from our MA study, we infer that the per-locus dinucleotide STR mutation rate is about twice as great in *C. briggsae*, as is the expected total number of dinucleotide STR mutations (0.24/gen as in *C. elegans* (0.12/gen). There are obviously many sources of uncertainty in those calculations. Nevertheless, the results are remarkably consistent with the body of evidence inferred indirectly from the cumulative phenotypic effects of MA.

Given the short average length of perfect dinucleotide STRs and their relative paucity in the *Caenorhabditis* genome (compare to the ~950,000 *orthologous* dinucleotide STR loci in the human/chimp genome), it is unlikely that differing properties of perfect dinucleotide STRs in the two species is the sole cause of the different cumulative effects of MA. Rather, the difference in the mutational properties of STR loci

is probably a byproduct of some more fundamental difference in the mutational input and/or output. One obvious possibility is that some property of the DNA repair machinery differs consistently between the two species. The two species are believed to have diverged on the order of 100 million generations ago, roughly on the timescale of the divergence of humans and rodents (Cutter 2008). Significant differences in various aspects of the DNA repair process are known to exist between humans and rodents, and between more closely-related taxa (Eisen and Hanawalt 1999), and it is certainly possible that differences exist between the two *Caenorhabditis* species. However, there is no concrete evidence for any such difference.

A second possibility, for which there is some evidence, albeit indirect, is that some aspect of oxygen free-radical metabolism differs between the two species, or at least between the strains included in this experiment. Reactive oxygen species (ROS) are normal byproducts of cellular metabolism, and oxidative stress has been implicated in microsatellite instability (Jackson and Loeb 2000; Lee et al. 2006). Howe and Denver (2008) have documented the presence of heteroplasmy for a deletion in the mitochondrial NADH-dehydrogenase 5 (ND5) gene. ND5 functions in ROS metabolism, and there is evidence that individuals with defective ND5 suffer increased ROS damage. The CGC stocks of HK104 and PB800 used by Howe and Denver both have low frequencies of the ND5 deletion, but the immediate ancestor of our HK104 MA lines apparently evolved a high frequency of the ND5 deletion during the inbreeding leading up to the MA experiment (D. Denver, personal communication).

Integrated over the genomic distribution of dinculeotide STRs, it is estimated that the genomic mutation rate at dinculeotide STRs in *C. briggsae* is roughly twice that of *C.*

elegans. This finding is entirely consistent with the body of evidence from the cumulative mutational effects on the phenotype, and provides one of the first direct demonstrations of a relationship between molecular and phenotypic mutational properties for a non-mutator genotype. Further, we find the indel spectrum differs between repeat types and species, in contrast to many (but certainly not all) previous studies that have found a general insertion bias at STR loci. Finally, we find that the per-locus mutation rate is significantly positively correlated with the standing genetic variation in both species. This result, which was also found in maize by Vigoroux et al. (2002), provides empirical justification for using the standing genetic variance as a proxy for the mutation rate.

Table 3-1. Summary of per-generation mutation rate μ for all loci assayed in *C. briggsae* and *C. elegans*. "Composite" repeats have > 1 repeat type in the locus in question); composite loci were omitted from the statistical analyses. μOBS is the mutation rate calculated from the raw data; μLS is the average mutation rate estimated by least squares mean; SEM in parentheses. See Methods for details of the general linear mixed model.

		C. briggsae		*C. elegans*	
Repeat Type		HK104	PB800	N2	PB306
AC	N loci ($\bar{n}$ lines)	11 (91)	11 (93)	13 (79)	16 (81)
	ave. repeat #	14.0	14.2	16.5	23.1
	μ_{OBS} x 10^5	6.28	5.57	3.08	8.20
	μ_{LS} x 10^5 (SEM)	6.65 (3.33)	4.55 (2.98)	2.43 (1.31)	7.09 (3.32)
AG	N loci ($\bar{n}$ lines)	30 (95)	30 (94)	15 (86)	14 (80)
	ave. repeat #	21.7	22.4	18.7	18.2
	μ_{OBS} x 10^5	10.6	11.6	4.85	2.99
	μ_{LS} x 10^5 (SEM)	9.79 (2.47)	9.86 (2.93)	4.34 (1.95)	0.82 (0.58)
AT	N loci ($\bar{n}$ lines)	14 (82)	12 (82)	8 (59)	9 (61)
	ave. repeat #	12.1	11.5	16.5	16.8
	μ_{OBS} x 10^5	1.96	3.57	2.98	1.78
	μ_{LS} x 10^5 (SEM)	2.18 (1.27)	4.49 (2.71)	1.59 (0.85)	0.89 (0.57)
Composite	N loci ($\bar{n}$ lines)	3 (87)	3 (92)	-	-
	ave. repeat #	21.0	21.0	-	-
	μ_{OBS} x 10^5	2.81	12.4	-	-

Table 3-2. Summary of the indel spectrum. "Composite" loci have > 1 repeat type; composite loci were omitted from the statistical analyses. For each repeat type, row headings are: N loci, the number of loci genotyped for a given repeat type with the mean number of lines genotyped ($\bar{n}$) in parentheses; Deletion (Insertion) > 1 is the number of mutant alleles > 1 repeat unit (2 bp) shorter (longer) than the ancestral allele; Deletion (Insertion) = 1 is the number of mutant alleles 1 repeat unit (2 bp) shorter (longer) than the ancestral allele. Indel Bias q is the proportion of mutations that are deletions.

		C. briggsae		*C. elegans*	
Repeat		HK104	PB800	N2	PB306
AC	N loci ($\bar{n}$ lines)	11 (91)	11 (93)	13 (79)	16 (81)
	Deletion > 1	3	3	2	3
	Deletion = 1	8	9	2	8
	Insertion > 1	0	0	2	4
	Insertion = 1	4	2	3	10
	Indel bias q	0.733	0.857	0.444	0.400
AG	N loci ($\bar{n}$ lines)	30 (95)	30 (94)	15 (86)	14 (80)
	Deletion > 1	3	7	0	0
	Deletion = 1	59	49	4	1
	Insertion > 1	1	6	1	2
	Insertion = 1	12	19	11	3
	Indel bias q	0.827	0.700	0.250	0.167
AT	N loci ($\bar{n}$ lines)	14 (82)	12 (82)	8 (59)	9 (61)
	Deletion > 1	0	0	0	1
	Deletion = 1	1	0	0	1
	Insertion > 1	3	4	0	0
	Insertion = 1	1	6	3	0
	Indel bias q	0.200	0	0	1
Composite	N loci ($\bar{n}$ lines)	3 (87)	3 (92)	-	-
	Deletion > 1	0	2	-	-
	Deletion = 1	2	7	-	-
	Insertion > 1	0	0	-	-
	Insertion = 1	0	0	-	-

Table 3-3. Inferred genome-wide mutational properties. Regression parameters are determined from the linear regression of per-locus mutation rate against repeat number. E(μ) is the expected genome-wide mutation rate; Total E(μ) is the weighted mean of the three repeat-type-specific values.

Species	Repeat type	~*N* loci (%)	Mean repeat number	Slope	Intercept	E(*μ*)	Expected # of mutations/gen
C. briggsae	AC	1340 (25.5)	7.3	6.91×10^{-6}	-4.0×10^{-5}	9.8×10^{-6}	0.06
	AG	3056 (58.1)	10.2	2.34×10^{-6}	6.0×10^{-5}	8.4×10^{-5}	0.14
	AT	714 (13.6)	7.9	2.87×10^{-9}	3.1×10^{-5}	3.1×10^{-5}	0.03
	CG	152(2.9)	-	-	-	-	-
	Total	5262	-	-	-	5.7×10^{-5}	0.24
C. elegans	AC	1930 (34.6)	7.3	1.41×10^{-6}	3.1×10^{-5}	4.1×10^{-5}	0.04
	AG	1955 (35.0)	7.6	2.68×10^{-6}	-1.0×10^{-5}	1.0×10^{-5}	0.04
	AT	1439 (25.8)	7.7	5.87×10^{-6}	-7.0×10^{-5}	4.5×10^{-5}	0.03
	CG	262 (4.7)	-	-	-	-	-
	Total	5586	-	-	-	2.3×10^{-5}	0.12

