Spring 2016

# The rate, spectrum and effects of spontaneous mutation in bacteria with multiple chromosomes

Marcus M. Dillon
*University of New Hampshire, Durham*

Recommended Citation
Dillon, Marcus M., "The rate, spectrum and effects of spontaneous mutation in bacteria with multiple chromosomes" (2016). *Doctoral Dissertations*. 2247.
https://scholars.unh.edu/dissertation/2247

# THE RATE, SPECTRUM AND EFFECTS OF SPONTANEOUS MUTATION IN BACTERIA WITH MULTIPLE CHROMOSOMES

BY

MARCUS MICHAEL DILLON

DISSERTATION

Submitted to the University of New Hampshire
in Partial Fulfillment of
the Requirements for the Degree of

Doctor of Philosophy
in
Microbiology

May, 2016

This dissertation has been examined and approved in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Microbiology by:

Dissertation Director, Cooper, Vaughn S., Associate Professor of Molecular, Cellular, and Biomedical Sciences; Professor of Microbiology and Molecular Genetics, University of Pittsburgh School of Medicine

Thomas, W. Kelley, Professor of Molecular, Cellular, and Biomedical Sciences

Culligan, Kevin M., Assistant Professor of Molecular, Cellular, and Biomedical Sciences

Lynch, Michael, Distinguished Professor of Biology, University of Indiana

Lang, Greg I., Assistant Professor of Biological Sciences, Lehigh University

On April 1$^{st}$ 2016

Original approval signatures are on file with the University of New Hampshire Graduate School.

# ACKNOWLEDGEMENTS

In my five years at the University of New Hampshire I have received enormous support from a variety of sources, both inside and outside the university. First, I would like to thank Dr. Vaughn Cooper for his unconditional support in all my academic endeavors and for giving me the freedom to pursue research questions that were not always within the scope of the mainstream research in his lab. I would also like to thank Dr. Kelley Thomas, Dr. Kevin Culligan, Dr. Michael Lynch, and Dr. Greg Lang for serving on my committee and providing instrumental feedback with this dissertation. Lastly, I would like to thank Dr. Way Sung, who helped design the bioinformatics approach described herein and made numerous contributions to my dissertation.

I am also grateful for the efforts of several undergraduate students who helped perform the research presented in this dissertation. Brian Van Dam provided crucial technical support for the duration of the initial eight-month mutation accumulation experiment. Both Chelsea Jones and Nicholas Rouillard subsequently helped with the design and completion of the fitness assays that were conducted on all of my mutants. Several other Cooper Lab members, particularly Kenneth Flynn and Sean Buskirk, also provided valuable feedback on the research presented herein.

Funding for this research was provided by the Multidisciplinary University Research Initiative Award from the US Army Research Office (to ML) and the National Science Foundation Career Award (to VC). In addition, I have been supported by the University of New Hampshire on two Summer TA Fellowships (2012, 2013) and the Dissertation Year Fellowship (2015-2016). I am extremely grateful to all of these funding sources, which were instrumental for the completion of my dissertation research.

## TABLE OF CONTENTS

# LIST OF TABLES

CHAPTER I:

CHAPTER II:

CHAPTER III:

CHAPTER IV:

APPENDICES:

# LIST OF FIGURES

CHAPTER I:

CHAPTER II:

CHAPTER III:

CHAPTER IV:

APPENDICES:

# ABSTRACT

## THE RATE, SPECTRUM AND EFFECTS OF SPONTANEOUS MUTATION IN BACTERIA WITH MULTIPLE CHROMOSOMES

by

Marcus Michael Dillon

University of New Hampshire, May, 2016

Despite their essentiality for evolutionary change and role in many diseases, spontaneous mutations remain understudied because of both biological and technical barriers. Prokaryotic mutation biases are especially understudied and no studies have been conducted on bacteria with multiple chromosomes, leaving major gaps in our understanding of the role of genome content and structure on mutation. The application of mutation accumulation lines to whole-genome sequencing offers the opportunity to study spontaneous mutations in a wide range of prokaryotic organisms. Here, we present a genome-wide view of molecular mutation rates and spectra in *Burkholderia cenocepacia*, *Vibrio fischeri*, and *Vibrio cholerae*, three bacterial species that harbor multiple chromosomes but differ dramatically in %GC-content. We demonstrate both general and species specific biases in spontaneous mutation rates and spectra, while also highlighting how some mutational biases vary within within genomes. We then study the distribution of effects of spontaneous mutations in *B. cenocepacia*, illustrating that most mutations have little or no effect on fitness and those that do are mostly deleterious across multiple environments. Overall, this body of work offers unprecedented insight into the rate, spectrum, and fitness effects of spontaneous mutations in three prokaryotic organisms whose genomes harbor multiple circular chromosomes, a common but underappreciated bacterial genome architecture.

# CHAPTER I

# RATE AND MOLECULAR SPECTRUM OF SPONTANEOUS MUTATIONS IN THE GC-RICH MULTI-CHROMOSOME GENOME OF *BURKHOLDERIA CENOCEPACIA*

## INTRODUCTION

As the ultimate source of genetic variation, mutation is implicit in every aspect of genetics and evolution. However, as a result of the genetic burden imposed by deleterious mutations, remarkably low mutation rates have evolved across all of life, making detection of these rare events technologically challenging and accurate measures of mutation rates and spectra exceedingly difficult (Kibota and Lynch 1996; Lynch and Walsh 1998; Sniegowski *et al.* 2000; Lynch 2011; Fijalkowska *et al.* 2012; Zhu *et al.* 2014). Until recently, most estimates of mutational properties have been derived indirectly using comparative genomics at putatively neutral sites (Graur and Li 2000; Wielgoss *et al.* 2011) or by extrapolation from small reporter-construct studies (Drake 1991). Both of these methods are subject to potentially significant biases, as many putatively neutral sites are subject to selection and mutation rates can vary substantially among different genomic regions (Lynch 2007).

To avoid the potential biases of these earlier methods, pairing classic mutation accumulation (MA) with whole-genome sequencing (WGS) has become the preferred method for obtaining direct measures of mutation rates and spectra (Lynch *et al.* 2008; Denver *et al.* 2009; Ossowski *et al.* 2010; Lee *et al.* 2012; Sung *et al.* 2012a; b, 2015; Heilbron *et al.* 2014; Foster *et al.* 2015; Long *et al.* 2015). Using this strategy, a single clonal ancestor is used to initiate several replicate lineages that are subsequently passaged through repeated single-cell bottlenecks for several thousand generations. The complete genomes of each evolved lineage are then sequenced and compared with the other lines to identify *de novo* mutations occurring over the course of the experiment. The bottlenecking regime minimizes the ability of natural selection to eliminate deleterious mutations, and the parallel sequencing provides a large enough body of information to yield a nearly unbiased picture of the natural mutation spectrum of the study organism (Lynch *et al.* 2008).

The MA-WGS method has now been used to examine mutational processes in several model eukaryotic and prokaryotic species, yielding a number of apparently generalizable conclusions about mutation rates and spectra. For example, a negative scaling between base-substitution mutation rates and both effective population size ($N_E$) and the amount of coding DNA supports the hypothesis that the refinement of replication fidelity that can be achieved by selection is determined by the power of random genetic drift among phylogenetic lineages (Lynch 2011; Sung *et al.* 2012a). This “drift-barrier hypothesis” therefore predicts that organisms with very large population sizes such as some bacteria should have evolved very low mutation rates (Lee *et al.* 2012; Sung *et al.* 2012a; Foster *et al.* 2013). Strong transition and G:C>A:T biases have also been observed in nearly all non-mutator MA studies to date (Lind and Andersson 2008; Lynch *et al.* 2008; Denver *et al.* 2009; Ossowski *et al.* 2010; Lee *et al.* 2012; Sung *et al.* 2012a; b), corroborating previous findings using indirect methods (Hershberg and Petrov 2010; Hildebrand *et al.* 2010). However, several additional characteristics of mutation spectra vary among species (Lynch *et al.* 2008; Denver *et al.* 2009; Ossowski *et al.* 2010; Lee *et al.* 2012; Sung *et al.* 2012a; b), and examining the role of genome architecture, size, and lifestyle in producing these idiosyncrasies will require a considerably larger number of detailed MA-WGS studies. Among bacterial species that have been subjected to mutational studies, genomes with high %GC-content are particularly sparse and no studies have been conducted on bacteria with multiple chromosomes, a genome architecture of many important bacterial species (e.g *Vibrio*, *Brucella*, *Burkholderia*).

*Burkholderia cenocepacia* is a member of the *Burkholderia cepacia* complex, a diverse group of bacteria with important clinical implications for patients with cystic fibrosis (CF), where they can form persistent lung infections and highly resistant biofilms (Coenye *et al.* 2004; Mahenthiralingam *et al.* 2005; Traverse *et al.* 2013). The core genome of *B. cenocepacia* HI2424 has a high %GC-content (66.80%) and harbors three chromosomes, each containing rDNA operons (LiPuma *et al.* 2002), although the third chromosome can be eliminated under certain conditions (Agnoli *et al.* 2012). The primary chromosome (chr1) is ≈ 3.48 Mb and contains 3253 genes; the secondary chromosome (chr2) is ≈ 3.00 Mb and contains 2709 genes; and the tertiary chromosome (chr3) is ≈ 1.06 Mb and contains 929 genes. In addition, *B. cenocepacia* HI2424 contains a 0.16 Mb plasmid, which contains 157 genes and lower %GC-content than the core genome (62.00%). Although the %GC-content is consistent across the three core chromosomes, the proportion of coding DNA declines from chr1 to chr3, while the synonymous and non-synonymous substitution rates increase from Chr1 to chr3 (Cooper *et al.* 2010; Morrow and Cooper 2012). Whether this variation in evolutionary rate is driven by variation in non-adaptive processes like mutation bias or variation in the relative strength of purifying selection remains a largely unanswered question in the evolution of bacteria with multiple chromosomes.

Here, I applied whole-genome sequencing to 47 MA lineages derived from *B. cenocepacia* HI2424 that were evolved in the near absence of natural selection for over 5550 generations each. I identified a total of 291 mutations spanning all three replicons and the plasmid, enabling a unique perspective of inter-chromosomal variation in both mutation rate and spectra, in a bacterium with the highest %GC-content studied with MA-WGS to date.

## MATERIALS AND METHODS

**Mutation accumulation.** Seventy-five independent lineages were founded by single cells derived from a single colony of *Burkholderia cenocepacia* HI2424, a soil isolate that had only previously been passaged in the laboratory during isolation (Coenye and LiPuma 2003). Independent lineages were then serially propagated every 24 hours onto fresh high nutrient Tryptic Soy Agar (TSA) plates (30 g/L Tryptic Soy Broth (TSB) Powder, 15 g/L Agar). Two lineages were maintained on each plate at 37°, and the isolated colony closest to the base of each plate half was chosen for daily re-streaking. Following 217-days of MA, frozen stocks of all lineages were prepared by growing a final colony per isolate in 5 ml TSB (30 g/L TSB) overnight at 37°, and freezing in 8% DMSO at -80°.

Daily generation times were estimated each month by placing a single representative colony from each line in 2 ml of Phosphate Buffer Saline (80 g/L NaCl, 2 g/L KCl, 14.4 g/L $Na_2HPO_4 \bullet 2H_2O$, 2.4 g/L $KH_2PO_4$), serially diluting to $10^{-3}$ and spread plating 100 ul on TSA. By counting the colonies on the resultant TSA plate, I calculated the number of viable cells in a single colony and the number of generations between each transfer. The average generation time across all lines was then calculated and used as the daily generation time for that month. These generation-time measurements were used to evaluate potential effects of declining colony size over the course of the

MA experiment as a result of mutational load, a phenotype that was observed (Figure A.1). Final generation numbers per line were estimated as the sum of monthly generation estimates, which were derived by multiplying the number of generations per day in that month by the number of days between measurements (Figure A.1).

**DNA extraction and sequencing.** Genomic DNA was extracted from 1 ml of overnight culture inoculated from 47 frozen derivatives of MA lines and the ancestor of the MA experiments using the Wizard Genomic DNA Purification Kit (Promega Inc.). Following library preparation, sequencing was performed using the 151-bp paired-end Illumina HiSeq platform at the University of New Hampshire Hubbard Center for Genomic Studies with an average fragment size between paired-end reads of ≈ 386 bps. All forward and reverse reads for each isolate and the ancestor were individually mapped to the reference genome of *Burkholderia cenocepacia* HI2424 (LiPuma *et al.* 2002), with both the Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009) and Novoalign (www.novocraft.com), producing an average sequence depth of 43x.

**Base-substitution mutation identification.** To identify spontaneous base-substitution mutations (bpsms), I used SAMtools to convert the SAM alignment files to mpileup format (Li *et al.* 2009), and in-house perl scripts to produce the forward and reverse read alignments for each position in each line. A three-step process was then used to detect putative bpsms. First, pooled reads across all lines were used to generate an ancestral consensus base at each site in the reference genome. This allows me to correct for any differences that may exist between the reference genomes and the ancestral colony of each my MA experiments. Second, a lineage specific consensus base was generated at each site in the reference genome for each individual MA lineage using only the reads from that line. Here, a lineage specific consensus base was only called if the site was covered by at least two forward and two reverse reads and at least 80% of the reads identified the same base. Otherwise, the site was not analyzed. Third, each lineage specific consensus base that was called was compared to the overall ancestral consensus of the MA experiment and a putative bpsm was identified if they differed. This analysis was carried out independently with the alignments generated by BWA and Novoalign, and putative bpsms were considered genuine only if both pipelines independently identified the bpsm and they were only identified in a single lineage.

Using relatively lenient criteria for identifying lineage specific consensus bases, I was able to analyze the majority of the genome in all lineages, but increase my risk of falsely identifying bpsms at low coverage sites. Therefore, I generated a supplementary dataset for all genuine bpsms identified in this study, which includes the read coverage and consensus at each site where a bpsm was identified (Table A.1). I do not see clusters of bpsms at the lower limits of my coverage or consensus requirements. In fact, the vast majority of bpsms in my MA experiment were covered by more than 25 reads and were supported by more than 95% of the reads that covered the site. Furthermore, I verified that none of the bpsms that I identified were present in the ancestral *B. cenocepacia* HI2424 strain that I sequenced, so I am confident that nearly all of the bpsms identified in this study were genuine spontaneous bpsms that arose during the mutation accumulation experiments.

**Insertion-deletion mutation identification.** For insertion-deletion mutations (indels), inherent difficulties with gaps and repeat elements can reduce agreement in the alignment of single reads using short-read alignment algorithms, even in the case of true indels. Simple-sequence repeats (SSRs) are an especially difficult challenge, as reads that are not anchored on both sides of the SSR will just align to the next repeated sequence of bases and fail to identify SSR variants, even if they are genuine. Therefore, I employed more lenient criteria to extract all putative indels from the raw alignments, requiring that the indel was covered by at least two forward and two reverse reads, and that 30% of those reads identified the exact same indel (size and motif). Among these putative indels, all indels that were independently identified by both BWA and Novoalign, where 80% of the reads identified the exact same indel were considered genuine indels. For indels where only 30-80% of the reads identified the exact same indel, I parsed out only reads that had bases on both the upstream and downstream regions of the SSR (if the indel was in an SSR), and on both the upstream and downstream regions of the indel itself (if the indel was not in an SSR). Using only this subset of reads, I reassessed the number of reads that identified the exact same indel (size and motif), and considered these initially low confidence putative indels genuine if more than 80% of these sub-reads identified the exact same indel. In addition, I passaged the alignment output through the pattern-growth algorithm PINDEL to identify any large genuine indels using paired-end information (Ye *et al.* 2009). Here, I required a total of 20 reads, with at least 6 forward and 6 reverse reads, and 80% of the reads to identify the exact same indel for the indel to be considered genuine. This summative collection of indels was then compared to the analysis of the ancestral *B. cenocepacia* HI2424 strain, and any indel that was identified in the ancestor or more than 50% of the other MA lineages was excluded from subsequent analyses.

My initial filters for indels were even more lenient than those for bpsms, which may have lead to false positive indel identification in the putative indel phase. However, I subsequently required at least 80% consensus for all genuine indels identified in this study among reads that had bases at both the upstream and downstream regions of putative indels that were not in SSRs and among reads that had bases at both the upstream and downstream regions of the SSR of putative indels that were in SSRs. Further, I verified that all indels identified were not present in the ancestral *B. cenocepacia* HI2424 strain and were not identified in more than 50% of the other lineages analyzed in the same MA experiment. As with my bpsms, I generated a supplementary dataset containing all genuine indels analyzed in this study, which includes the read coverage and consensus at each site where an indel was identified, as well as the read coverage and consensus among reads with bases covering both the upstream and downstream regions of the indel or SSR if the initial consensus among reads covering the indel was below 80% (Table A.2). Thus, I am confident that nearly all indels identified in this study were genuine spontaneous indels that arose during the mutation accumulation process.

**Mutation-rate analysis.** Once a complete set of mutations had been identified in each lineage, I calculated the substitution and indel mutation rates for each line using the equation:

$$\mu = m/nT,$$

where $\mu$ represents the mutation rate ($\mu_{bs}$ for bpsms, $\mu_{indel}$ for indels), $m$ represents the number of mutations observed, $n$ represents the number of sites that had sufficient depth and consensus to analyze, and $T$ represents the total generations over the course of the MA study for an individual line. The standard error of the mutation rate for each line was measured as described previously (Denver *et al.* 2004, 2009), with the equation:

$$SE_x = \sqrt{\mu/nT}.$$

The final $\mu_{bs}$ and $\mu_{indel}$ for *B. cenocepacia* were calculated by taking the average mutation rates of all sequenced lineages, and the total standard error was calculated as the standard deviation of the mutation rates across all lines ($s$) divided by the square root of the number of lines analyzed (*N*):

$$SE_{pooled} = s/\sqrt{N}.$$

Specific base-substitution mutation rates were further divided into conditional rates for each substitution type, again using the equation:

$$\mu_{bs} = m/nT,$$

where $m$ is the number of substitutions of a particular type, and $n$ is the number of ancestral bases that can lead to each substitution with sufficient depth and consensus to analyze. The conditional substitution rates at seven MLST loci (*atpD*, *gltB*, *gyrB*, *lepA*, *phaC*, *recA*, and *trpB*) were calculated under the assumption that the most common nucleotide was the ancestral state and any deviation from that ancestral state occurred only once and spread through the population (Jolley and Maiden 2010). I then estimated conditional substitution rates as:

$$\mu_{bs} = m/n,$$

as described above.

**Calculation of $G_E$, $\pi_s$, and $N_E$.** Effective genome size ($G_E$) was determined as the total coding bases in the *B. cenocepacia* genome. Silent site diversity ($\pi_s$) was derived using the MLST loci described above, which were concatenated and aligned using BIGSdb (Jolley and Maiden 2010), and analyzed using DNAsp (Librado and Rozas 2009). Using the value of $\mu_{bs}$ obtained in this study, $N_E$ was estimated by dividing the value of $\pi_s$ by $2\mu_{bs}$ ($\pi_s = 2N_E\mu_{bs}$) (Kimura 1983).

**Statistical analyses.** All statistical analyses were performed in R Studio Version 0.99.489 using the Stats analysis package (R Development Core Team 2013).

## RESULTS

A classic mutation-accumulation experiment was carried out for 217 days with 75 independent lineages all derived from the same ancestral colony of *B. cenocepacia* HI2424 (LiPuma *et al.* 2002). This method founds a new population each day by a single cell, which limits the efficiency with which natural selection can purge deleterious and enrich beneficial mutations. Measurements of generations of growth per day were taken monthly and varied from 26.2 ± 0.12 to 24.9 ± 0.14 (mean ± 95% CI of highest and lowest measurements, respectively) (Figure A.1), resulting in an average of 5554 generations per line over the course of the MA experiment. Thus, across the 47 lines whose complete genomes were sequenced, I was able to visualize the natural mutation spectrum of *B. cenocepacia* HI2424 over 261,047 generations of MA.

From the comparative sequencing data, I identified 245 bpsms, 42 indels, and four plasmid-loss events spanning the entire genome. With means of 5.21 bpsms and 0.89 indels per line, the distribution of bpsms and indels across individual lines did not differ significantly from a Poisson distribution (bpsms: $\chi^2 = 1.81$, p = 0.99; indels: $\chi^2 = 0.48$, p = 0.92), indicating that mutation rates did not vary over the course of the MA experiment (Figure A.2).

Mutation-accumulation experiments rely on the basic principle that when $N_E$ is sufficiently reduced, the efficiency of selection is minimized to the point at which nearly all mutations become fixed by genetic drift with equal probability (Kibota and Lynch 1996). $N_E$ in this mutation accumulation study was calculated to be 12.86, using the harmonic mean of the population size over 24 hours of colony growth (Hall *et al.* 2008). The threshold selective coefficient below which genetic drift will overpower natural selection is:

$$N_E \bullet s = 1 \text{ (Lynch 2007).}$$

Thus, only mutations conferring adaptive or deleterious effects of s>0.078 would be subject to the biases of natural selection in this study, which is expected to be a very small fraction of mutations (Kimura 1983; Elena *et al.* 1998; Zeyl and de Visser 2001; Hall *et al.* 2008).

Given the codon usage and %GC-content of synonymous and non-synonymous sites in *B. cenocepacia* HI2424, 27.80% of coding substitutions are expected to be synonymous in the absence of natural selection. The observed percentage of synonymous substitutions (25.51%) did not differ significantly from this null-expectation ($\chi^2$ = 0.54, df = 1, p = 0.46). Further, I found limited evidence of positive selection since parallel evolution among base-substitution mutations was rare in this study; no gene was hit more than twice across any of the 47 independently derived lineages, excluding indel hotspots in SSRs (Table A.1; Table A.2). Although both bpsms ($\chi^2$ = 4.20, df = 1, p = 0.04) and indels ($\chi^2$ = 21.3, df = 1, p < 0.0001) were biased to non-coding DNA, evidence exists that mismatch repair preferentially repairs damage in coding regions, which can create artificial signatures of selection in MA experiments (Lee *et al.* 2012). Thus, my overall observations are consistent with this MA experiment inducing limited selection on the mutation spectra; at least as far as bpsms are concerned.

**Low base-substitution and indel mutation rates.** The preceding results imply that bpsm and indel rates for *B. cenocepacia* are 1.33 (0.08) • $10^{-10}$ /bp/generation and 2.19 (0.30) • $10^{-11}$ /bp/generation (SEM), respectively. Based on the 7.70 Mb genome size, these per-bp mutation rates correspond to a genome-wide bpsm rate of only 0.001/genome/generation, and a genome-wide indel rate of only 0.0002/genome/generation. Although the ratio of synonymous to non-synonymous substitutions is consistent with negligible influence of selection on base-substitution mutations in this study, it is impossible to know for certain whether the lack of non-coding indels was generated by purifying selection or non-adaptive mutation biases, but their scarcity could reflect some selective loss of genotypes with loss-of-function mutations (Foster *et al.* 2013; Heilbron *et al.* 2014; Zhu *et al.* 2014; Dettman *et al.* 2016).

**Base-substitution mutations are not AT-biased.** One of the central motivations for studying the molecular mutation spectrum of *B. cenocepacia* was its high %GC-content (66.80%). The vast majority of wild-type MA-WGS studies to date have demonstrated a mutation bias in the direction of AT (Table 1), and a similar bias has also been inferred in comparative analyses of several bacterial species, including *Burkholderia pseudomallei* (Lynch *et al.* 2008; Denver *et al.* 2009, 2012; Keightley *et al.* 2009; Hershberg and Petrov 2010; Hildebrand *et al.* 2010; Lynch 2010a; Ossowski *et al.* 2010; Sung *et al.* 2012a; b; Lee *et al.* 2012; Schrider *et al.* 2013; Zhu *et al.* 2014). Thus*,* biased gene conversion and selection have been invoked to explain the high %GC-content realized in many genomes (Lynch *et al.* 2008; Duret and Galtier 2009; Raghavan *et al.* 2012; Zhu *et al.* 2014; Lassalle *et al.* 2015). These data for *B. cenocepacia* are inconsistent with prior published studies showing a mutation bias in the direction of AT (Table 1), but also suggest that biased gene conversion and/or selection must have mostly generated the realized %GC-content of *B. cenocepacia*, which is substantially higher than expected based on mutation pressure alone.

**Table 1. *Burkholderia cenocepacia* AT-mutation bias comparison against seven other organisms.** The strength of the AT-mutation bias is calculated as the ratio of the conditional mutation rates in the in the G:C>A:T direction to the conditional mutation rates in the A:T>G:C direction, which is substantially higher in all other species than it is in *B. cenocepacia*.

| ORGANISM (%GC)[a] | TRANSITIONS | | TRANSVERSIONS | | | | **AT - BIAS** |
|---|---|---|---|---|---|---|---|
| | A:T>G:C | G:C>A:T | A:T>T:A | G:C>T:A | A:T>C:G | G:C>C:G | |
| *B. cenocepacia (0.67)* | 6.88 | 7.51 | 2.67 | 2.44 | 5.35 | 2.38 | **0.81** |
| *E. coli (0.51)*[b] | 8.74 | 13.71 | 2.80 | 5.08 | 6.64 | 2.88 | **1.22** |
| *M. florum (0.27)*[b] | 50.93 | 640.83 | 15.67 | 360.14 | 11.75 | 185.36 | **15.97** |
| *H. sapiens (0.45)* | 429.00 | 961.00 | 129.00 | 258.00 | 152.00 | 295.00 | **2.10** |
| *D. melanogaster (0.42)*[b] | 101.13 | 513.19 | 98.06 | 130.76 | 48.01 | 74.52 | **4.32** |
| *S. cerevisiae (0.38)*[b] | 7.13 | 17.86 | 3.03 | 9.69 | 5.30 | 7.82 | **2.22** |
| *A. thaliana (0.36)*[b] | 104.54 | 896.30 | 43.56 | 139.08 | 60.98 | 123.63 | **6.26** |
| *C. elegans (0.35)*[b] | 16.47 | 57.23 | 17.50 | 44.09 | 7.72 | 16.89 | **4.19** |

[a] Data was obtained from the following studies: *E. coli - Lee et al., 2012; M. florum - Sung et al., 2012a; H. sapiens - Lynch, 2010; D. melanogaster - Schrider et al., 2013; S. cerevisiae - Zhu et al., 2014; A. thaliana* - *Ossowski et al., 2010; C. elegans - Denver et al., 2012*.

[b] Conditional mutation rates (×$10^{11}$) are calculated as the number of each mutation type, divided by the product of the number of generations and the total A:T or G:C sites in each respective reference genome if the raw data is not directly available in the cited reference (Wei *et al.* 2014).

In comparing the relative rates of G:C>A:T transition and G:C>T:A transversion mutations with those of A:T>G:C transitions and A:T>C:G transversions, corrected for the ratio of G:C to A:T sites analyzed in this study, I found that substitutions in the G:C direction were 17% more frequent than substitutions in the A:T direction per base pair, although the conditional rates were not significantly different ($\chi^2$ = 0.91, df = 1, p = 0.33). The lack of mutational bias in the A:T direction can largely be attributed to A:T>C:G transversions occurring at significantly higher rates than any other transversion type, most notably the G:C>T:A transversions ($\chi^2$ = 8.68, df = 1, p = 0.0032). However, A:T>G:C transitions also occurred at nearly the same rate as G:C>A:T transitions, the latter of which have been the most commonly observed substitution in other studies, putatively due to deamination of cytosine or 5-methyl-cytosine (Figure 1) (Lee *et al.* 2012; Sung *et al.* 2012b; Zhu *et al.* 2014).

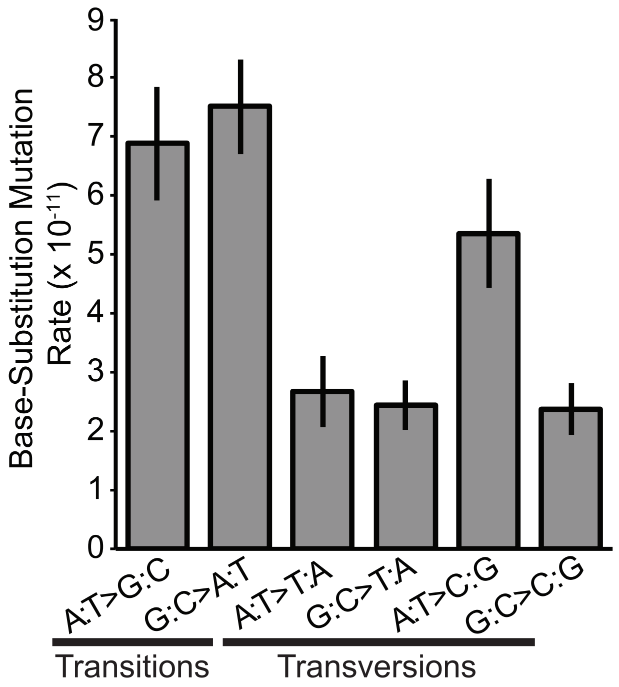

**Figure 1. Conditional base-substitution mutation (bpsm) rates of *Burkholderia cenocepacia* mutation accumulation (MA) lines across all three chromosomes.** Conditional base-substitution mutation rates per conditional base-pair per generation, estimated by dividing the number of observed mutations by the product of the analyzed sites capable of producing a given mutation and the number of generations of mutation accumulation. Error bars indicate one standard error of the mean.

Using the ratio of the conditional rate of mutation in the G:C direction to that in the A:T direction (x), the expected %GC-content under mutation-drift equilibrium is x/(1+x) = 0.539 (0.043) (SEM). Therefore, although mutation pressure in *B. cenocepacia* does not favor AT-bases, it is clear that the observed mutation bias is not sufficient to elicit the realized %GC-content of 66.80%. Thus, either the *B. cenocepacia* genome is still moving towards mutation-drift equilibrium, or GC-biased gene conversion and/or natural selection are contributing substantively to the observed %GC-content (Lynch *et al.* 2008; Duret and Galtier 2009; Raghavan *et al.* 2012; Zhu *et al.* 2014; Lassalle *et al.* 2015).

**Deletion bias favors genome-size reduction and AT composition.** Although my lower bound estimates of the insertion and deletion mutation rates are both ≈ 10-fold lower than the base-substitution mutation rate, many indels affect more than one base. Specifically, the 21 deletions observed in this study result in the deletion of a total of 414 bases, while the 21 insertions result in a gain of 164 bases. Therefore, the number of bases that are impacted by indels in this study is more than twice the number impacted by bpsms, indicating that indels may still play a central role in the genome evolution of *B. cenocepacia* if they are not purged by natural selection.

Although I observed the exact same number of deletions and insertions in this study, the per base-pair deletion rate (1.97 (0.86) • $10^{-10}$/bp/generation (SEM)) was substantially higher than the insertion rate (6.11 (1.90) • $10^{-11}$/bp/generation (SEM)), since the average size of deletions was greater than the average size of insertions. Thus, there is a net deletion rate of 1.34 • $10^{-10}$/bp/generation (Table 2). Although no indels >150 bps were observed in this study, examining the depth of coverage of the *B. cenocepacia* HI2424 plasmid relative to the rest of the genome revealed that the plasmid was lost at a rate of 1.53 × $10^{-5}$ per cell division, while gains in plasmid copy number were not observed (Table 2).

**Table 2. Parameters of insertion-deletion mutations (indels) in the *Burkholderia cenocepacia* mutation accumulation experiment.**

| Parameter | Deletions | Insertions |
|---|---|---|
| Events Observed | 21 | 21 |
| Total Nucleotides Affected | 414 | 164 |
| Total GC Bases Affected | 311 | 110 |
| Total AT Bases Affected | 103 | 54 |
| Proportion of GC Bases Affected | 75.12 | 67.07 |
| Plasmid Copy Number Loss/Gain | 4 | 0 |

The base composition of deletions was also biased, with GC bases being deleted significantly more than expected based on the genome content ($\chi^2$ = 12.92, df = 1, p = 0.0003). In contrast, no detectable bias was observed towards insertions of GC over AT bases ($\chi^2$ = 0.55 • $10^{-2}$, df = 1, p = 0.9408) (Table 2). Thus, indels in *B. cenocepacia* are expected to reduce genome wide %GC-content, further supporting the need for other population-genetic processes to account for the composition of high-GC genomes (Lynch *et al.* 2008; Duret and Galtier 2009; Raghavan *et al.* 2012; Zhu *et al.* 2014; Lassalle *et al.* 2015). Overall, the observed mutation spectra in this study suggest that the natural indel spectrum of *B. cenocepacia* causes both genome-size reduction and increased %AT-content.