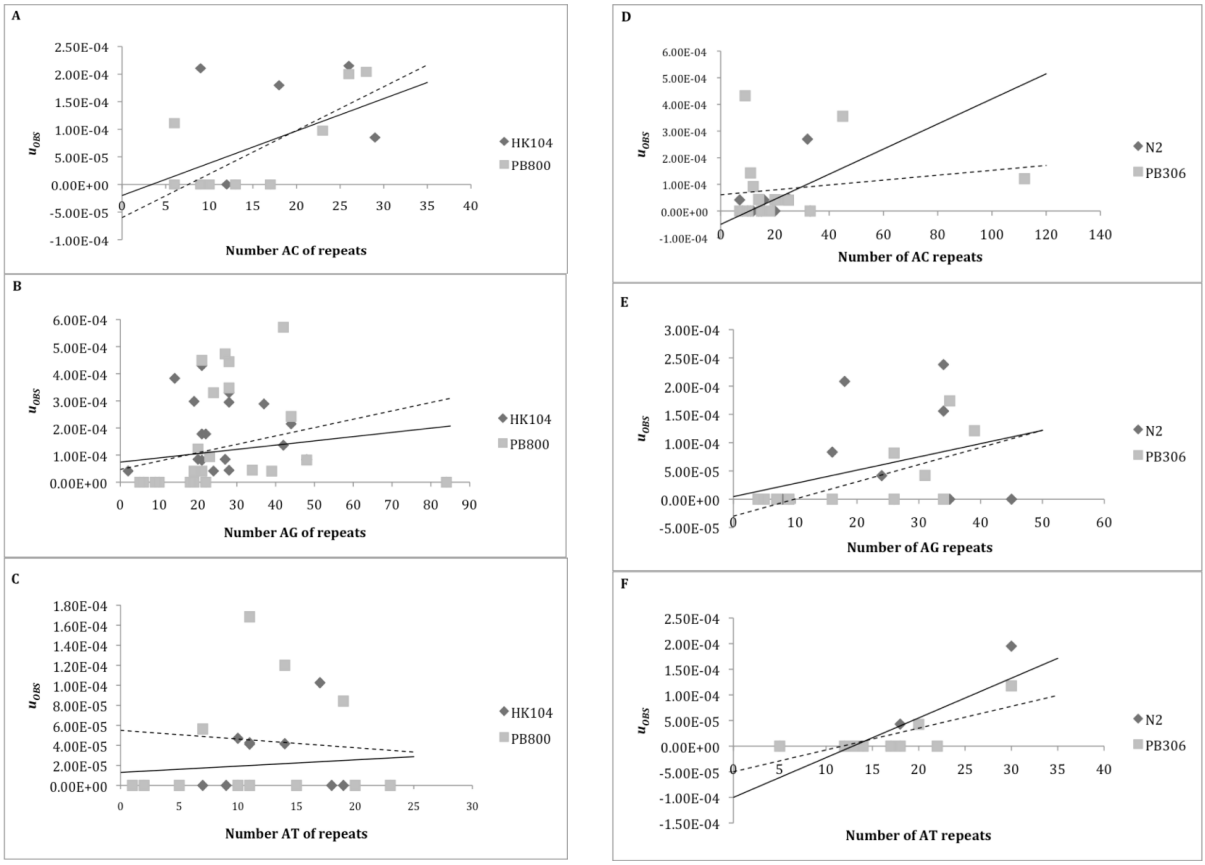

Figure 3-1. Relationship between observed mutation rate (μ_{OBS}; dependent variable) and number of perfect repeats (independent variable) for each repeat type in the two strains *C. briggsae* and *C. elegans*. Panels A-C are of AC, AG, and AT repeats in the HK104 (diamonds) and PB800 (squares) strains of *C. briggsae*. Solid and dashed lines are the best-fit linear regression of μ_{OBS} on repeat number in HK104 and PB800, respectively. Panels D-F are of AC, AG, and AT repeats in the N2 (diamonds) and PB306 (squares) strains of *C. elegans*. Solid and dashed lines are the regression of μ_{OBS} on repeat number in N2 and PB306, respectively.

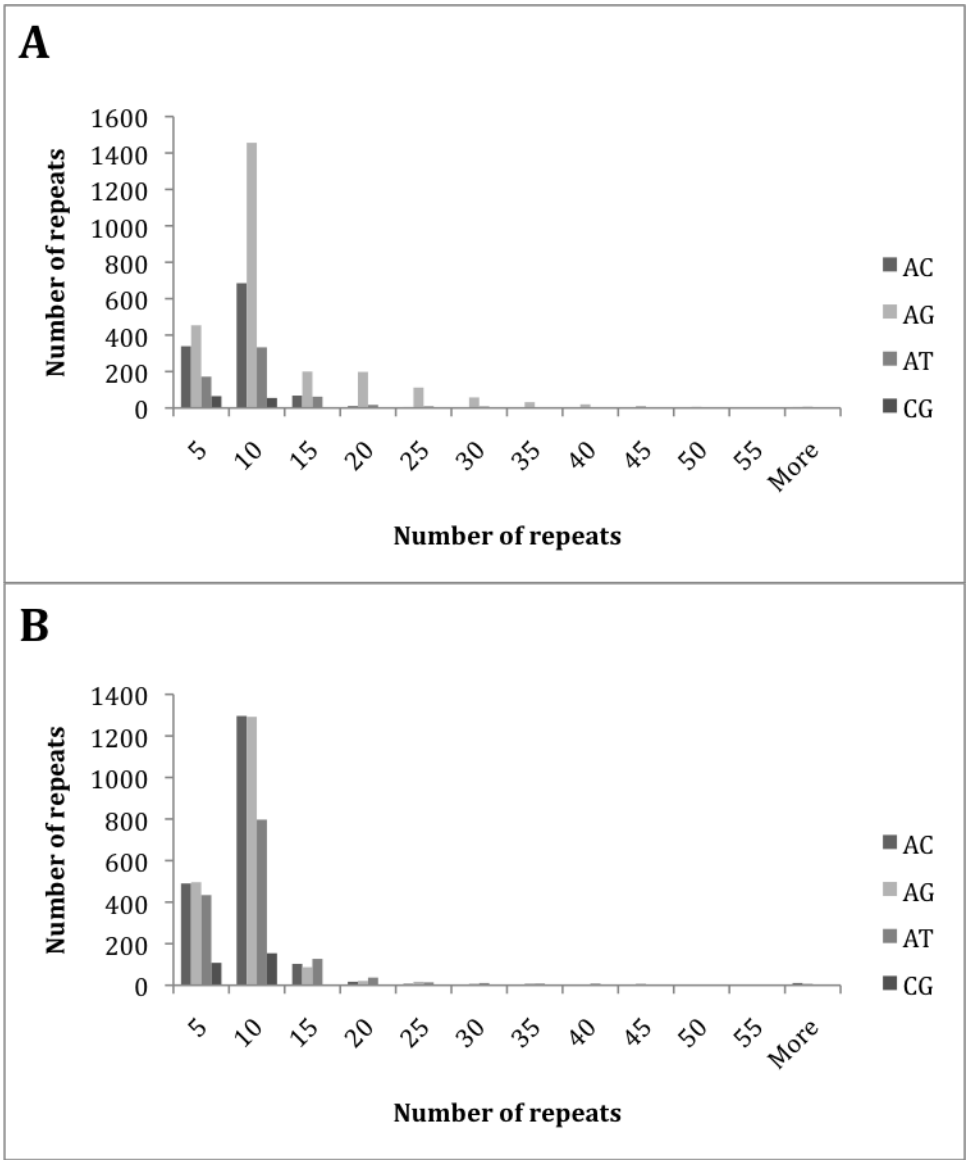

Figure 3-2. Genome-wide distribution of dinucleotide STR loci in the *C. elegans* and *C. briggsae* genomes. Only perfect repeats are reported. Y-axis is the number of repeats. X-axis is divided into bins of five repeats, with all loci greater than 60 repeats pooled. Panel A is the distribution of STR dinucleotide loci in the *C. briggsae* genome, Panel B is the *C. elegans* genome.

CHAPTER 4
THE RATE AND SPECTRUM OF DI-NUCLEOTIDE AUTOSOMAL AND X CHROMOSOME MICROSATELLITE MUTATIONS IN TWO SPECIES OF CAENORHABDITID NEMATODE WORMS THAT DIFFER IN REPRODUCTIVE STRATEGY

To what degree natural selection has shaped the rate of spontaneous mutations among different taxa remains an unresolved question in evolutionary biology. While mutation rates are known to vary among and within taxa (Drake et al. 1998; Lynch 2010), the relative importance of natural selection versus non-adaptive and physiological processes has yet to be determined (Lynch 2008).

Theoretical investigations into the evolution of mutation rates provide two general predictions (Kondrashov 1988, 1995; Drake et al. 1998). First, selection to reduce the deleterious mutation rate should be stronger in asexual and selfing organisms than in sexual ones. In a diploid asexual organism, the strength of selection favoring an allele that modifies the mutation rate by a factor of ΔU is approximately ΔU, whereas in an outcrossing sexual taxon the strength of selection on such a modifier allele is approximately $h\bar{s}\,\Delta U$, where $h\bar{s}$ is the average strength of selection against deleterious mutations in heterozygotes. Second, selection to reduce the mutation rate should be stronger in an obligately selfing sexual organism than in an obligate outcrossing organism. In an obligately selfing sexual taxon, the strength of selection acting on a modifier of the mutation rate is $(1/2)\Delta U$ if deleterious mutations are not completely recessive. Thus, the strength of selection on modifiers of the mutation rate in selfing taxa relative to outcrossing taxa is greater by a factor of about $2h\bar{s}$. However, the possibility of an increase in mutation rate may exist if a mutator of mutation rate rises in frequency along with a beneficial mutation through hitchhiking.

Similar reasoning applies to modifiers of sex-linked and autosomal mutation rates in taxa with defined sex chromosomes (McVean and Hurst 1997). Partially recessive deleterious alleles at loci on the hemizygous chromosome impose a greater fitness cost than autosomal alleles with the same effect. Therefore, selection on modifiers of the sex-chromosomal mutation rate is stronger than selection on modifiers of the autosomal mutation rate.

This theory leads to several clear empirical predictions: *all else equal*, (1) asexual taxa should evolve a lower genomic mutation rate than related sexual taxa, (2) obligately (or predominantly) selfing taxa should have a lower mutation rate than related obligately (or predominantly) outcrossing taxa, and (3) the X-linked mutation rate should be lower than the autosomal mutation rate in taxa with X/Y sex determination. While the theoretical predictions are clear, to what extent do these theoretical predictions hold in real life? Have obligately selfing species evolved lower mutation rates than their outcrossing relatives?

Members of the nematode genus *Caenorhabditis* provide an ideal system to examine questions of how mutation rates vary among closely related species with different reproductive strategies. The ancestral reproductive state within the genus *Caenorhabditis* is outcrossing (gonochorism), however self-fetilization (hermaphroditism) has evolved independently at least three times in the genus, with species *C. elegans*, *C. briggsae* (Kiontke et al. 2004) and in the newly described species *C. "species 11"* (C. Braendle per. comm.) being hermaphroditic.

Here we present results from comparisons of di-nucleotide microsatellite mutation rates for two autosomes and the X-chromosome in the outcrossing species *C. remanei*

and in a strain of *C. elegans* that was maintained by male-female mating. We demonstrate that the overall di-nucleotide mutation rate in the historically outcrossing species, *C. remanei*, is indeed greater than that in the historically selfing species *C. elegans*. Our results are consistent with the theoretical prediction that natural selection to reduce the spontaneous mutation rate for di-nucleotide microsatellite repeats has been stronger in the selfing species.

Material and Methods

Mutation Accumulation Lines

All MA lines used in this study were previously described in Baer et al. (2010). Briefly, highly inbred stocks of each strain were replicated and perpetuated by transferring a single female and male worm each generation for ~ 100 consecutive generations. This protocol results in a genetic effective population size of $N_e \approx 2$ (the approximation is the result of occasionally having to use backup stocks of worms when the original worm did not survive, see Baer et al. 2010 for details on N_e), thereby minimizing the efficiency of natural selection and ensuring all but the most deleterious mutations behave according to neutral dynamics. In addition, the MA lines used in the present study were subjected to additional generations of MA past those assayed in Baer et al. (2010) as noted below.