**Non-uniform chromosomal distribution of mutations.** Another major goal of this study was to investigate whether mutation rates and spectra vary among chromosomes and chromosomal regions. The three core chromosomes of *B. cenocepacia* vary in size and content but are sufficiently large to have each accumulated a considerable number of mutations in this study (Morrow and Cooper 2012). Chromosome 1 (chr1) is the largest chromosome (both in size and in gene count), with more essential and highly expressed genes than either chromosome 2 (chr2) or 3 (chr3) (Figure A.3). Expression and number of essential genes are second highest on chr2 and lowest on chr3 (Cooper *et al.* 2010; Morrow and Cooper 2012). In contrast, average non-synonymous and synonymous variation among orthologs shared by multiple strains of *B. cenocepacia,* as well as fixed variation among *Burkholderia* species (dN and dS), are highest on chr3 and lowest on chr1 (Figure A.3) (Cooper *et al.* 2010; Morrow and Cooper 2012).

The overall bpsm rates of the three core chromosomes differ significantly based on a chi-square proportions test, where the null expectation was that the number of substitutions would be proportional to the number of sites covered on each chromosome ($\chi^2 = 6.77$, df = 2, p = 0.0340) (Figure 2A; Figure 3A). Specifically, bpsm rates are highest on chr1, and lowest on chr2, which is the opposite of observed evolutionary rates on these chromosomes (Figure A.3) (Cooper *et al.* 2010). In addition, a second chi-squared test was performed to test whether the observed bpsm rates differed from the conditional bpsm rates expected on each chromosome given their respective nucleotide contents, which are similar (%GC: Chr1-66.8%; Chr2-66.9%; Chr3-67.3%). Here, the null expectation for the total number of bpsms on each chromosome was calculated as the product of the number of GC bases covered, the total number of generations across lines, and the overall GC bpsm rate across the genome, added to the product of the same calculation for AT bpsms. The differences in the bpsm rates of the three core chromosomes remained significant when this test was performed ($\chi^2 = 6.88$, df = 2, p = 0.0320), indicating that the intra-chromosomal heterogeneity in bpsm rates cannot be explained by variation in nucleotide content.

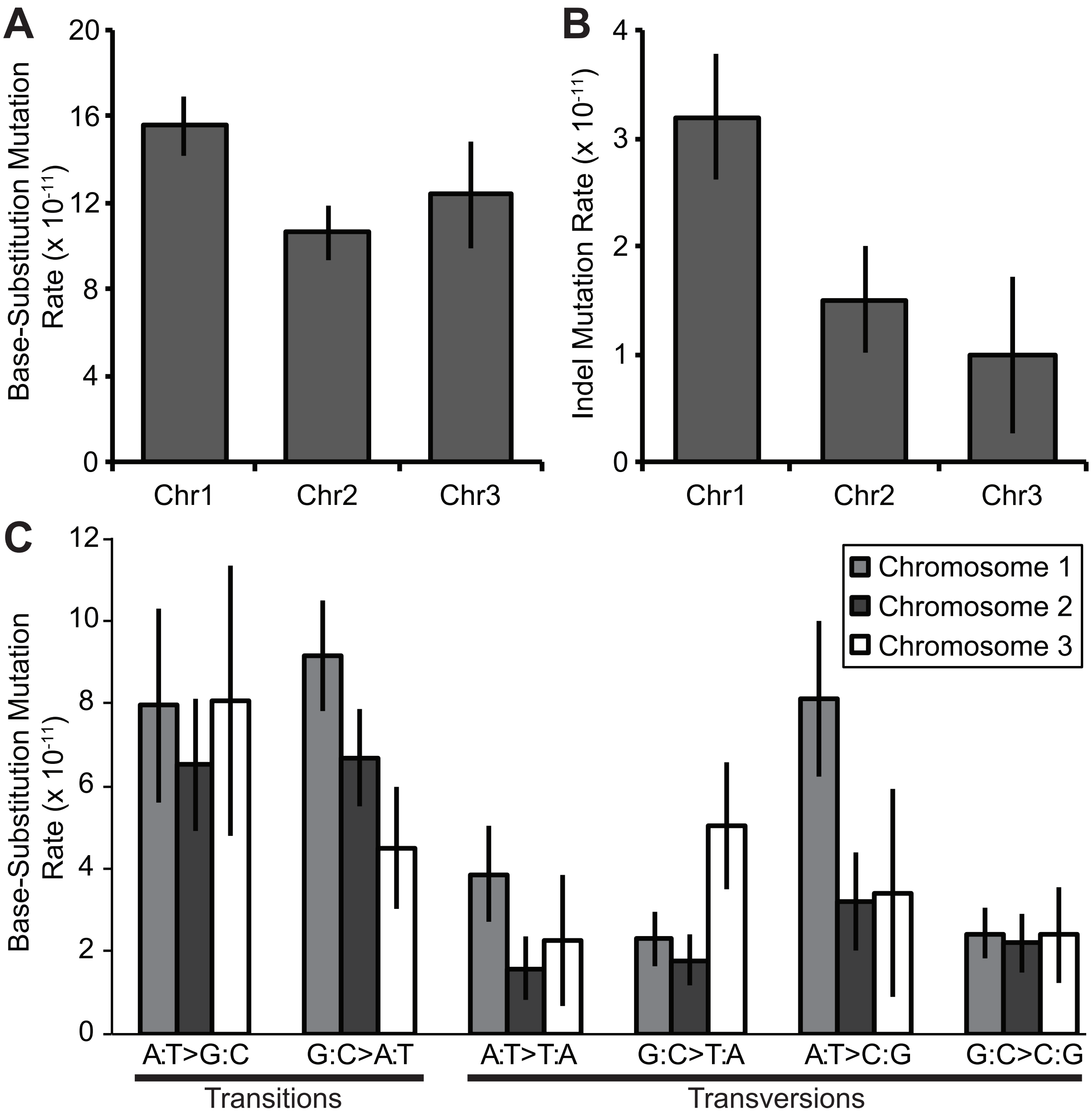

**Figure 2**. **Base-substitution (bpsm) and insertion-deletion (indel) mutation rates for the three chromosomes of *Burkholderia cenocepacia*.** (A, B) Overall bpsm and indel rates. C) Conditional bpsm rates for each chromosome of *B. cenocepacia* estimated by dividing the number of observed mutations on each chromosome by the product of the analyzed sites capable of producing a given mutation on each chromosome and the number of generations of mutation accumulation. Error bars indicate one standard error of the mean.

The conditional bpsm spectra were also significantly different in all pairwise chi-squared proportions tests between chromosomes (chr1/chr2: $\chi^2$=14.32, df=5, p=0.0141;

chr1/chr3: $\chi^2$=17.02, df=5, p=0.0043; chr2/chr3: $\chi^2$=13.44, df=5, p=0.0201) (Figure 2C). These comparisons further illustrate that the significant variation in conditional bpsm rates is mostly driven by a few bpsm types that occur at higher rates on particular chromosomes. Specifically, although their individual differences were not quite statistically significant, G:C>T:A transversions seem to occur at the highest rate on chr3 ($\chi^2$ = 5.94, df = 2, p = 0.0511) and A:T>C:G transversions occur at the highest rate on chr1 ($\chi^2$ = 5.67, df = 2, p = 0.0590) (Figure 2C).

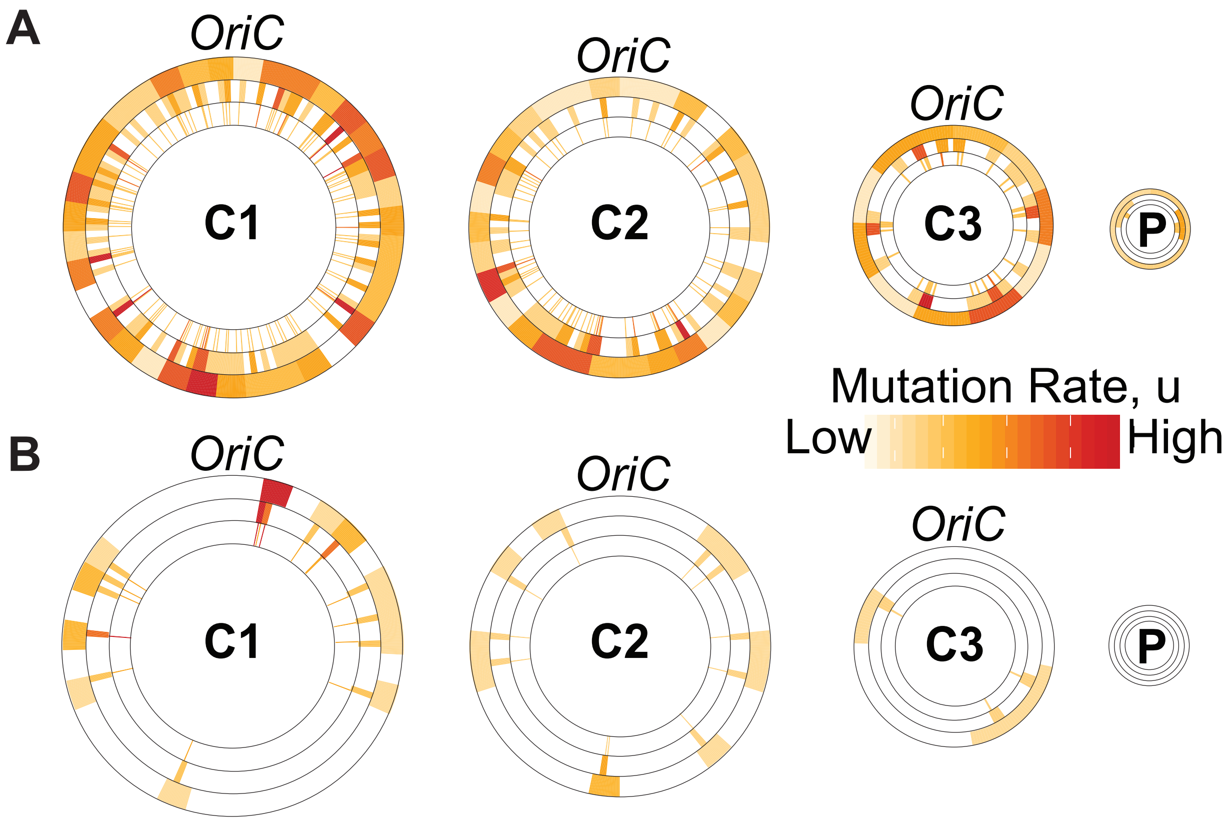

**Figure 3. Intra-chromosomal variation in the base-substitution (bpsm) and insertion-deletion (indel) mutation rates from the *Burkholderia cenocepacia* mutation accumulation (MA) experiment.** Overall bpsm (A) and indel (B) mutation rates are separated into 100 Kb (outer), 25 Kb (middle), and 5 Kb (inner) intervals extending clockwise from the origin of replication (*oriC*) to reveal broad and local properties of variation in mutation rates. Mutation rates were analyzed independently for each interval length, so color shades in shorter intervals do not directly compare to the same color shades in longer intervals. The 0.164 Mb plasmid is not to scale.

Studies in *Vibrio cholerae* have suggested that in bacteria with multiple chromosomes, smaller secondary chromosomes delay their replication until there remains approximately the same number of bases to be replicated on larger chromosomes (Rasmussen *et al.* 2007; Cooper *et al.* 2010). This ensures synchrony of replication termination between chromosomes of different sizes, despite the fact that their replication proceeds at the same rate. To test whether this replication timing gradient is partially responsible for the patterns I observe in base-substitution mutation spectra between chromosomes, I binned chr1 and chr2 into late and early replicating regions, where the early replicating regions represent bases presumed to replicate prior to chr3 initiation, and the late replicated regions represent bases presumed to replicate following chr3 initiation (the last 1.06 Mb replicated).

In support of this model, G:C>T:A transversions also occur at a slightly higher rate in late replicated regions of chr1 and chr2 than they do in early replicated regions of chr1 and chr2 (Figure 4A). However, when mutations are binned by overall replication timing (combining late replicating regions on chr1 and chr2 with chr3 and comparing them to early replicating regions on chr1 and chr2), the rate of G:C>T:A transversions is not significantly higher than it is in early replication-timing regions, likely due to small sample sizes ($\chi^2$ = 2.52, df = 1, p = 0.1127). A:T>C:G transversions occur at slightly higher rates in early replicated regions of chr1 and chr2 than they do in late replicated regions (Figure 4B), but again the difference is not statistically significant ($\chi^2$ = 1.26, df = 1, p = 0.2621). Together, these findings suggest that late replicating DNA is predisposed to incur more G:C>T:A transversions and early replicating DNA is predisposed to incur more A:T>C:G transversions, but a larger collection of mutations will be necessary to fully address this question.

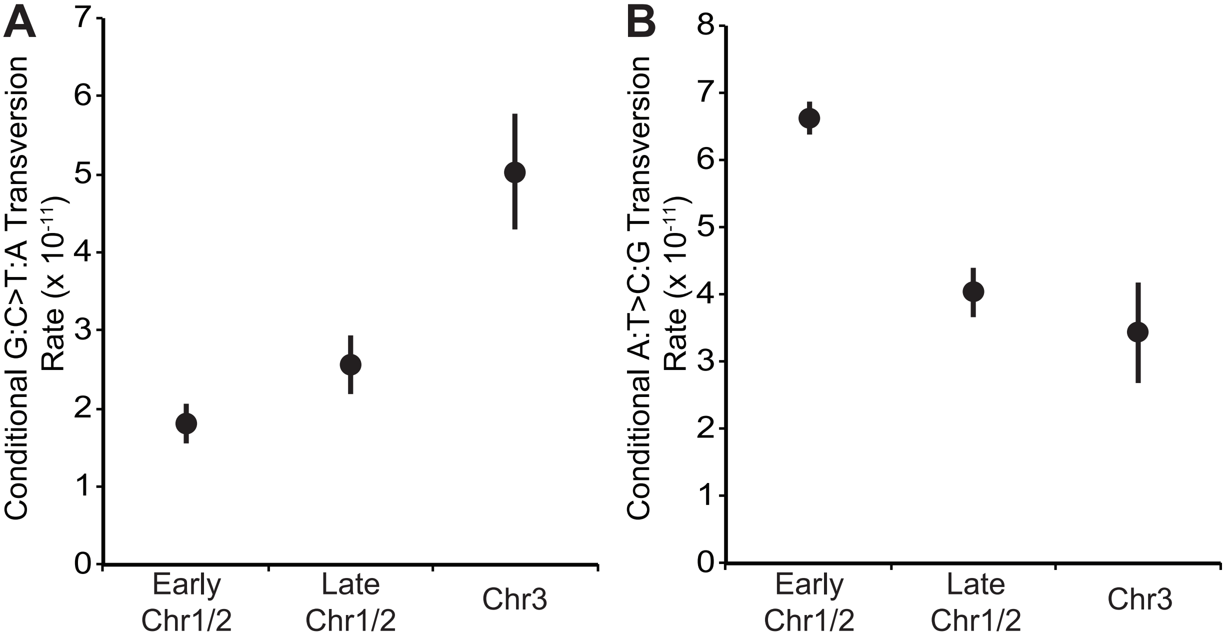

**Figure 4. Conditional G:C>T:A and A:T>C:G transversion rates in the genome of *Burkholderia cenocepacia*, separated by replication timing regions.** Conditional G:C>T:A (A) and A:T>C:G (B) transversion rates, normalized for base-composition as described in Figures 1 and 2, were calculated for regions on the primary and secondary chromosomes that are replicated prior to initiation of replication of the third chromosome (Early Chr1/2), regions on the primary and secondary chromosomes that are replicated simultaneously with the third chromosome (Late Chr1/2), and the third chromosome itself (Chr3), based on models from (Rasmussen *et al.* 2007). Error bars indicate one standard error of the mean.

Like bpsm rates, indel rates also varied significantly among chromosomes, as indel rates were highest on chr1, intermediate on chr2, and lowest on chr3 ($\chi^2$ = 8.2652, df = 2, p = 0.0160), (Figure 2B; Figure 3B). No indels were observed on the 0.16 Mb plasmid, but as noted above, four plasmid loss events were observed. The latter events involve the loss of 157 genes, and are expected to have phenotypic consequences. The relative rarity of indels observed in this study limits my ability to analyze their intra-chromosomal biases in great detail, but the repeated occurrence of indels in short 5kb regions, and particularly within microsatellites (57.6% of all indels) suggests that replication slippage is a common cause of indels in the *B. cenocepacia* genome (Figure 3B).

## DISCUSSION

Despite their relevance to both evolutionary theory and human health, the extent to which generalizations about mutation rates and spectra are conserved across organisms remains unclear. Because of their diverse genome content, bacterial genomes are particularly amenable to studying these issues (Lynch 2007). In measuring the rate and molecular spectrum of mutations in the high-GC, multi-replicon genome of *B. cenocepacia*, I have corroborated some prior findings of MA studies in model organisms, but also demonstrated idiosyncrasies in the *B. cenocepacia* spectrum that may extend to other organisms with high %GC-content and/or multiple chromosomes. Specifically, *B. cenocepacia* has a relatively low mutation rate and is consistent with a universal deletion bias in prokaryotes (Mira *et al.* 2001). However, the lack of AT-mutation bias is inconsistent with all previous findings in mismatch-repair proficient organisms (Lynch *et al.* 2008; Denver *et al.* 2009; Hershberg and Petrov 2010; Hildebrand *et al.* 2010; Ossowski *et al.* 2010; Lee *et al.* 2012; Sung *et al.* 2012b, 2015), and both mutation rates and spectra differed significantly among chromosomes, in a manner suggesting greater oxidative damage or more inefficient repair in late replicated regions.

As a member of a species complex with broad ecological and clinical significance, *B. cenocepacia* is a taxon with rich genomic resources that enable comparisons between the *de novo* mutations reported here and extant sequence diversity. With 7050 genes, *B. cenocepacia* HI2424 has a large amount of coding DNA ($G_E$) (6.8 • $10^6$ base pairs), and a high average nucleotide heterozygosity at silent-sites ($\pi_s$) (6.57 • $10^{-2}$) relative to other strains (Watterson 1975; Mahenthiralingam *et al.* 2005). By combining this $\pi_s$ measurement and the bpsm rate from this study, I estimate that the $N_E$ of *B. cenocepacia* is approximately 2.47 • $10^8$, which is in the upper echelon among species whose $N_E$ has been derived in this manner (Figure A.4). Under the drift-barrier hypothesis, high target size for functional DNA and high $N_E$ increase the ability of natural selection to reduce mutation rates (Lynch 2010b, 2011; Sung *et al.* 2012a). Thus, given the large proteome and $N_E$ of *B. cenocepacia*, it is unsurprising that *B. cenocepacia* has relatively low base-substitution and indel mutation rates when compared to other organisms (Sung *et al.* 2012a). However, the low base-substitution and indel mutation rates observed in this study need not imply limited genetic diversity among species of the *Burkholderia cepacia* complex. Rather, because of their high $N_E$ and evidently frequent lateral genetic transfer, species of the *Burkholderia cepacia* complex are remarkably diverse (Baldwin *et al.* 2005; Pearson *et al.* 2009), demonstrating that low mutation rates need not imply low levels of genetic diversity.

*Burkholderia* genomes also tend to be large in comparison to other Proteobacteria, but this is evidently not the product of more frequent insertions. Rather, insertions and deletions occurred at identical rates but deletions were larger than insertions, and plasmids were lost relatively frequently, which together add to the general model that bacterial genomes are subject to a deletion bias (Mira *et al.* 2001; Kuo and Ochman 2009). Ultimately, this dynamic has the potential to drive the irreversible loss of previously essential genes during prolonged colonization of a host and may enable host dependence to form more rapidly in prokaryotic organisms than in eukaryotes, which do not have a strong deletion bias (Denver *et al.* 2004; Kuo and Ochman 2009; Dyall *et al.* 2014). Consistent with this dynamic, host-restricted *Burkholderia* genomes evolving at lower $N_E$ are indeed substantially smaller than free-living genomes (Mahenthiralingam *et al.* 2005).

A lack of mutational bias towards AT bases was also recently observed in the %GC-rich bacteria *Deinococcus radiodurans* (Long *et al.* 2015), but had not been previously observed in non-mutator MA lineages of any kind (Lind and Andersson 2008; Lynch *et al.* 2008; Denver *et al.* 2009; Keightley *et al.* 2009; Ossowski *et al.* 2010; Lee *et al.* 2012; Sung *et al.* 2012a; b). Yet even though bpsms are not AT-biased in *B. cenocepacia*, selection and/or biased gene conversion must still be invoked to explain their high %GC-content (Hershberg and Petrov 2010; Hildebrand *et al.* 2010). Of these two explanations, selection favoring %GC-content may be the more influential force, given that there is no evidence for increased %GC-content in recombinant genes of *Burkholderia*, despite its prevalence in other bacteria (Lassalle *et al.* 2015). It is also notable that similar substitution biases can be observed at polymorphic sites of several MLST loci shared across *B. cenocepacia* isolates (Jolley and Maiden 2010). Specifically, A:T>C:G transversions are more common than G:C>T:A transversions, and the rates of G:C>A:T and A:T>G:C transitions are nearly indistinguishable at six of the seven loci (Figure A.5). However, the evolutionary mechanism of these substitution biases are uncertain given the potential for ongoing recombination and/or natural selection to influence polymorphisms at these sites in conserved housekeeping genes (Lynch *et al.* 2008; Duret and Galtier 2009; Raghavan *et al.* 2012; Zhu *et al.* 2014).

In principle, a decreased rate of G:C>A:T transition mutation relative to other bacteria could be achieved by an increased abundance of uracil-DNA-glycosylases (UDGs), which remove uracils from DNA following cytosine deamination (Pearl 2000), or by a lack of cytosine methyltransferases, which methylate the C-5 carbon of cytosines and expose them to increased rates of cytosine deamination (Kahramanoglou *et al.* 2012). However, *B. cenocepacia* HI2424 does not appear to have an exceptionally high number of UDGs, and it does contain a cytosine methyltransferase homolog, suggesting that active methylation of cytosines does occur in *B. cenocepacia*. Extending these methods to more genomes with high %GC-content will be required to determine whether a lack of AT-mutation bias is a common feature of GC-rich genomes.

Perhaps the most important finding from this study is that both mutation rates and spectra vary significantly among the three autonomously replicating chromosomes that make up the *B. cenocepacia* genome (Figure 2). My data demonstrate that base-substitution mutation rates vary significantly among chromosomes, but not in the direction predicted by comparative studies (Mira and Ochman 2002; Cooper *et al.* 2010; Lang and Murray 2011; Agier and Fischer 2012; Morrow and Cooper 2012). Specifically, I find that base-substitution mutation rates are highest on the primary chromosome (Figure 2A,B), where evolutionary rates are lowest. Thus, purifying selection must be substantially stronger on the primary chromosome to offset the effect of an elevated mutation rate.

The spectra of base-substitutions also differed significantly among chromosomes. Specifically, A:T>C:G transversions are more than twice as likely to occur on chr1 as elsewhere, and G:C>T:A transversions are more than twice as likely to occur on the chr3 (Figure 2C). One possible explanation for the increased rate of G:C>T:A transversions on chr3 is that they can arise through oxidative damage (Michaels *et al.* 1992; Lee *et al.* 2012) and may be elevated late in the cell cycle when intracellular levels of reactive oxygen species are high (Mira and Ochman 2002; Stamatoyannopoulos *et al.* 2009; Chen *et al.* 2010). Thus, because tertiary chromosomes are expected to be replicated late in the cell cycle (Rasmussen *et al.* 2007), I would expect these elevated rates of G:C>T:A transversions on chr3. Of course, if this explanation were accurate, I would also observe and increased rate of G:C>T:A transversions in late-replicated regions of chr1 and chr2. Although the low number of total G:C>T:A transversions observed in this study prevents me from statistically distinguishing G:C>T:A transversion rates between late and early replicated regions of chr1 and chr2, the rate of G:C>T:A transversions is higher in late replicated regions of chr1 and chr2 (Figure 4A), a remarkable finding considering that early replicated genes on chr1 and chr2 are expressed more, which has been shown to induce G:C>T:A transversions independent of replication (Klapacz and Bhagwat 2002; Kim and Jinks-Robertson 2012; Alexander *et al.* 2013). Thus, I suggest that late replicating DNA is inherently predisposed to increased rates of G:C>T:A transversions, possibly due to increased exposure to oxidative damage (Michaels *et al.* 1992), variation in nucleotide-pool composition (Kunkel 1992; Zhang and Mathews 1995), or variation in DNA-repair mechanisms (Hawk *et al.* 2005; Courcelle 2009).

A mechanism of an increased A:T>C:G transversion mutation rate on the primary chromosome is less clear, but a decreased rate of A:T>C:G transversions in a late replicating reporter relative to that on an intermediate replicating reporter has been demonstrated previously in *Salmonella enterica* (Hudson *et al.* 2002). Thus, it is possible that the rate of this form of transversion is increased in early replicating DNA, or that it is primarily caused by other forms of mutagenesis (Klapacz and Bhagwat 2002). A:T>C:G transversion rates in early replicating regions of chr1 and chr2 support the former hypothesis, as early replicated regions of chr1 and chr2 experience the highest rates of A:T>C:G transversions (Figure 4B). The alternative mechanism of transcriptional mutagenesis seems less likely as A:T>C:G transversions occurred frequently in non-coding DNA relative to other substitution types (Figure A.6).

In summary, this study has demonstrated that the GC-rich genome of *B. cenocepacia* has a relatively low mutation rate, with a mutation spectrum that lacks an AT-bias and is biased toward deletion. Moreover, both the rate and types of base-substitution mutations that occur most frequently vary by chromosome, likely related to replication dynamics, the cell cycle, and transcription (Klapacz and Bhagwat 2002; Cooper *et al.* 2010; Merrikh *et al.* 2012). Although this study has broadened our understanding of mutation rates and spectra beyond that of model organisms, whether the observed mutational biases are common to all GC-rich genomes with multiple replicons, or are merely species-specific idiosyncrasies will require a more thorough investigation across a more diverse collection of GC-rich and multi-replicon bacterial genomes. Ultimately, by better understanding the core mutational processes that generate the variation on which evolution acts, we can aspire to develop true species-specific null hypotheses for molecular evolution, and by extension, enable more accurate analyses of the role of all evolutionary forces in driving genome evolution.

# CHAPTER II

## GENOME-WIDE BIASES IN THE RATE AND MOLECULAR SPECTRUM OF SPONTANEOUS MUTATIONS IN *VIBRIO CHOLERAE* AND *VIBRIO FISCHERI*

## INTRODUCTION

Spontaneous mutations generate the raw genetic variation on which evolution proceeds, but our knowledge of the biases associated with spontaneous mutations has been restricted because of technological challenges associated with accumulating and identifying mutations on a genome-wide scale. Mutation accumulation (MA) experiments paired with whole-genome sequencing (WGS) in microbes now offer an unprecedented opportunity to study genome-wide mutation rates and spectra in a diverse array of culturable microbes. Using the MA-WGS method, a single clonal ancestor is used to initiate many replicate lineages that are passaged through hundreds of single cell bottlenecks before each is subjected to WGS. The bottlenecking regime minimizes the efficiency of natural selection to operate on mutations and when the genome sequences are compared, we obtain a nearly unbiased picture of the natural mutation rates and spectra of the ancestor. A growing body of MA-WGS studies in microbes have began to reveal some general properties of spontaneous mutation, but unique properties of the spontaneous mutation rates and spectra in some taxa emphasize the importance of conducting detailed MA-WGS studies in a more comprehensive collection of species (Lynch *et al.* 2008; Denver *et al.* 2009; Ossowski *et al.* 2010; Sung *et al.* 2012a; b, 2015; Lee *et al.* 2012; Schrider *et al.* 2013; Heilbron *et al.* 2014; Zhu *et al.* 2014; Long *et al.* 2014, 2015; Dillon *et al.* 2015; Dettman *et al.* 2016).

Base-substitution mutations (bpsms) can be categorized into six different types, which include A:T>G:C and G:C>A:T transitions, and A:T>T:A, G:C>T:A, A:T>C:G, and G:C>C:G transversions. Differences in the relative rates of these bpsms may exert pressure on genome-wide %GC content and cause different sites to experience different bpsm rates. Evidence from indirect methods and MA experiments have supported nearly universal transition and AT biases in bpsm spectra across cellular life (Lind and Andersson 2008; Lynch *et al.* 2008; Denver *et al.* 2009; Hershberg and Petrov 2010; Hildebrand *et al.* 2010; Ossowski *et al.* 2010; Lee *et al.* 2012; Sung *et al.* 2012b; Dillon *et al.* 2015; Dettman *et al.* 2016), although there are exceptions in wild-type and mutator MA experiments (Dillon *et al.* 2015; Long *et al.* 2015; Dettman *et al.* 2016). Furthermore, bpsms have been shown to be context-dependent, where neighboring nucleotides affect site-specific bpsm rates (Long *et al.* 2014, 2015; Sung *et al.* 2015; Dettman *et al.* 2016). Insertion-deletion mutations (indels) can be categorized into insertions or deletions, but can also be grouped based on their size. The indel spectra of bacterial genomes are thought to be universally biased towards deletion (Mira *et al.* 2001; Kuo and Ochman 2009) but results have been mixed in wild-type and mutator MA experiments (Lee *et al.* 2012; Sung *et al.* 2012a; Long *et al.* 2014; Dillon *et al.* 2015; Dettman *et al.* 2016), possibly because it is the size rather than number of deletions that is greater (Dillon *et al.* 2015). Evidence is also mounting that the majority of indels involve the loss or gain of a single nucleotide and occur predominantly in simple-sequence repeats (SSRs), where the number of repeats scales positively with the indel rate (Lee *et al.* 2012; Long *et al.* 2014; Dettman *et al.* 2016). These SSRs have gained attention not only because they vary sufficiently among strains to enable rapid genotyping (van Belkum *et al.* 1998; Danin-Poleg *et al.* 2007; Ghosh *et al.* 2008) but also because in some species they associate with variable heritable expression of genes related to host colonization and disease, begging the question of whether these mutation-prone sequences have indirectly evolved to enable this plasticity (Moxon *et al.* 1994, 2006; Field *et al.* 1999).

The mismatch repair system (MMR) helps detect and repair DNA replication errors by excising and resynthesizing the DNA, reducing both bpsm and indel rates (Kunkel and Erie 2005; Reyes *et al.* 2015). Consequently, strains defective in MMR will experience substantially higher bpsm and indel rates, allowing for an expanded collection of replication errors in a shorter time span. While these bpsm and indel rates and spectra are unlikely to be representative of the wild-type bpsm and indel rates and spectra, they can reveal biases that are generated by MMR and elucidate subtle polymerase error biases with the enhanced statistical power (Lee *et al.* 2012; Long *et al.* 2014). Specifically, MA studies with MMR-deficient strains have shown that the majority of replication errors corrected by MMR are transitions and single-nucleotide indels because of the predominance of these types of mutations in MMR-deficient MA experiments (Lee *et al.* 2012; Long *et al.* 2014, 2015; Foster *et al.* 2015; Dettman *et al.* 2016). Furthermore, bpsms have been shown to vary in a mirrored wave-like pattern on the two opposing replichores of bacteria with a single circular chromosome (Foster *et al.* 2013; Long *et al.* 2014; Dettman *et al.* 2016), an analysis that is only enabled when a large number of bpsms that are distributed across the genome are available. Context-dependent analyses, which have shown that bpsm rates are impacted by upstream and downstream base-pairs (Long *et al.* 2014; Sung *et al.* 2015; Dettman *et al.* 2016), have also benefited from MMR-deficient MA experiments because robust quantification of bpsm rates in 64 triplets requires a substantial number of bpsms.

*Vibrio fischeri* and *Vibrio cholerae* are globally significant bacterial species because of their roles in marine symbiosis and pathogenesis, respectively (Goldberg and Murphy 1983; Thompson *et al.* 2004; Ruby *et al.* 2005). The core genome of *V. fischeri* has a low %GC content (38.35%) and harbors two chromosomes. Chromosome 1 (chr1) is ≈ 2.90 Mb, containing 2584 protein coding genes with a %GC content of 38.96%. Chromosome 2 (chr2) is smaller (≈ 1.33 Mb), has fewer protein coding genes (1174) and a slightly lower %GC content (37.02%). Furthermore, the *V. fischeri* ES114 strain used in this study has a ≈ 45.85 Kb plasmid, which contains 57 genes and has a %GC content of 38.44% (Ruby *et al.* 2005). The core genome of *V. cholerae* is also divided into two chromosomes and has a relatively neutral %GC content (47.58%). Chr1 is ≈ 2.99 Mb, containing 2605 protein coding genes and a %GC content of 47.85%. As was the case in *V. fischeri*, chr2 is smaller (≈ 1.10 Mb), has fewer protein coding genes (1001), and a slightly lower %GC content (46.83%). Importantly, replication of chr2 in bacteria with multiple circular chromosomes is initiated by different molecules and is delayed until late in the cell cycle so that chr1 and chr2 terminate replication synchronously, despite replicating at the same rate (Egan and Waldor 2003; Duigou *et al.* 2006; Rasmussen *et al.* 2007). Despite their broad relevance and peculiar genome architectures, no detailed MA-WGS studies have been performed on any *Vibrio* species.

Here, I performed a detailed MA-WGS experiment using 48 lineages derived from *V. fischeri* ES114 and 48 lineages derived from *V. cholerae* 2740-80. Each *V. fischeri* wild-type lineage was evolved in the near absence of natural selection for 5187 generations, while each *V. cholerae* wild-type lineage was evolved for 6453 generations. I then engineered MMR-deficient strains of *V. cholerae* and *V. fischeri*, creating *V. fischeri* ES114 *ΔmutS* and *V. cholerae* 27040-80 *ΔmutS*, and performed a second pair of MA-WGS experiments using 19 lineages derived from *V. fischeri* ES114 *ΔmutS* and 22 lineages derived from *V. cholerae* 27040-80 *ΔmutS*. Because of their elevated mutation rates, these MA experiments were only carried out for 810 and 1254 generations per lineage, for *V. fischeri* ES114 *ΔmutS* and *V. cholerae* 27040-80 *ΔmutS* respectively. I identified a total of 439 wild-type mutations and 5990 *ΔmutS* mutations distributed across both chromosomes of *V. fischeri* and *V. cholerae*, enabling a unique perspective into the bpsm and indel biases in wild-type and MMR-deficient strains of these two significant bacterial species.

## MATERIALS AND METHODS

**Bacterial strains and culture conditions.** The two wild-type MA experiments were founded from a single clone derived from *V. fischeri* ES114 and *V. cholerae* 2740-80, respectively. All MA experiments with *V. fischeri* were carried out on tryptic soy agar plates supplemented with NaCl (TSAN) (30 g/liter tryptic soy broth powder, 20 g/liter NaCl, 15 g/liter agar) and were incubated at 28°. Frozen stocks of each MA lineage were prepared at the end of the experiment by growing a single colony overnight in 5ml of tryptic soy broth supplemented with NaCl (TSBN) (30 g/liter tryptic soy broth powder, 20 g/liter NaCl) at 28° and freezing in 8% DMSO at -80°. For *V. cholerae*, all MA experiments were carried out on tryptic soy agar plates (TSA) (30 g/liter tryptic soy broth powder, 15 g/liter agar) and were incubated at 37°. Similarly, frozen stocks were prepared by growing a single colony from each lineage overnight in 5ml of tryptic soy broth (TSB) (30 g/liter tryptic soy broth powder) at 37° and were stored in 8% DMSO at -80°.

Mutator strains of *V. fischeri* ES114 and *V. cholerae* 2740-80 were generated by replacing the *mutS* gene in each genome with an erythromycin resistance cassette, as described previously (Datsenko and Wanner 2000; Heckman and Pease 2007; Val *et al.* 2012). Briefly, I used splicing by overlap extension (PCR-SOE) to generate two erythromycin resistance cassettes, one of which was flanked by ≈ 750 bps of the upstream and downstream regions of the *mutS* gene in *V. fischeri* ES114, while the second was flanked by ≈ 750 bps of the upstream and downstream regions of the *mutS* gene in *V. cholerae* 2740-80 (Heckman and Pease 2007). Both the *V. fischeri* ES114 and *V. cholerae* 2740-80 Δ*mutS* fragments were then cloned into the R6K γ-*ori*-based suicide vector pSW7848, which contains a *ccdB* toxin gene that is arabinose-inducible and glucose-repressible ($P_{BAD}$) (Val *et al.* 2012). Both of these pSW7848 plasmids, henceforth referred to as pSW7848-*Vf*Δ*mutS* and pSW7848-*Vc*Δ*mutS*, were transformed into *Escherichia coli* pi3813 chemically competent cells and stored at -80° (Datsenko and Wanner 2000).

Conjugal transfer of the pSW7848-*Vf*Δ*mutS* and pSW7848-*Vc*Δ*mutS* plasmids was performed using a tri-parental mating with the *E. coli* pi3813 cells as the donors (Val *et al.* 2012), *E. coli* DH5α-pEVS104 as the helper (Stabb and Ruby 2002), and *V. fischeri* ES114 and *V. cholerae* 2740-80 as the respective recipients. For *V. fischeri* ES114, the chromosomally inserted pSW7848-*Vf*Δ*mutS* plasmid resulting from a single crossover at the Δ*mutS* gene was selected on LBS plates (Graf *et al.* 1994) containing

1% glucose and 1 ug/ml chloramphenicol at 28°. Selection for loss of the plasmid backbone from a second recombination step was then performed on LBS plates containing 0.2% arabinose at 28°, which induces the $P_{BAD}$ promoter of the *ccdB* gene and ensures that all cells that have not lost the integrated plasmid will die (Val *et al.* 2012). For *V. cholerae*, the chromosomally inserted pSW7848-*Vc*Δ*mutS* plasmid was selected on LB plates (Sambrook *et al.* 1989) containing 1% glucose and 5 ug/ml chloramphenicol at 30°. Selection for loss of the plasmid backbone was performed on LB plates with 0.2% arabinose at 30°. Replacement of the *mutS* gene in *V. fischeri* ES114 and *V. cholerae* 2740-80 were verified by conventional sequencing, and *V. fischeri* ES114 Δ*mutS* and *V. cholerae* 2740-80 Δ*mutS* were used to found the two mutator MA experiments, under identical conditions to those described above for the wild-type MA experiments.

**Ancestral reference genomes.** Prior to this study, the genome of *V. fischeri* ES114 was already in completed form and annotated, consisting of three contigs representing chr1, chr2, and the 45.85 Kb plasmid (Ruby *et al.* 2005). Further, the location of the *oriC* on both chromosomes was available in dOriC 5.0, a database for the predicted *oriC* regions in bacterial and archaeal genomes (Gao *et al.* 2013). Fortunately, the *oriC* region on both chromosomes had been placed at coordinate zero, allowing me to proceed with this *V. fischeri* ES114 reference genome for all subsequent *V. fischeri* analyses. In contrast, when I initiated these MA experiments, the *V. cholerae* 2740-80 genome was still in draft form, consisting of 257 scaffolds with unknown chromosome association. Therefore, to reveal inter-chromosomal variation and assess the effects of genome location on bpsm and indel rates, I used single molecule, real-time (SMRT) sequencing to generate a complete assembly separated into the two contigs of *V. cholerae* 2740-80.

The Pacific Biosciences RSII sequencer facilitates the completion of microbial genomes by producing reads of multiple kilobases that extend across repetitive regions and allow whole-genomes to be assembled at a relatively limited cost (Koren and Phillippy 2015). Genomic DNA (gDNA) was prepared using the Qiagen Genomic-Tip Kit (20/G) from overnight cultures of *V. cholerae* 2740-80 grown in LB at 37° using manufacturers instructions. Importantly, this kit uses gravity filtration to purify gDNA, which limits shearing and increases the average fragment size of the resulting gDNA sample. Long insert library preparation and SMRT sequencing was performed on this *V. cholerae* 2740-80 gDNA at the Icahn School of Medicine at Mount Sinai according to the manufacturer's instructions, as described previously (Beaulaurier *et al.* 2015). Briefly, libraries were size selected using Sage Science Blue Pippin 0.75% agarose cassettes to enrich for long-reads, and were assessed for quantity and insert size using an Agilent DNA 12,000 gel chip. Primers, polymerases, and magnetic beads were loaded to generate a completed SMRTbell library, which was run in a single SMRT cell of a Pacific Biosciences RSII sequencer at a concentration of 75 pM for 180 minutes.

As expected, the long insert SMRT sequencing library generated mostly long reads, with an average sub-read length of 8,401 bps and an N50 of 11,480 bps. I used the hierarchical genome-assembly process workflow (HGAP3) to generate a completed assembly of *V. cholerae* 2740-80 and polished the assembly using the Quiver algorithm (Chin *et al.* 2013). The resultant assembly consisted of two contigs representing chr1 and chr2, with an average coverage of 128x. I annotated this assembly using prokka (v1.11), specifying *Vibrio* as the genus (Seemann 2014). I then identified the location of the *oriC* on both contigs using Ori-finder, which applies analogous methods to those used by dOriC 5.0 to identify *oriC* regions in bacterial genomes (Gao and Zhang 2008; Gao *et al.* 2013). Of course, these *oriC* regions were not located at coordinate zero of the *V. cholerae* 2740-80 reference genome, so I re-formatted the reference genome to place each *oriC* region at the beginning of the chr1 and chr2 contigs, then stitched the contigs back together and re-polished the genome using Quiver. Prokka was then run a second time to update the location of all genes, and this re-formatted *V. cholerae* 2740-80 genome was used as the ancestral reference genome for all subsequent *V. cholerae* analyses.

**MA-WGS Process.** For the two wild-type MA experiments, seventy-five independent lineages were founded by single cells derived from a single colony of *V. fischeri* ES114 and *V. cholerae* 2740-80, respectively. Each of these lineages was then independently propagated every 24 hours onto fresh TSAN for *V. fischeri* and fresh TSA for *V. cholerae*, and this cycle was repeated for a total of 217 days. For the two mutator MA experiments, forty-eight independent lineages were founded and propagated as described above from a single colony each of *V. fischeri* ES114 Δ*mutS* and *V. cholerae* 2740-80 Δ*mutS*, respectively. However, because of their higher mutation rates, these lineages were only propagated for a total of 43 days. At the conclusion of the four MA experiments, each lineage was grown overnight in the appropriate liquid broth at the appropriate temperature (see above), and stored at -80° in 8% DMSO.

Daily generations were estimated monthly for the wild-type lineages and bi-monthly for the mutator lineages by calculating the number of viable cells in a representative colony from 10 lineages per MA experiment following 24 hours of growth. During each measurement, the representative colonies were placed in 2 ml of phosphate buffer saline (80 g/liter NaCl, 2 g/liter KCl, 14.4 g/liter $Na_2HPO_4 \bullet 2H_2O$, 2.4 g//liter $KH_2PO_4$), serially diluted, and spread plated on TSAN or TSA  for *V. fischeri* and *V. cholerae*, respectively. These plates were then incubated for 24 hours at 28° or 37°, and the daily generations per colony were calculated from the number of viable cells in each representative colony. The average daily generations were then calculated for each time-point using the average of the ten representative colonies, and the total generations elapsed between each measurement were calculated as the product of the average daily generations and the number of days before the next measurement. The total of number of generations elapsed during the duration of the MA experiment per lineage was then calculated as the sum of these totals over the course of each MA study (Figure B.1).

At the conclusion of each of the four MA experiments, gDNA was extracted using the Wizard Genomic DNA Purification Kit (Promega) from 1 ml of overnight culture (TSBN at 28° for *V. fischeri*; TSB at 37° for *V. cholerae*) inoculated from 50 representative stored lineages for *Vf*-wt and *Vc*-wt experiments, and all 48 stored lineages for the *Vf*-mut and *Vc*-mut experiments. For the wild-type MA experiments, gDNA from the ancestral *V. fischeri* ES114 and *V. cholerae* 2740-80 strains was also extracted. All libraries were prepared using a modified Illumina Nextera protocol designed for inexpensive library preparation of microbial genomes (Baym *et al.* 2015).

Sequencing of the *Vf*-wt and *Vc*-wt lineages and their respective ancestors was performed using the 101-bp paired-end Illumina HiSeq platform at the Beijing Genome Institute (BGI), while sequencing of the *Vf*-mut and *Vc*-mut lineages was performed using the 151-bp paired-end Illumina HISeq platform at the University of New Hampshire Hubbard Center for Genomic Studies.

The raw fastQ reads were analyzed using fastQC, and revealed that 48 *Vf*-wt lineages, 49 *Vc*-wt lineages, 19 *Vf*-mut lineages, and 22 *Vc*-mut lineages were sequenced at sufficient depth to accurately identify bpsm and indel mutations. The failure to successfully sequence a high proportion of *Vf*-mut and *Vc*-mut lineages was mostly generated by a poorly normalized library, leading to limited sequence data for several of the mutator lineages. For the successfully sequenced lineages, all reads were mapped to their respective reference genomes with both the Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009) and Novoalign (www.novocraft.com). The average depth of coverage across the successfully sequenced lineages of each MA experiment was 100x for *Vf*-wt, 96x for *Vc*-wt, 124x for *Vf*-mut, and 92x for *Vc*-mut.

**Base-substitution mutation identification.** For all four MA experiments, bpsms were identified as described in the methods of Chapter 1. Briefly, a three-step process was used to identify bpsms. First, I identified an ancestral consensus base at each site in the reference genome using pooled reads across all lines from each MA experiment. Second, I identified a lineage specific consensus base at each site in the reference genome for each individual lineage using only the reads from that MA line. Here, I required a minimum of two forward and two reverse reads, and 80% consensus among those reads. Third, lineage specific consensus bases for each lineage were compared to the overall ancestral consensus of the MA experiment to identify putative bpsms. Putative bpsms were considered genuine if they were independently identified by both the BWA and Novoalign alignments, and they were only identified in a single lineage. Any sites at which I did not identify an ancestral and lineage specific consensus base were not analyzed for mutations. As was the case in Chapter 1, I generated a supplementary dataset for all genuine bpsms identified in this study (Table B.1), demonstrating that the vast majority of bpsms in all four MA experiments from this study were covered by more than 50 reads and were supported by more than 95% of the reads that covered the site. Further, none of the bpsms identified in this study were present in any of the MA ancestral strains, which were also sequenced and analyzed. Thus, I am confident that all bpsms identified in this study represent true spontaneous bpsms that arose during the MA experiments.

**Insertion-deletion mutation identification.** All indels identified in this study were also identified as described in the methods of Chapter 1. Briefly, I started by extracting putative indels using lenient filters so that I would not rule out genuine indels in long SSRs. These putative indels were extracted from both the BWA and Novoalign alignments as long as they were covered by at least two forward and two reverse reads, and 30% of those reads identified the exact same indel (size and motif). All putative indels that were called by more than 80% of the reads that covered the site and were independently identified by BWA and Novoalign were considered genuine indels. For putative indels that were supported by 30-80% of the reads that covered the site in both the BWA and Novoalign alignments, I parsed out only reads that had bases on both the upstream and downstream regions of the SSR (if the indel was in an SSR), and on both the upstream and downstream region of the indel (if the indel was not in an SSR). If the indel was called by more than 80% of these sub-reads in both the BWA and Novoalign alignments, it was also considered genuine. Lastly, as described in Chapter 1, I employed PINDEL to all MA lineages to identify large genuine indels that went undetected with the short-read aligners (Ye *et al.* 2009), requiring a total of 20 reads (6 forward and 6 reverse) and 80% consensus (size and motif). A supplementary dataset for all genuine indels identified in the four MA experiments described in this study is provided in Table B.2, which highlights that nearly all indels were covered by more than 50 reads, with at least 80% consensus. Further, as with the bpsms, none of the indels that were identified in this study were present in the ancestral strains, so I am confident that nearly all of these indels represent genuine indels that arose during the MA experiments.

**Mutation-rate analyses.** Overall bpsm and indel rates were calculated for each lineage using the equation:

$$\mu = m/nT,$$

where $\mu$ represents the mutation rate, $m$ represents the number of mutations observed, $n$ represents the number of ancestral sites analyzed, and $T$ represents the total number of generations elapsed per lineage. Conditional bpsm rates for each lineage were calculated using the same equation, but with $m$ representing the number of bpsms of the focal bpsm type, and $n$ representing the number analyzed ancestral sites that could generate the focal bpsm type. All summative bpsm and indel rates presented for each MA experiment were calculated as the average mutation rate across all analyzed lineages, while summative standard errors were calculated as the standard deviation of the mutation rate across all lines ($s$), divided by the square root of the total number of lines in the corresponding MA experiment ($N$):

$$SE_{pooled} = s/\sqrt{N}.$$

For my interval analysis of bpsm and indel rates within chromosomes, I divided each chromosome into 100 kb intervals, starting at the origin of replication and extending bi-directionally to the replication terminus. Bpsm rates in each interval were measured by dividing the total number of bpsms or indels from this study by the product of the total number of analyzed sites in each interval across all lines and the number of generations per line, using the same formula described above for genome wide mutation rates:

$$\mu = m/nT.$$

Because none of the chromosomes were exactly divisible by 100 kb, the terminal intervals on each replichore were always less than 100 kb, but their mutation rates were calibrated to the number of of bases analyzed in those intervals.

**Statistical analyses.** All statistical analyses were performed in R Studio Version 0.99.489 using the Stats analysis package (R Development Core Team 2013).

## RESULTS

Four MA experiments were carried out in this study using daily single-cell bottlenecks that limit the efficiency of natural selection to purge deleterious and enrich beneficial mutations. For the two wild-type (wt) experiments, *V. fischeri* ES114 (*Vf*-wt) and *V. cholerae* 2740-80 (*Vc*-wt) colonies were used to found 75 MA lineages, each of which was propagated for 217 days. For the two mutator (mut) experiments, *V. fischeri* ES114 (*Vf*-mut) and *V. cholerae* 2740-80 (*Vc*-mut) strains lacking a *mutS* gene were used to found 48 MA lineages, each of which was propagated for 43 days. The parameters of each MA experiment and the mutations that were identified from all of the final isolates are summarized in Table 1. In all four experiments, generations of growth per day declined over the course of the MA experiment, particularly in the mutator lineages, as a result of the fitness cost of bearing the acquired mutations (Figure B.1).

**Table 1. Parameters and observed mutations in the four *Vibrio fischeri* and *Vibrio cholerae* mutation accumulation experiments.**

| MA Lines | Sequenced Lines | Gen. per line | Gen. total | No. of bpsm | No. of indels | Bpsm rate per nucleotide[a] | Bpsm rate per genome[b] | Indel rate per nucleotide[a] | Indel rate per genome[b] |
|---|---|---|---|---|---|---|---|---|---|
| Vf_wt | 48 | 5187 | 248976 | 219 | 60 | $2.07 \cdot 10^{-10}$ | $8.85 \cdot 10^{-4}$ | $5.68 \cdot 10^{-11}$ | $2.43 \cdot 10^{-4}$ |
| Vc_wt | 49 | 6453 | 316197 | 138 | 22 | $1.07 \cdot 10^{-10}$ | $4.38 \cdot 10^{-4}$ | $1.71 \cdot 10^{-11}$ | $6.98 \cdot 10^{-5}$ |
| Vf_mut | 19 | 810 | 15390 | 4313 | 382 | $6.57 \cdot 10^{-8}$ | $2.81 \cdot 10^{-1}$ | $5.82 \cdot 10^{-9}$ | $2.49 \cdot 10^{-2}$ |
| Vc_mut | 22 | 1254 | 27588 | 1022 | 273 | $9.09 \cdot 10^{-9}$ | $3.72 \cdot 10^{-2}$ | $2.43 \cdot 10^{-9}$ | $9.93 \cdot 10^{-3}$ |

[a] Bpsm and indel mutation rates/nucleotide/generation are calculated as the number of observed mutations, divided by the product of sites analyzed and number of generations per lineage. The above estimates represent the average rate across all sequenced lineages.

[b] Bpsm and indel mutation rates/genome/generation are calculated by multiplying the mutation rate/nucleotide/generation in each lineage by the genome size. The above estimates represent the average rate across all sequenced lineages.

The properties of my MA experiments allowed me to assume that few mutations were subject to the biases of natural selection. The threshold selective coefficient (*s*) below which genetic drift will overpower natural selection is determined by:

$$N_E \bullet s = 1 \text{ (Lynch 2007),}$$

where $N_E$ is the effective population size, estimated here as the harmonic mean of the population size (*N*) (Hall *et al.* 2008). By calculating $N_E$ for each MA experiment, I estimate that only mutations conferring an adaptive or deleterious effect ($s$) greater than 0.083, 0.067, 0.106, and 0.069 for *Vf*-wt, *Vc*-wt, *Vf*-mut, and *Vc*-mut, respectively, were subject to the biases of natural selection, which is expected to be a very small fraction of mutations (Kimura 1983; Hall *et al.* 2008). Furthermore, if I exclude indels that were identified at the same site or SSR, only four genes were hit more than once and no genes were hit more than twice across all wild-type lineages, suggesting that positive selection acting on common traits was minimal in these experiments.

Other metrics that have been used to test that the efficiency of purifying selection is minimized in MA experiments include the ratio of coding to non-coding mutations and the ratio of synonymous to nonsynonymous bpsms. However, both of these tests are problematic as preferential mismatch repair in coding regions (Lee *et al.* 2012), context-dependent mutation biases (Sung *et al.* 2015), and a non-uniform distribution of mutation rates and spectra across the genome (Foster *et al.* 2013; Dillon *et al.* 2015; Dettman *et al.* 2016) can generate artificial signatures of natural selection. These issues were evident in my MA experiments, where chi-square tests comparing my observed mutations with the expected ratios of coding to non-coding DNA and synonymous to nonsynonymous sites in each genome were at times inconsistent.

For each MA experiment, the expected ratio of coding to non-coding mutations was determined directly from each ancestral reference genome, and the expected ratio of synonymous to nonsynonymous bpsms was calculated from each ancestral reference genome, after accounting for codon usage and %GC content at synonymous and nonsynonymous sites. In the *Vf*-wt lines, I observed an excess of non-coding indels and bpsms (Bpsm: $\chi^2$ = 4.01, d.f. = 1, p = 0.0451, Indels: $\chi^2$ = 61.43, d.f. = 1, p < 0.0001), while the ratio of nonsynonymous to synonymous bpsms did not differ significantly from the null expectation ($\chi^2$ = 0.91, d.f. = 1, p = 0.3410). In the *Vc*-wt lines, non-coding bpsms were again in excess ($\chi^2$ = 8.74, d.f. = 1, p = 0.0028), while the ratio of coding to non-coding indels ($\chi^2$ = 1.48, d.f. = 1, p = 0.2240) and nonsynonymous to synonymous bpsms ($\chi^2$ = 1.47, d.f. = 1, p = 0.2262) did not differ from the null expectation. The excess of non-coding indels and bpsms could imply that selection played a small role in eradicating coding mutations or that mismatch repair is more active in coding regions, while all other lines of evidence support that minimal selection was operating in my wild-type MA experiments.

In the *Vf*-mut lines, I observed an excess of coding bpsms ($\chi^2$ = 39.08, d.f. = 1, p < 0.0001) and non-coding indels ($\chi^2$ = 206.82, d.f. = 1, p < 0.0001), but the ratio of nonsynonymous to synonymous bpsms did not differ from the null expectation ($\chi^2$ = 2.53, d.f. = 1, p = 0.1113). In the *Vc*-mut lines, I observed an excess of non-coding indels ($\chi^2$ = 123.71, d.f. = 1, p < 0.0001) and synonymous bpsms ($\chi^2$ = 5.60, d.f. = 1, p = 0.0182), while the ratio of coding to non-coding substitutions did not differ from the null expectation ($\chi^2$ = 2.354, d.f. = 1, p = 0.125). The excess of coding bpsms and non-coding indels observed in the *Vf*-mut MA experiment suggest contradictory forms of selection, since an excess of coding bpsms is a signature of positive selection, and an excess of non-coding indels is a signature of purifying selection. Both the excess of non-coding indels and synonymous bpsms observed in the *Vc*-mut MA experiment suggest the operation of purifying selection. Altogether, these observations suggest that selection may have played a small role in eradicating some mutations during the mutator MA experiments, however, we should not rule out the possibility that these signatures of selection were generated by non-adaptive mutational biases (Foster *et al.* 2013; Dillon *et al.* 2015; Sung *et al.* 2015; Dettman *et al.* 2016).

**Wild-type base-substitution mutation rates and spectra.** The wild-type bpsm rates observed in this study are among the lowest per generation rates that have been observed in any organism (Figure B.2). The bpsm rate for *V. fischeri* is 2.07 (0.207) • $10^{-10}$ /bp/generation (SEM), which is approximately twice the rate of bpsm in *V. cholerae*, 1.07 (0.094) • $10^{-10}$ /bp/generation (SEM). Based on genome sizes of 4,273,718 bps for *V. fischeri* ES114 and 4,088,961 bps for *V. cholerae* 2740-80, these per-base bpsm rates correspond to genome-wide bpsm rates of 0.0009 /genome/generation and 0.0004 /genome/generation, respectively (Table 1). When bpsms are separated by chromosome, I find that chr2 has a significantly higher bpsm rate than chr1 in *V. fischeri* where the null expectation was that the number of bpsms would be proportional to the number of sites analyzed on each chromosome ($\chi^2$ = 4.80, d.f. = 1, p = 0.0282), but I find no such difference in *V. cholerae* ($\chi^2$ = 0.56, d.f. = 1, p = 0.4562) (Figure 1A). The increased bpsm rate on chr2 of *V. fischeri* cannot be explained by the relative nucleotide contents (%GC: chr1, 39.0%; chr2, 37.0%). In fact, when I conduct a chi-square test that accounts for AT-biased mutation in *V. fischeri*, the difference between the expected and observed bpsms between the two chromosomes is increased ($\chi^2$ = 5.75, d.f. = 1, p = 0.0171). Here, I use the product of the number of GC bps analyzed, the genome wide GC mutation rates, and the number of generations across all lines, added to the product of the same variables for AT bps to generate the null expectation for the ratio of bpsms between chromosomes. Correcting for AT-biased mutation in *V. cholerae* (%GC: chr1, 47.9%; chr2, 46.8%) does not result a significant difference between the bpsm rates of chr1 and chr2 ($\chi^2 = 0.12$, d.f. = 1, $p = 0.7260$).

Base-substitution mutation spectra in both *V. fischeri* and *V. cholerae* were AT-biased, as the combined rates of G:C>A:T transitions and G:C>T:A transversions, corrected for the ratio of G:C to A:T sites analyzed, were significantly higher than the combined rates of A:T>G:C transitions and A:T>C:G transversions (*Vf*: $\chi^2 = 108.09$, d.f. = 1, $p < 0.0001$ ; *Vc*: $\chi^2 = 28.74$, d.f. = 1, $p < 0.0001$) (Figure 1C). Interestingly, the AT-bias is stronger in *V. fischeri*, which has the lower genome wide %GC-content. However, consistent with previous prokaryotic MA studies (Lind and Andersson 2008; Lee *et al.* 2012; Sung *et al.* 2012a, 2015; Dillon *et al.* 2015; Dettman *et al.* 2016), AT-mutation bias alone fails to explain realized genome-wide %GC-contents. For *V. fischeri*, the expected %GC-content under mutation-drift equilibrium is 0.20 ± 0.03 (SEM), 0.18 lower than the genome-wide %GC-content (0.38). This AT-bias is generated by both G:C>A:T transitions and G:C>T:A transversions, but it is the relative G:C>T:A transversion rate that is especially high in comparison to previous MA studies. With a rate of $9.19 \bullet 10^{-9}$ /bp/generation, G:C>T:A transversions occur at a higher rate than any other bpsm, generating a transition/transversion ratio ($T_S/T_V$) of 0.85. Similarly, the expected %GC content of *V. cholerae* under mutation-drift equilibrium is 0.28 ± 0.04 (SEM), 0.20 lower than the genome-wide %GC content (0.48). However, in *V. cholerae*, the AT-bias is generated mostly by G:C>A:T transitions rather than G:C>T:A transversions, resulting in a $T_S/T_V$ of 1.59.

To test whether the bpsm spectra on chr1 were significantly different from those on chr2, I used the conditional bpsm rates on chr1 to generate probabilities for each bpsm and tested these against the observed bpsm spectra on chr2, after correcting for %GC-content (Figure 1C). Neither the bpsm spectra of *V. fischeri* or *V. cholerae* varied significantly between chromosomes (*Vf*: $\chi^2$ = 7.80, d.f. = 5, p = 0.1681; *Vc*: $\chi^2$ = 6.52, d.f. = 5, p = 0.2594). However, the G:C>T:A transversion rate was significantly higher on chr2 of *V. fischeri* (Welch's two tailed t-test; t = 2.35, df = 71.95, p = 0.0221) and the A:T>G:C transition rate was significantly lower on chr2 of *V. cholerae* (Welch's two tailed t-test; t = -2.16, df = 95.75, p = 0.0340) (Figure 1C). Interestingly, late replicating regions of chr1 in *V. fischeri* (terminal 1,330,333 bp, equal to the size of chr2) also had elevated G:C>T:A transversion rates and late replicating regions of chr1 in *V. cholerae* (terminal 1,101,931 bp, equal to the size of chr2) had reduced A:T>G:C transition rates, relative to early replicating regions, though neither of these intra-chromosomal differences were significant (Welch's two tailed t-test; *Vf* G:C>T:A $T_V$: t = 1.79, df = 81.65, p = 0.0767; *Vc* A:T>G:C $T_S$: t = -0.72, df = 95.49, p = 0.4730). However, the rates of bpsm for other forms of bpsm in late replicating regions of chr1 do not appear to conform to the rates of the same bpsm of chr2 (Table B.3), suggesting that not all bpsms are impacted by replication timing.

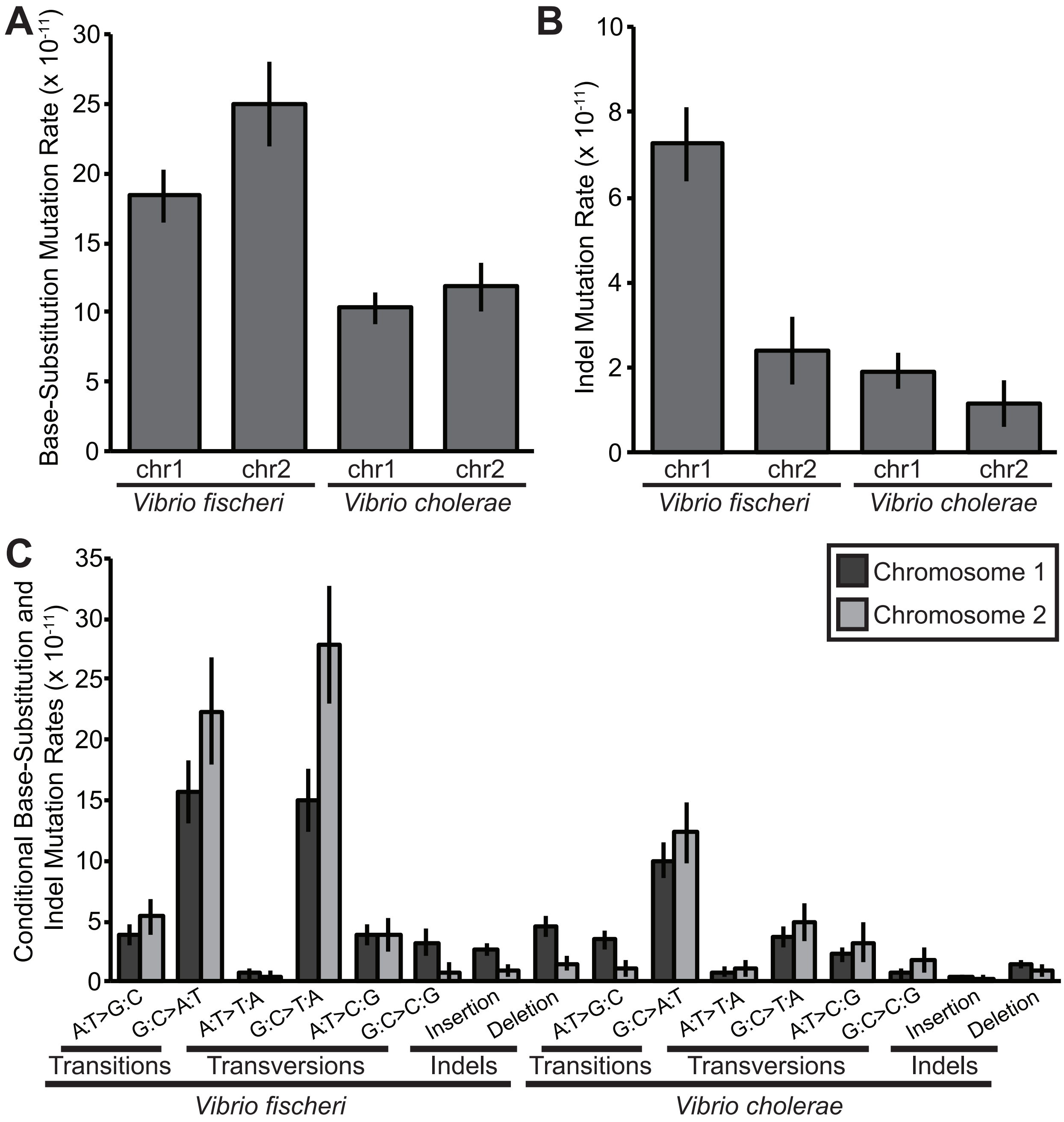

**Figure 1. Wild-type base-substitution (bpsm) and insertion-deletion (indel) mutation rates and spectra for the two chromosomes of wild-type *Vibrio fischeri* and *Vibrio cholerae*.** (A and B) Overall bpsm and indel mutation rates per base-pair per generation. (C) Conditional base-substitution and indel mutation rates per conditional base-pair per generation, estimated by dividing the number of observed mutations by the product of the analyzed sites capable of producing a given mutation and the number of generations of mutation accumulation. Error bars indicate one standard error of the mean.