Two sets of the MA lines reported in Baer et al. (2010) were used in this study. To represent an outcrossing reproductive system, a set of MA lines derived from the PB2282 strain of *Caenorhabditis remanei* that had undergone 122 generations of MA were used. In addition, a strain of *C. elegans* in which the capability for self-fertilization was blocked was also included. This strain, hereafter referred to as *C. elegans fog-2*, was constructed by introgressing the *fog-2* mutation (Schedl and Kimble 1988) into the

canonical N2 genetic background (Vassilieva and Lynch 1999; Baer et al. 2005) via backcrossing for 12 generations (see Baer et al. 2010 for details). The *fog-2* mutation eliminates the hermaphrodite's ability to generate sperm and thus renders a hermaphroditic individual into a functional female. The *C. elegans fog-2* MA lines were subject to a total of 200 generations of MA. The greater number of generations in the *C. elegans fog-2* MA lines resulted from the slowing of generation time in the *C.remanei* MA lines as the MA process progressed.

Selection of Microsatellite Loci

The goal of this study was to compare how mutational properties differ with respect to reproductive system and chromosome type and thus we choose to select loci that would maximize the probability of observing enough mutations to make such comparisons. From previous work, we have shown that di-nucleotide microsatellite mutation rates in *Caenorhabditis* are dependent on both repeat type and repeat length (Phillips et al. 2009). Therefore, we chose to focus on AG di-nucleotide microsatellite repeats in both species and to match loci with respect to repeat length between the two species as much as possible. Furthermore, AG di-nucleotide repeats are the most abundant repeat type in all five of the currently available *Caenorhabditis* genome assemblies (Figure 1). In addition, to test if the rate and spectrum of di-nucleotide microsatellite mutations differed with respect to chromosome type between *C. remanei* and *C.elegans fog-2,* loci from two autosomes were selected to compare to loci selected from the X chromosome. For the two autosomes, we chose to include the IV chromosome, which is known to be highly heterozygous in a different inbred strain derived from the same wild isolate of *C. remanei* (Barriere et al. 2009), as well as an

additional autosome (chromosome II) selected randomly from the remaining four autosomes.

All perfect di-nucleotide microsatellites with 5 repeats (10 bp in length) or greater were identified in the published genomes of *C. elegans* (build WS205), and *C. reamnei* (build WS205) using the PHOBOS algorithm version 3.3.12 (Christoph Mayer, Ruhr-Universität Bochum). The PHOBOS parameters for searches in both species were: -M imperfect, -u minimum repeat length = 2, -U maximum repeat length = 2, -m mismatch score = -6, -r recursion depth = 7, -s minimum length score = 8, -f number of bases flanking a repeat = 250 bp. These results were further filtered by repeat perfection and length using a set of custom Perl scripts. At present the draft genome build of *C. remanei* does not include assignment of contigs to individual chromosomes and therefore the chromosomal context of each microsatellite locus had to be assigned based on sequence similarity to the most closely related species for which chromosome information is available, *C. briggsae*. Two criteria were used to assign the *C. remanei* loci to a putative chromosome number based on sequence similarity to the *C. briggsae* genome assembly. First, BLAST searches were used to identify the single best hit for the 250 bp of flanking sequence surrounding each *C. remanei* locus to the *C. briggsae* (build WS205) genome. All BLAST parameters were left as defaults. Second, the single best BLAST hit for each of the *C. remanei* WS205 contigs was found in the *C. briggsae* genome assembly. From these two BLAST searches, all di-nucleotide microsatellite loci identified in *C. remanei* were assigned to a putative chromosome number if: (1) the flanking sequence surrounding each locus and (2) the contig from

which the locus was identified both had a best single hit BLAST match to the same *C. briggsae* chromosome.

Microsatellite mutation rates are known to positively correlated with the number of repeats (Ellgran 2004) and therefore, to increase the probability of observing a mutation at a given locus, we selected microsatellite loci from two different size classes based on the 90th to 95th, and 96th to 99th percentiles of the length distribution of each species.

To amplify each locus, PCR primers were designed using the Primer 3 software (Rozen and Skaletsky 2000) for each locus (plus 250 bp on either side of the locus) to generate an *in silico* predicted "PCR fragment" using the default parameters. PCR fragment size was constrained to between 100 bp and 400 bp for all loci. Predicted "PCR products" were then screened for the presence of any other repeat motif, other than the focal perfect AG di-nucleotide repeat, with a period size of 1 bp to 100 bp using the PHOBOS algorithm version 3.3.12 (Christoph Mayer, Ruhr-Universität Bochum). The PHOBOS parameters for searches in both species were: -M imperfect, -u minimum repeat length = 2, -U maximum repeat length = 100, -m mismatch score = -6, -r recursion depth = 7, -s minimum length score = 8, -f number of bases flanking a repeat = 250 bp. All predicted "PCR products" containing repeats other than the focal AG di-nucleotide repeat were removed from the list. From the remaining list, 10 loci from both the 90th to 95th percentile bin and the 96th to 99th percentile bin for autosomes II and IV were selected. For the X chromosome, 15 loci were selected per bin. PCR primer information for all loci is presented in Table 1.

Genotyping

Genomic DNA was extracted from two replicate cultures of each MA line, as well as the ancestral control stock, using the Qiagen 96 well DNeasy Blood and Tissue kit

following the manufacturers protocol (Qiagen, USA). We employed a nested PCR strategy with fluorescently tagged primers via a modification of the "three-primer" method of (Schuelke 2000). Multiplexed PCR reactions of 15 µl were performed in 96-well plates, using 1 ul of DNA template, 60 pmol of selective primer, 6 pmol of M13-tail primer, 60 pmol of labeled M13 primer, and 7.5 ul of Qiagen Type-it Microsatellite PCR kit master mix (Qiagen, USA). Four to five loci were amplified together per multiplex reaction with ~ 50 bp of spacing separating each locus. Reactions were initially run for 10 cycles of 40 seconds denaturing at 94°C, 40 seconds annealing at 60°C, and 40 seconds extension at 72°C; the annealing temperature was then decreased to 48°C and the reaction was continued for an additional 20 cycles.

Two different fluorescent labels were used (FAM and NED), with only a single label used for a given multiplex. PCR products were analyzed using an Applied Biosystems 3730XL DNA analyzer (Interdisciplinary Center for Biotechnology Research, University of Florida, USA). Fragment length was established relative to a known size-standard ladder (GeneScan 600, Applied Biosystems, USA). All genotypes were manually inspected using the GeneMarker version 1.6 software (SoftGenetics, USA). We employed an iterative binning procedure to identify putative mutants. In the first iteration we calculated the mean fragment length for all replicates. Fragments that deviated by > 1.5 bases from the mean were removed from the dataset and the mean recalculated. Alleles that differed by > 1.5 bases from the re-calculated mean were scored as putative mutants.

Genotyping of Natural Isolates

Seventeen wild isolates of *C. remanei* were included in this study to estimate levels of standing genetic variation for di-nucleotide microsatellite repeats present in

nature. Stocks of *C. remanei* strains JU1087, JU1086, JU1084, JU1082, JU724, MY37, MY32, MY31, MY28, PB4641, PB229, PB228, PB227, PB219, PB212, PB206, and SB146 were obtained from the *Caenorahbditis* Genetics Center and collection information is available on their web site (http://www.cbs.umn.edu/CGC/; University of Minnesota, USA). Upon receipt of these strains, genomic DNA was isolated and each strain was genotyped as described above for the MA lines.

Data Analysis

Variation in Mutation Rates between Species

Mutation rates were estimated by modeling the allelic state as a binomially-distributed random variable X with two states, state 0 = wild-type and state 1 = variant (length differences among variant alleles were not considered). The null hypothesis is that no difference among *C. remanei* and *C. elegans* in the binomial parameter p = Pr(X=1) exists. The per-locus mutation rate μ can be approximated by p/t, where t is the number of generations of MA. Loci were treated as a random effect and nested within chromosome which in turn is nested within species to account for the fact that loci were not chosen based on sequence homology and thus are not matched between the two species by evolutionary origin. A generalized linear mixed model was used to fit the model using the PROC GLIMMIX procedure in SAS version 9.2 with a logit link function. A residual pseudo-likelihood F-test method was used to test for significance of fixed effects (Wolfinger and O'Connell 1993), with degrees of freedom calculated by the Kenward-Rogers method. Differences between the two species was modeled as p = species + *repeat_number* + *chromosome* + error. The code used to fit the above model is presented in Table 2.

Mutational Spectrum

Mutations were characterized as insertions or deletions (without respect to length), and the proportion of deletions *q* was calculated for each locus in each strain, where *q* = #deletions/total # of mutations. Allelic state (insertion or deletion) is modeled as a binomially-distributed random variable, with the null hypothesis of no difference among groups in the binomial parameter *q*. Variation in the indel spectrum between species was modeled similarly as above, with the final model being *q* = *species* + *repeat_number* + all interaction terms, with locus treated as a random effect.

Correlation of Per-Locus Mutation Rate to Standing Genetic Variation

The relationship between the mutation rate at a given locus and the standing genetic variation at the same locus was estimated for 56 loci in 17 wild isolates of *C. remanei*. The locus-specific effective number of alleles was calculated using the Microsatellite Analysis software package (Dieringer and Schlotterer 2003) under both the stepwise mutation model (SMM) (Kimura and Ohta 1964) and infinite alleles model (IAM) (Kimura and Crow 1964). The Spearman's correlation was calculated between the per-locus mutation rate and the SMM and IAM effective number of alleles at each locus using the PROC CORR procedure in SAS v.9.2. In addition, we controlled for the effect of different repeat lengths at each locus by accounting for the repeat number at each locus and calculated the partial Spearman's correlation using the PROC CORR procedure in SAS v.9.2.

Results

Variation between Species in Di-nucleotide Mutation Rate

Estimates of mutation rates between species are summarized in Table 3. There is a significant difference in overall di-nucleotide microsatellite mutation rate between

C.remanei and *C. elegans fog-2* ($P < 0.01$), with *C. remanei* having an average mutation rate that is approximately seven-fold higher than that of *C. elegans fog-2* (8.0×10^{-5} /generation vs. 1.2×10^{-5} /generation). There is a significant positive association between repeat length and mutation rate in both species, with the effect being greater in *C. elegans fog-2*.

Variation between Autosomes and X-Chromosome

While mutation rates did differ between species, there was no significant effect of chromosome type with mutation rates of both autosomes and the X chromosome not differing significantly ($P = 0.18$) (summarized in Table 3).

Indel Spectrum

The indel spectrum does not differ significantly between the two species ($P = 0.37$). For both *C.remanei* and *C. elegans fog-2* there was an excess of insertions. In addition, the data provide a good fit to the strict stepwise mutation model, with only two mutations deviating from an increase or decrease of a single dinucleotide repeat unit (summarized in Table 4).