**Wild-type insertion-deletion mutation rates.** Indels occurred at approximately one-fifth the rate of bpsm and occurred predominantly in simple sequence repeats. Cumulative indel mutation rates for *V. fischeri* and *V. cholerae* were 5.68 (0.691) • $10^{-11}$ /bp/generation and 1.71 (0.337) • $10^{-11}$ /bp/generation (SEM), respectively (Table 1). These rates correspond to genome-wide indel rates of 0.0002 /genome/generation for *V. fischeri* and 0.00007 /genome/generation for *V. cholerae* (Table 1). In both species, indel spectra were biased towards deletions, with deletions occurring at approximately twice the rate of insertions in *V. fischeri* ($\chi^2$ = 4.27, d.f. = 1, p = 0.0391), and three-times the rate of insertions in *V. cholerae* ($\chi^2$ = 6.55, d.f. = 1, p = 0.0112) (Figure 1). In *V. fischeri*, indels occurred more frequently than expected on chr1 based on the total sites analyzed on each chromosome ($\chi^2$ = 9.07, d.f. = 1, p = 0.0027). However, although indels were also slightly more common than expected on chr1 in *V. cholerae*, the difference in indel rates between chr1 and chr2 was not significant ($\chi^2$ = 0.12, d.f. = 1, p = 0.7260) (Figure 1B).

In contrast with previous reports, I find many multi-nucleotide indels in both the *Vf*-wt and *Vc*-wt mutation accumulation experiments. In *Vf*-wt, only 20.00% of indels involved the insertion or deletion of a single-nucleotide, while 36.36% of the *Vc*-wt indels involved a single-nucleotide. The analytic challenges associated with identifying indels greater than 10-bps in length and the biases generated by identifying a single particularly long indel make the average indel lengths in each dataset misleading. However, the distribution of the number of indels identified for each length below 10-bps demonstrates that short multi-nucleotide indels were relatively common, particularly in *V. fischeri* (Figure 2).

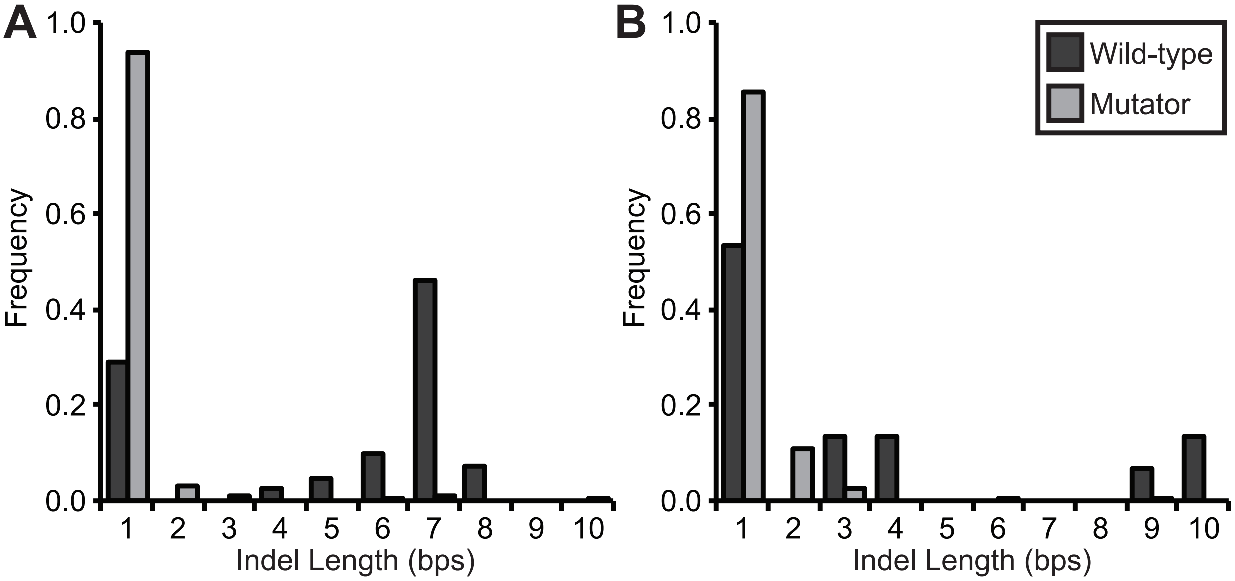

**Figure 2. Relative frequency of indels of different lengths observed in the wild-type (wt) and MMR repair deficient (mut) strains of *Vibrio fischeri* (A) and *Vibrio cholerae* (B).** The overall indel rates of *Vf*-mut and *Vc*-mut are 102-fold and 142-fold higher than the wild-type rates, respectively, but it is the relative frequencies of different indel lengths that are represented here.

Of the 60 indels that were observed in *V. fischeri*, 41 (68.33%) occurred in simple-sequence repeats (SSRs) with three or more repeats, which is significantly more than I expected based on the frequency of bases in SSRs of three or more in the *V. fischeri* ES114 genome ($\chi^2 = 82.92$, d.f. = 1, $p < 0.0001$). Interestingly, the indel rate in SSRs scaled positively with the repeat length, varying over orders of magnitude and differing significantly from the null expectation based on the frequency of each repeat type in the genome ($\chi^2 = 2.12 \bullet 10^5$, d.f. = 8, $p < 0.0001$) (Figure 3). The number of repeats in a SSR also scaled positively with the indel rate, differing significantly from the null expectation that the SSR indel rate would be proportional to the number of bases analyzed within each repeat length category (Chi-square test, repeat numbers 3-10; $\chi^2 = 5.59 \bullet 10^2$, d.f. = 7, $p < 0.0001$). Furthermore, a few SSRs were especially mutagenic, with the same SSR being mutated independently in multiple lineages (Table B.2). I cannot ascertain whether similar SSR biases exist in *V. cholerae* because only 22 indels were observed and only 8 of those were in SSRs with three or more repeats (36.36%). However, as was the case in *V. fischeri*, the occurrence of indels in SSRs is significantly higher than expected based on the frequency of SSRs with three or more copies in *V. cholerae* 2740-80 ($\chi^2$ = 4.55, d.f. = 1, p = 0.0329).

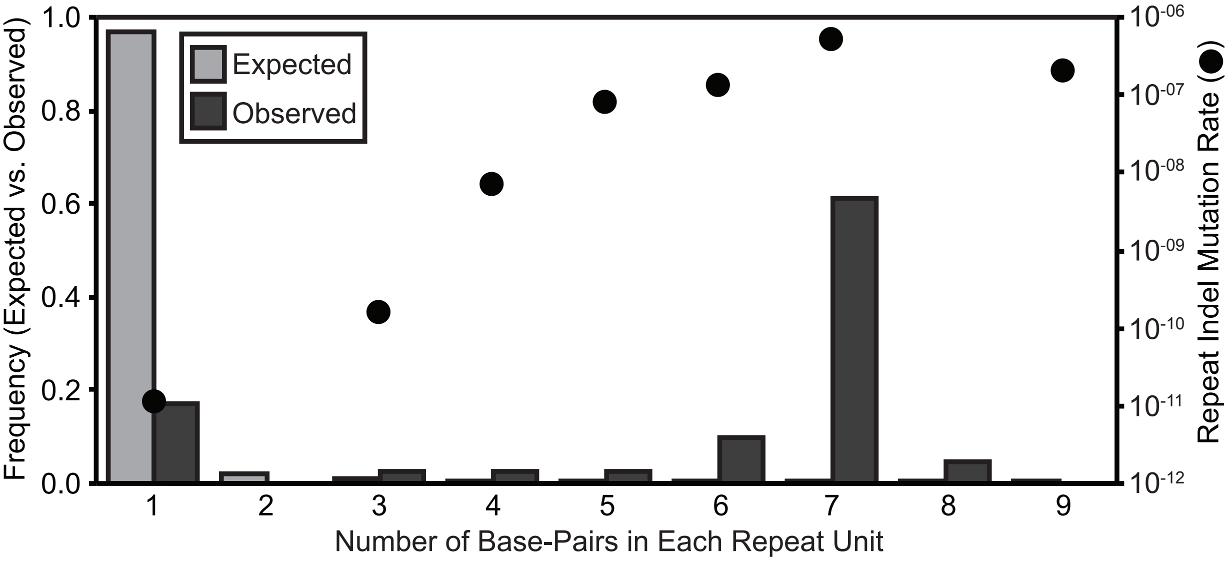

**Figure 3. Wild-type insertion-deletion mutation (indel) rates per run per generation and frequencies in simple-sequence repeats containing three or more repeats in *Vibrio fischeri*.** Indel rates per run per generation were calculated as the number of observed indels in that simple-sequence repeat type, divided by the product of the occurrence of that simple-sequence repeat type in the genome, the number of generations, and the number of MA lineages analyzed. Expected frequencies were calculated based on the relative occurrence of each simple-sequence repeat type in the *Vibrio fischeri* ES114 genome.

**Effects of losing DNA mismatch repair.** The deletion of the *mutS* gene results in a faulty MMR system and is expected to have dramatic consequences on the rates and spectra of bpsms and indels. In *V. fischeri*, the deletion of the *mutS* gene resulted in a

317-fold increase in the bpsm rate and a 102-fold increase in the indel rate (Table 1). Furthermore, the deletion of the *mutS* gene abolished the chromosomal biases in both bpsm and indel rates observed in the *Vf*-wt lineages (bpsm: $\chi^2$ = 0.11, d.f. = 1, p = 0.7413; indels: $\chi^2$ = 2.08, d.f. = 1, p = 0.1503). In *V. cholerae*, the deletion of the *mutS* gene resulted in an 85-fold increase in the bpsm rate and a 142-fold increase in the indel rate (Table 1). Although *Vc*-mut lineages still show no indel bias between chromosomes, slightly more bpsm occurred on chr1 than expected in this MA experiment (bpsm: $\chi^2$ = 4.54, d.f. = 1, p = 0.0331; indels: $\chi^2$ = 0.04, d.f. = 1, p = 0.8402).

The conditional bpsm spectra of both *V. fischeri* and *V. cholerae* also changed dramatically as a result of the loss of a functional MMR system (Figure 4). Interestingly, despite having widely different bpsm spectra when the MMR system was intact, the bpsm spectra of the *Vf*-mut and *Vc*-mut lineages are remarkably similar. Both mutator bpsm spectra are dominated by A:T>G:C and G:C>T:A transitions, generating an expected %GC-content of 0.482 ± 0.016 (SEM) in *Vf*-mut and 0.475 ± 0.011 (SEM) in *Vc*-mut at mutation-drift equilibrium. This slight AT-bias was significant for *Vf*-mut, but not for *Vc*-mut (*Vf*-mut: $\chi^2$ = 6.08, d.f. = 1, p = 0.0136; *Vc*-mut: $\chi^2$ = 2.59, d.f. = 1, p = 0.1071). In addition, while *V. fischeri* continues to have a slight deletion bias in the absence of MMR, insertions actually occur at a slightly higher rate than deletions in the *Vc*-mut lineages, although not significantly so (*Vf*-mut: $\chi^2$ = 6.54, d.f. = 1, p = 0.011; *Vc*-mut: $\chi^2$ = 0.297, d.f. = 1, p = 0.5860). Lastly, the bpsm spectra do not vary significantly between chromosomes for either MMR deficient species (*Vf*-mut: $\chi^2$ = 8.93, d.f. = 1, p = 0.1120; *Vc*-mut: $\chi^2$ = 5.24, d.f. = 1, p = 0.3872).

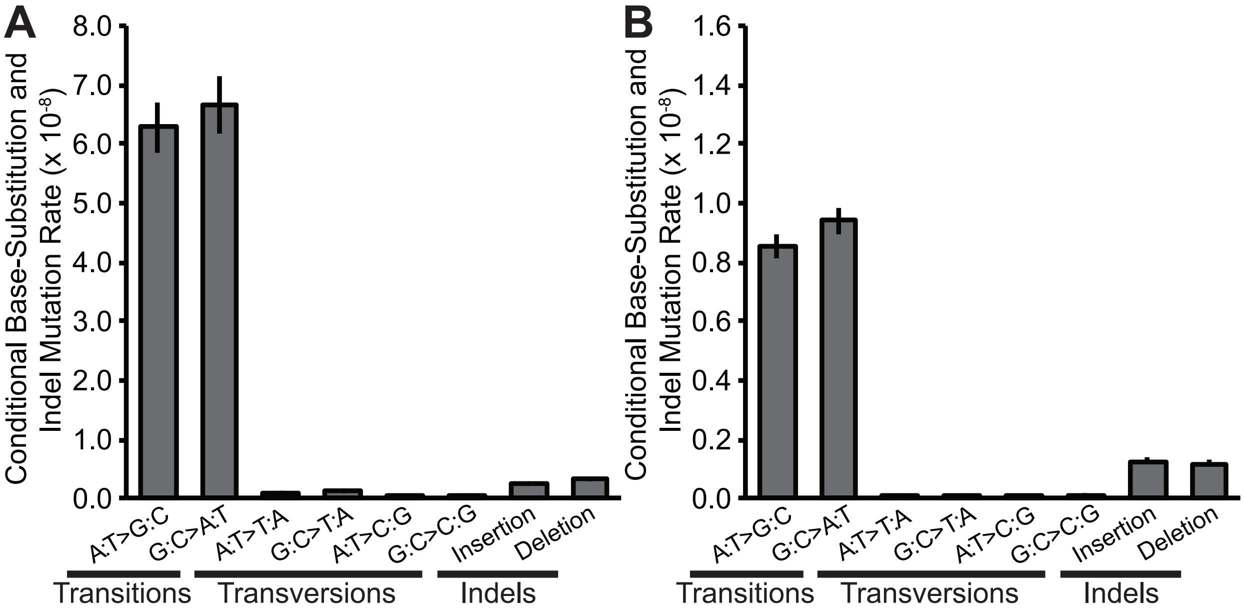

**Figure 4. Mismatch repair deficient conditional base-substitution (bpsm) and insertion-deletion (indel) rates per conditional base-pair per generation for *Vibrio fischeri ΔmutS* and *Vibrio cholerae ΔmutS* mutation accumulation lines**. Conditional bpsm and indel rates were estimated by dividing the number of observed mutations by the product of the analyzed sites capable of producing a given mutation and the number of generations of mutation accumulation. Error bars indicate one standard error of the mean.

In contrast to the wild-type MA lines, single-nucleotide indels represent the vast majority of the indels observed in both the *Vf*-mut (93.62%) and *Vc*-mut (85.66%) MA experiments (Figure 2). While the single-nucleotide indel rate increased 473-fold as a result of the loss of MMR in *V. fischeri*, the multi-nucleotide indel rate increased only 13-fold. The impact of the loss of a functional MMR system was similar in *V. cholerae*, as the single-nucleotide indel rate increased 334-fold and the multi-nucleotide indel rate increased 64-fold. Most of the increases in the multi-nucleotide mutation rate are the result of di- and tri-nucleotide indels, which were rarely observed in the wild-type MA experiments (Figure 2). As expected, the number of single-nucleotide indels observed in both *Vf*-mut and *Vc*-mut is significantly higher than their occurrence in the wild-type experiments, based on the null-expectation that the observed number of indels would be proportional to the product of the total number of sites analyzed across all lineages and the number of generations in the wild-type and mutator experiments (*Vf*-mut: $\chi^2$ = 5.46 • $10^3$, d.f. = 1, p < 0.0001; *Vc*-mut: $\chi^2$ = 2.57 • $10^3$, d.f. = 1, p < 0.0001). However, all indels that were observed in the mutator MA experiments up to 10-bps in length were significantly over-represented in mutator MA experiments (Table B.4), so the loss of a functional MMR system also had consequences for indels involving more than a single bp.

Most single-nucleotide indels in *Vf*-mut and *Vc*-mut lines occurred in homopolymeric runs. Therefore, in the mutator lines, I can focus on these runs and show that the ancestral repeat number is positively correlated with the single-nucleotide indel mutation rate in both *Vf*-mut and *Vc*-mut (Figure 5), differing significantly from the null expectation that the SSR indel rate would be proportional to the number of bases analyzed in each homopolymer length category in the *V. fischeri* ES114 and *V. cholerae* 2740-80 genomes, respectively (Chi-square test, repeat numbers 3-11; *Vf*-mut: $\chi^2$ = 3.19 • $10^4$, d.f. = 8, p < 0.0001; *Vc*-mut: $\chi^2$ = 6.85 • $10^4$, d.f. = 8, p < 0.0001). In conjunction with SSR biases of the *Vf*-wt indels, this suggests that both the repeat length and number of repeats are positively correlated with indel rates.

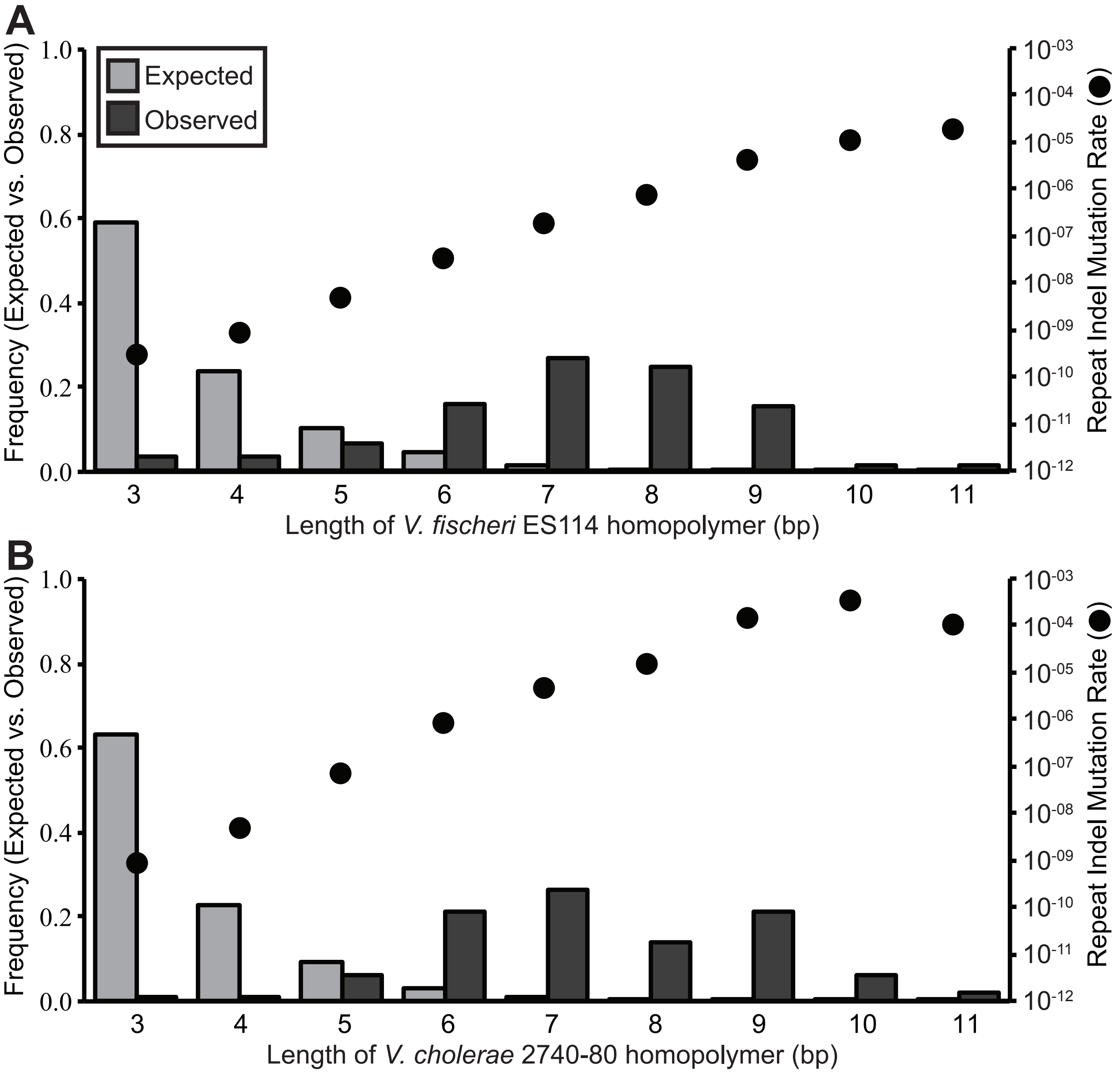

**Figure 5. Mismatch repair deficient insertion-deletion mutation (indel) rates per run per generation and frequencies in homopolymer repeats containing three or more repeats for *Vibrio fischeri ΔmutS* (A) and *Vibrio cholerae ΔmutS* (B).** Indel rates per run per generation were calculated as the number of observed indels in each homopolymer length category, divided by the product of the occurrence of homopolymers of that length in the genome, the number of generations, and the number of MA lineages analyzed. Expected frequencies were calculated based on the relative occurrence of each simple-sequence repeat (SSR) type in the *V. fischeri* ES114 and *V. cholerae* 2740-80 genomes, corrected for the number of bases in differently sized SSRs.

**Genomic distribution of spontaneous mutations.** To this point I have only discussed inter-chromosomal differences in mutation rates between two autonomously replicating chromosomes. However, the distribution of bpsms and indels in *V. fischeri* and *V. cholerae* may also vary among regions within these circular chromosomes. Therefore, I also analyzed the genome-wide distribution of bpsms by dividing each chromosome into 100 kb intervals extending bi-directionally from the origin of replication. Despite apparent intra-chromosome variation in the bpsm rate of both the *Vf*-wt and *Vc*-wt MA experiments (Figure 6A,B), the observed number of bpsms in 100kb intervals does not differ significantly from the null expectation that they would reflect the number of sites analyzed in each interval on chr1 or chr2 (*Vf*-wt Chr1: $\chi^2$ = 31.25, d.f. = 29, p = 0.3541; *Vf*-wt Chr2: $\chi^2$ = 17.19, d.f. = 15, p = 0.3076; *Vc*-wt Chr1: $\chi^2$ = 29.61, d.f. = 29, p = 0.4341; *Vc*-wt Chr2: $\chi^2$ = 12.83, d.f. = 11, p = 0.3050). However, the observed distribution of bpsms does differ significantly from this null expectation on chr1 of the *Vf*-mut and *Vc*-mut experiments (*Vf*-mut: $\chi^2$ = 132.97, d.f. = 29, p < 0.0001; *Vc*-mut: $\chi^2$ = 102.42, d.f. = 29, p < 0.0001) (Figure 6A,B). In these mutator lineages, both *Vf*-mut and *Vc*-mut lines experience similar bpsm patterns on chr1 (Figure 6A,B), where bpsm rates appear to be mirrored on the right and left replichores. However, these mirror images are not exact in either *Vf*-mut or *Vc*-mut, as the observed number of bpsms in each 100 kb interval of the left replichore of chr1 does vary significantly from the expected distribution of bpsm predicted by the concurrently replicated intervals on the right replichore of chr1 (*Vf*-mut: $\chi^2$ = 46.68, d.f. = 14, p < 0.0001; *Vc*-mut: $\chi^2$ = 53.52, d.f. = 14, p < 0.0001). Unlike chr1, the observed distribution of bpsms on chr2 of the mutator lineages does not differ from the null expectation generated by the number of sites analyzed per interval (*Vf*-mut: $\chi^2$ = 17.46, d.f. = 15, p = 0.2920; *Vc*-mut: $\chi^2$ = 10.98, d.f. = 11, p = 0.4451), suggesting that bpsm rates are more consistent across chr2 than they are on chr1 in the mutator lineages (Figure 6A, B).

Indels occur predominantly in SSRs and are thus also expected to vary significantly between genome regions. I also analyzed the distribution of indel rates across the 100kb intervals (Figure 6 C,D) and found that the observed distribution differed significantly from the null expectation in all MA experiments except *Vc*-wt, likely because only 22 indels were observed in this experiment (*Vf*-wt: $\chi^2$ = 83.08, d.f. = 43, p < 0.0001; *Vc*-wt: $\chi^2$ = 42.22, d.f. = 41, p = 0.4180; *Vf*-mut: $\chi^2$ = 96.38, d.f. = 43, p < 0.0001; *Vc*-mut: $\chi^2$ = 157.24, d.f. = 41, p < 0.0001). However, I see no conserved patters of indel rates on opposing replichores of either chromosome, with regions of high indel rates occurring at different genome locations in each MA experiment.

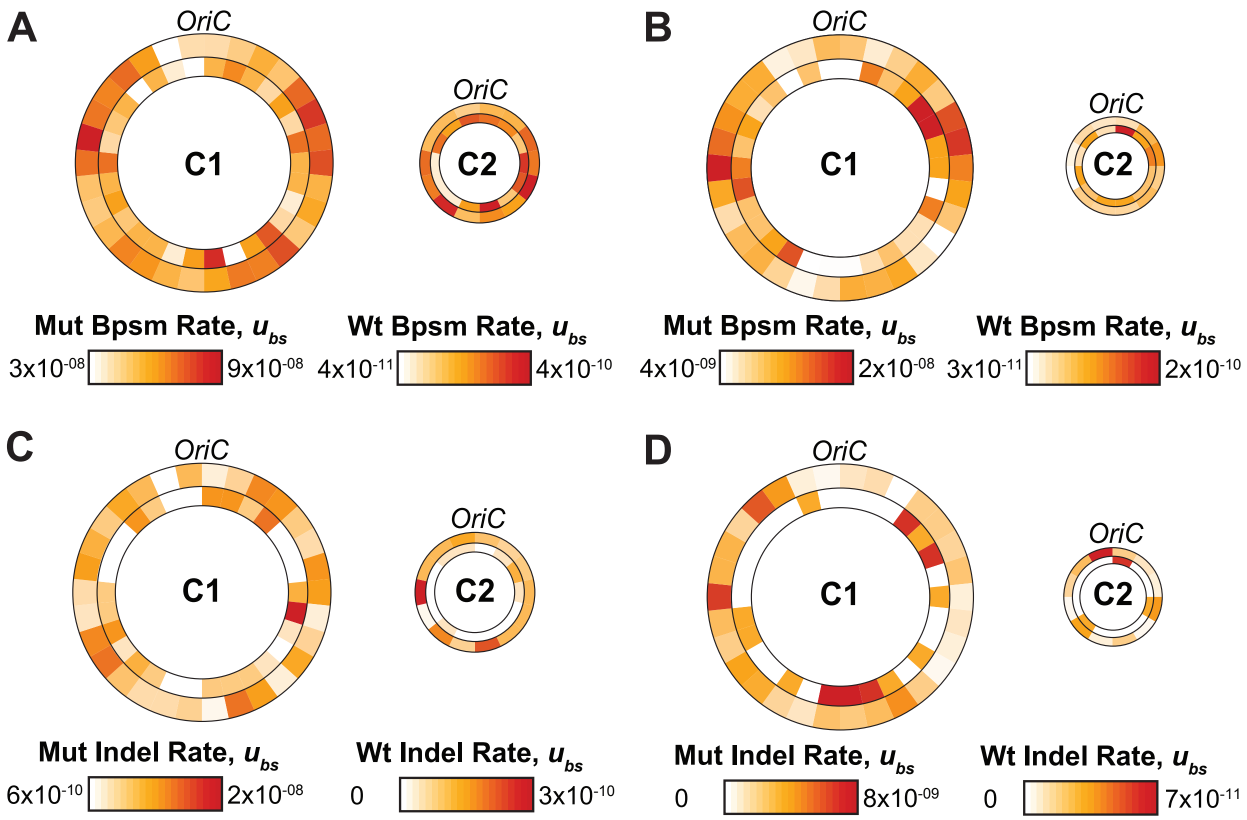

**Figure 6. Base-substitution (bpsm) and insertion-deletion (indel) mutation rates per base-pair per generation in 100kb intervals extending bi-directionally from the origin of replication (*OriC*) for all *Vibrio fischeri* and *Vibrio cholerae* mutation accumulation (MA) experiments.** Outer rings on each chromosome represent the mutator MA experiment and inner rings represent the corresponding wild-type MA experiment. (A) Bpsm rates of *Vf*-wt and *Vf*-mut; (B) Bpsm rates of *Vc*-wt and *Vc*-mut; (C) Indel rates of *Vf*-wt and *Vf*-mut; (D) Indel rates of *Vc*-wt and *Vc*-mut.

## DISCUSSION

A complete understanding of the biases generated by spontaneous mutation and their generality will inevitably require the extension of MA-WGS studies to a far greater collection of species. However, with each successive MA-WGS experiment, the general properties of spontaneous mutation become clearer and help produce more powerful null hypotheses for molecular evolution in diverse organisms. The major conclusions of my MA-WGS study of *V. fischeri* and *V. cholerae* are: (a) Base-substitution and insertion-deletion mutation rates are low, consistent with other bacterial species; (b) Base-substitution mutation biases contribute to, but don’t fully explain genome-wide %GC content; (c) Both the length of repeat units and the number of repeated units in simple-sequence repeats correlate positively with insertion-deletion rates; (d) Loss of a proficient mismatch repair system generates convergent mutation biases dominated by transitions and short insertions; and (e) Base-substitution mutations in strains deficient in mismatch repair vary in a mirrored wave-like pattern on opposing replichores on chromosome 1, but variation is limited on chromosome 2.

*Vibrio* species are abundant in marine environments worldwide and have high intraspecies genetic and phenotypic diversity (Thompson *et al.* 2004; Sawabe *et al.* 2009). Both mutation and horizontal-gene transfer (HGT) are expected to contribute to *Vibrio* biodiversity, but evidence exists that mutation is the primary force driving diversification within *Vibrio* clades (Thompson *et al.* 2004; Sawabe *et al.* 2009; Vos and Didelot 2009). Therefore, it has been tempting to invoke high mutation rates in *Vibrio* species to explain their high genetic diversity. However, I show here that both bpsm and indel rates in *V. fischeri* and *V. cholerae* are low, even for bacteria (Figure B.2). In fact, *V. cholerae* has the lowest recorded genome-wide rates of bpsms (0.0004 /genome/generation) and indels (0.00007 /genome/generation) of any bacterial species (Sung *et al.* 2012a, 2016). I suggest that the high genetic diversity and low mutation rates in these *Vibrio* species can be reconciled by the drift-barrier hypothesis, which states generally that any trait, including replication fidelity, may be refined by natural selection only to the point at which further improvement becomes overwhelmed by the power of genetic drift (Lynch 2010b, 2011; Sung *et al.* 2012a). Natural selection is most powerful in large populations of organisms with genomes composed of a high amount of coding sequence. Although *Vibrio* genomes are not exceptionally large, most sites are coding and potentially subject to selection. Furthermore, synonymous sites in both *Vibrio* genomes exhibit high diversity (Wollenberg and Ruby 2012; Sung *et al.* 2016), which is consistent with very large effective population size. Thus, both high amounts of coding sequence and high effective population size increase the ability of natural selection to reduce both bpsm and indel rates (Lynch 2010b, 2011; Sung *et al.* 2012a), yet yield enormous allelic diversity at any given time in both of these *Vibrio* species (Thompson *et al.* 2004; Vos and Didelot 2009).