Correlation to Standing Genetic Variation

The raw correlation between the per-locus mutation rate and the effective number of alleles at that locus present in the 17 wild isolates was significant (Spearman's $r=0.43$, $P < 0.05$). However, when repeat number is taken into consideration the partial correlation between the mutation rate at a given locus and standing genetic variation is not significant (Spearman's $r=0.34$, $P=0.12$).

Discussion

Comparison of Mutation Rate between Species

The goal of this study was to contrast the mutational properties of di-nucleotide microsatellite repeats in two closely related species of *Caenorhabditis* that differ in reproductive strategies in order to test if mutation rates differ between reproductive strategies as predicted by theory. Two previous studies have looked at the mutational properties of closely related species that differ in reproductive strategy, however these two studies have provided conflicting results. The first study to directly test the role of reproductive strategy in closely related species was that of Schoen (2005). In this study, the author established two sets of MA lines from members of the plant genus *Amsincka*, the outcrossing species *A. douglasiana* and the predominately selfing species *A. gloriosa*. After 11 generations (~4 years) of MA, Schoen found no significant difference in the deleterious mutation rate (*U*) for several quantitative characters, suggesting that the rate of deleterious mutations did not differ between the selfing and outcrossing species of *Amsincka* in a detectable way. Contrary to this, Baer et al. (2010) established sets of MA lines from four outcrossing species of nematode worms in the genus *Caenorhabditis*. After ~ 100 generations of MA, Baer et al. (2010) observed a roughly four fold decline in fitness between the outcrossing species as compared to a historically selfing control. This result is consistent with the interpretation that the outcrossing species of *Caenorhabditis* have a greater deleterious mutation rate for fitness than their selfing relative. However, the possibility that a significant amount of residual heterozygosity may have existed in the progenitors of the outcrossing species MA lines makes it difficult to rule out the effects of inbreeding depression on the decline in fitness observed in these lines (Barriere et al. 2008). In addition, estimating

mutational properties from quantitative traits in MA experiments is limited to those mutations of intermediate effect and mutations of small effect may readily escape detection (Halligan and Keightley 2009).

Here we present molecular level data on mutational properties that are in good agreement to those previously obtained for the fitness. We have previously shown that differences between closely related species in di-nucleotide microsatellite mutation rates are consistent with the decline in fitness of the same MA lines (Phillips et al. 2009). In the current study, the average di-nucleotide microsatellite mutation rate in the outcrossing species *C. remanei* was found to be approximately six times that of the historically selfing species *C. elegans fog-2,* a very similar pattern to the fitness data. Our estimates for *C. elegans fog-2* mutation rates are consistent with those from a previous study of di-nucleotide microsatellite mutation rates in the N2 geneotype of *C. elegans* (1.2×10^{-5} / generation in the current study versus 1.98×10^{-5} /generation) (Phillips et al. 2009).

Comparison of Mutation Rate among Chromosomes

While a significant difference in overall di-nucleotide microsatellite mutation rate was detected between the two species, no significant effect of chromosome type (autosome vs. X-chromosome) was detected in either species. There are two plausible reasons for this observation, one experimental and one biological. First, assuming a difference truly exists between chromosome types, it is possible that the present study lacks the statistical power to detect any difference between chromosome types. Given the rarity of mutations at any particular locus, more loci on both types of chromosome are needed to confidently reject the possibility of differential mutation rates between chromosome types. Alternatively, it is possible that a modifier of the X-chromosome

mutation rate has not had sufficient time to operate in *C. elegans* given the 0.32 X 10^6 to 23.3 X 10^6 generation since the establishment of selfing in this species (Cutter 2006). Asexual, as well as predominately selfing taxa, tend to be short-lived linages compared to related outcrossing taxa. This reduced longevity is thought to be a result of the accumulation of deleterious mutations via Müller's Ratchet, resulting in an ever increasing mutational load that, in turn, leads to an inevitable decline in fitness and the eventual extinction of the lineage (Lynch and Gabriel 1983). Therefore it is possible that there has not been sufficient time for *C. elegans* populations to reach the necessary equilibrium required for the effects of a modifier of X-chromosome mutation rates to become detectable.

Comparison of Indel Spectrum between Species

In general, studies of the mutational properties of microsatellite repeats commonly observe a bias towards mutations that lead to the expansion of the number of repeats through insertions (Ellegren 2004). Two previous studies of the mutational properties of di-nucleotide microsatellites in *C. elegans* have shown a significant bias towards insertions (Seyfert et al 2008; Phillips et al. 2009). The results of this study support the conclusion that *C. elegans* does indeed have an insertion bias at di-nucleotide microsatellite loci. Similarly, *C. remanei* had a greater number of insertions compared to deletions. The insertion bias observed for both species in this study is contradictory to the deletion bias shown in substitution patterns within *mariner* transposons between *C. elegans* and *C. briggsae* (Witherspoon and Robertson 2003). In addition, Philips et al. (2009) observed a strong deletion bias in di-nucleotide microsatellite loci in *C. briggsae*, the sister species to *C. remanei*. The discrepancy between patterns of indel bias accumulated over long evolutionary time spans, such as substitutions in

pseudogenes, and more recent patterns observed in MA experiments, raises an interesting possibility that indel bias has recently changed in *C. elegans* and *C. remanei*. There may be two possible explanations for such a shift. First, it is possible that the selection pressure has shifted in such a way to no longer favor deletions, but instead to favor insertion. Denver et al. (2004) suggested that such a shift in selection might explain the insertion bias that was observed in random nuclear loci they sequenced in a set of *C. elegans* MA lines. Alternatively, one of the central models of genome size evolution states that a mutational equilibrium will exist when the amount of DNA loss by frequent small deletions equals the input of DNA sequence by more, rare, large-scale insertions (Petrov 2002). It is possible that both *C. elegans* and *C. remanei* are not, at present, in an equilibrium state between the rate of insertions and deletions.

Correlation with Standing Genetic Variation

The interpretation of the relationship between mutation rates and standing genetic variation is difficult. The raw correlation between the per-locus mutation rate and the effective number of alleles at that locus among the 17 wild isolates was significant. This pattern is consistent with the previously observed correlation between the per-locus mutation rate and standing genetic variation in *C. elegans* and *C. briggsae* (Phillips et al. 2009). However, when repeat number is taken into consideration no significant relationship remains. One possible explanation is the uncertainty of the exact repeat length for each locus in the PB2282 genotype used in this study. All loci were designed from a different geneotype of *C. remanei* (strain EM4641). Alternatively, it is possible that we simply lack the statistical power to detect a moderate positive partial correlation of the magnitude of the point estimate from this study (and from that in *C. briggsae*, Phillips et al. 2009). We are currently sequencing several complete genomes of the

PB2282 genotype used in this study using Illumina technology. From this data we will be better able to get an exact repeat size for each locus. Until then, there is reason to believe, based on the similarity of the *C. remanei* data to that of *C . elegans* and *C. briggsae*, that a positive relationship between the per-locus mutation rate and levels of standing genetic variation does exist.

Conclusions

The current study provides empirical support for the theoretical prediction that natural selection plays a greater role in shaping the rate of spontaneous mutations across the genomes of selfing than in outcrossing species. In addition, these results are consistent with our previously observed patterns of the mutational properties at the phenotypic level in the same set of MA lines. This provides further evidence for a close relationship between the rates of spontaneous mutations estimated at both the phenotypic and molecular levels in *Caenorhabditis*. Conversely, it appears that natural selection has not led to a reduced spontaneous mutation rate between autosomes and the X-chromosome in either of the two *Caenorhabditis* species included in this study, although the statistical power of this study may simply be insufficient to detect a difference. While mutation rates did differ between species, the indel spectrum did not, with both species showing an excess of insertion mutations. However, the current study utilized a relatively small number of loci, and future studies using higher throughput, whole genome DNA sequencing methods will be needed to definitively determine if mutation rates differ by chromosome type in *Caenorhabditis*.

Table 4-1. PCR primers used to amplify all loci in *Caenorhabditis elegans* and *C. remanei*. Chr = chromosome number in *C. elegans* and reference contig number with assigned chromosome number in parentheses for *C. remanei*. Start = the genome coordinates of the beginning of the microsatellite repeat. Repeat number is the number of di-nucleotide repeats, i.e. repeat number of 8 = 16bp. PCR product sizes are indicated in the last column.

Species	Locus	Chr.	Start	Repeat number	Forward primer	Reverse primer	PCR product
Cel	Cel_1	II	7227732	8	CCGAATAAAAGGGAACGGAG	CAATGACGTGGCAAAAGAGA	324
	Cel_2	II	12961355	8	CCACCCCAAAATGACCATAG	ATTGTCACATTTCGCTGCTG	169
	Cel_3	II	7850841	9	GGTAACATCAAATGTCCGGG	TTGAGCAAGTGTGGCTGTTC	396
	Cel_4	II	3838988	9	CAGAAAAATAGGCGGACCAA	AACTCCTCTACTGCGCCTCA	351
	Cel_5	II	3952501	8	CAGACACTCACAGCGTTGGT	CTCCGGTTCCGAATTATCAA	329
	Cel_6	II	4391268	9	TTTCATCAGAGCACGATTGC	ATGCGATGTTTGGTTCATCA	397
	Cel_7	II	10366804	9	TCCCATGTTTCTTGTGGTCA	TTGGTTGACCAATTTTTAGGTG	208
	Cel_8	II	10743591	10	GTGCCAAACCACAACATGAA	GCCCACTTACTCTCGTCTCG	247
	Cel_9	II	10612895	10	TCCCCTTCTCCCTCTTCATT	GGATGGGAGGAGCACAAATA	257
	Cel_10	II	11890727	10	TTCCCGCAATACCAAATCTC	GGGCCACCTCAACTAAACAA	230
	Cel_11	II	1989829	11	TTGAAAAGCCGATTGGAAAC	TTCAGTGCACGGAGAGTCTG	353
	Cel_12	II	15194037	11	AGAGCAGCACACACAAATCG	CTCGTAATCCTCTCTCCCCC	270
	Cel_13	II	10549040	12	TCTTCCACACGCTCACAGAC	CCCCATCTTCTCCATTTCAA	363
	Cel_14	II	14765429	13	ATAGTTAGGCCTGACGGAGC	TTAATGTTCCCGCGAAAAAG	230
	Cel_15	II	10547953	15	GTGCATCGGTGGGAGACTAT	TCCTTTGTTATGGCACCCTC	200
	Cel_16	II	9497222	9	CCATCTAGGTACGCAATTCCA	GATGAGTTGGAGCCCTTCAG	177
	Cel_17	II	11391680	2	GGAATTGAGGAAGCGAACAG	CCGCTTCAACATTCACATTG	392
	Cel_18	II	12052795	22	TGGCCCTTCAACTGAACTTT	ATCTGGTGAGAACCTGGTGG	338
	Cel_19	II	15264823	24	CCGCGCAAATATTGACTTTT	TTTCCCGTTGAGATACAGGC	321
	Cel_20	II	5322006	25	AAATCCACAGTGCTTTTGGG	CTCCCTTTCTCTGTACCCCC	285
	Cel_21	IV	12947009	9	AACAGCGCTGAGCTATTTCG	CTCACGATTTCGTGGGTCTT	239
	Cel_22	IV	6676681	8	CCCAGACTTCCCAACTCAAA	CTGGCCCGTTAGACCAATAA	109
	Cel_23	IV	5172466	9	TCTCCGATTGTGAGCATCTG	TGGGCGAAAACTATTGAACTTT	374