Bpsm and indel rates and spectra varied among genomic regions of *V. fischeri* and *V. cholerae*, which can generate local sequence biases. In the *Vf*-wt lines, both G:C>A:T transitions and G:C>T:A transversions occurred at higher rates on chr2, although only the rate of G:C>T:A transversions was significantly higher (Figure 1C). It is also noteworthy that late replicating regions of chr1 also experience elevated G:C>T:A transversion rates. These elevated G:C>A:T and G:C>T:A rates are expected to enhance AT-biased mutation on chr2. Although I lack a sufficient number of mutations per lineage to confidently estimate %GC-content at mutation-drift equilibrium on individual chromosomes, I can use experiment-wide estimates of each conditional bpsm rate to estimate genome-wide, chr1, and chr2 %GC-content at mutation-drift equilibrium. Using experiment-wide bpsm rates, I estimate an overall %GC-content of 0.18 for *V. fischeri* at mutation-drift equilibrium. Surprisingly, the %GC-content of chr2 is expected to be 0.16 at mutation-drift equilibrium, 0.04 lower than expectations for chr1

(0.20). The actual %GC content of chr2 of the *V. fischeri* ES114 genome is also lower than chr1 (chr1: 0.39; chr2: 0.37) suggesting that mutation biases have contributed to this pattern. However, AT-mutation bias alone fails to explain realized %GC-contents on chr1 and chr2 in *V. fischeri*. Even stronger biases differentiating the chromosomes are seen in the *Vc*-wt lines, driven by significantly higher A:T>G:C transition rates on chr1 and by non-significant increases in G:C>A:T and G:C>T:A rates on chr2 (Figure 1C). These spectra predict %GC-content of 0.29 for chr1 and 0.20 for chr2 at mutation-drift equilibrium and likely contribute to the lower realized %GC content of chr2 in *V. cholerae* 2740-80 (chr1: 0.48; chr2: 0.47). Overall, these findings suggest that bpsm pressures contribute to genome-wide and intra-genome variation in %GC contents, but indel biases (Dillon *et al.* 2015), selection (Hershberg and Petrov 2010; Hildebrand *et al.* 2010), and/or biased gene conversion (Hershberg and Petrov 2010; Lassalle *et al.* 2015) must also contribute to produce the realized %GC content in *V. fischeri* and *V. cholerae*.

The most surprising indel biases found in *V. fischeri* was their size-distribution and their propensity to occur in SSRs. Given that the vast majority of indels observed in previous bacterial MA-WGS studies have been single-nucleotide indels (Lee *et al.* 2012; Long *et al.* 2014; Dettman *et al.* 2016), I was amazed by the high relative occurrence of indels between four and eight base-pairs in *V. fischeri* (Figure 2). One possible reason for this discrepancy might be an increased occurrence of SSRs with long repeated units in the *V. fischeri* ES114 genome (Ruby *et al.* 2005), which I find to be highly mutagenic (Figure 3). There are 100 SSRs of three or more units in the *V. fischeri* ES114 genome where the repeated unit is at least 4-bps in length, and I find that both the length of the repeated unit and the number of repeats in a SSR scale positively with the indel rate. A second possibility is that larger indels have gone undetected by prior MA-WGS analyses focused on MMR-deficient strains, in which single-nucleotide indels are evidently more common (Table B.2). I emphasize that this experiment demonstrates that the loss of MMR shifts the spectrum of indel mutations from a bias towards SSR's within longer repeats towards single nucleotides in homopolymeric runs, a shift with potentially broad phenotypic consequences. Lastly, longer indels, especially those in SSRs, are subject to increased false-negative rates due to the nature of short-read sequencing used in MA-WGS experiments. The majority of multi-nucleotide indels were supported with very low consensus in the initial alignments because of reads that only partly covered the SSR. Only when I filtered out reads that were not anchored by bps on both sides of the SSR did I achieve high consensus for these indels (Methods; Table B.2). It will be interesting to apply these sensitive detection methods for long SSR-associated indels to future experiments to see whether other species also experience elevated indel rates in SSRs with longer repeat units. I lack a sufficient number of SSR-associated indels to test for this correlation in *V. cholerae*, but it is worth noting that there are only 40 SSRs in the *V. cholerae* 2740-80 genome of three or more units where the repeated unit is at least 4-bps in length, which is less than half the number of long repeat SSRs in *V. fischeri* ES114.

Taken together, these long indels in SSRs generate localized hyper-mutation that may serve as contingency loci, enabling plasticity in both function and expression of the affected genes, and more accurate genotyping of closely related strains (van Belkum *et al.* 1998; Moxon *et al.* 2006; Danin-Poleg *et al.* 2007; Ghosh *et al.* 2008). These loci have received fairly little attention for genotyping *V. fischeri* strains, which have mostly been analyzed by multi-locus sequence typing of housekeeping genes (Mandel *et al.* 2009; Wollenberg and Ruby 2012). However, with a growing body of *V. fischeri* sequences becoming available, more fine-scale evolutionary relationships could be established using the highly-mutable long repeat SSRs identified in this study. Specifically, a number of intergenic regions in *V. fischeri* are especially mutagenic, with the same SSR being mutated in multiple independent lineages (Table B.2). Genotypic analyses using long SSRs are more common in *V. cholerae*, where establishing the evolutionary relationships between strains enables enhanced assessment of the epidemic risk (Danin-Poleg *et al.* 2007; Ghosh *et al.* 2008). Interestingly, of the nine long repeat SSRs that have been used as genotypic markers in these studies, only one, located in a secreted microbial collagenase, was mutated in this study. However, given that I observe only twenty-two total indels in the *Vc*-wt MA lineages, it is impossible to quantify whether these nine long repeat SSRs experience particularly high indel rates.

The mismatch repair pathway is a primary DNA repair pathway in diverse organisms across the tree of life (Kunkel and Erie 2005), but strains lacking MMR are not uncommon in nature (Hazen *et al.* 2009), chronic infections (Hall and Henderson-Begg 2006; Mena *et al.* 2008; Oliver 2010; Marvig *et al.* 2013), or long-term evolution experiments (Sniegowski *et al.* 1997). Loss of a functional MMR system can elevate mutation rates anywhere from 5 to 1000-fold, depending on both the defective component of the pathway and the genetic background (Lyer *et al.* 2006; Long *et al.* 2015; Reyes *et al.* 2015). The primary proteins involved in the MMR pathway in bacteria include the MutS protein, which binds mismatches to initiate repair, the MutL protein, which coordinates multiple steps of MMR synthesis, and the MutH protein, which nicks the unmethylated strand to remove the replication error (Kunkel and Erie 2005). The removal of the *mutS* gene in this study resulted in a 317-fold increase in the bpsm rate and a 102-fold increase in the indel rate in *V. fischeri* ES114. The removal of the *mutS* gene in *V. cholerae* had a less dramatic effect on the bpsm rate (85-fold increase), but a more dramatic effect on the indel rate (142-fold increase). Overall, this suggests that MMR is more central to the repair of bpsms in *V. fischeri*, while it is more important for the repair of indels in *V. cholerae*.

Despite the relatively wide range in the consequences of losing a functional MMR system for bpsm and indel rates, the changes in mutation spectra that result from MMR-deficiency are relatively conserved. Namely, nearly all bpsms observed in the *Vf*-mut and *Vc*-mut MA lineages are transitions, and nearly all indels involve only a single-nucleotide. Furthermore, an even higher proportion of single-nucleotide indels in both MMR-deficient MA experiments occur in homopolymers, where their rates scaled positively with the length of the homopolymer (Figure 5). These observations are consistent with previous reports in other bacterial MA experiments using MMR-deficient strains (Lee *et al.* 2012; Long *et al.* 2014; Dettman *et al.* 2016), and exert substantially stronger mutation biases at specific sites than genotypes with functional MMR. Importantly, this suggests that strong site-specific biases in the mutation spectra generated by the loss of MMR, along with the overall increase in mutation rates, may help to explain their evolutionary success. Couce et al. have found that mutator alleles can modify the distribution of fitness effects of individual beneficial mutations by enriching a specific spectrum of spontaneous mutations, and impact the evolutionary trajectories of different strains (Couce *et al.* 2013, 2015). The strong spectra biases generated by loss of a functional MMR system in *V. fischeri*, *V. cholerae*, and other species (Lee *et al.* 2012; Long *et al.* 2014; Dettman *et al.* 2016) would inevitably generate such biases, which may permit more rapid access to specific beneficial mutations and impact the evolutionary trajectories of MMR-deficient strains in clinical and environmental settings (Hall and Henderson-Begg 2006; Mena *et al.* 2008; Hazen *et al.* 2009; Oliver 2010; Marvig *et al.* 2013). Furthermore, MMR-deficiency will enrich for polymorphism at traditionally more stable sites, which will affect evolutionary analyses that combine highly mutable long SSR's with more stable mononucleotide repeats (Danin-Poleg *et al.* 2007).

The loss of MMR also helps reveal subtle mutation biases associated with the replicative polymerase that cannot be observed using the low number of mutations generated in wild-type MA-WGS experiments (Lee *et al.* 2012; Sung *et al.* 2015; Dettman *et al.* 2016). A mirrored wave-like pattern of bpsm rates on opposing replichores has now been observed in multiple MMR-deficient species studied by MA-WGS, although the exact shape of the pattern varies between species (Foster *et al.* 2013; Long *et al.* 2014; Dettman *et al.* 2016). I find the same mirrored wave-like pattern of bpsm rates on the opposing replichores of chr1 in MMR-deficient *V. fischeri* and *V. cholerae* (Figure 6A, B), which suggests that bpsm rates are impacted by genome location and that regions replicated at similar times on opposing replichores experience similar bpsm rates, at least in MMR-deficient strains. However, I do not observe any significant variation in the bpsm rates on chr2 (Figure 6A, B). I suggest that bpsm rates are less variable on chr2 because of their delayed replication. Specifically, chr2 replication is not initiated until a large portion of chr1 has already been replicated (Egan and Waldor 2003; Duigou *et al.* 2006; Rasmussen *et al.* 2007), which means that chr2 is not replicated during the primary peaks in the bpsm rate on the opposing replichores of chr1, and thus experience more consistent bpsm rates across the chromosome.

Mutation accumulation paired with whole-genome sequencing enables an unprecedented view of genome-wide mutation rates and spectra, revealing the underlying biases of spontaneous mutation. These underlying biases can explain why some genome regions evolve more rapidly than others and why the coding content of different genome regions varies. Moreover, the loss of a functional mismatch repair pathway can generate an entirely new spectrum of spontaneous mutations with different biases, which can have important consequences for our understanding of the evolutionary relationships between strains. As we continue to generate data on the properties of spontaneous mutation in diverse microbes, we can begin to assess the generality of mutational biases and more accurately evaluate the role of mutation bias in the molecular evolution.

# CHAPTER III

# REPLICATION TIMING GENERATES CONSERVED BASE-SUBSTITUTION MUTATION RATES IN CONCURRENTLY REPLICATED REGIONS OF MISMATCH REPAIR DEFICIENT BACTERIAL GENOMES

## INTRODUCTION

Although genome architecture varies markedly across the tree of life, some level of spatiotemporal organization of the genome is essential for all organisms because DNA must be compact but also available for gene expression, DNA replication, and chromosome segregation (Lynch 2007; Herrick 2011; Dorman 2013). Patterns of spatiotemporal organization are thus likely to influence a number of cellular processes (Dame *et al.* 2011; Sobetzko *et al.* 2012; Dorman 2013), and may impact mutation rates (Baer *et al.* 2007; Warnecke *et al.* 2012). Along these lines, a series of comparative studies in multicellular eukaryotes (Stamatoyannopoulos *et al.* 2009; Chen *et al.* 2010; Mugal *et al.* 2010; Herrick 2011), unicellular eukaryotes (Herrick 2011; Agier and Fischer 2012), archaea (Flynn *et al.* 2010), and bacteria (Mira and Ochman 2002; Cooper *et al.* 2010; Martincorena *et al.* 2012) have shown that substitution rates vary across the genome, correlating positively with replication timing. However, this correlation could result from higher base-substitution mutation (bpsm) rates or weaker purifying selection in late replicating regions (Ochman 2003; Cooper *et al.* 2010). Direct studies of bpsm rates using reporter constructs have also shown that mutation rates vary across the genomes of eukaryotes (Lang and Murray 2011) and bacteria (Hudson *et al.* 2002), but the latter study suggested that bpsm rates were highest at intermediate replicating regions. This within genome variation in mutation rates may have important implications for molecular clocks (Baer *et al.* 2007; Herrick 2011), horizontal gene transfer (Dorman 2013), and cellular disease (Schuster-Böckler and Lehner 2012; Donley and Thayer 2013; Lawrence *et al.* 2013; Liu *et al.* 2013), but a complete understanding of the effects of genome architecture on mutation rates will require direct estimates of genome-wide rates in a considerably more diverse array of organisms

Mutation accumulation (MA) experiments paired with whole-genome sequencing (WGS) offer a unique perspective into genome-wide biases in mutation rates and spectra. By initiating several replicate lineages from a single clonal ancestor and passaging each lineage through hundreds of single cell bottlenecks, the lineages will accrue mutations in the near absence of natural selection. All of these mutations can then be identified by whole-genome sequencing, and their genome-wide distributions can be characterized to compare local mutation rates. Interestingly, a collection of bacterial MA-WGS studies that have been carried out in mismatch repair (MMR)-deficient strains of *Escherichia coli* (Foster *et al.* 2013), *Pseudomonas fluorescens* (Long *et al.* 2014), *Bacillus subtilis* (Sung *et al.* 2015), and *Pseudomonas aeruginosa* (Dettman *et al.* 2016) have corroborated indirect evidence that bpsm rates are non-uniformly distributed across the genome. Yet another MA-WGS study found that bpsms were uniformly distributed across the genome in MMR-deficient yeast (Lang *et al.* 2013a), and most MMR-proficient MA experiments lack a sufficient number of bpsms to determine whether bpsm rates are non-uniform across their genomes (Ossowski *et al.* 2010; Denver *et al.* 2012; Lee *et al.* 2012; Zhu *et al.* 2014; Dillon *et al.* 2015).

The most remarkable feature of the MMR-deficient bacterial MA-WGS studies that have found that bpsm rates are non-uniformly distributed across the genome is that the pattern of bpsm rate variation across these genomes is very similar (Foster *et al.* 2013; Long *et al.* 2014; Dettman *et al.* 2016). Specifically, each of these genomes consists of a single circular chromosome and replication is initiated bi-directionally from a single origin of replication (*oriC*). These two opposing replichores eventually meet at the replication terminus, at which point the daughter chromosomes are segregated prior to cell division (Ochman 2002; Egan *et al.* 2005). Bpsm rates are always low near the *oriC* and reach intermediate peaks at approximately the same distance from the *oriC* on each replichore. Bpsm rates then decline into valleys on each replichore, before rising again as they approach the replication terminus. Although the magnitudes of these bpsm rate peaks and valleys vary between species, they are near mirror images of one another on the opposing replichores within each genome (Foster *et al.* 2013; Long *et al.* 2014; Dettman *et al.* 2016). It follows then that concurrently replicated regions in bacterial chromosomes appear to experience similar bpsm rates, either directly because of their concurrent replication or because other genomic features that impact bpsm rates are correlated with replication timing. Replication timing itself has long been expected to impact mutation rates because of the use of error prone polymerases (Courcelle 2009), error prone repair pathways (Lang and Murray 2011), and/or inconsistent nucleotide pools during late replication (Zhang and Mathews 1995; Cooper *et al.* 2010). However, other genomic features like compaction of the bacterial nucleoid, binding of nucleoid associated proteins (NAPs), and transcription may also impact mutation rates (Dame *et al.* 2011; Sobetzko *et al.* 2012; Warnecke *et al.* 2012; Dorman 2013). Interestingly, a number of sigma factors and NAPs are temporally regulated and their activity appears to be mirrored on the right and left replichores of *E. coli* (Dame *et al.* 2011; Sobetzko *et al.* 2012), which suggests that they are correlated with replication timing.

Although most well-studied bacterial genomes consist of a single circular chromosome, more complex bacterial genome architectures that consist of multiple circular chromosomes are not uncommon (Ochman 2002; Egan *et al.* 2005; Cooper *et al.* 2010; Val *et al.* 2014). Setting aside the distinction between chromosomes and megaplasmids (Agnoli *et al.* 2012), the *V. cholerae* and *V. fischeri* genomes are composed of two chromosomes, while the *B. cenocepacia* genome is composed of three. In all three of these species, the first chromosome (chr1) is largest, harbors the most essential genes and is expressed at the highest levels (Cooper *et al.* 2010; Morrow and Cooper 2012; Dillon *et al.* 2015). Yet, secondary chromosomes (chr2, chr3) do share similar structure, as they are also circular, initiate replication from a single origin of replication and are replicated bi-directionally on two replichores (Egan and Waldor 2003; Rasmussen *et al.* 2007; Val *et al.* 2014). However, while secondary chromosomes are replicated at the same rate as the first chromosome, their origins of replication (*oriCII*) have distinct initiation requirements from those of chr1 origins (*oriCI*) (Egan *et al.* 2005; Duigou *et al.* 2006). Importantly, chr2 (or chr3) replication is delayed relative to chr1, which ensures that replication of all chromosomes will terminate synchronously (Rasmussen *et al.* 2007; Baek and Chattoraj 2014; Val *et al.* 2014). The delayed initiation of chr2 replication means that the genome region near the origin of chr1 is always replicated prior to the replication of secondary chromosomes, while late replicated regions of chr1 are replicated concurrently with chr2. This general replication timing pattern in bacteria with multiple circular chromosomes begs the question of whether secondary chromosomes experience similar mirrored patterns of bpsm rate to the first chromosome, or are more similar to concurrent late replicating regions.

Here, I analyze the genome-wide distribution of spontaneous bpsms generated by MA-WGS experiments with MMR-deficient strains of *V. fischeri* (4313 bpsms) and *V. cholerae* (1022 bpsms), and spontaneous bpsms generated by MA-WGS experiments with MMR-proficient strains of *V. fischeri* (219 bpsms), *V. cholerae* (138 bpsms), and *B. cenocepacia* (245 bpsms). The bpsm rates in MMR-deficient MA-WGS experiments reveal that the patterns of bpsm rates on chr1 share the mirrored wave-like pattern described in previous MMR-deficient MA-WGS experiments. Although the distribution of bpsms on chr2 is not significantly heterogeneous, their patterns of bpsm rates are similar to those of late replicated regions of chr1, suggesting that concurrently replicated regions on different chromosomes also experience similar bpsm rates. However, with nearly an order of magnitude fewer bpsms available in the MMR-proficient MA-WGS experiments, I do not find that bpsm rates vary significantly within any of the wild-type chromosomes.

## MATERIALS AND METHODS

**Bacterial strains and culture conditions.** The founding strains of the five MA experiments conducted in this study were *V. fischeri* ES114 *ΔmutS* (*Vf*-mut), *V. cholerae* 2740-80 *ΔmutS* (*Vc*-mut), *V. fischeri* ES114 wild-type (*Vf*-wt), *V. cholerae* 2740-80 wild-type (*Vc*-wt), and *B. cenocepacia* HI2424 wild-type (*Bc*-wt). The mutator ancestors were generated by replacing the *mutS* gene in *V. fischeri* ES114 and *V. cholerae* 2740-80 with an erythromycin resistance cassette, as described previously (Datsenko and Wanner 2000; Stabb and Ruby 2002; Heckman and Pease 2007; Val *et al.* 2012). Completed genomes for *V. fischeri* ES114 and *B. cenocepacia* HI2424 were downloaded from NCBI (LiPuma *et al.* 2002; Ruby *et al.* 2005) and the location of *oriCI*, *oriCII*, and *oriCIII* (if applicable) were downloaded from the dOriC 5.0 database (Gao *et al.* 2013). Because the *V. cholerae* 2740-80 genome was only in draft form, the *V. cholerae* 2740-80 ancestor was sequenced using a long-insert library on single SMRT cell of a Pacific Biosciences RSII sequencer at the Icahn School of Medicine at Mount Sinai (Beaulaurier *et al.* 2015). I then assembled the *V. cholerae* 2740-80 genome anew into two contigs, representing chr1 and chr2, using HGAP3, and polished the assembly with Quiver (Chin *et al.* 2013). The *oriCI* and *oriCII* regions of the resultant *V. cholerae* 2740-80 assembly were identified using Ori-finder (Gao and Zhang 2008; Gao *et al.* 2013). More detailed methods for *ΔmutS* mutant construction and *V. cholerae* 2740-80 genome assembly are provided in Chapter 2.

MA experiments with *Vf*-mut and *Vf*-wt were carried out on tryptic soy agar plates supplemented with NaCl (TSAN) (30 g/liter tryptic soy broth powder, 20 g/liter NaCl, 15 g/liter agar) and were incubated at 28°. MA experiments with *Vc*-mut, *Vc*-wt, and *Bc*-wt were carried out on tryptic soy agar plates (TSA) (30 g/liter tryptic soy broth powder, 15 g/liter agar) and were incubated at 37°. Frozen stocks of each MA lineage were prepared at the end of the experiment by growing a single colony overnight in 5ml of tryptic soy broth supplemented with NaCl (TSBN) (30 g/liter tryptic soy broth powder, 20 g/liter NaCl) at 28° for *V. fischeri*, and in 5ml of tryptic soy broth (TSB) (30 g/liter tryptic soy broth powder) at 37° for *V. cholerae* and *B. cenocepacia*. Final isolates from all MA lineages in each experiment were stored in 8% DMSO at -80°.

**MA-WGS Process.** For each of the mutator MA experiments, forty-eight independent lineages were founded from single colonies of *V. fischeri* ES114 *ΔmutS* and *V. cholerae* 2740-80 *ΔmutS*. These lineages were independently propagated every 24 hours onto fresh media for 43 days and daily generations were estimated bi-monthly. For each of the three wild-type MA experiments, seventy-five independent lineages were founded from single colonies of *V. fischeri* ES114, *V. cholerae* 2740-80, and *B. cenocepacia* HI2424. These lineages were independently propagated every 24 hours onto fresh media for 217 days and daily generations were estimated monthly. To estimate daily generation times, ten representative colonies following 24 hours of growth were placed in 2 ml of phosphate buffer saline, serially diluted, and spread plated on the appropriate media (see above) to calculate the number of viable cells in each colony. The number of generations elapsed over 24 hours of growth was then calculated, and the average number of generations across the ten representative colonies was used as the experiment-wide daily generations for each lineage at that time-point. The total generations elapsed between each measurement was calculated as the product of the average daily generations and the number of days before the next measurement, and the total of number of generations elapsed during the entire MA experiment, per lineage, was calculated as the sum of these totals. At the conclusion of each MA experiment, all lineages were stored at -80° in 8% DMSO and later revived for WGS.

Genomic DNA was extracted from 1 ml of overnight culture (TSBN at 28° for *V. fischeri*; TSB at 37° for *V. cholerae* and *B. cenocepacia*) in 50 representative lineages from the three wild-type experiments and all 48 lineages from the two mutator experiments using the Wizard Genomic DNA Purification Kit (Promega). All libraries were prepared using a modified Illumina Nextera protocol designed for inexpensive library preparation of microbial genomes (Baym *et al.* 2015). Sequencing of the *Vf*-mut, *Vc*-mut, and *Bc*-wt lineages was performed using the 151-bp paired-end Illumina HISeq platform at the University of New Hampshire Hubbard Center for Genomic Studies, while sequencing for the *Vf*-wt and *Vc*-wt lineages was performed using the 101-bp paired-end Illumina HiSeq platform at the Beijing Genome Institute (BGI). In sum, I analyzed 19 *Vf*-mut lineages, 22 *Vc*-mut lineages, 48 *Vf*-wt lineages, 49 *Vc*-wt lineages, and 47 *Bc*-wt lineages, as fastQC revealed that the depth of coverage for the other sequenced lineages was insufficient for accurate detection of polymorphism (Andrews 2010). The reads from each of these lineages were mapped to their respective reference genomes with the Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009) and Novoalign (www.novocraft.com). The average depth of coverage was 124x for *Vf*-mut, 92x for *Vc*-mut. 100x for *Vf*-wt, 96x for *Vc*-wt, and 43x for *Bc*-wt.

**Base-substitution mutation identification.** For each MA experiment, bpsms were identified as described in Chapters 1 and 2. Briefly, I used SAMtools to convert the SAM alignment files from each lineage to mpileup format (Li *et al.* 2009), then in-house perl scripts to produce the forward and reverse read alignments for each position in each line. A three-step process was then used to detect putative bpsms. First, pooled reads across all lines were used to generate an ancestral consensus base at each site in the reference genome. This allows me to correct for any differences that may exist between the reference genomes and the ancestral colony of each my MA experiments.

Second, a lineage specific consensus base was generated at each site in the reference genome for each individual MA lineage using only the reads from that line. Here, a lineage specific consensus base was only called if the site was covered by at least two forward and two reverse reads and at least 80% of the reads identified the same base. Otherwise, the site was not analyzed. Third, each lineage specific consensus base that was called was compared to the overall ancestral consensus of the MA experiment and a putative bpsm was identified if they differed. This analysis was carried out independently with the alignments generated by BWA and Novoalign, and putative bpsms were considered genuine only if both pipelines independently identified the bpsm and they were only identified in a single lineage. All genuine bpsms analyzed in this study are summarized in Table C.1, which shows that nearly all bpsms were identified with high-confidence and were not clustered at my lower limits of detection.

**Base-substitution mutation rate analysis at different interval lengths.** Genomes were divided into intervals of 10 Kb, 25 Kb, 50 Kb, 100 Kb, 250 Kb, and 500 Kb, and bpsms were categorized into those intervals based on their location. On chr1, these intervals started at the origin of replication (*oriCI*), and extended bi-directionally to the replication terminus to mimic the progression of the two replication forks. Therefore, in each analysis, the final interval on each replichore of chr1 was smaller than the rest, since each half of chr1 is not exactly divisible by any of the intervals lengths. For example, chr1 of *V. fischeri* is 2,897,536 bps in length, so there are 1,448,768 bps on each replichore. Therefore, when I analyze bpsm rates in 100 Kb intervals from *oriCI*, the first fourteen intervals on each replichore are 100 Kb in length, while the final interval on each replichore is only 48.768 Kb in length.

On secondary chromosomes, bpsm rates were analyzed with the same interval sizes as chr1, where intervals were still measured relative to the initiation of replication of *oriCI*, rather than *oriCII* (Figure 1). This ensured that I could make direct comparisons between concurrently replicated intervals on chr1 and chr2, where the limits of each interval relative to *oriCI* initiation were identical. I achieved this by starting intervals bi-directionally from the replication terminus in the opposite direction of the replication forks, starting with the smaller intervals used in the corresponding interval analysis on chr1. All intervals were thus exactly the same length as those of the concurrently replicated intervals on chr1, with the exception of the final two intervals on each replichore, which extended into the concurrently replicated interval of chr1 until they reached their own origin of replication (*oriCII*). In the example presented above for *V. fischeri*, chr2 is 1,330,333 bps in length. Therefore, for 100 Kb intervals, the final interval on each replichore of chr2 was 48.768 Kb in length, followed by six 100 Kb intervals extending away from the terminus on each replichore, and finishing with a single 16.398 Kb interval on each replichore, until the two replichores meet at *oriCII*. Bpsm rates in each interval were calculated as the number of mutations observed in each interval, divided by the product of the total number of sites analyzed in that interval across all lines and the number of generations of mutation accumulation per line, so even the bpsm rates in the smaller intervals near the origin on chr2 could be directly compared to the larger concurrently replicated interval on chr1.

**Statistical analyses.** All statistical analyses were performed in R Studio Version 0.99.489 using the Stats analysis package (R Development Core Team 2013).

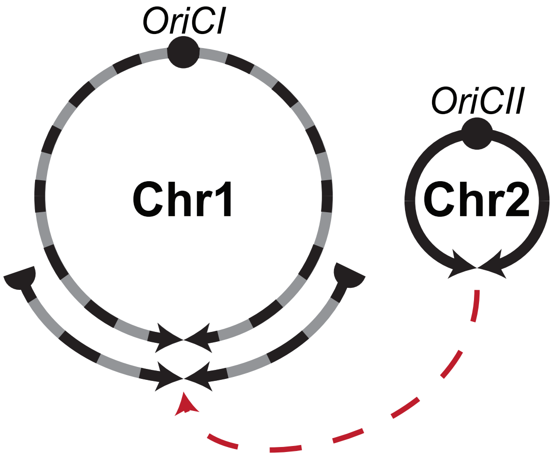

**Figure 1. Interval analysis allowing direct comparisons of base-substitution mutation (bpsm) rates of concurrently replicated regions on chromosome 1 (chr1) and chromosome 2 (chr2)**. Chr2 is split at it's origin of replication (*oriCII*), and mapped directly to concurrently replicated intervals in late replicating regions of chr1, since the two chromosomes terminate replication synchronously. All intervals on both chromosomes are thus relative to the initiation of replication of *oriCI*, and the boundaries of the intervals are at identical locations.

## RESULTS

Two MMR-deficient (mutator) and three MMR-proficient (wild-type) MA-WGS experiments were founded by five different ancestral strains: a) *V. fischeri* ES114 *ΔmutS* (Vf-mut), b) *V. cholerae* 2740-80 *ΔmutS* (*Vc*-mut), c) *V. fischeri* ES114 wild-type (*Vf*-wt), d) *V. cholerae* 2740-80 wild-type (*Vc*-wt), and e) *B. cenocepacia* HI2424 wild-type (*Bc*-wt). Forty-eight independent MA lineages were propagated for 43 days in the two mutator experiments and seventy-five MA lineages were propagated for 217 days in the three wild-type experiments. In total, I performed successful WGS on evolved clones of 19 *Vf*-mut lineages, 22 *Vc*-mut lineages, 48 *Vf*-wt lineages, 49 *Vc*-wt lineages, and 47 *Bc*-wt lineages. However, despite the fact that the mutator experiments were shorter and involved fewer lineages, the vast majority of bpsms were generated in the *Vf*-mut and *Vc*-mut lineages, as their bpsm rates are 317-fold and 85-fold greater than those of their wild-type counterparts, respectively (Table 1). Consequently, I can study the effects of genomic position on bpsm rates in much greater detail in the mutator lineages, where I observe a reasonable number of bpsms distributed across the genome at intervals as low as 10 Kb in length (Table 1), the approximate length of bacterial microdomains (Dorman 2013).

**Table 1. Average number of base-substitution mutations (bpsms) in each of the interval lengths analyzed and corresponding standard errors (SEM) across all intervals of that length for all five MA experiments.**

| MA | 500 Kb | | 250 Kb | | 100 Kb | | 50 Kb | | 25 Kb | | 10 Kb | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lines | Avg. | SEM | Avg. | SEM | Avg. | SEM | Avg. | SEM | Avg. | SEM | Avg. | SEM |
| *Vf*-mut | 499.0 | 35.8 | 253.5 | 13.8 | 101.1 | 3.1 | 50.5 | 1.3 | 25.3 | 0.5 | 10.1 | 0.2 |
| *Vc*-mut | 141.5 | 21.1 | 65.3 | 6.2 | 25.5 | 1.5 | 12.5 | 0.6 | 6.2 | 0.3 | 2.5 | 0.1 |
| *Vf*-wt | 22.3 | 3.3 | 12.3 | 1.4 | 5.0 | 0.4 | 2.5 | 0.2 | 1.2 | 0.1 | 0.5 | 0.0 |
| *Vc*-wt | 18.0 | 3.1 | 8.8 | 1.0 | 3.4 | 0.3 | 1.7 | 0.1 | 0.8 | 0.1 | 0.3 | 0.0 |
| *Bc*-wt | 15.9 | 1.4 | 7.6 | 0.6 | 3.3 | 0.2 | 1.6 | 0.1 | 0.8 | 0.1 | 0.3 | 0.0 |

To analyze the variation in bpsm rates on chr1 of the *Vf*-mut and *Vc*-mut lineages, I independently analyzed the bpsm rates in 10 Kb, 25 Kb, 50 Kb, 100 Kb, 250 Kb, and 500 Kb intervals, with intervals extending bi-directionally from the *oriCI*, like the replication forks do when the chromosome is being replicated. For secondary chromosomes, I independently analyzed bpsm rates using the same interval sizes as chr1, but these chromosomes were analyzed relative to the initiation of replication of *oriCI* (See methods; Figure 1). This ensures that intervals on secondary chromosomes can be directly compared to late replicating intervals on chr1, allowing me to assess whether concurrently replicated regions on different chromosomes experience similar bpsm rates. The bpsm rates of different intervals may vary because of nucleotide context, methylation, heterogeneous DNA synthesis and repair, nucleoid structure, NAPs, or transcription, but focusing on genome-wide concurrently replicated regions allows me to narrow my focus to the genomic features that contribute most substantively to global patterns of bpsm rate variation.

**Patterns of mutator base-substitution mutation rate variation within single chromosomes.** In the *Vf*-mut MA experiment, there was no significant difference in the bpsm rate between chr1 and chr2 based on the ratio of bpsm rates to the number of sites analyzed on each chromosome (Chi-square test; $\chi^2$ = 0.11, d.f. = 1, p = 0.7410). However, chr1 did experience a slightly higher bpsm rate than chr2 in the *Vc*-mut MA experiment (Chi-square test; bpsm: $\chi^2$ = 4.54, d.f. = 1, p = 0.0331). On a finer scale, I can reject the null hypothesis that bpsms were uniformly distributed across 500 Kb, 250 Kb, 100 Kb, 50 Kb, and 10 Kb intervals on chr1 in both the *Vf*-mut and *Vc*-mut MA experiments (Table 2). Yet I cannot reject this null hypothesis on chr2 for either MA experiment, suggesting that although bpsm rates vary within chr1, they do not vary on chr2 (Table 2).