Table 4-1 Continued

Species	Locus	Chr.	Start	Repeat number	Forward primer	Reverse primer	PCR product
	Cel_24	IV	12754725	8	AAGGGGATGGGAGAGAAGAA	ACGAGGAAATGCGTATGGAC	218
	Cel_25	IV	7983643	8	CGCAGCCGTTTTAATTTTGT	GGTTTCGGAGAGTTTGACCA	185
	Cel_26	IV	7390527	9	GGACCCAGACTGGCATAAAA	CGGGAAATGAAGGAACTTGA	154
	Cel_27	IV	4490492	9	TAAAACCGGAAGACCACTGC	AGAGACGTCGCTTCTTCTGC	198
	Cel_28	IV	13469567	9	GCTATATGAAACTTGGCGGG	CCCTTCCGTCCAATACCTTT	370
	Cel_29	IV	1523670	10	TTTCAGCCAGCTGGAATTTT	CACTTGGCAAAGAGATGGGT	178
	Cel_30	IV	2753761	10	ATTTTCCCGGAGGAGTTACG	GTATCACGGCGAGAGGAGAG	128
	Cel_31	IV	2080792	11	ACGTGGACGGAAATTGTGTT	TTTTAAAGTTCCGGCAAACC	381
	Cel_32	IV	11364280	12	TTCGTTTGACTGTCTGCGTC	CGGAAGCTTTTTGTGTCTCC	193
	Cel_33	IV	1616951	13	TATCAACCGCCCGTCTTATC	AGCCAACCGGTGAGTATTTG	343
	Cel_34	IV	3742521	14	TTCCCTATGCTGCCTCTCAT	AGAGACCGACGACGAGATGT	316
	Cel_35	IV	14547429	16	AGTTCCCGTGAAGCTTGAAA	AACCAATCAGCGCCTTCTAA	335
	Cel_36	IV	11899513	16	CTTTTATTCCGCCTCCAGTG	ATTACACAGGAAACAGGCGG	364
	Cel_37	IV	5561641	19	CTTTTCAACGACCTCGACAG	TATTTCCCTCAATTCGTCGG	180
	Cel_38	IV	7279971	25	GCGAAACGGAAATGTTGAAT	ATGGACAAACGGAATAAGCG	346
	Cel_39	IV	67795	26	TGCACTCCAGTTTCTTGTGC	TCCGAAATCCTGGAGATTCA	309
	Cel_40	IV	4454515	30	TGCTTCTTCTCCTTCACCGT	TTTGCATTCTTCCGCTCTCT	186
	Cel_41	X	11249733	8	ATCACGCGGAACACATTACA	ACTTGCAATTTCGCTTTCGT	321
	Cel_42	X	2920621	8	TTGTGTCAGCTCCAATCTGC	TTTGTGCAGCCATCATAAGG	222
	Cel_43	X	12670382	9	GGGCTTCTATGGCAACACAT	TCTCAACTGTCACTCGCACC	356
	Cel_44	X	13940928	8	GCTTCTCCGAATATGCCAAA	TTTGGTTCGATACCAGTAGCA	285
	Cel_45	X	14561516	8	ATGCACGTCAGTTCCCTTCT	CGGAGAAATGGGATCGAGT	309
	Cel_46	X	2641220	8	AAAAGAAGGCTGCTCCACAA	TCATAAGCCTTGTCACATCCA	220
	Cel_47	X	3355405	8	CGTACGACACTTTTCGGGTT	AAATTTGTCGGGGGAAGAGT	281
	Cel_48	X	4506943	10	CCGAGGTGTGGAAATAACTTG	TCAAAATCGGGCATTCTAGG	309
	Cel_49	X	12080930	10	GAAGGCGTCTTTCTGTCTGC	GTACCGCCGAAAGGTAATGA	340

Table 4-1 Continued

Species	Locus	Chr.	Start	Repeat number	Forward primer	Reverse primer	PCR product
	Cel_50	X	15340980	9	GATCGTCTAACCTTTTGCCG	AATTGTTTGGGGGAGGAAAC	273
	Cel_51	X	5733980	9	AATGAGCCTGGGAATGTCTG	CCCTACCTCACCTAAAGCCC	177
	Cel_52	X	1199052	10	ATCATGGGCACTAAAGCTGG	TCCCACCTCTGCTCTCATCT	241
	Cel_53	X	10808511	9	TCCTCCAGTTCCTGGTGTTC	GCGGTCAAATAAACTGGCAT	284
	Cel_54	X	17514630	10	CCAGGTGGAAATAGGGGTTT	TCGATTCCGGATGCTTATTC	288
	Cel_55	X	12013797	10	GCTTTCTACCGCCAATATGC	AAAAATGGGCGGAGCTTAAC	279
	Cel_56	X	13264748	10	CCTCCGCTCTACCAACTGAG	GGCACTTTTCACGCTTCTTC	160
	Cel_57	X	3867976	11	GGACAAGAGTGGCGATGAAT	CCTTCTTCTACATGGGGCAA	310
	Cel_58	X	940771	12	AAGGTGAAACGAGGTTGTGG	GGAAATGGCGAAATTGAAAA	370
	Cel_59	X	10629304	12	GTAGGTTAGTCGTGGTCCGC	TCCCATTTGGTCTCTCCATC	162
	Cel_60	X	16684711	13	GGTGCAAGGAATTTGACGTT	TCGAATTACGGGACAGAACC	335
	Cel_61	X	7379218	13	TTTCCAATAAATCCCACCCA	CCACTATTTCGGGCTTCAAA	246
	Cel_62	X	1520216	14	CCCAGCAAACCCACTAGGTA	CCCTCCGAACACACTTGAAT	157
	Cel_63	X	16339303	16	CCAGTCACGTGGAAACGTC	AAATGTCTCTCGGCTCTCCA	209
	Cel_64	X	984580	19	TTCCGTGGCTTTGAAAACTT	TACAGCTGCGGATTCTGATG	241
	Cel_65	X	7527688	20	TTTGGAAAAATGGGTATCGC	TCGAGCCAATAAAACCCATC	181
	Cel_66	X	1224249	21	GAAAGAGCGACATACGGAGC	ATTTACCGCACACCCATGAT	342
	Cel_67	X	4319206	24	CTCCGAGAGCACCGTAGTTC	CACTACACGAGGAAGTGCGA	287
	Cel_68	X	9225274	26	TGTGGACCATCAAACCTCAA	TACCGGTGGGGTATGGTAGA	323
	Cel_69	X	11913941	28	CACGGTTTTTGTTCCGTTCT	CCTCAATTCCTGCGTTGAAT	326
	Cel_70	X	4982083	28	TCATCCACCATTCTTACGCA	ACGACCAGAACGAAATACGG	221
Crem	Crem_1	Contig1(II)	1505667	8	ACATTGGCAGATCTTCCGTC	TCATACCTGTGGATTGCGAA	398
	Crem_2	Contig6(II)	148970	8	AGTTCCCATGAGATGCCAAA	TCACTCCTCCTCGTTTCACC	224
	Crem_3	Contig6(II)	1652084	8	GCGACTTTTGGAGAATGAGC	TGACTCCATCACATCCCAGA	198
	Crem_4	Contig1(II)	470444	8	TGAGAGGAGATGGAGAGGGA	GGCTTTATTCTGCTTGCCTG	350
	Crem_5	Contig17(II)	361211	8	AATTGGCTTCCATTTTCGTG	CAGCACCAACAACTTCCAGA	333

Table 4-1 Continued

Species	Locus	Chr.	Start	Repeat number	Forward primer	Reverse primer	PCR product
	Crem_6	Contig17(II)	63610	9	TTGGGTGGGTTCTCCTACTG	ATATGGTTCGAACGGCAAAG	200
	Crem_7	Contig101(II)	82047	10	GACCGACGGTAAGTGATGCT	CCCCGTAAATCTACCCCAGT	215
	Crem_8	Contig1(II)	1356960	10	TCTCCGGTCTATTGCACACA	CCATCATTCCAACCAGTTCC	193
	Crem_9	Contig242(II)	86558	10	GCCCCCATCCATATATAGCC	CATCACGATGTCTGCGTCTC	291
	Crem_10	Contig176(II)	119754	11	GTGGCAGCAGTGACAGAAAA	TCTCTCGCTCAAGGGATTGT	182
	Crem_11	Contig95(II)	223194	11	AGGCTCAGCAAAAACCAAGA	TTTCCAAAGAAGGGCAAGTG	190
	Crem_12	Contig1(II)	2064941	11	GAAGAGCGGATAAAAGCGTG	CGCTTGGATATGCGAAATCT	211
	Crem_13	Contig1(II)	1017753	12	GTCAGAGAAAAGTGGGACGC	AGAGAGCATGAGACCGGAAA	299
	Crem_14	Contig1(II)	1955413	12	CTTTCTCCGGATTCATTCCA	TGAGGATCATTTTCGAACCC	196
	Crem_15	Contig23(II)	959355	18	TTAGACACCCCCGTCACTTC	GGAGGGAGTGAACGGTACAA	160
	Crem_16	Contig423(II)	6744	18	GCCAACCTGGTAACCGTCTA	CGCCCTTCTCAATTTCTCAG	268
	Crem_17	Contig1(II)	179644	21	ATTTGCCAAAACCGAAACTG	TTTCAGAACACCTCCGGAAC	346
	Crem_18	Contig39(II)	731944	11	CGGAGAAGGAGTGTGTTGGT	TAATGTTCGATGGCTCCTCC	247
	Crem_19	Contig14(II)	515033	15	TCCGGTAGGTCTTCATGGAC	GACTCACTCCGCCTCAACAT	273
	Crem_20	Contig57(II)	137400	33	AAGTTCCAAACGAATCGTCG	ATTGATTTTATAGGCCGGGG	330
	Crem_21	Contig18(IV)	1255264	8	CGCCTTTTCATAATCGCAA	TTTGGGACGTTCGAGTTTTC	349
	Crem_22	Contig817(IV)	4150	8	GCGAATGAAGTGTGTGATGG	AACTGATCGCATCTGGTTCC	355
	Crem_23	Contig1538(IV)	4365	8	TGGTTGGTGGATGAACTGAA	GAGGTGGATGGATATGGTCG	292
	Crem_24	Contig88(IV)	82980	9	AAATTGGCGTAGAATGTCGC	TTTCGTAGAACGTCACAGCG	218
	Crem_25	Contig18(IV)	364230	9	GGAGGGAGACTCAAAAAGGG	TACCAACCAAATCCCCTTGA	372
	Crem_26	Contig62(IV)	238124	10	CCAACTACCTGTGGCCATCT	GAAGACACTTGCTTTTCGGC	179
	Crem_27	Contig96(IV)	209462	10	AGCAAACTTTCGGAGCAAAA	TCTGTGGGGAGAGAAAAGGA	307
	Crem_28	Contig652(IV)	18340	10	GGTAATTGAATCGACACGGG	TCGAATACCGCAAGCTTTTT	248
	Crem_29	Contig18(IV)	161029	10	TGCTTTTGAGCACATTTTCTG	CGGTTCGAACTCGAAATCAT	326
	Crem_30	Contig18(IV)	405201	11	ACGTGTCGTGTGCTCTTTTT	TCGATGCAACTGAACCTGTC	348
	Crem_31	Contig126(IV)	36257	10	TGTAGGCGTGATGGTGAGAG	AGGGAGAAACCTCGGTCAAT	205

Table 4-1 Continued

Species	Locus	Chr.	Start	Repeat number	Forward primer	Reverse primer	PCR product
	Crem_32	Contig19(IV)	130079	11	GGCTCTCTGTACACTTCCCG	AGGCAAACGAAGGGAAAGAT	239
	Crem_33	Contig383(IV)	33924	11	GGTAAGCAGCTCCGAAAGTG	AAAATAGTGAGCCCCGTCCT	365
	Crem_34	Contig1203(IV)	9684	11	TCTTCAATTCAATTTCCGGTG	TTCAAGGAATTCCGTTTTCG	186
	Crem_35	Contig275(IV)	10682	13	AAATAAATAGTGCGCGTGGG	TCAGCACATTTTCCCACAAA	212
	Crem_36	Contig1062(IV)	2230	13	ATCCAATTCAATCCTCACCG	GGGACGTCTTCAGACGAGAG	128
	Crem_37	Contig127(IV)	202882	21	AATGGACACCTCAATTCCCA	GTCTAACAGTCTGCGCTCCC	267
	Crem_38	Contig88(IV)	204840	23	GGAAACGGCGACATAAGGTA	CTGTCGTCGTCGTTGTCAGT	209
	Crem_39	Contig1397(IV)	7906	30	TTCCGTTATCGCATCCTTTC	TTTGCACTCCGAATTGACCT	339
	Crem_40	Contig18(IV)	825684	31	CCCCTCCTGTCTTCATTCAA	ATGCATGGAGGCATTTTCTC	237
	Crem_41	Contig408(IV)	2373	43	TGTTTTTAGTTGGTTCCGCC	GTCAATTGAGAAACGCGTCA	313
	Crem_42	Contig0(X)	2394041	8	CTAACAAATTTGGCGTGCCT	GCTTCTCCACCCATTTTCAA	396
	Crem_43	Contig0(X)	4331840	8	CCGGTTGTCACTTCCAATCT	ATATGGCTGCCCTTTACACG	236
	Crem_44	Contig0(X)	354475	9	TAATTGACGAGCACTTTGCG	TTCTTCCTTTGCCATCATCC	386
	Crem_45	Contig110(X)	5708	9	CATGGCCAAACATAGCATTG	GCTTTGAGACCATTTCGAGC	240
	Crem_46	Contig171(X)	67196	8	AATGGCTGTGATCCTGAACC	ATAGGACAGGACCGACGATG	376
	Crem_47	Contig50(X)	43231	8	TTCGCGAGAGAGGAAGAGAG	CCGGGAAAAAGTGGTAGTGA	253
	Crem_48	Contig63(X)	33361	8	TGTTGGTTTTCCCACATTAGG	TCCGCGAAAAGTTCTTTGAT	378
	Crem_49	Contig0(X)	2245153	9	CGTTGGTTACGCATCATTTG	GGCCCTCTTTCATTTTCTCC	270
	Crem_50	Contig27(X)	729555	9	TCGGAAAACTCTCGAAAACG	GGTGGAGGGTATAGGGGAAA	149
	Crem_51	Contig27(X)	362881	8	CTGCCTGGAAAATTGAGGAA	ATTTCGGCCACTCACAAAAC	286
	Crem_52	Contig0(X)	3582349	9	GGGTATGCGAACTTCACTTCA	TTCAAGAAAATGAGCCACCC	305
	Crem_53	Contig0(X)	2286485	8	TCACAAATGTGGCGTTTGAT	ATCCTTCGTCGTTTCACGAT	254
	Crem_54	Contig24(X)	446000	8	TTTCTATGTGCCCCGTTAGG	GAGAAAACGCACAAAGAGCC	290
	Crem_55	Contig27(X)	602465	9	AAATGAATCCCGGGGAATAG	TATGAGCAAAACGTAGGGGG	308
	Crem_56	Contig50(X)	31793	9	GGAAAACCACTGACCCGATA	GCCGTGATACGGAACAGATT	262
	Crem_57	Contig15(X)	952829	10	ATTACGCCCGTAAAACTCCC	CGTCGTAAACTTCGTGAGCA	400

Table 4-1Continued

Species	Locus	Chr.	Start	Repeat number	Forward primer	Reverse primer	PCR product
	Crem_58	Contig31(X)	88169	10	AGCCGTTCCTGAGCAACTAA	CGGCAGTCTTTTTCCAAATC	380
	Crem_59	Contig31(X)	466824	9	GAACCTGTCATGTGCCAGAA	CAGCGGACACTCAGAAAACA	362
	Crem_60	Contig93(X)	87084	10	TGCACTTCCTTATGGATGGA	ATTCCCACACCTTTTCTCCC	160
	Crem_61	Contig93(X)	317705	10	GTGTGCTCCGCAGACAGTAA	AAAGAAACGAAATGCGATGG	271
	Crem_62	Contig110(X)	6088	11	CATTCCGCGAAATTTCTGAT	GCGATTTTTCTGCATTTTCC	324
	Crem_63	Contig31(X)	217239	11	TTTTTCTCCACACCACCCTC	ATCGCGCTCTTCACATTTTT	178
	Crem_64	Contig0(X)	3553370	11	CTCACGGATCCCTTTGTGTT	GCCCATTGTGTTTTTCGAGT	345
	Crem_65	Contig63(X)	106732	10	ACACATTCGCATCACCAGAA	CTTTACACTTTTCGCGCTCC	377
	Crem_66	Contig208(X)	113813	11	TGAGAAACGCATTACGTGGA	TCTCCGCTGCAATAACACAC	318
	Crem_67	Contig31(X)	592786	11	GGGACTGACTTGCTTTCGTC	GTTTCCGCGTTAGCAGAGAG	206
	Crem_68	Contig0(X)	2916438	12	GACGTGTGCCTCTCTTTTCC	GCGCTACACGTCCTTCTTTC	298
	Crem_69	Contig31(X)	470920	11	GTGGTGTCTTCTCTCTCGGC	CGAAACGAATAGGAATCGGA	168
	Crem_70	Contig0(X)	3864041	14	TTTATCCCAAAACACCGCAT	CTATTGCCACAAAAACCCGT	301
	Crem_71	Contig24(X)	445852	18	CCCATCCCACTCCACATATC	CCTAACGGGGCACATAGAAA	304
	Crem_72	Contig145(X)	228387	34	TTTTTGCCACATAGCCACCT	CCTGCTGGTGACCAAGGTAT	268

Table 4-2. SAS code used to fit the generalized linear model to test for variation in mutation rates between *C. remanei* and *C. elegans fog-2*. Bold terms are SAS code key words and lowercase words are the parameters specified for this model. Variables are, species, the number of repeats at a locus, chromosome number, and locus.

```
PROC GLIMMIX data=Outcross_msats pconv=0.001;
NLOPTIONS maxiter=40;
CLASS species locus chromosome;
MODEL p/N = species|repeat_number|chromosome / SOLUTION dist=binomial
LINK=logit ddfm=kenwardroger;
RANDOM _resid_ / subject=locus group=chromosome(species);
LSMEANS species|chromosome /ilink;
RUN;
```

Table 4-3. Summary of the per-generation mutation rate for all loci in *Caenorhabditis elegans* and *C. remanei*. μ_{OBS} is the average mutation rate estimated from the raw data. μ_{LS} is the mutation rate estimated by the least-squares means method. MA lines of *C. elegans fog-2* were genotyped at 200 generation of MA, while the *C. remanei* MA lines were genotyped at 122 generations of MA.

Chromosome		*C. elegans*		*C. remanei*	
II	*N* loci	15		16	
	Ave. *N* lines	64		46	
	Ave. repeat #	12.85		12.1	
	μ_{OBS} (SEM)	0.52E-05	(0.52E-05)	3.32E-05	(2.41E-05)
	μ_{LS} (SEM)	0.42E-05	(0.44E-05)	3.50E-05	(3.30E-05)
IV	*N* loci	18		14	
	Ave. *N* lines	64		46	
	Ave. repeat #	13.05		14.33	
	μ_{OBS} (SEM)	3.98E-05	(2.77E-05)	6.34E-05	(2.36E-05)
	μ_{LS} (SEM)	2.50E-05	(1.80E-05)	15.7E-05	(6.50E-05)
X	*N* loci	27		25	
	Ave. *N* lines	64		46	
	Ave. repeat #	12.93		10.1	
	μ_{OBS} (SEM)	1.72E-05	(0.76E-05)	2.89E-05	(1.36E-05)
	μ_{LS} (SEM)	1.80E-05	(1.00E-05)	9.40E-05	(6.50E-05)

Table 4-4. Summary of the indel spectrum. For each repeat type, row headings are: N loci, the number of loci genotyped for a given repeat type; the mean number of lines genotyped; Deletion (Insertion) > 1 is the number of mutant alleles > 1 repeat unit (2 bp) shorter (longer) than the ancestral allele; Deletion (Insertion) = 1 is the number of mutant alleles 1 repeat unit (2 bp) shorter (longer) than the ancestral allele. Insertion bias is the proportion of mutations that are insertions.