**Table 2. Chi-square test statistics testing the null hypothesis that the observed number of bpsms in each interval simply reflects the number of sites analyzed in that interval, which would suggest that bpsms are uniformly distributed across the genome.** Chi-square tests were conducted for each chromosome at each interval length for each MA experiment.

| MA Lines | Interval Length | Chr1 | | | Chr2 | | |
|---|---|---|---|---|---|---|---|
| | | $\chi^2$ | df | p | $\chi^2$ | df | p |
| *Vf*-mut | 500 Kb | 25.89 | 5 | <0.0001 | 7.53 | 3 | 0.057 |
| | 250 Kb | 89.27 | 11 | <0.0001 | 8.59 | 5 | 0.127 |
| | 100 Kb | 132.97 | 29 | <0.0001 | 17.46 | 15 | 0.292 |
| | 50 Kb | 172.15 | 57 | <0.0001 | 28.19 | 27 | 0.401 |
| | 25 Kb | 223.93 | 115 | <0.0001 | 52.56 | 53 | 0.491 |
| | 10 Kb | 432.48 | 289 | <0.0001 | 110.89 | 133 | 0.919 |
| *Vc*-mut | 500 Kb | 43.28 | 5 | <0.0001 | 7.18 | 3 | 0.066 |
| | 250 Kb | 76.02 | 11 | <0.0001 | 7.82 | 5 | 0.166 |
| | 100 Kb | 102.42 | 29 | <0.0001 | 10.98 | 11 | 0.445 |
| | 50 Kb | 139.26 | 59 | <0.0001 | 17.84 | 23 | 0.766 |
| | 25 Kb | 208.58 | 119 | <0.0001 | 35.50 | 45 | 0.844 |
| | 10 Kb | 398.74 | 299 | <0.0001 | 90.72 | 111 | 0.921 |
| *Vf*-wt | 500 Kb | 5.02 | 5 | 0.414 | 5.24 | 3 | 0.155 |
| | 250 Kb | 13.62 | 11 | 0.255 | 6.00 | 5 | 0.306 |
| | 100 Kb | 31.25 | 29 | 0.354 | 17.19 | 15 | 0.308 |
| | 50 Kb | 64.18 | 57 | 0.240 | 34.38 | 27 | 0.155 |
| | 25 Kb | 121.33 | 115 | 0.325 | 49.86 | 53 | 0.597 |
| | 10 Kb | 308.56 | 289 | 0.205 | 120.31 | 133 | 0.777 |
| *Vc*-wt | 500 Kb | 8.39 | 5 | 0.136 | 4.77 | 3 | 0.190 |
| | 250 Kb | 14.43 | 11 | 0.210 | 5.18 | 5 | 0.394 |
| | 100 Kb | 29.61 | 29 | 0.434 | 12.83 | 11 | 0.305 |
| | 50 Kb | 56.36 | 59 | 0.574 | 19.31 | 23 | 0.683 |
| | 25 Kb | 113.63 | 119 | 0.622 | 35.06 | 45 | 0.857 |
| | 10 Kb | 269.45 | 299 | 0.889 | 106.40 | 111 | 0.606 |
| *Bc*-wt[a] | 500 Kb | 6.32 | 7 | 0.503 | 13.39 | 7 | 0.063 |
| | 250 Kb | 9.86 | 13 | 0.706 | 16.16 | 13 | 0.241 |
| | 100 Kb | 27.33 | 35 | 0.819 | 35.30 | 31 | 0.272 |
| | 50 Kb | 58.42 | 69 | 0.814 | 63.11 | 61 | 0.402 |
| | 25 Kb | 153.51 | 139 | 0.189 | 129.01 | 121 | 0.292 |
| | 10 Kb | 411.34 | 349 | 0.012 | 327.15 | 301 | 0.144 |

[a] *B. cenocepacia* also has a third chromosome and in all cases I cannot reject the null hypothesis that bpsms are uniformly distributed across the chromosome for any interval length.

Interestingly, although bpsm rates are non-uniform on chr1 for both the *Vf*-mut and *Vc*-mut MA experiments, they both follow a remarkably similar wave-like pattern of bpsm rates extending bi-directionally from the origin of replication (Figure 2A, B). Bpsm rates are low at the *oriC*, increase to bpsm rate peaks approximately 600 Kb from the *oriC* on both replichores, then decline into another valley before rising again as they approach the replication terminus. Furthermore, while the bpsm rate increases as the two replichores approach the replication terminus in both *Vf*-mut and *Vc*-mut, a narrow terminal valley in bpsm rate may also exist directly at the replication terminus (Figure 2A, B), although this is only apparent in the 100 Kb analyses. This and other features of the mirrored wave-like pattern of bpsm rates become somewhat dampened when interval lengths exceed 100 Kb, while interval lengths lower than 100 Kb often obscure features of the pattern because of background noise. Thus I focus the majority of my analyses on 100 Kb intervals, where I achieve the best balance between the total number of intervals and the number of bpsms in each interval.

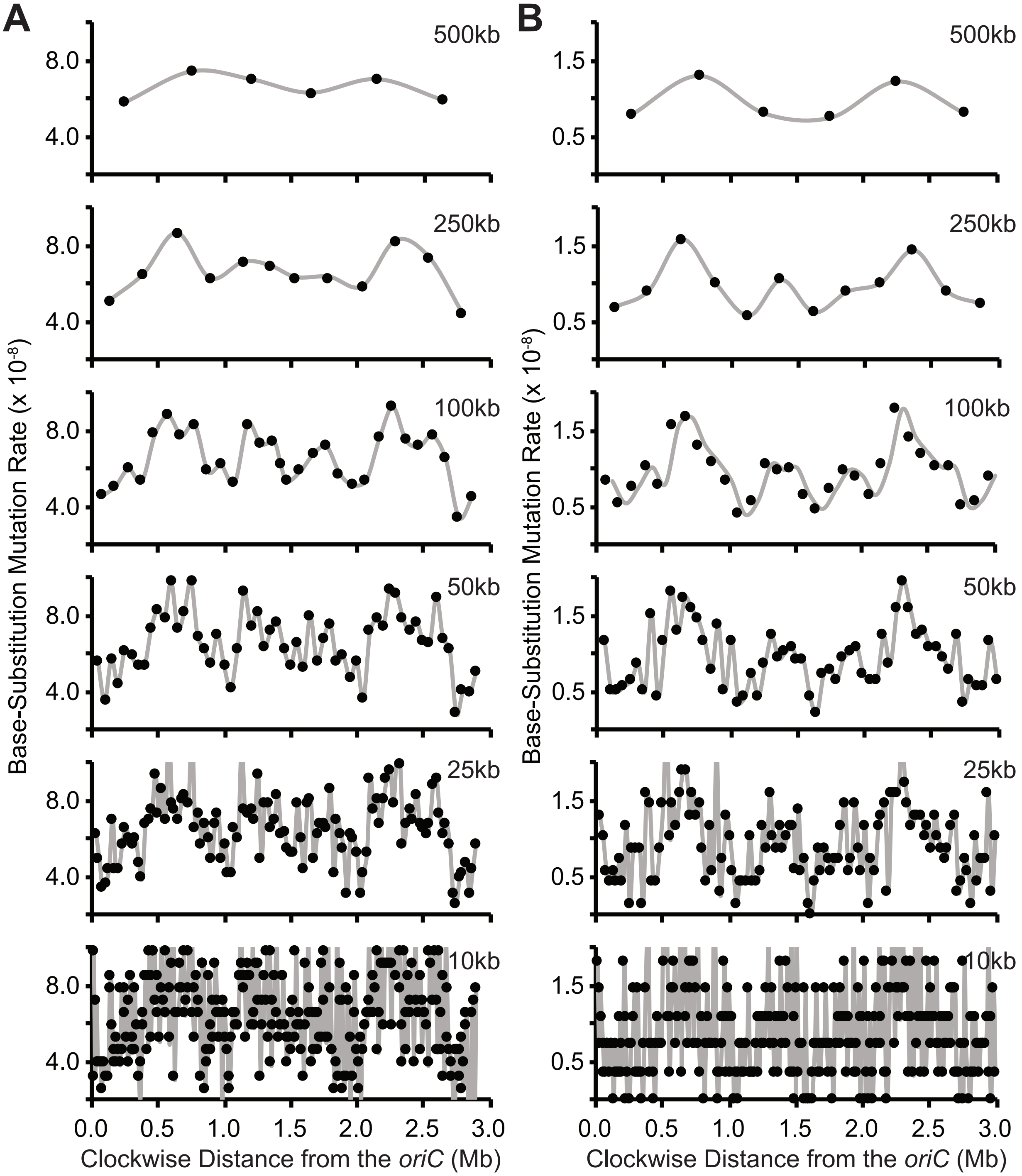

**Figure 2. Patterns of base-substitution mutation (bpsm) rates at various size intervals extending clockwise from the origin of replication (*oriC*) in MMR-deficient mutation accumulation lineages of *Vibrio fischeri* (A) and *Vibrio cholerae* (B) on chromosome 1.** Bpsm rates are calculated as the number of mutations observed within each interval, divided by product of the total number of sites analyzed within that interval across all lines and the number of generations of mutation accumulation.

As expected from the mirrored wave-like pattern of bpsm rates on opposing replichores of chr1, I find a significant positive relationship between the bpsm rates of right replichore intervals and the bpsm rates of concurrently replicated left replichore intervals at an interval length of 100 Kb in both the *Vf*-mut and *Vc*-mut experiments (Figure 3 A, B) (Linear regression; *Vf*-mut: F = 10.98, df = 13, p = 0.0060, $r^2$ = 0.46; *Vc*-mut: F = 6.76, df = 13, p = 0.0221, $r^2$ = 0.34). This relationship is also positive in analyses at all other interval lengths, but is only significantly positive at intervals of 250 Kb, 100 Kb, 50 Kb, 25 Kb, and 10 Kb for *Vf*-mut, and intervals of 100 Kb, 50 Kb, and 10 Kb for *Vc*-mut (Table C.2; Table C.3). In contrast, I find no relationship between right replichore bpsm rates and concurrently replicated left replichore bpsm rates on chr2 at an interval lengths of 100 Kb for *Vf*-mut or *Vc*-mut (Figure 3A, B), nor is there a significant relationship in analyses at any other interval lengths for chr2 (Table C.2; Table C.3). Concurrently replicated regions on primary bacterial chromosomes and all regions on secondary bacterial chromosomes therefore appear to experience similar bpsm rates, either directly because of their concurrent replication or because other genomic features that impact bpsm rates are correlated with replication timing.

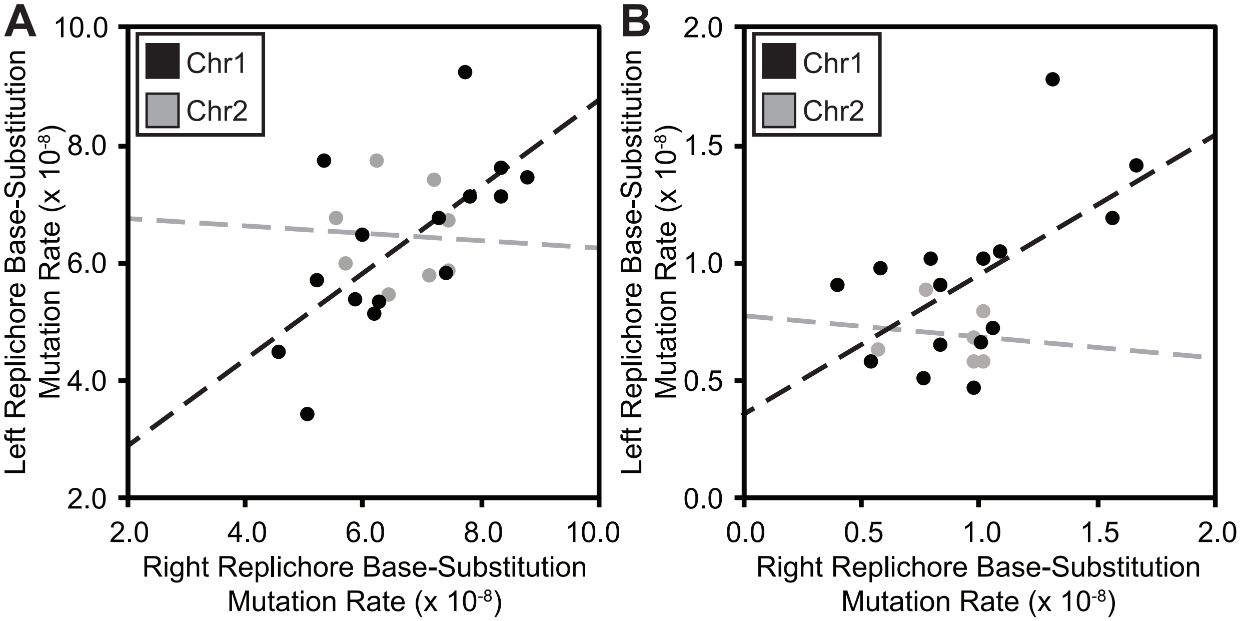

**Figure 3. Relationship between base-substitution mutation (bpsm) rates in 100 Kb intervals on the right replichore with concurrently replicated 100 Kb intervals on the left replichore in MMR-deficient *Vibrio fischeri* (A) and *Vibrio cholerae* (B).** Bpsm rates were calculated as the number of mutations observed within each interval, divided by product of the total number of sites analyzed within that interval across all lines and the number of generations of mutation accumulation. The range of bpsm rates between intervals on chromosome 1 is greater for both *V. fischeri* and *V. cholerae* and both linear regressions are significant (*V. fischeri*: F = 10.98, df = 13, p = 0.0060, $r^2$ = 0.46; *V. cholerae*: F = 6.76, df = 13, p = 0.0221, $r^2$ = 0.34), while neither linear regressions on chromosome 2 are significant (*V. fischeri*: F = 0.02, df = 6, p = 0.8911, $r^2$ = 0.03 • $10^{-1}$ ; *V. cholerae*: F = 0.06, df = 4, p = 0.8141, $r^2$ = 0.02).

**Patterns of mutator base-substitution mutation rates of concurrently replicated regions on different chromosomes.** Late replicated regions of chr1 may experience similar bpsm rates to chr2 because like the opposing replichores of chr1, they are replicated concurrently. To study this relationship, I mapped the patterns of bpsm rates in 100 Kb intervals on chr2 to those of late replicated 100 Kb intervals on chr1 for both *Vf*-mut and *Vc*-mut (Figure 1). Interestingly, because of its smaller size and delayed replication, chr2 narrowly avoids the high bpsm rate peaks on the right and left replichores of chr1 in both *Vf*-mut and *Vc*-mut (Figure 4A, B). The patterns of bpsm rates on chr2 appear similar to the bpsm rates of the late replicated regions on chr1 in both species (Figure 4A, B), but the overall variation in bpsm rates between different intervals is relatively limited in late replicated regions. Consequently, there is not a significant correlation between chr2 bpsm rates and the bpsm rates of late replicated regions of chr1, despite their apparent similarity (Linear regression; *Vf*-mut: F = 0.62, df = 14, p = 0.4443, $r^2$ = 0.04; *Vc*-mut: F = 0.072, df = 10, p = 0.7938, $r^2$ = 0.01).

Although the bpsm rates of chr2 and of late replicated regions on chr1 were poorly correlated, I also wanted to test whether the bpsm rates in 100 Kb intervals on chr2 were more similar to the concurrently replicated intervals on chr1 than they were to any other chr1 regions. Therefore, I mapped chr2 bpsm rates to all possible interval combinations on the right and left replichores of chr1, starting with the concurrently replicated terminal intervals and shifting back one interval per replichore until the origin of replication of chr1 was reached. I then calculated the sum of the residuals in each analysis to identify the best fit for chr2 bpsm rates. For *Vf*-mut, the lowest sum of the residuals (14.01 • $10^{-8}$) occurs when the chr2 intervals were mapped to the concurrently late replicated intervals on chr1 (Figure 4A; Table C.4). Similarly, the lowest sum of the residuals in *Vc*-mut (2.53 • $10^{-8}$) occurs when the chr2 intervals are mapped to their concurrently replicated intervals on chr1 (Figure 4B; Table C.4). This suggests that although I do not find a significant relationship between concurrently replicated 100 Kb intervals on chr1 and chr2, bpsm rates on chr2 are more similar to the bpsm rates of the late replicating concurrently replicated regions on chr1 than any other location on chr1.

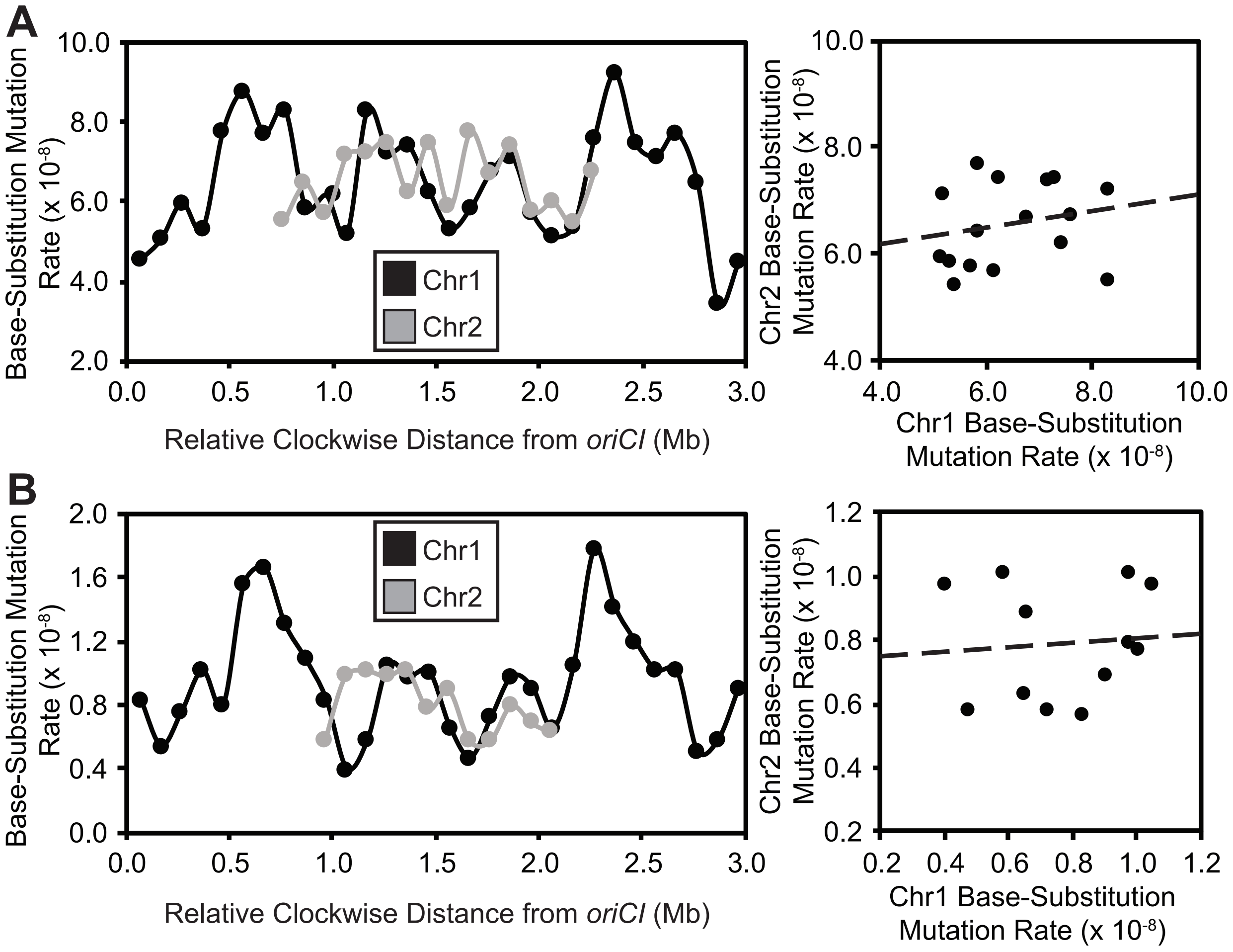

**Figure 4. Patterns of base-substitution mutation (bpsm) rates in 100 Kb intervals extending clockwise from the origin of replication (*oriCI*) on chromosome 1 (chr1) and patterns of bpsm of concurrently replicated 100 Kb intervals on chromosome 2 (chr2) for MMR-deficient *Vibrio fischeri* (A) and *Vibrio cholerae* (B).** Bpsm rates were calculated as the number of mutations observed within each interval, divided by product of the total number of sites analyzed within that interval across all lines and the number of generations of mutation accumulation. Patterns of bpsm rates on chr2 appear to map to those of concurrently replicated regions on chr1 in both species, but the variance in bpsm rate between intervals is not sufficient to produce significant linear regressions between concurrently replicated intervals on chr1 and chr2 in either *V. fischeri* or *V. cholerae* (*V. fischeri*: F = 0.62, df = 14, p = 0.4443, $r^2$ = 0.04; *V. cholerae*: F = 0.07, df = 10, p = 0.7938, $r^2$ = 0.01).

**Base-substitution mutation rates in wild-type MA lineages.** Although the total number of bpsms in the mutator MA experiments gives me substantially more statistical power to evaluate intra-genome variation in bpsm rates, these rates do not necessarily reflect rates in the wild-type lineages. Therefore, I also attempted to analyze the effects of replication timing on the bpsm rates of chr1 and chr2 of the three wild-type MA experiments. Here, I have considerably fewer bpsms at all interval lengths and cannot confidently reject the null hypothesis that bpsms are uniformly distributed across chr1, chr2, or chr3 (for *Bc*-wt) in the *Vf*-wt, *Vc*-wt, or *Bc*-wt MA experiments (Table 2), although there is slightly more variation in bpsm rates than expected based on the number of sites analyzed in the 10 Kb intervals on chr1 of *Bc*-wt (Chi-square test; $\chi^2$ = 411.34, d.f. = 349, p = 0.0122).

The lack of significant variation in bpsm rates within chromosomes in wild-type MA lineages also means that mirrored wave-like patterns of bpsm rates were not observed (Figure 5A, B, C). Moreover, there is no significant relationship between the bpsm rates of 100kb intervals on the right replichore and concurrently replicated left replichore intervals of chr1 for *Vf*-wt, *Vc*-wt, or *Bc*-wt (Linear regression; *Vf*-wt: F = 0.82, df = 13, p = 0.3810, $r^2$ = 0.06; *Vc*-wt: F = 0.11, df = 13, p = 0.7480, $r^2$ = 0.01; *Bc*-wt: F = 0.86, df = 16, p = 0.3670, $r^2$ = 0.05). There does appear to be some similarity between the bpsm rates on chr2 and late replicating regions on chr1 (Figure 5), but a linear regression analysis found that this relationship was not significant (Linear regression; *Vf*-wt: F = 0.16, df = 14, p = 0.7001, $r^2$ = 0.01; *Vc*-wt: F = 2.72, df = 10, p = 0.1300, $r^2$ = 0.21; *Bc*-wt: F = 0.32, df = 30, p = 0.5760, $r^2$ = 0.01). It is nevertheless important to recognize that even at interval lengths of 100 Kb, I observe an average of only 4.65 (0.38), 3.29 (0.28), and 3.08 (0.22) (SEM) bpsms per interval for the *Vf*-wt, *Vc*-wt, and *Bc*-wt MA lineages, respectively. Consequently, I emphasize that I may not have found any statistically significant relationships between the bpsm rates of concurrently replicated regions in the wild-type MA experiments because the patterns are truly a phenomenon specific to MMR-deficient genotypes, or because I lack the statistical power to analyze true within-chromosome variation in bpsm rates in the wild-type MA lineages.

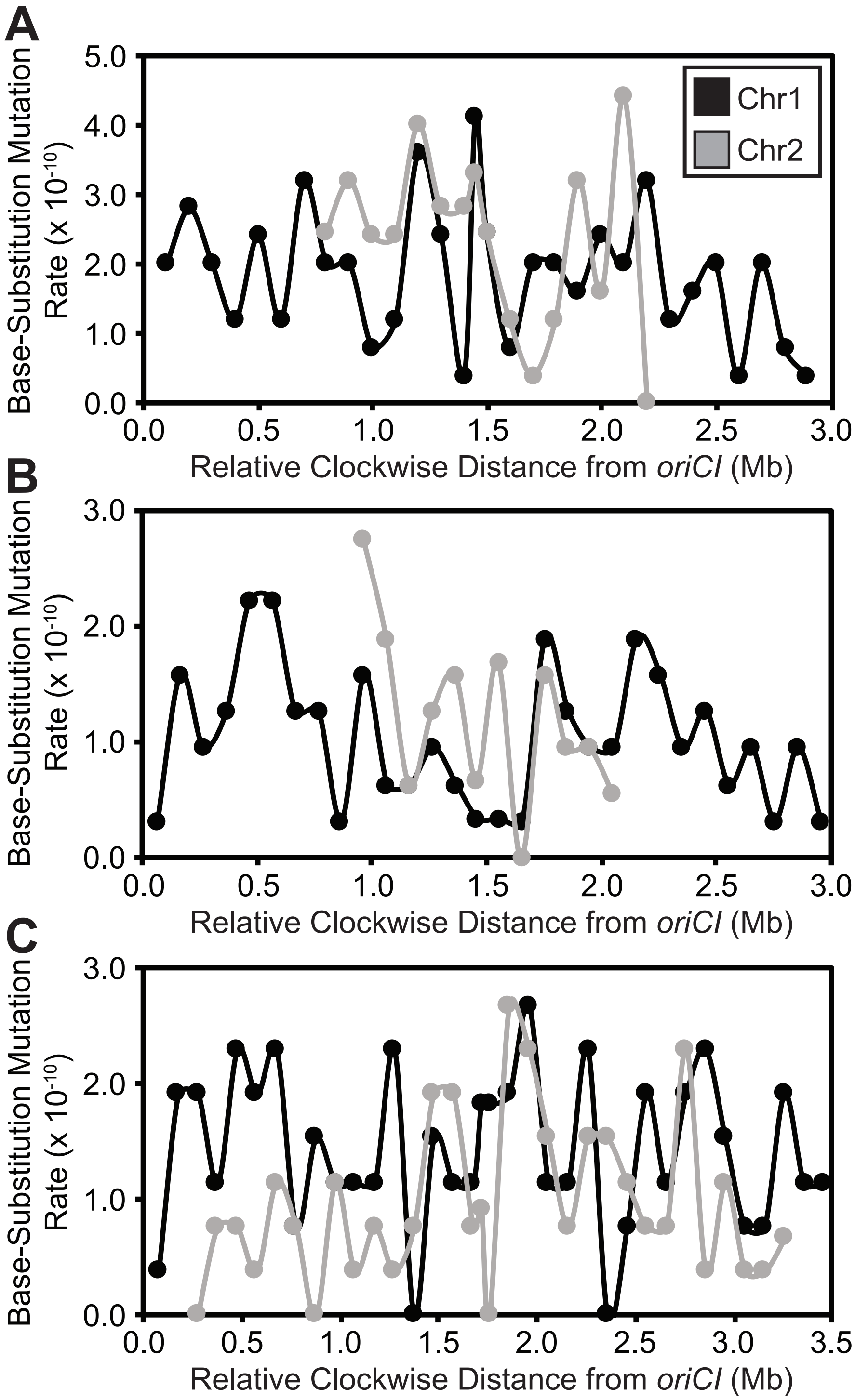

**Figure 5. Patterns of base-substitution mutation (bpsm) rates in 100 Kb intervals extending clockwise from the origin of replication (*oriC*) on chromosome 1 (chr1) and concurrently replicated intervals of chromosome 2 (chr2) for MMR-proficient *Vibrio fischeri* (A), *Vibrio cholerae* (B), and *Burkholderia cenocepacia* (C).** Bpsm rates were calculated as described in the MMR-deficient MA experiments. *B. cenocepacia* also has a third chromosome, which is not shown.

## DISCUSSION

Non-uniform genome-wide mutation rates have important implications for molecular evolution (Baer *et al.* 2007; Herrick 2011), genome organization (Dame *et al.* 2011; Sobetzko *et al.* 2012; Dorman 2013), and cellular disease (Schuster-Böckler and Lehner 2012; Donley and Thayer 2013; Lawrence *et al.* 2013; Liu *et al.* 2013), but incorporating calibrated bpsm rates into evolutionary models requires that we understand the true patterns of non-uniform bpsm rates and the genomic features that govern them. Remarkably, there appears to be a relatively conserved pattern of bpsm rates in the genomes of single-chromosome bacteria that are MMR-deficient, where bpsm rates vary in a bimodal wave that is mirrored on each replichore (Foster *et al.* 2013; Long *et al.* 2014; Dettman *et al.* 2016). I find MMR-deficient multi-chromosome bacteria display similar patterns of bpsm rates on chr1 (Figure 2A, B), and although I cannot reject the null hypothesis that bpsm rates are uniform chr2, the patterns of bpsm rates on chr2 appear to map to those of concurrently late replicating regions on chr1 (Figure 4A, B). Elucidating the true patterns of bpsm in wild-type bacteria will require a larger collection of bpsms, as I was unable to rule out uniform bpsm rates in any of the three MMR-proficient MA experiments.

In principle, concurrently replicated genome regions within and between chromosomes may experience similar bpsm rates for a number of reasons. First, nucleotide context can generate heterogeneous bpsm rates because certain nucleotides or nucleotide sequences are more prone to incur bpsms than others (Baer *et al.* 2007; Dettman *et al.* 2014; Long *et al.* 2014; Sung *et al.* 2015). However, there is only minimal variation in the nucleotide composition between the intervals analyzed in this study and I observe nearly all G:C>A:T and A:T>G:C transitions, occurring at statistically indistinguishable rates, in both the *Vf*-mut and *Vc*-mut experiments. Therefore, it is unlikely that nucleotide content generates substantial variation in the bpsm rates of the different genome intervals analyzed in this study. Second, the replication machinery itself may generate heterogeneous bpsm rates because of the use error prone polymerases (Courcelle 2009), error prone repair pathways (Lang and Murray 2011), or nucleotide pool inconsistency late in the replication cycle (Zhang and Mathews 1995; Cooper *et al.* 2010). These mechanisms have all been invoked to explain why substitution rates scale positively with replication timing (Mira and Ochman 2002; Stamatoyannopoulos *et al.* 2009; Chen *et al.* 2010; Cooper *et al.* 2010; Flynn *et al.* 2010; Mugal *et al.* 2010; Herrick 2011; Lang and Murray 2011; Agier and Fischer 2012; Martincorena *et al.* 2012), but it is difficult to imagine how they might create the mirrored wave-like patterns of bpsm rates observed in bacterial chromosomes. Lastly, a number of genomic features that are indirectly related to replication, like compaction of the bacterial nucleoid, binding of NAPs, and transcription may also generate heterogeneous bpsm rates (Schmidt *et al.* 2006; Dame *et al.* 2011; Sobetzko *et al.* 2012; Warnecke *et al.* 2012; Dorman 2013). Indeed, negative DNA superhelicity does correlate positively with the mirrored wave-like patterns of bpsm rates on opposing replichores of *E. coli* (Foster *et al.* 2013), and patterns of extant sequence variation are impacted by NAPs in a growth phase-specific manner (Warnecke *et al.* 2012). Furthermore, sigma factors, DNA gyrase, and a number of NAPs have mirrored patterns of activity on the right and left replichores in the single chromosome of *E. coli* (Sobetzko *et al.* 2012), possibly resulting from their concurrent replication. Although transcription is also expected to impact bpsm rates through gene expression and replication-transcription conflicts (Martincorena *et al.* 2012; Merrikh *et al.* 2012), oscillations in expression patterns and gene density are not consistent with concurrently replicated regions experiencing similar expression levels (Allen *et al.* 2006), and expression was poorly correlated with the patterns of bpsm rates in *E. coli* (Foster *et al.* 2013). It seems likely that several of these factors contribute to non-uniform mutation rates within genomes, but evidence suggesting that DNA superhelicity and NAP activity are conserved among concurrently replicated regions is a particularly compelling explanation for the global patterns of bpsm rates observed in this study.