Chromosome		*C. elegans*	*C. remanei*
II	*N* loci	15	16
	Ave. *N* lines	64	46
	Deletion > 2bp	0	1
	Deletion = 2bp	0	0
	Insertion > 2bp	0	0
	Insertion = 2bp	1	2
	Insertion bias	1.00	0.67
IV	*N* loci	18	14
	Ave. *N* lines	64	46
	Deletion > 2bp	0	0
	Deletion = 2bp	2	0
	Insertion > 2bp	1	1
	Insertion = 2bp	6	2
	Insertion bias	0.78	1.00
X	*N* loci	27	25
	Ave. *N* lines	64	46
	Deletion > 2bp	0	1
	Deletion = 2bp	2	0
	Insertion > 2bp	0	0
	Insertion = 2bp	3	3
	Insertion bias	0.60	0.75

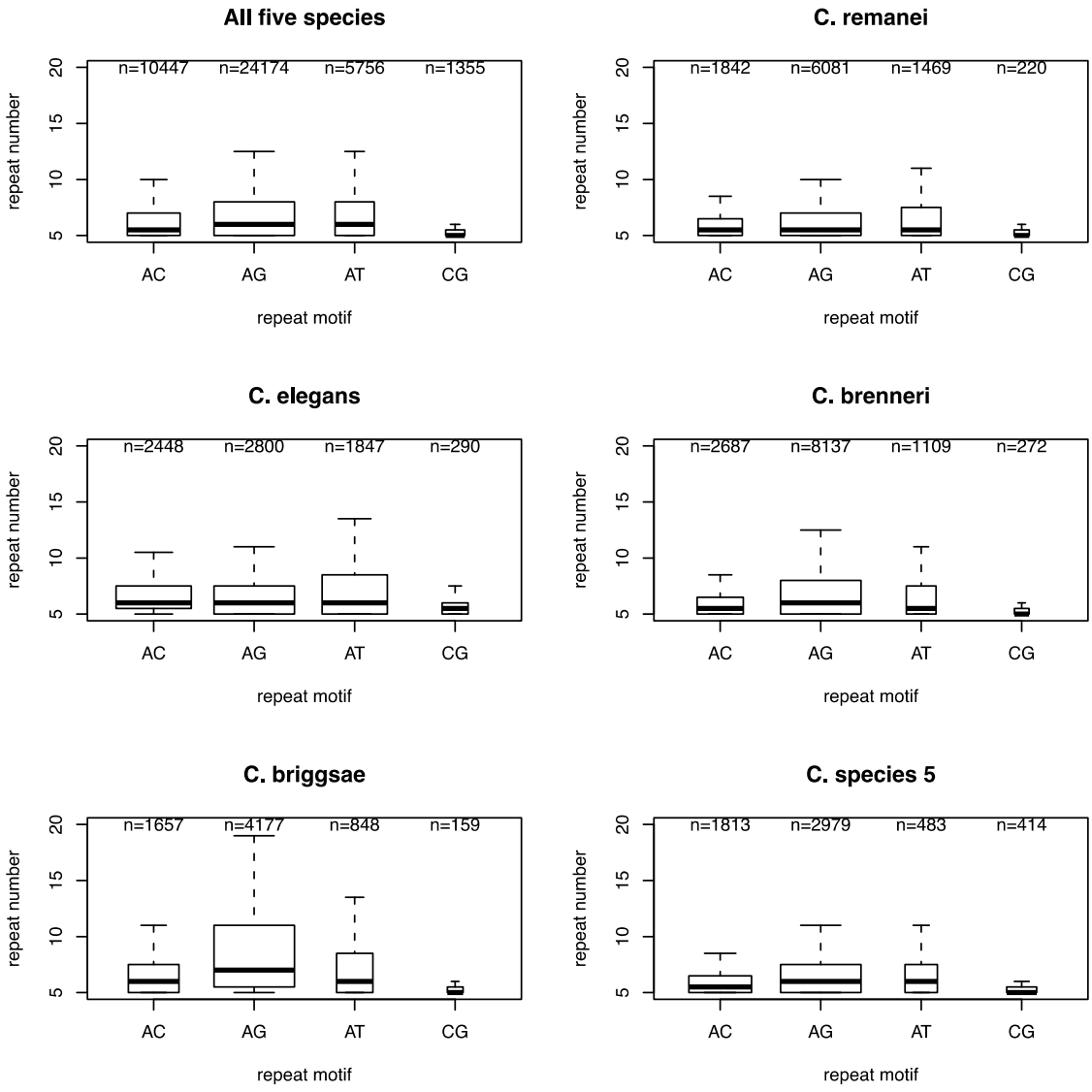

Figure 4-1. Boxplots of di-nucleotide microsatellite distributions for five species of *Caenorhabditis*. Among all five species, AG reapeats are the most abundant. The width of the boxes corresponds to the variance in repeat number within a given group. Extreme values have been removed for clarity.

CHAPTER 5
SUMMARY

Do differences in the mutational properties of different taxa influence the levels of standing genetic variation observed in nature? The overall results of the three studies described here suggest they do. First, when examining how differences in mutational properties influence levels of standing genetic variation and averaged over traits, species, and populations within species, the relationship between V_G and V_M is quite stable and consistent with the hypothesis that differences among groups in standing variance can be explained by differences in mutational input. With one exception, the variance present in a worldwide sample of these species is similar to the variance present within a sample from a single locale. These results are consistent with species-wide MSB and uniform purifying selection, but genetic draft (hitchhiking) is a plausible alternative possibility. The results of this study illustrated that differences in the mutational properties are indeed reflected in levels of standing genetic variation at the phenotypic level. Additionally, the observed levels of standing genetic variation were consistent with the hypothesis of mutation-selection balance acting to maintain variation in fitness and body size in both species.

Similar to the phenotypic data, the di-nucleotide microsatellite mutation rate of *C. briggsae* was indeed approximately two times that of *C. elegans*. Furthermore, the mutation rate and spectrum differed significantly between the two species, with the repeat motif AG having the highest mutation rate. Additionally, the indel bias was significant between the two species with *C. elegans* showing a significant insertion bias. Integrated over the whole genome the mutation rate of *C. briggsae* is about twice that of *C. elegans*, consistent with the cumulative mutational effects on fitness observed

previously. The per-locus mutation rate is significantly positively correlated with the standing genetic variation at the same locus in both species, providing justification for the common practice of using the standing genetic variance as a surrogate for the mutation rate.

Finally, the third study provides support that natural selection has shaped the rate of spontaneous mutations to different degrees in species that differ in their reproductive strategies. The findings of this study agree with the previously observed phenotypic patterns assayed in the same set of MA lines. While the results of this study are consistent with what is predicted by theory at the species level, no support for natural selection shaping the rate of spontaneous mutations at the chromosomal level was found.

With the rapid advances in DNA sequencing technology that are occurring at an ever-increasing rate, the ability to more accurately quantify the mutational process is just around the corner. In the near future, the use of highly parallel DNA sequencing methods will allow for the interrogation of the mutational process at the scale of the whole genome. In addition, it soon will be possible to extend the genomic techniques now reserved for model organism to a greater diversity of organisms, thus allowing for a more complete picture of the degree of variation in mutational properties across the tree of life. Armed with these data, a clearer picture of the processes that underlie the origin and maintenance of genetic variation will begin to emerge and, undoubtedly, the role of different mutation processes among taxa will be found to play a major role.

LIST OF REFERENCES

Amos, W, Flint J, Xu X. 2008. Heterozygosity increases microsatellite mutation rate, linking it to demographic history. BMC Genet. 9:72.

Amos, W, Sawcer S, Feakes R, Rubinsztein D. 1996. Microsatellites show mutational bias and heterozygote instability. Nat Genet. 13:390-391.

Arndt, PF, Hwa T, Petrov DA. 2005. Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. J of Mol Evol. 60:748-763.

Azevedo, R, Keighdey P, Lauren-Maatta C, Vassilieva L, Lynch M, Leroi A. 2002. Spontaneous mutational variation for body size in *Caenorhabditis elegans*. Genetics. 162:755-765.

Baer, CF, Joyner-Matos J, Ostrow D, Grigaltchik V, Salomon MP, Upadhyay A. 2010. Rapid decline in fitness of mutation accumulation lines of gonochoristic (outcrossing) Caenorhabditis nematodes. Evolution. 64:3242-3253.

Baer, CF. 2008. Does mutation rate depend on itself. PLoS Biology. 6:52.

Baer, C, Miyamoto M, Denver D. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat. Rev Genet. 8:619-631.

Baer, C, 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, C, Shaw F, Steding C, et al. 2005. Comparative evolutionary genetics of spontaneous mutations affecting fitness in rhabditid nematodes. Proc Nat Acad Sci U S A. 102:5785-5790.

Barriere, A, Yang S-P, Pekarek E, Thomas CG, Haag ES, Ruvinsky I. 2008. Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes. Genome Res. 19:470-480.

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

Bégin, M, Schoen DJ. 2006. Low impact of germline transposition on the rate of mildly deleterious mutation in Caenorhabditis elegans. Genetics. 174:2129-2136.

Byers, D, Waller D. 1999. Do plant populations purge their genetic load? Effects of population size and mating history on inbreeding depression. Annu Rev Ecol Syst. 30:479-513.

Crow, J. 1958. Some possibilities for measuring selection intensities in man. Am. Anthro. 60:1-13.

Crow, J. 1993. Mutation, mean fitness, and genetic load. Oxford Surveys in Evolutionary Biology 9:3-42.

Cutter, AD. 2006. Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer C*aenorhabditis elegans*. Genetics. 172:171-184.

Cutter, AD, Félix M-A, Barrière A, Charlesworth D. 2006. Patterns of nucleotide polymorphism distinguish temperate and tropical wild isolates of *Caenorhabditis briggsae*. Genetics. 173:2021-2031.

Cutter, AD. 2008. Divergence times in Caenorhabditis and Drosophila inferred from direct estimates of the neutral mutation rate. Mol Biol Evol. 25:778-786.

Cutter, AD, Wasmuth, JD, Washington, NL. 2008. Patterns of molecular evolution in *Ceanorhabditis* preclude ancient origins of selfing. Genetics. 178:2093-2104.