Bpsm rates of the *Vf*-wt, *Vc*-wt, and *Bc*-wt MA-WGS experiment did not vary significantly on either chromosome and concurrently replicated regions do not appear to experience similar bpsm rates in these studies. However, the small number bpsms that were obtained in these studies may inhibit me from elucidating the true patterns of bpsm rates across these wild-type genomes. Alternatively, it is possible that even with more comprehensive wild-type bpsm datasets, bpsm rates are truly uniform across the genomes of multi-chromosome bacteria. Evidence already exists that MMR preferentially repairs errors in coding regions (Lee *et al.* 2012; Dillon *et al.* 2015) and is more concentrated at replication forks (Lopez De Saro *et al.* 2006), so it is possible that the patterns of bpsm rates observed in MMR-deficient MA lineages are mitigated by a functional MMR pathway. Clearly, more robust wild-type MA-WGS studies will need to be carried out to elucidate the true patterns of bpsm rates in MMR-proficient bacterial genomes.

One important implication of this work is that both the sign and the strength of the relationship between bpsm rates and replication timing changes depending on the genome location. Specifically, starting at *oriCI* in *Vf*-mut and *Vc*-mut, bpsm rates increase with replication timing until the mirrored bpsm rate peaks at approximately 600 Kb. Following these peak bpsm rates, however, bpsm rates decrease with replication timing until approximately 1100 Kb from *oriCI*. Bpsm rates are relatively stable in late replicating regions of chr1 and chr2, which suggests that replication timing as no affect on bpsm rates (Figure 4A, B). Consequently, studies that focus only on the relationship between replication timing and mutation rates in sub-regions of the genome may not reflect the genome-wide relationship. Although it is becoming relatively straightforward to study genome-wide mutation rates in smaller bacterial genomes, analyzing genome-wide mutation rates in the larger genomes of higher organisms remains a substantial challenge. Therefore, until these genomes can be sequenced and assembled in a high-throughput and cost-effective manner, reporter studies that analyze the relationship between replication timing and mutation rates at a number of sites will be most informative.

An enhanced understanding on the patterns of bpsm rates also has important implications for models of molecular evolution (Baer *et al.* 2007; Herrick 2011), genome organization (Dame *et al.* 2011; Sobetzko *et al.* 2012; Dorman 2013), and cancer biology (Schuster-Böckler and Lehner 2012; Donley and Thayer 2013; Lawrence *et al.* 2013; Liu *et al.* 2013). First, traditional models of molecular evolution used to measure evolutionary rates and generate molecular clocks often assume uniform mutation rates. Yet even orthologous genes may be evolving at variable rates in different lineages because they are located in genome regions exposed to different mutation rates (Ochman 2003; Baer *et al.* 2007; Herrick 2011). These models could be improved by a better understanding of how and why bpsm rates vary across genomes. Second, bacterial genomes are highly organized entities, and evidence now exists that one of the most conserved properties of this genome organization is their relative distance from the origin of replication and replication terminus (Sobetzko *et al.* 2012). Although gene regulation by various NAPs has been invoked to explain this spatial organization, heterogeneous mutation rates may also be key contributors to genome organization. Ideally, housekeeping genes would be organized in regions with low mutation rates while accessory genes suffer the burden of locations where mutation rate is elevated (Taxis *et al.* 2005; Donley and Thayer 2013; Dorman 2013). This may be particularly important in bacteria, where genes are often acquired horizontally and their evolutionary success can depend on whether they are inserted into regions of high or low expression and mutation rate (Dorman 2013). Third, if genomes are organized in such a way that optimizes gene expression and mutational burden, molecular modifications that change genome wide-patterns of expression and mutation rates may have dramatic consequences. Interestingly, alteration of the replication timing program appears to be a very early step in carcinogenesis and a number of other disease states (Donley and Thayer 2013). These alterations likely affect gene regulation and may lead to increased genomic instability in essential genes. A better understanding of the genomic features that govern patterns of genome-wide mutation rates in healthy cells may thus be a useful tool for forecasting disease risk.

We have shown that in MMR-deficient lineages of bacteria with multiple chromosomes, bpsm rates are non-uniformly distributed on chr1, varying in a mirrored wave-like pattern extending bi-directionally from the origin of replication. In contrast, bpsm rates on chr2 are more constant, like those of late replicating regions on chr1. These observations suggest that concurrently replicated regions of bacterial genomes experience similar bpsm rates prior to MMR, which could be governed by a number of temporally regulated cellular processes. Differentiating the relative impact of these cellular processes on bpsm rates and identifying whether these patterns also apply to wild-type genomes will undoubtedly shed new light on the nature of intra-genome variation in bpsm rates. However, the generality of the mirrored wave-like pattern in bpsm rates in all MMR-deficient MA-WGS experiments to date (Foster *et al.* 2013; Long *et al.* 2014; Dettman *et al.* 2016) suggests that the underlying biases of bpsm rates prior to MMR are highly conserved in bacteria and that concurrently replicated regions experience similar bpsm rates.

# CHAPTER IV

# DISTRIBUTION OF FITNESS EFFECTS OF SPONTANEOUS MUTATIONS IN *BURKHOLDERIA CENOCEPACIA*

## INTRODUCTION

The extent to which spontaneous mutations contribute to evolutionary change largely depends on their rates and fitness effects. Both parameters are fundamental to several evolutionary problems, including the preservation of genetic variation (Charlesworth *et al.* 1993, 2009; Charlesworth and Charlesworth 1998), the evolution of recombination (Muller 1964; Kondrashov 1988; Otto and Lenormand 2002; Roze and Blanckaert 2014), the evolution of mutator alleles (Sniegowski *et al.* 1997; Tenaillon *et al.* 1999), and deleterious mutation accumulation in small populations (Lande 1994; Lynch *et al.* 1995, 1999; Schwander and Crespi 2009). However, while a number of studies have now obtained direct and robust estimates of mutation rates and spectra across diverse organisms, our understanding of the distribution of fitness effects of spontaneous mutations remains limited to mostly indirect estimates in classic model organisms (Eyre-Walker and Keightley 2007).

Mutation accumulation (MA) experiments offer an exceptional opportunity to perform detailed analyses of the fitness effects of spontaneous mutations that have not been exposed to the sieve of natural selection. Specifically, MA experiments limit the efficiency of natural selection by passaging replicate lineages through repeated single cell bottlenecks. These lineages accumulate mutations independently over several thousand generations, and the magnitude and variance in fitness between lineages can be used to estimate several properties of the distribution of fitness effects (Halligan and Keightley 2009). MA studies have been used to characterize the distribution of fitness effects in *Drosophila melanogaster* (Bateman 1959; Mukai 1964a; Keightley 1994; Fry *et al.* 1999), *Arabidopsis thaliana* (Schultz *et al.* 1999; Shaw *et al.* 2000, 2002), *Caenorhabditis elegans* (Keightley and Caballero 1997; Vassilieva *et al.* 2000; Estes *et al.* 2004; Katju *et al.* 2015), *Saccharomyces cerevisiae* (Wloch *et al.* 2001; Zeyl and de Visser 2001; Dickinson 2008; Jasmin and Lenormand 2015), *Escherichia coli* (Kibota and Lynch 1996; Trindade *et al.* 2010), and other microbes (Heilbron *et al.* 2014; Kraemer *et al.* 2015). These studies have at times been inconsistent, but the majority of results suggest that most spontaneous mutations have mild effects (Eyre-Walker and Keightley 2007; Halligan and Keightley 2009; Agrawal and Whitlock 2012; Heilbron *et al.* 2014), that deleterious mutations far outnumber beneficial mutations (Keightley and Lynch 2003; Silander *et al.* 2007; Eyre-Walker and Keightley 2007), and that the distribution of effects of deleterious mutations is complex and multimodal (Zeyl and de Visser 2001; Eyre-Walker and Keightley 2007).

A more powerful approach for studying the distribution of fitness effects of spontaneous mutations is to pair MA experiments with whole-genome sequencing (MA-WGS), so that both the genetic basis and fitness effects of a collection of mutations can be known. MA-WGS studies have been conducted in a diverse array of bacteria, generating a growing database of naturally accumulated mutations that has dramatically improved estimates of mutation rates and spectra (Lee *et al.* 2012; Sung *et al.* 2012a, 2015; Heilbron *et al.* 2014; Long *et al.* 2014, 2015; Dillon *et al.* 2015; Foster *et al.* 2015; Dettman *et al.* 2016). Yet, only one of these studies has also characterized the fitness of MA-WGS lines (Heilbron *et al.* 2014), and that study was conducted with mutator lineages, which are known to have altered base-substitution and indel biases (Lee *et al.* 2012; Sung *et al.* 2015). Our understanding of the distribution of fitness effects of spontaneous mutations would benefit greatly from more direct estimates of fitness derived from MA lineages that harbor known mutational load.

Here, I measured the relative fitness of forty-three fully sequenced MA lineages derived from *B. cenocepacia* HI2424 in three laboratory environments after they had been evolved in the near absence of natural selection for 5554 generations. Following the MA experiment, each lineage harbored a total mutational load of between two and fourteen spontaneous mutations, including base-substitution mutations (bpsms), insertion-deletion mutations (indels), and whole-plasmid deletions. By correlating the relative fitness of these MA lineages with the particular mutations that they harbor, I present a comprehensive picture of the fitness effects of spontaneous mutations, and precise estimates of deleterious mutation rates and fitness effects in *B. cenocepacia*.

## MATERIALS AND METHODS

**Bacterial strains and culture conditions.** All MA experiments were founded from a single colony of *Burkholderia cenocepacia* HI2424, which was isolated from the soil and only passaged in the laboratory during isolation (Coenye and LiPuma 2003). As a member of the diverse *B. cepacia* complex, *Burkholderia cenocepacia* can form highly resistant biofilms and has been associated with persistent lung infections in patients with cystic fibrosis (Mahenthiralingam *et al.* 2005; Traverse *et al.* 2013). The genome of *B. cenocepacia* HI2424 has been fully sequenced and is composed of three chromosomes (Chr1: 3.48-Mb, 3253 genes; Chr2: 3.00-Mb, 2709 genes; Chr3: 1.06-Mb, 929 genes) and a plasmid (0.164-Mb, 157 genes), though the third chromosome can be eliminated under some conditions (Agnoli *et al.* 2012). To facilitate relative fitness assays, I competed all *B. cenocepacia* HI2424 strains derived from the MA experiments with a *B. cenocepacia* HI2424 Lac+ strain, which is isogenic to *B. cenocepacia* HI2424, except for the introduction of the *lacZ* gene at the attTn7 site, which causes colonies to turn blue when exposed to 5-bromo-4-chloro-indolyl-β-galactopyranoside (X-gal) (Choi *et al.* 2005).

MA experiments were conducted on tryptic soy agar plates (TSA) (30 g/liter tryptic soy broth powder, 15 g/liter agar) and were incubated at 37°. At the conclusion of the MA experiment, frozen stocks were prepared by growing a single colony from each lineage overnight in 5ml of tryptic soy broth (TSOY) (30 g/liter tryptic soy broth powder) at 37° and freezing at -80° in 8% DMSO. All relative fitness assays were conducted in 18 x 150mm glass capped tubes with 5ml of liquid medium and were maintained at 37° in a roller drum (30 rpm). Relative fitness of each lineage was assayed in three different environments. First, I conducted relative fitness assays in TSOY, a medium that mimics the conditions of the MA experiment and is expected to be very permissive. Second, I conducted relative fitness assays in M9 Minimal Medium supplemented with 0.3% casamino acids (M9MM+CAA) (3 g/liter casamino acid powder, 1 g/liter glucose, 6 g/liter sodium phosphate dibasic anhydrous, 3 g/liter potassium phosphate monobasic, 1 g/liter ammonium chloride, 0.5 g/liter sodium chloride, 0.1204 g/liter magnesium sulfate, 0.0147 g/liter calcium chloride), a medium that is more nutrient restrictive than TSOY, but contains all essential amino acids except tryptophan. Lastly, I conducted relative fitness assays in M9 Minimal Medium (M9MM) (1 g/liter glucose, 6 g/liter sodium phosphate dibasic anhydrous, 3 g/liter potassium phosphate monobasic, 1 g/liter ammonium chloride, 0.5 g/liter sodium chloride, 0.1204 g/liter magnesium sulfate, 0.0147 g/liter calcium chloride), which is a fully defined medium that is more restrictive than either TSOY or M9MM+CAA. Serial passaging during fitness assays was performed using 100-fold dilutions, so all relative fitness assays were conducted over the same number of generations, despite the moderately different carrying capacities of these mediums. All dilutions were performed using phosphate buffer saline (PBS) (80 g/liter NaCl, 2 g/liter KCl, 14.4 g/liter $Na_2HPO_4 \bullet 2H_2O$, 2.4 g//liter $KH_2PO_4$) in 96-well plates.

**MA-WGS process.** The mutation accumulation experiment that generated the mutational load for this study was reported in Chapter 1. Briefly, seventy-five independent lineages were founded from a single colony of *B. cenocepacia* HI2424 and independently propagated every 24 hours onto a fresh TSA plate for 217 days. Daily generations were estimated monthly by taking a single representative colony from each lineage following 24 hours of growth, placing it in 2 ml of PBS, then serially diluting and spread plating it on TSA. The number of viable cells in each colony was then used to calculate the number of generations elapsed between each transfer, and the average number of generations across all lineages was used as the number of generations per day for that entire month. By multiplying the number of generations per day for each month by the number of days in that month, then summing these totals over the course of the experiment, I calculated the total number of generations elapsed per MA lineage over the course of the MA experiment.

Genomic DNA was extracted from 1 ml of overnight TSB culture founded by forty-seven of the *B. cenocepacia* isolates that were stored at the conclusion of my MA experiment. I used the Wizard Genomic DNA Purification kit for DNA extraction (Promega), all libraries were prepared using a modified Illumina Nextera protocol (Baym *et al.* 2015), and sequencing was performed with the 151-bp paired end platform on the Illumina HiSeq at the Hubbard Center for Genomic Studies at the University of New Hampshire (Chapter 1). Following fastQC analysis, all reads were mapped to the the *B. cenocepacia* HI2424 reference genome (LiPuma *et al.* 2002) with both the Burrows-Wheeler aligner (BWA) (Li and Durbin 2009) and Novoalign (www.novocraft.com). The average depth of coverage across all forty-seven lines was 43x, but only forty-three of these lines were used in this study, and the average depth of coverage across these lines was 46x.

**Spontaneous mutation identification.** All bpsms were identified as described in Chapter 1. Briefly, after using a combination of SAMtools and in house perl scripts to produce all read alignments for each position in each line (Li *et al.* 2009), a three step process was used to detect putative bpsms. First, pooled reads across all lines were used to generate an ancestral consensus base at each site in the reference genome, allowing me to correct differences between the published reference genome and the ancestral colony of this MA experiment. Second, reads from the individual lines were used to generate a lineage specific consensus base at each site in the reference genome for each lineage, as long as the site was covered by at least two forward and two reverse reads, and at least 80% of the reads identified the same base. Sites that did not meet these criteria were not analyzed in the respective lineage. Third, lineage specific consensus bases for each lineage were compared to the ancestral consensus base at each site, and a putative bpsm was identified if they differed. This three step process was carried out independently using both the BWA and Novoalign alignments, and putative bpsms were considered genuine only if both pipelines independently identified the bpsm. Despite these lenient criteria, all of the bpsms that were identified at analyzed sites in this study had considerably greater coverage and consensus than the minimal criteria (Table D.1), demonstrating that they are unlikely to be false positives in low coverage regions. The frequency of sites that were not analyzed in each lineage varied from 0.01 to 0.18. Putative bpsms in these regions were estimated by multiplying the number of unanalyzed sites in each lineage by the overall *B. cenocepacia* bpsm rate calculated in this study (1.31 (0.08) • $10^{-10}$ /bp/generation) and the number of generations of mutation accumulation in each lineage (5554) (Table D.2).

Indels are inherently more difficult to identify than bpsms because gaps and simple sequence repeats (SSRs) reduce the accuracy of short-read alignment algorithms. To overcome these issues, I extracted all putative indels where at least 30% of the reads that covered the site identified the exact same indel (size and motif), as long as the site was covered by at least two forward and two reverse reads. These putative indels were then subject to a series of more strenuous filters based on consensus between the putative indels identified by the BWA alignment, the Novoalign alignment, and PINDEL (Ye *et al.* 2009; Dillon *et al.* 2015). Specifically, all putative indels where more than 80% of the reads identified the exact same indel in both the BWA and Novoalign alignments were considered genuine indels. For putative indels where only 30-80% of the reads identified the exact same indel, I parsed out only reads that had bases covering both the upstream and downstream regions of the indel (if it was not in an SSR), and both the upstream and downstream regions of the SSR (if it was in an SSR). Using this subset of reads, I reassessed the frequency of reads that identified the exact same indel, allowing more accurate identification of indels involving the gain or loss of a single repeat within a SSR. These indels were considered genuine if more than 80% of the parsed reads identified the exact same indel and were discarded if they did not. Putative indels were also extracted using PINDEL, and were considered genuine if they were covered by at least six forward and six reverse reads, and at least 80% of the reads identified the exact same indel (Ye *et al.* 2009). Lastly, I analyzed the distribution of coverage between chromosomes and the 0.164-Mb plasmid to detect any plasmid copy number variants. As with bpsms, putative indels in regions that were not analyzed were estimated by multiplying the number of unanalyzed sites in each lineage by the overall indel rate calculated in this study (2.39 (0.34) • $10^{-11}$ /bp/generation) and the number of generations of MA in each lineage (5554) (Table D.2). All indels identified in this study are also summarized in Table D.1.

**Quantifying relative fitness.** To quantify the selection coefficients of each of the forty-three derived MA lineages, I conducted three-day competitions between each MA lineage and the *B. cenocepacia* HI2424 Lac+ strain. These competitions were carried out independently in TSOY, M9MM+CAA, and M9MM, with four replicates being conducted for each lineage in each environment. The MA ancestral *B. cenocepacia* HI2424 strain was also competed against *B. cenocepacia* HI2424 Lac+ as a control, with four replicates for each environment. Selection coefficients were estimated as described previously, using the relative growth of the focal MA lineage and the *B.*

*cenocepacia* HI2424 Lac+ reference strain, normalized by the number of generations elapsed by the reference strain ($G$) (Chevin 2011; Perfeito *et al.* 2014). First, the difference in growth ($\Delta r_{ab}$) between the two strains was estimated as:

$$\Delta r_{ab} = \ln\left(\frac{N_{fa}}{N_{ia}}\right) - \ln\left(\frac{N_{fb}}{N_{ib}}\right)$$

where $N_{ia}$ and $N_{ib}$ were the initial numbers of test and reference bacteria, respectively, and $N_{fa}$ and $N_{fb}$ were the final numbers of test and reference bacteria, respectively. Selection coefficients ($s_{ab}$) were then calculated as:

$$s_{ab} \;=\; \Delta r_{ab}\;/\;G$$

where $G$, generations elapsed by the reference strain, is equal to:

$$G \;=\; log_2\left(\frac{N_{fb}}{N_{ib}}\right)$$

For each replicate, all forty-three derived MA lineages, *B. cenocepacia* HI2424, and *B. cenocepacia* Lac+ were resurrected from frozen culture by inoculating them into 5 ml of TSOY broth and incubating overnight in a roller drum at 30 rpm. Depending on which environment was being assayed, each strain was then transferred to fresh TSOY, M9MM+CAA, or M9MM via a 10,000-fold dilution and acclimated for 24 hours at 37° and 30 rpm. Following acclimation, forty-four competitions (forty-three MA lineages + control) were generated in the appropriate fresh medium at a 1:1 ratio via 100-fold dilution, and 30 ul from each was extracted to quantify the initial frequency of each competitor ($N_{ia}$, $N_{ib}$). Competitions were then incubated for 72 hours at 37° and 30 rpm, being transferred to fresh media every 24 hours via a 100-fold dilution. At the conclusion of the 72 hour competition, 30 ul of the final culture was extracted to quantify the final frequency of each competitor ($N_{fa}$, $N_{fb}$).

To measure the initial frequency of each competitor, the extracted culture was diluted in PBS and 100 ul of the diluted sample was plated on a TSA + X-Gal plate. Specifically, in the TSOY competitions the samples were diluted to $10^{-4}$ and 1/3, in the M9MM+CAA competitions the samples were diluted to $10^{-4}$ and 1/2, and in the M9MM competitions the samples were diluted to $10^{-4}$. Following a 48 hour incubation, the number of white and blue colonies were quantified and used to calculate $N_{ia}$ and $N_{ib}$, respectively, after accounting for the dilutions. Final frequencies were measured in the same way, except that an additional $10^{-2}$ dilution was required for each competition because the cultures were at carrying capacity. In addition, to calculate $N_{fa}$ and $N_{fb}$, I had to account for the dilutions that were conducted prior to plating and the two 100-fold dilutions that were conducted during the three-day competition. Importantly, the selection coefficient of the *B. cenocepacia* HI2424 MA ancestor was not significantly different from 0 in any of environments (TSOY: *s* = -0.0002 (0.0020), M9MM+CAA: *s* = +0.0075 (0.0030), M9MM: *s* = -0.0016 (0.0034) (SEM)).

**Statistical analysis.** All statistical analyses were performed in R Version 0.98.1091 using the Stats analysis package (R Development Core Team 2013). For independent two-tailed t-tests, all p-values were corrected for multiple comparisons using a Benjamini-Hochberg correction (Table D.3), which ensures that my false positive rate remains below 5%, despite testing whether the selection coefficient differed significantly from 0 for forty-three lineages in each environment (Benjamini and Hochberg 1995). Corrected p-values that were below a threshold of 0.05 were considered significant. Linear regressions were used to evaluate the correlation between the number of mutations in a lineage and its selection coefficient, as well as the correlation between the selection coefficients of lineages in different environments. Lastly, to test for effects of replicate, genotype, environment, and genotype*environment interaction on the fitness of each lineage, I performed an analysis of variance (ANOVA) on the cumulative dataset.

## RESULTS

We previously reported the rate and molecular spectrum of spontaneous mutations in wild-type *B. cenocepacia*, as determined from the cumulative results of a MA-WGS experiment involving forty-seven replicate lineages derived from *B. cenocepacia* HI2424 (Chapter 1). Each lineage was passaged through daily single-cell bottlenecks for 217 days, resulting in 5554 generations of MA per lineage. The average number of generations of growth per day within a colony declined from 26.16 (0.06) to 24.92 (0.07) (SEM) over the course of the 5554 generations of MA, suggesting that some of the accumulated mutations had deleterious fitness effects. Here, I present a detailed picture of the distribution of fitness effects of the spontaneous mutations accumulated during this MA-WGS experiment using forty-three of replicate lineages, as the remaining four lineages were discarded because of a lack of sufficient coverage in the WGS data.

A detailed analysis of the mutations from the forty-three MA-WGS lineages used in this study is consistent with constant mutation rates and limited selection over the course of my MA-WGS experiment. Specifically, neither the distribution of bpsms or indels differed significantly from a Poisson distribution (bpsms: $\chi^2$ = 3.463, p = 0.943;

indels: $\chi^2$ = 0.280, p = 0.964), signifying that mutation rates did not vary across the forty-three MA lineages. Limited purifying selection was supported by he fact that the ratio of synonymous to nonsynonymous bpsms did not differ from the expected ratio based on the codon-usage and %GC content at synonymous and non-synonymous sites in *B. cenocepacia* HI2424 ($\chi^2$ = 0.776, d.f. = 1, p = 0.378), while limited positive selection was supported by the lack genetic parallelism in the bpsm spectra across lineages (Table D.1). Both bpsms and indels were observed more frequently than expected in non-coding DNA (bpsms: $\chi^2$ = 2.194, d.f. = 1, p = 0.139; indels: $\chi^2$ = 45.816, d.f. = 1, $p < 0.0001$), but this pattern could be generated by selection against coding mutations, preferential mismatch repair in coding regions, or the mutation prone nature of repetitive DNA in non-coding regions, so has the potential to be misleading (Lee *et al.* 2012; Heilbron *et al.* 2014; Dillon *et al.* 2015). In any event, I estimate that the threshold selection coefficient below which genetic drift will overpower natural selection, as determined by $N_E \times s = 1$ in haploid organisms, is 0.078 (Chapter 1). Thus, while a small class of adaptive or deleterious mutations with effects in excess of *s* = +/- 0.078 may be subject to the biases of natural selection (Kimura 1983; Elena *et al.* 1998; Zeyl and de Visser 2001; Hall *et al.* 2008), the vast vast majority of mutations that were observed in this study likely fixed irrespective of their fitness effects.

**Genetic basis of spontaneous mutations.** The spontaneous bpsms and indels reported here are similar to those reported previously (Chapter 1), with two exceptions. First, I allowed for bpsms to be called in more than one lineage, resulting in the addition of two bpsms. These bpsms are assumed to have occurred in the ancestral colony, but their presence in each lineage must be documented to accurately quantify the relationship between the fitness of each lineage and the mutations it harbors. Second, I did not analyze four of the lineages from the MA-WGS experiment in *B. cenocepacia* because less than 80% of their genomes had sufficient coverage to be analyzed for the presence of bpsms and indels (see Methods). This low coverage would render me blind to a considerable portion of the mutational load in these lineages, which warranted their exclusion.

In sum, I have identified 233 bpsms, 42 short indels, and 4 plasmid-loss events distributed across the forty-three MA lineages analyzed in this study (Figure 1). The most common class of bpsms were missense bpsms (141), followed by synonymous bpsms (49), intergenic bpsms (37), and nonsense bpsms (6). Among indels, coding indels involving only a single gene (22) were slightly more common than intergenic indels (20), while loss of the 0.16-Mb plasmid, which encodes 157 genes, was observed in 4 lineages. Furthermore, I estimated false negative rates in each lineage as the number of sites that were not analyzed for mutations, multiplied by the product of experiment-wide bpsm and indel rates per base-pair per generation and the number of generations experienced by each lineage. However, because I cover the majority of each genome with sufficient depth to analyze both bpsms and indels, I estimate that an average of only 0.25 (0.03) additional bpsms and 0.05 (0.01) (SEM) additional indels would have been identified per lineage if the entire genome were analyzed (Table D.2). Overall, mutations were not uniformly distributed across the forty-three MA lineages, allowing me to analyze which mutation types are most likely to have fitness effects.

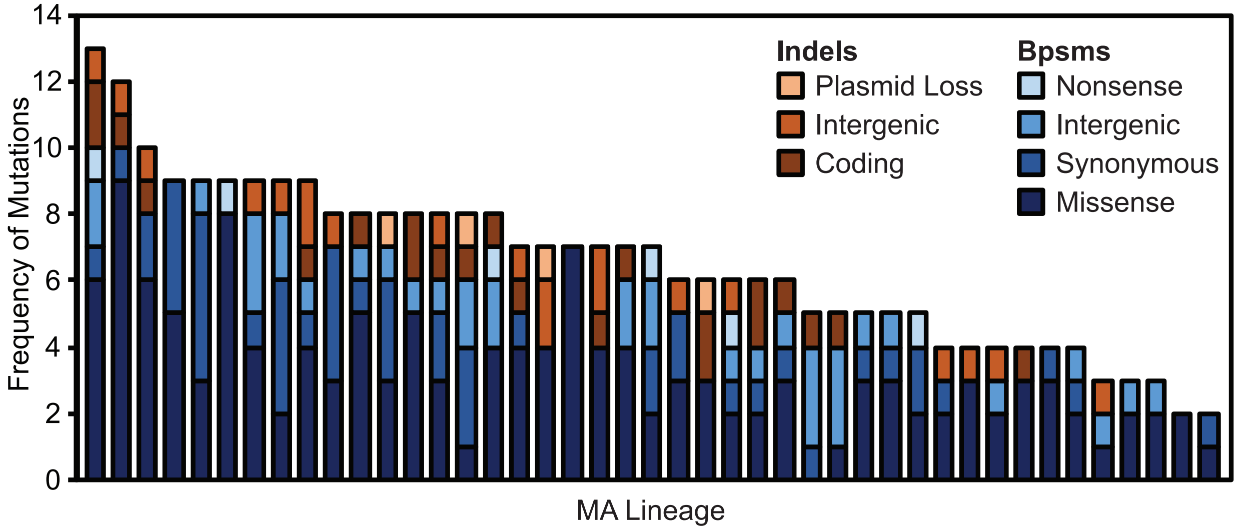

**Figure 1. Distribution of base-substitution mutations (bpsms) and insertion-deletion mutations (indels) across the forty-three *Burkholderia cenocepacia* mutation accumulation lineages analyzed in this study.**

**Distribution of fitness effects. I** measured the fitness effects of mutational load in each lineage following 5554 generations of MA by measuring their selection coefficients relative to the ancestral *B. cenocepacia* HI2424 strain in three different broth culture conditions. TSOY broth is a very permissive medium used to mimic the conditions of the MA experiment, M9MM+CAA is an amino-acid supplemented minimal medium that is less permissive than TSOY but more permissive than a strictly minimal medium, and M9MM is a fully defined minimal medium that is the least permissive of the three environments. In TSOY, 17 lineages had significantly reduced fitness and no lineages had significantly increased fitness (Figure 2A). The average fitness across all MA lineages in TSOY was -0.024 (0.005) (SEM), with a range of -0.111 to +0.037. Similarly, 13 lineages had significantly reduced fitness in M9MM+CAA and none had significantly increased fitness (Figure 2B). The average fitness decline and the range across all MA lineages in M9MM+CAA were also similar to those observed in TSOY (Average: -0.020

(0.005) (SEM); Range: -0.116 to +0.006). Lastly, I observed 13 lineages with significantly reduced fitness in M9MM, but here, 4 other lineages had significantly increased in fitness (Figure 2C). Thus, the average fitness decline of the MA lineages in M9MM was only -0.013 (0.005) (SEM), despite having a similar overall range to TSOY and M9MM+CAA (-0.090 to +0.026). Overall, most MA lineages did not significantly decline in fitness in any of the environments after seven months of evolution under greatly minimized selection, despite accumulating substantial and variable mutational load (Figure 1).

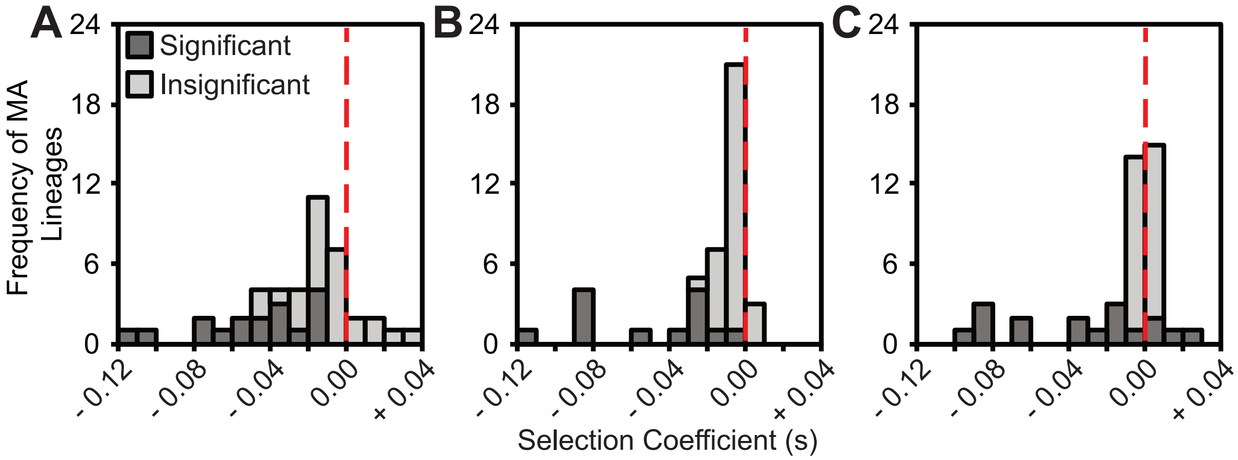

**Figure 2. Distribution of the selection coefficients of each *Burkholderia cenocepacia* MA lineage relative to the ancestral *B. cenocepacia* HI2424 strain in tryptic soy broth (A), M9 minimal medium supplemented with casamino acids (B), and M9 minimal medium (C).** Significance was determined from independent two-tailed t-tests on four replicate fitness assays for each lineage. P-values were corrected for multiple comparisons using a Benjamini-Hochberg correction, and corrected p-values that remained below 0.05 were considered significant.