Davies, EK, Peters AD, Keightley PD. 1999. High frequency of cryptic deleterious mutations in *Caenorhabditis elegans*. Science. 285:1748-1751.

Denver, DR, Morris K, Thomas WK. 2003. Phylogenetics in *Caenorhabditis elegans*: an analysis of divergence and outcrossing. Mol Biol Evol. 20:393-400.

Dieringer, D, Schlotterer C. 2003. MICROSATELLITE ANALYSER (MSA): a platform independent analysis tool for large microsatellite data sets. Mol Ecol Notes. 3:167-169.

Dolgin, ES, Charlesworth B, Baird SE, Cutter AD. 2007. Inbreeding and outbreeding depression in Caenorhabditis nematodes. Evolution. 61:1339-1352.

Dolgin, ES, Felix M-A, Cutter AD. 2008. Hakuna Nematoda: genetic and phenotypic diversity in African isolates of *Caenorhabditis elegans* and *C briggsae*. Heredity. 100:304-315.

Drake, J, Charlesworth B, Charlesworth D, Crow J. 1998. Rates of spontaneous mutation. Genetics. 148:1667-1686.

Eisen, JA, Hanawalt PC. 1999. A phylogenomic study of DNA repair genes, proteins, and processes. Mut Res. 435:171-213.

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 5:435-445.

Gillespie, J. 2000. Genetic drift in an infinite population: The pseudohitchhiking model. Genetics. 155:909-919.

Halligan, DL, Keightley, PD. 2009. Spontaneous mutation accumulation studies in evolutionary genetics. 40:151-172.

Hardison, RC, Roskin KM, Yang S, et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res. 13:13-26.

Hill, WG, Robertson A. 2009. The effect of linkage on limits to artificial selection. Genet Res. 8:269-294.

Hodgkin, J, Doniach T. 1997. Natural variation and copulatory plug formation in *Caenorhabditis elegans*. Genetics. 146:149-164.

Houle, D. 1989. Allozyme-associated heterosis in *Drosophila Melanogaster*. Genetics. 123:789-801.

Houle, D. 1992. Comparing evolvability and variability of quantitative traits. Genetics. 130:195-204.

Houle, D, Morikawa B, Lynch M. 1996. Comparing mutational variabilities. Genetics. 143:1467-1483.

Howe, DK, Denver DR. 2008. Muller's Ratchet and compensatory mutation in Caenorhabditis briggsae mitochondrial genome evolution. BMC Evol Biol. 8:62.

Jackson, A. 2000. Microsatellite instability induced by hydrogen peroxide in Escherichia coli. Mut Res Mol Mech Mut. 447:187-198.

Keightley, P, Eyre-Walker A. 1999. Terumi Mukai and the riddle of deleterious mutation rates. Genetics. 153:515-523.

Keightley, P, Hill W. 1988. Quantitative genetic variability maintained by mutation-stabilizing selection salance in finite populations. Genet Res. 52:33-43.

Kelkar, YD, Tyekucheva S, Chiaromonte F, Makova KD. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res. 18:30-38.

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

Kimura, M. 1964. The number of alleles that can be maintained in a finite population. Genetics. 49:725-738.

Kimura, M, Ota T. 1975. Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. Proc Nat Acad U S A. 72:2761-2764.

Kiontke, K, Gavin N, Raynes Y, Roehrig C, Piano F, Fitch D. 2004. Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss. Proc Nat Acad U S A. 101:9003-9008.

Knapp, S, Bridges W, Yang M. 1989. Nonparametric confidence-interval estimators for heritability and expected selection response. Genetics. 121:891-898.

Kondrashov, A. 1988. Deleterious mutations and the evolution of sexual reproduction. Nature. 336:435-440.

Kondrashov, AS. 1995. Modifiers of mutation-selection balance-general approach and the evolution of mutation rates. Genet Res. 66:53-69.

Kondrashov, F, Ogurtsov A. 2006. Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. J Theor Biol. 240:616-626.

Lee, DH, Esworthy RS, Chu C, Pfeifer GP, Chu FF. 2006. Mutation accumulation in the Intestine and colon of mice deficient in two intracellular glutathione peroxidases. Cancer Res. 66:9845-9851.

Lercher, MJ, Chamary JV, Hurst LD. 2004. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. Genome Res. 14:1002-1013.

Lewis, JA, Fleming, JT. 1995. Basic culture methods. Pp 4-29 in Epstein, HF and Shakes, DC eds. Methods in Cell Biology, Vol. 48. Caenorhabditis elegans: modern biological analysis of an organism. Academic Press, London.

Lichtenauer-Kaligis, EGR, Dijke lvdV, Dulk Hd, Putte Pvd, Giphart-Gassler M, Tasseron-de JG. 1993. Genomic position influences spontaneous mutagenesis of an integrated retroviral vector containing the hprt cDNA as target for mutagenesis. Hum Mol Genet. 2:173-182.

Lynch, M. 2010. Evolution of the mutation rate. Trend Genet. 26:345-352.

Lynch, M. 2008. The cellular, developmental and population-genetic determinants of mutation-rate evolution. Genetics. 180:933-943.

Lynch, M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Nat Acad U S A. 104:8597-8604.

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

Lynch, M, Walsh B. 1998. Genetics and analysis of quantitative traits. Sinauer, Sunderland, MA.

Lynch M, Gabriel, W. 1983. Phenotyoic evolution and parthenogenesis. Am Nat. 122:745-764.

Matsuoka, Y, Vigouroux Y, Goodman MM, Sanchez GJ, Buckler E, Doebley J. 2002. A single domestication for maize shown by multilocus microsatellite genotyping. Proc Nat Acad U S A. 99:6080-6084.

McVean, G, Hurst L. 1997. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. Nature. 386:388-392.

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. Nature. 246:96-98.

Ostrow, D, Phillips N, Avalos A, Blanton D, Boggs A, Keller T, Levy L, Rosenbloom J, Baer C. 2007. Mutational bias for body size in rhabditid nematodes. Genetics. 176:1653-1661.

Paun, O, Hörandl E. 2006. Evolution of hypervariable microsatellites in apomictic polyploid lineages of *Ranunculus carpaticola*: directional bias at dinucleotide loci. Genetics. 174:387-398.

Petrov, D. 2002. Mutational equilibrium model of genome size evolution. Theor Pop Biol. 61:531-544.

Petrov, DA. 2001. Evolution of genome size: new approaches to an old problem. Trend Genet. 17:23-28.

Phillips, N, Salomon MP, Custer A, Ostrow D, Baer CF. 2009. Spontaneous mutational and standing genetic (co)variation at dinucleotide microsatellites in *Caenorhabditis briggsae* and *Caenorhabditis elegans*. Mol Biol Evol. 26:659-669.

Prendergast, JGD, Campbell H, Gilbert N, Dunlop MG, Bickmore WA, Semple CAM. 2007. Chromatin structure and evolution in the human genome. BMC Evol Biol. 7:72.

Rozen, S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. Humana Press, Totowa, NJ.

Sabeti, PC. 2006. Positive natural selection in the human lineage. Science. 312:1614-1620.

Schedl, T, Kimble J. 1988. fog-2, a germ-line-specific sex determination gene required for hermaphrodite spermatogenesis in *Caenorhabditis elegans*. Genetics. 119:43-61.

Schoen, DJ. 2005. Deleterious mutation in related species of the plant genus Amsinckia with contrasting mating systems. Evolution. 59:2370-2377.

Schuelke, M. 2000. An economic method for the fluorescent labeling of PCR fragments. Nat Biotech. 18:233-234.

Seyfert, AL, Cristescu MEA, Frisse L, Schaack S, Thomas WK, Lynch M. 2008. The rate and spectrum of microsatellite mutation in *Caenorhabditis elegans* and *Daphnia pulex.* Genetics. 178:2113-2121.

Sivasundar, A, Hey J. 2003. Population genetics of *Caenorhabditis elegans*: The paradox of low polymorphism in a widespread species. Genetics. 163:147-157.

Sniegowski, P, Gerrish P, Johnson T, Shaver A. 2000. The evolution of mutation rates: separating causes from consequences. BioEssays 22:1057-1066.

Sokal, R, Rohlf F. 1981. Biometry: the principles and practice of statistics in biological research. W. H. Freeman, New York, NY.

Sturtevant, A. 1937. Essays on evolution I On the effects of selection on mutation rate. Quart Rev Biol. 12:464-467.

Tian, D, Wang Q, Zhang P, Araki H, Yang S, Kreitman M, Nagylaki T, Hudson R, Bergelson J, Chen J-Q. 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature. 455:105-108.

Vassilieva, L, Hook A, Lynch M. 2000. The fitness effects of spontaneous mutations in Caenorhabditis elegans. Evolution. 54:1234-1246.

Vassilieva, L, Lynch M. 1999. The rate of spontaneous mutation for life-history traits in *Caenorhabditis elegans*. Genetics. 151:119-129.

Vigouroux, Y, Jaqueth J, Matsuoka Y, Smith O, Beavis W, Smith J, Doebley J. 2002. Rate and pattern of mutation at microsatellite loci in maize. Mol Biol Evol. 9:1251-1260.

Williams, B, Schrank B, Huynh C, Shownkeen R, Waterston R. 1992. A genetic mapping system in *Caenorhabditis elegans* based on polymorphic sequence tagged sites. Genetics. 131:609-624.

Witherspoon, D, Robertson H. 2003. Neutral evolution of ten types of mariner transposons in the genomes of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. J Mol Evol. 56:751-769.

Wolfinger, R, Oconnell M. 1993. Generalized linear mixed models - a pseudo-likelihood approach. J Stat Comp Sim. 48:233-243.

Wade, MJ. 2006. Natural selection. Pp 49-64 in Evolutionary genetics : concepts and case studies. edited by Fox, CW and Wolf, JB. Oxford University Press, Oxford, NY.

Wood, W. 1988. The Nematode Caenorhabditis elegans. Cold Spring Harbor Laborartory, Cold Spring Harbor, NY.

BIOGRAPHICAL SKETCH

Matt was born in Los Angeles California in 1975. Matt began his college career at Glendale Community College in Glendale California and completed a Bachelor of Science in biology at the University of La Verne in La Verne California. After completing his undergraduate degree at La Verne, Matt entered the Master of Biology program at the California State University Northridge where he worked under the guidance of Larry Allen as part of Dr. Allen's Nearshore Marine Fish Research Program. After graduating from California State University Northridge, Matt entered the PhD program in the Department of Zoology at the University of Florida working under Charles F. Baer.