Significant correlations between the selection coefficients of individual MA lineages across environments suggests that experiment-wide mutational load was not especially pleiotropic (Figure 3). Specifically, linear regressions between the selection coefficients of each lineage in TSOY and both M9MM+CAA and M9MM demonstrated significantly positive relationships (TSOY-M9MM+CAA: F = 17.180, df = 41, p = 0.0002,

$r^2 = 0.2953$; TSOY-M9MM: F = 8.613, df = 41, p = 0.0054, $r^2 = 0.1736$). The relationship between selection coefficients of each lineage in M9MM+CAA and M9MM was also significant and explained a greater fraction of the variance than either of the TSOY regressions (F = 124.00, df = 41, p < 0.0001, $r^2 = 0.7515$), which was expected given that these environments were more similar to each other than either is to TSOY (Figure 3). However, there were examples of MA lineages that declined significantly in fitness in one environment but not others, suggesting that some mutations did have observable pleiotropic fitness effects in this study. A total of nine MA lineages only significantly reduced fitness in a single environment (four TSOY, two M9MM+CAA, three M9MM), eight MA lineages significantly reduced fitness in two of the environments, and six MA lineages significantly reduced fitness in all three environments. An experiment-wide ANOVA revealed significant effects of replicate (df = 3, SS = 0.0037, F = 10.0220, P < 0.0001), genotype (df = 42, SS = 0.3302, F = 63.7212, P < 0.0001), environment (df = 2, SS = 0.0190, F = 76.9015, P < 0.0001), and genotype x environment interaction (df = 84, SS = 0.1166, F = 11.2485, P < 0.0001). These findings highlight the environmental dependence of the fitness effects of several individual spontaneous mutations, despite the fact that the general properties of the distribution of fitness effects in my MA lineages are similar across the three environments that I tested (Figure 2; Figure 3).

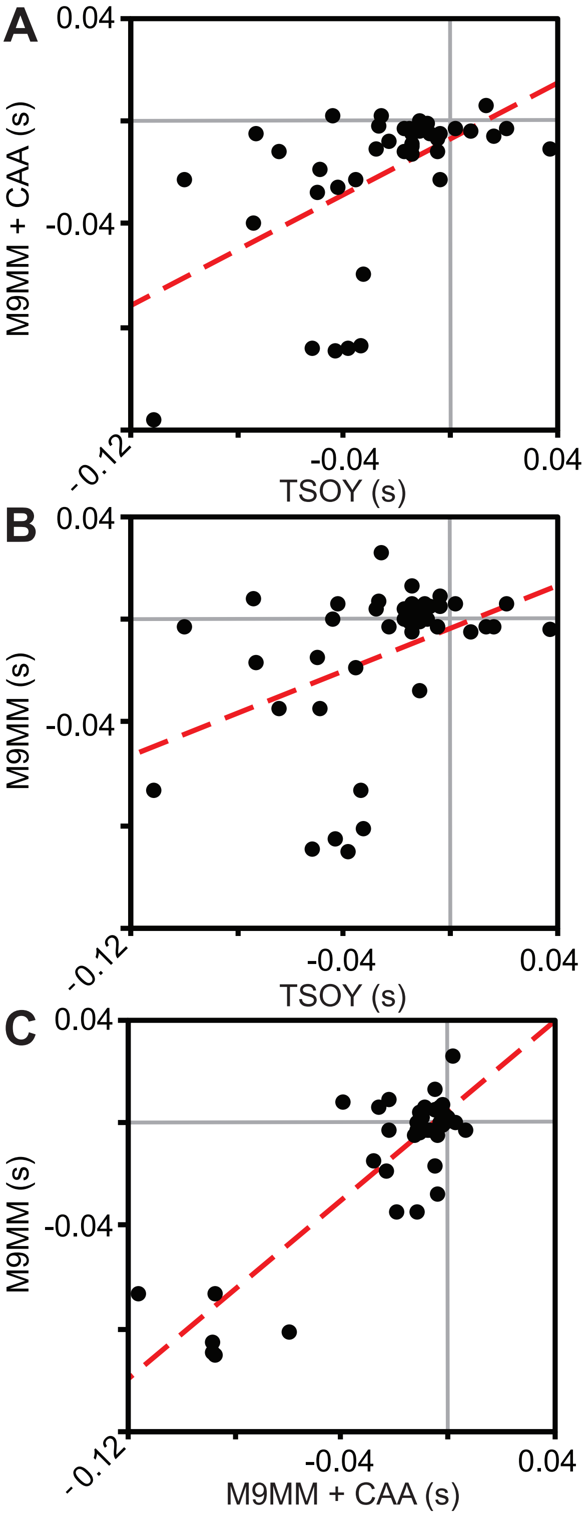

**Figure 3. Relationship between selection coefficients of all *Burkholderia cenocepacia* MA lineages in each of the different pairs of environments.** All linear regressions are significant, but much of the variance is unexplained (A: F = 17.18, df = 41, p = 0.0002, $r^2$ = 0.2953; B: F = 8.613, df = 41, p = 0.0054, $r^2$ = 0.1736; C: F = 124.00, df = 41, p < 0.0001, $r^2$ = 0.7515).

Despite harboring multiple mutations, the majority of MA lineages did not have significantly different fitness from the ancestral reference strain and there was not a significant correlation between the number of spontaneous mutations in a line and their absolute selection coefficients in any environment (TSOY: F = 1.401, df = 41, p = 0.2434, $r^2$ = 0.0330; M9MM+CAA: F = 1.354, df = 41, p = 0.2513, $r^2$ = 0.0320; M9MM: F = 2.957, df = 41, p = 0.0930, $r^2$ = 0.0673) (Figure D.1). After adding the 11 additional mutations presumed to have been missed in the unanalyzed regions across all of my MA lines, I estimate that I accumulated 290 spontaneous mutations, with an average of 6.73 (0.36) (SEM) mutations per lineage. Because each lineage harbored multiple mutations, this suggests that the majority of spontaneous mutations had undetectable fitness effects. Specifically, by dividing the average selection coefficient of each line by the number of mutations that it harbors, I estimate that the average fitness effect (*s*) of a single mutation was -0.0040 (0.0008) in TSOY, -0.0031 (0.0007) in M9MM+CAA, and -0.0017 (0.0007) (SEM) in M9MM.

The lack of significant fitness declines in many lineages harboring multiple mutations and the lack of significant correlation between the number of mutations in a lineage and its fitness suggest that most of the losses and gains in fitness were caused by rare, single spontaneous mutations with significant fitness effects. Therefore, I estimate that only 17/290 mutations significantly affected fitness in TSOY, 13/290 mutations significantly affected fitness in M9MM+CAA, and 17/290 mutations significantly affected fitness in M9MM. Based on these estimates, the deleterious mutation rates ($U_d$) are 7.12 × $10^{-5}$ /genome/generation in TSOY, 5.44 × $10^{-5}$ /genome/generation in M9MM+CAA, and 5.44 × $10^{-5}$ /genome/generation in M9MM.

Furthermore, in analyzing only lineages that experienced significant fitness declines, I estimate that the average effects of significantly deleterious mutations are -0.048 (0.007) in TSOY, -0.053 (0.011) in M9MM+CAA, and -0.048 (0.009) in M9MM (SEM). Although I observe no significantly beneficial mutations in TSOY or M9MM+CAA, my data suggest that the beneficial mutation rate ($U_b$) is $1.68 \times 10^{-5}$ /genome/generation and the average significantly beneficial mutation has a selection coefficient of 0.013 (0.005) in M9MM, assuming that all gains in fitness were driven by a single beneficial mutation.

**Molecular basis of the distribution of fitness effects.** Without sequencing and measuring fitness at intermediate time-points or genetically engineering *B. cenocepacia* HI2424 strains that harbor only single spontaneous mutations, it is difficult to pinpoint which mutations generate the fitness declines in my MA lineages. However, I can examine whether there is any relationship between the forms of mutational load harbored by each lineage and the fitness of those lineages. Specifically, I tested whether any of the mutation types from Figure 1 were overrepresented in lineages that had significantly reduced fitness (Table 1). Interestingly, the only mutation that was significantly overrepresented in lineages with reduced fitness in TSOY was the loss of the 0.164-Mb plasmid ($\chi^2$ = 6.118, d.f. = 1, p = 0.013). All four of the MA lineages that had lost the plasmid had significantly reduced fitness in TSOY, resulting in an average selection coefficient of -0.060 (0.007) (SEM). The four MA lineages that had lost the 0.164-Mb plasmid were also significantly deleterious in M9MM ($\chi^2$ = 9.231, d.f. = 1, p = 0.002), with an average selection coefficient of -0.043 (0.018) (SEM), but the deleterious effects of plasmid loss appear to be mitigated in M9MM+CAA, where only one of the lineages that had lost the plasmid had significantly reduced fitness.

Although no other mutation types were significantly over-represented among lineages that had reduced fitness, it is worth noting that there were more coding indels, nonsense bpsms, and missense bpsms than expected in the lineages with reduced fitness for all three environments, with the exception of coding indels in M9MM+CAA (Table 1). In contrast, intergenic bpsms and indels appear to be evenly distributed between lineages with significantly reduced fitness and those where *s* was not significantly different from 0, suggesting that few if any intergenic mutations from this study have deleterious fitness effects. Similarly, the synonymous bpsms observed in this study do not appear to have deleterious effects, as they are observed less than expected in lineages with reduced fitness in TSOY and are evenly distributed between neutral and reduced fitness lineages in M9MM+CAA and M9MM (Table 1). Overall, these results support the notion that coding indels, nonsense bpsms, and missense mutations are more likely to have deleterious effects than intergenic and synonymous mutations.

**Table 1. Chi-test statistics comparing the observed number of mutations in lineages with significantly reduced fitness to the expected number of mutations in lineages with significantly reduced fitness, as determined by the proportion of total lineages that had significantly reduced fitness.**

| Mutation Type | Environment | Observed | Expected | $\chi^2$ | df | p |
|---|---|---|---|---|---|---|
| Intergenic Bpsm | TSOY | 13 | 14.63 | 0.2996 | 1 | 0.5841 |
| Synonymous Bpsm | TSOY | 16 | 19.37 | 0.9708 | 1 | 0.3245 |
| Missense Bpsm | TSOY | 60 | 55.74 | 0.5374 | 1 | 0.4635 |
| Nonsense Bpsm | TSOY | 4 | 2.37 | 1.8477 | 1 | 0.1741 |
| Intergenic Indel | TSOY | 8 | 7.91 | 0.0018 | 1 | 0.9661 |
| Coding Indel | TSOY | 12 | 8.70 | 2.0736 | 1 | 0.1499 |
| Plasmid Loss | TSOY | 4 | 1.58 | 6.1176 | 1 | 0.0134 |
| Total Mutations | TSOY | 117 | 110.30 | 0.6726 | 1 | 0.4121 |
| Intergenic Bpsm | M9MM+CAA | 12 | 11.19 | 0.0849 | 1 | 0.7708 |
| Synonymous Bpsm | M9MM+CAA | 15 | 14.81 | 0.0033 | 1 | 0.9539 |
| Missense Bpsm | M9MM+CAA | 47 | 42.63 | 0.6427 | 1 | 0.4227 |
| Nonsense Bpsm | M9MM+CAA | 3 | 1.81 | 1.1115 | 1 | 0.2917 |
| Intergenic Indel | M9MM+CAA | 6 | 6.05 | 0.0005 | 1 | 0.9819 |
| Coding Indel | M9MM+CAA | 6 | 6.65 | 0.0914 | 1 | 0.7624 |
| Plasmid Loss | M9MM+CAA | 1 | 1.21 | 0.0519 | 1 | 0.8198 |
| Total Mutations | M9MM+CAA | 90 | 84.35 | 0.5427 | 1 | 0.4613 |
| Intergenic Bpsm | M9MM | 10 | 11.19 | 0.1803 | 1 | 0.6712 |
| Synonymous Bpsm | M9MM | 15 | 14.81 | 0.0033 | 1 | 0.9539 |
| Missense Bpsm | M9MM | 50 | 42.63 | 1.8274 | 1 | 0.1764 |
| Nonsense Bpsm | M9MM | 2 | 1.81 | 0.0274 | 1 | 0.8686 |
| Intergenic Indel | M9MM | 6 | 6.05 | 0.0005 | 1 | 0.9819 |
| Coding Indel | M9MM | 8 | 6.65 | 0.3921 | 1 | 0.5312 |
| Plasmid Loss | M9MM | 4 | 1.21 | 9.2308 | 1 | 0.0024 |
| Total Mutations | M9MM | 95 | 84.35 | 1.9278 | 1 | 0.1650 |

## DISCUSSION

Combining high-throughput fitness measurements with MA-WGS experiments can dramatically advance our understanding of the distribution of fitness effects of spontaneous mutations in diverse organisms. My study combined the results of a MA-WGS experiment in *B. cenocepacia* with fitness measurements in three environments, allowing me to quantify the impacts of different forms of mutational load on fitness. I find that many lineages in TSOY, M9MM+CAA, and M9MM did not have selection coefficients that were significantly different from $s = 0$ and that most lineages that did had reduced fitness. Given that each lineage harbors between two and fourteen spontaneous mutations, the most likely explanation for this observation is that the vast majority of spontaneous mutations have minimal effects on fitness across this range of environments. Under the assumption that the significant reductions in lineage fitness were driven mostly by single deleterious mutations (Davies *et al.* 1999; Heilbron *et al.* 2014), I also obtain new estimates of the average effect of spontaneous mutations ($s$), the deleterious mutation rate ($U_D$), and the average effect of deleterious mutations ($s_D$) in all three environments, which suggests that the general features of the distribution of fitness effects do not differ significantly between these conditions. I also provide evidence that loss of the 0.164-Mb plasmid consistently reduces fitness in TSOY and M9MM but not M9MM+CAA, while nonsense bpsms, missense bpsms, and coding indels are more likely to have contributed to the deleterious mutational load than synonymous bpsms, intergenic bpsms, and intergenic indels.

Although a few select studies have claimed that a substantial fraction of spontaneous mutations are beneficial under certain conditions (Shaw *et al.* 2002; Silander *et al.* 2007; Dickinson 2008), evidence from diverse sources strongly suggests that the effect of most spontaneous mutations is to reduce fitness (Kibota and Lynch 1996; Keightley and Caballero 1997; Fry *et al.* 1999; Vassilieva *et al.* 2000; Wloch *et al.* 2001; Zeyl and de Visser 2001; Keightley and Lynch 2003; Trindade *et al.* 2010; Heilbron *et al.* 2014). My data on the selection coefficients of forty-three MA lineages in TSOY supports the notion that the majority of spontaneous mutations are neutral or deleterious. Specifically, among lineages whose selection coefficients are significantly different from 0 in TSOY, all of them are negative, with selection coefficients ranging from *s* = -0.112 to *s* = -0.014. Selection coefficients in TSOY across these lineages also do not appear to have a clear mode, but whether this is the result of a complex and multimodal distribution of deleterious mutations (Zeyl and de Visser 2001; Eyre-Walker and Keightley 2007) or a lack of ability to detect deleterious mutations with especially small and/or large effects is uncertain. Specifically, peaks in the distribution of deleterious mutations may exist outside the detection range of this study, either because deleterious effects in excess of *s* = -0.078 were exposed to the sieve of natural selection or because I lacked the statistical power to distinguish a number of small deleterious selection coefficients from neutrality. Among lineages whose selection coefficients are not significantly different from 0, most are clearly negative (Chi-square test; $\chi^2$ = 7.54, df = 1, p = 0.0060) (Figure 2), which suggests that at least some of these lineages do harbor moderately deleterious mutations and there may be a peak in the distribution of deleterious mutations where mutations have very small deleterious effects.

Whether the environment affects the distribution of fitness effects of spontaneous mutations has also been the subject of considerable debate. Specifically, some studies have shown that larger declines in fitness are experienced in harsher environments, while others have not (Martin and Lenormand 2006; Halligan and Keightley 2009; Kraemer *et al.* 2015). In M9MM+CAA and M9MM, I was able to statistically distinguish selection coefficients from *s* = 0 with greater precision (*s* < -0.03 in M9MM+CAA and *s* < -0.01 or *s*>0.01 in M9MM) because the formulations for these mediums are more defined than TSOY. These mediums are also expected to be harsher than TSOY because nutrients are more limited, but I found a similar distribution of effects among the selection coefficients of lineages in these environments (Figure 2). Specifically, most lineages that were not neutral had reduced fitness and there was no clear mode in the distribution of deleterious effects. However, in M9MM there are four lineages that have significantly increased fitness and there are as many lineages whose selection coefficients are non-significantly greater than 0 as there are lineage whose selection coefficients are non-significantly less than 0 (Figure 2). This suggests that fewer spontaneous mutations have deleterious effects on fitness in M9MM, possibly because a greater proportion of genes are unused when metabolizing only a single carbon substrate. Overall, these data support the notion that the environment can impact the fitness effects of some individual spontaneous mutations, despite the fact that the overall distribution of fitness effects is similar between the three environments assayed in this study.

Although I cannot fully discredit the possibility that lineages that did not experience significant declines in fitness simply contain both beneficial and deleterious mutations that cancel each other out, the most parsimonious explanation for these distributions is that most spontaneous mutations had very minimal affects on fitness, and a few, rare, large-effect mutations drove the significant fitness declines in some lineages (Davies *et al.* 1999; Heilbron *et al.* 2014). By dividing the selection coefficient in each lineage by the number of mutations that it harbors, I estimate that the mean fitness effect of mutations observed in this study was less than $s = 0.01$ in all three environments, and that the vast majority of mutations had near neutral affects on fitness. These estimates are remarkably similar to estimates from prior MA studies that harbored fully characterized mutational load in *P. aeruginosa* and *S. cereviseae* (Lynch *et al.* 2008; Heilbron *et al.* 2014), but are lower than estimates derived from unsequenced MA lineages, where the number and type of mutations is unknown (Halligan and Keightley 2009; Trindade *et al.* 2010). However, while this suggests that the vast majority of spontaneous mutations in *B. cenocepacia* have very low selection coefficients in the laboratory, it should not imply that all of these mutations are effectively neutral in natural conditions. In fact, sequence analyses in enteric bacteria have revealed that fewer than 2.8% of amino-acid changing mutations are evolving neutrally, and this may be an overestimate due to the presence of adaptive mutations (Charlesworth and Eyre-Walker 2006; Eyre-Walker and Keightley 2007).

Considering only lineages that have significantly reduced selection coefficients, I also estimated the deleterious mutation rates ($U_D$) and the mean fitness effects of deleterious mutations ($s_D$) in each environment. However, it is important to acknowledge three essential caveats to these estimates. First, I assume that only a single deleterious mutation per lineage contributed to the selection coefficient and that its effects are independent of the other mutations in the MA lineage. My data and a prior studies support that rare large-effect mutations will disproportionately drive the fitness declines in MA lineages, even when they harbor hundreds of mutations (Davies *et al.* 1999; Heilbron *et al.* 2014). The possibility of pervasive epistasis between spontaneous deleterious mutations does exist (Mukai 1964b; Dickinson 2008; Schaack *et al.* 2013), but a recent study in yeast showed that synergistic epistasis need not be invoked to explain accelerated fitness decline in MA experiments (Jasmin and Lenormand 2015).

Second, my inability to distinguish the fitness effects of a subset of spontaneous deleterious mutations from $s$ = 0 will generate a slight downward biased in my estimates of $U_D$ and $s_D$ because selection coefficients with small magnitudes are excluded. Third, highly deleterious and lethal mutations that had an s < -0.078 were subject to the biases of natural selection in my MA experiment, which will generate a slight downward bias in my estimates of $U_D$ and a slight upward bias in my estimates of $s_D$, because selection coefficients with high magnitudes are more likely to be purged by natural selection.

In spite of these potential biases, my estimates of $U_D$ and $s_D$ in all three environments are similar to prior estimates in *E. coli* (Kibota and Lynch 1996; Trindade *et al.* 2010) in all environments (TSOY: $U_D = 7.12 \times 10^{-5}$ /genome/generation, $s_D$ = -0.048; M9MM+CAA: $U_D = 5.44 \times 10^{-5}$, $s_D$ = -0.053; M9MM: $U_D = 5.44 \times 10^{-5}$, $s_D$ = -0.048). However, it is notable that my $U_D$ estimates are all slightly lower than the estimates from the previous studies and my $s_D$ estimates are of greater absolute magnitude, which is consistent with a failure to differentiate some moderately deleterious mutations from $s$ = 0. Significantly beneficial mutations were only observed in M9MM and I did not observe any lineages where $s$ was greater than +0.03. Although these limited observations prevent me from performing any detailed analyses on the rate and effects of beneficial mutations, they support the notion that beneficial mutations are rare relative to deleterious mutations (Keightley and Lynch 2003). Furthermore, they suggest that the majority of beneficial mutations likely provide moderate benefits, even though the beneficial mutations that often fix in experimental populations can have large beneficial effects (Lenski *et al.* 1991; Ostrowski *et al.* 2005; Lang *et al.* 2013b; Levy *et al.* 2015).

It is a well-established dogma in evolutionary biology that mutations that disrupt coding sequences are most likely to have fitness effects, but this has never been quantitatively tested with naturally accumulated mutations. Specifically, mutations that frequently generate non-functional proteins, like nonsense bpsms or coding indels, are expected to have the most deleterious effects, followed by missense bpsms that mostly generate modified proteins, then synonymous and non-coding mutations that do not alter protein sequences. The fitness effects of plasmid gain and loss are less certain, as the size and genetic contents of plasmids vary, but they may be energetically expensive to maintain (Smith and Bidochka 1998). Consequently, plasmids may be selectively lost in permissive laboratory environments where maintenance of the plasmid has a fitness cost (Lenski and Bouma 1987; Smith and Bidochka 1998). My data suggest that although loss of the 0.164-Mb plasmid in *B. cenocepacia* occurs at an appreciable rate in the absence of selection during my MA experiments, it is universally deleterious to lose the plasmid in TSOY and M9MM. However, these effects appear to be mitigated in M9MM+CAA, suggesting that these fitness loses are related to amino acid synthesis. Overall, these data suggest that in permissive laboratory conditions, the loss of some plasmids can be deleterious and not just beneficially more efficient as is widely presumed. Other mutation types were not significantly overrepresented in lineages with significantly reduced fitness (Table 1), but I do find that there are slightly more nonsense bpsms, missense bpsms, and coding indels than expected in lineages with significantly reduced fitness. This supports the assertion that these protein modifying mutations are more likely to affect fitness than synonymous or intergenic mutations, and that most synonymous and intergenic mutations do not measurably affect fitness, even though they can be under selective constraints (Eyre-Walker and Keightley 2007; Bailey *et al.* 2014).

The rate and distribution of fitness effects of spontaneous mutations are fundamental evolutionary quantities that will help explain a number of evolutionary phenomena, including the preservation of genetic variation (Charlesworth *et al.* 1993, 2009; Charlesworth and Charlesworth 1998), the evolution of recombination (Muller 1964; Kondrashov 1988; Otto and Lenormand 2002; Roze and Blanckaert 2014), the evolution of mutator alleles (Sniegowski *et al.* 1997; Tenaillon *et al.* 1999), and the mutational meltdown of small populations (Lande 1994; Lynch *et al.* 1995, 1999; Schwander and Crespi 2009). By measuring the fitness effects of MA lineages with fully characterized mutational load, I provide a uniquely systematic study of the rate and fitness effects of naturally accumulated mutations with known genetic bases, demonstrating that the vast majority of spontaneous mutations accumulated in *B. cenocepacia* MA lines are neutral or deleterious for fitness, and that the fitness of individual mutations can be environmentally dependent, even though the general features of the distribution of fitness effects are similar in different environments. In concert with data from several other species (Kibota and Lynch 1996; Keightley and Caballero 1997; Fry *et al.* 1999; Vassilieva *et al.* 2000; Wloch *et al.* 2001; Zeyl and de Visser 2001; Keightley and Lynch 2003; Trindade *et al.* 2010; Heilbron *et al.* 2014) and my own preliminary analyses of the fitness of the *Vibrio fischeri* and *Vibrio cholerae* MA-WGS from Chapter 2 (Figure D.2; Figure D.3), the deleterious nature of spontaneous mutations suggests a bleak outlook for asexual species at small population sizes (Sniegowski and Lenski 1995; Lynch *et al.* 1999). However, because the nature of a number of deleterious mutations may depend on the environment or the genetic background, we must continue to consider the effects of spontaneous mutations across multiple environments to understand their true nature.

# LIST OF REFERENCES

Agier N., Fischer G., 2012 The mutational profile of the yeast genome is shaped by replication. Mol. Biol. Evol. **29**: 905–913.

Agnoli K., Schwager S., Uehlinger S., Vergunst A., Viteri D. F., Nguyen D. T., Sokol P. A., Carlier A., Eberl L., 2012 Exposing the third chromosome of *Burkholderia cepacia* complex strains as a virulence plasmid. Mol. Microbiol. **83**: 362–378.

Agrawal A. F., Whitlock M. C., 2012 Mutation load: The fitness of individuals in populations where deleterious alleles are abundant. Annu. Rev. Ecol. Evol. Syst. **43**: 115–135.

Alexander M. P., Begins K. J., Crall W. C., Holmes M. P., Lippert M. J., 2013 High levels of transcription stimulate transversions at GC base pairs in yeast. Environ. Mol. Mutagen. **54**: 44–53.

Allen T. E., Price N. D., Joyce A. R., Palsson B., 2006 Long-range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organization. PLoS Comput. Biol. **2**: e2.

Andrews S., 2010 FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Baek J. H., Chattoraj D. K., 2014 Chromosome I controls chromosome II replication in *Vibrio cholerae*. PLoS Genet. **10**: e1004184.

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

Bailey S. F., Hinz A., Kassen R., 2014 Adaptive synonymous mutations in an experimentally evolved *Pseudomonas fluorescens* population. Nat. Commun. **5**: 4076.

Baldwin A., Mahenthiralingam E., Thickett K. M., Honeybourne D., Maiden M. C. J., Govan J. R., Speert D. P., Lipuma J. J., Vandamme P., Dowson C. G., 2005 Multilocus sequence typing scheme that provides both species and strain differentiation for the *Burkholderia cepacia* complex. J. Clin. Microbiol. **43**: 4665–4673.

Bateman A. J., 1959 The viability of near-normal irradiated chromosomes. Int. J. Radiat. Biol. **1**: 170–180.

Baym M., Kryazhimskiy S., Lieberman T. D., Chung H., Desai M. M., Kishony R., 2015 Inexpensive multiplexed library preparation for megabase-sized genomes. PLoS One **10**: e0128036.

Beaulaurier J., Zhang X., Zhu S., Sebra R., Rosenbluh C., Deikus G., Shen N., Munera D., Waldor M. K., Chess A., Blaser M. J., Schadt E. E., Fang G., 2015 Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. Nat. Commun. **6**: 7438.

Belkum A. van, Scherer S., Alphen L. van, Verbrugh H., 1998 Short-sequence DNA repeats in prokaryotic genomes. Microbiol. Mol. Biol. Rev. **62**: 275–293.

Benjamini Y., Hochberg Y., 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. **57**: 289–300.

Charlesworth B., Morgan M. T., Charlesworth D., 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134**: 1289–1303.

Charlesworth B., Charlesworth D., 1998 Some evolutionary consequences of deleterious mutations. Genetica **102-103**: 3–19.

Charlesworth J., Eyre-Walker A., 2006 The rate of adaptive evolution in enteric bacteria. Mol. Biol. Evol. **23**: 1348–1356.

Charlesworth B., Betancourt A. J., Kaiser V. B., Gordo I., 2009 Genetic recombination and molecular evolution. In: *Cold Spring Harbor Symposia on Quantitative Biology*,, pp. 177–186.

Chen C.-L., Rappailles A., Duquenne L., Huvet M., Guilbaud G., Farinelli L., Audit B., D'Aubenton-Carafa Y., Arneodo A., Hyrien O., Thermes C., 2010 Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res. **20**: 447–457.

Chevin L.-M., 2011 On measuring selection in experimental evolution. Biol. Lett. **7**: 210–213.

Chin C.-S., Alexander D. H., Marks P., Klammer A. A., Drake J., Heiner C., Clum A., Copeland A., Huddleston J., Eichler E. E., Turner S. W., Korlach J., 2013 Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods **10**: 563–569.

Choi K.-H., Gaynor J. B., White K. G., Lopez C., Bosio C. M., Karkhoff-Schweizer R. R., Schweizer H. P., 2005 A Tn7-based broad-range bacterial cloning and expression system. Nat. Methods **2**: 443–448.

Coenye T., LiPuma J. J., 2003 Population structure analysis of *Burkholderia cepacia* genomovar III: varying degrees of genetic recombination characterize major clonal complexes. Microbiology-Sgm **149**: 77–88.

Coenye T., Spilker T., Schoor A. Van, LiPuma J. J., Vandamme P., 2004 Recovery of *Burkholderia cenocepacia* strain PHDC from cystic fibrosis patients in Europe. Thorax **59**: 952–954.

Cooper V. S., Vohr S. H., Wrocklage S. C., Hatcher P. J., 2010 Why genes evolve faster on secondary chromosomes in bacteria. Plos Comput. Biol. **6**: e1000732.

Couce A., Guelfo J., Blazquez J., 2013 Mutational spectrum drives the rise of mutator bacteria. Plos Genet. **9**: e1003167.

Couce A., Rodrıguez-Rojas A., Blazquez J., 2015 Bypass of genetic constraints during mutator evolution to antibiotic resistance. Proc. R. Soc. London Ser. B-Biological Sci. **282**: 20142698.

Courcelle J., 2009 Shifting replication between IInd, IIIrd, and IVth gears. Proc. Natl. Acad. Sci. U. S. A. **106**: 6027–6028.

Dame R. T., Kalmykowa O. J., Grainger D. C., 2011 Chromosomal macrodomains and associated proteins: Implications for DNA organization and replication in gram negative bacteria. PLoS Genet. **7**.

Danin-Poleg Y., Cohen L. A., Gancz H., Broza Y. Y., Goldshmidt H., Malul E., Valinsky

L., Lerner L., Broza M., Kashi Y., 2007 *Vibrio cholerae* strain typing and phylogeny study based on simple sequence repeats. J. Clin. Microbiol. **45**: 736–746.

Datsenko K. A., Wanner B. L., 2000 One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc. Natl. Acad. Sci. U. S. A. **97**: 6640–6645.

Davies E. K., Peters  a D., Keightley P. D., 1999 High frequency of cryptic deleterious mutations in *Caenorhabditis elegans*. Science **285**: 1748–1751.

Denver D. R., Morris K., Lynch M., Thomas W. K., 2004 High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. Nature **430**: 679–682.

Denver D. R., Dolan P. C., Wilhelm L. J., Sung W., Lucas-Lledo J. I., Howe D. K., Lewis S. C., Okamoto K., Thomas W. K., Lynch M., Baer C. F., 2009 A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. Proc. Natl. Acad. Sci. U. S. A. **106**: 16310–16314.

Denver D. R., Wilhelm L. J., Howe D. K., Gafner K., Dolan P. C., Baer C. F., 2012 Variation in base-substitution mutation in experimental and natural lineages of *Caenorhabditis* nematodes. Genome Biol. Evol. **4**: 513–522.

Dettman J. R., Rodrigue N., Kassen R., 2014 Genome-Wide Patterns of Recombination in the Opportunistic Human Pathogen Pseudomonas aeruginosa. Genome Biol. Evol. **7**: 18–34.

Dettman J. R., Sztepanacz J. L., Kassen R., 2016 The properties of spontaneous mutations in the opportunistic pathogen *Pseudomonas aeruginosa*. BMC Genomics **17**: 27.

Dickinson W. J., 2008 Synergistic fitness interactions and a high frequency of beneficial changes among mutations accumulated under relaxed selection in *Saccharomyces cerevisiae*. Genetics **178**: 1571–1578.

Dillon M. M., Sung W., Lynch M., Cooper V. S., 2015 The rate and molecular spectrum of spontaneous mutations in the GC-rich multichromosome genome of *Burkholderia cenocepacia*. Genetics **200**: 935–946.

Donley N., Thayer M. J., 2013 DNA replication timing, genome stability and cancer Late and/or delayed DNA replication timing is associated with increased genomic instability. Semin. Cancer Biol. **23**: 80–89.

Dorman C. J., 2013 Genome architecture and global gene regulation in bacteria: making progress towards a unified model? Nat. Rev. Microbiol. **11**: 349–355.

Drake J. W., 1991 A constant rate of spontaneous mutation in DNA-based microbes. Proc. Natl. Acad. Sci. U. S. A. **88**: 7160–7164.

Duigou S., Knudsen K. G., Skovgaard O., Egan E. S., Lobner-Olesen A., Waldor M. K., 2006 Independent control of replication initiation of the two *Vibrio cholerae* chromosomes by DnaA and RctB. J. Bacteriol. **188**: 6419–6424.

Duret L., Galtier N., 2009 Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. **10**: 285–311.

Dyall S. D., Brown M. T., Johnson P. J., 2014 Ancient invasions : from endosymbionts to organelles. Science **304**: 253–257.

Egan E. S., Waldor M. K., 2003 Distinct replication requirements for the two *Vibrio cholerae* chromosomes. Cell **114**: 521–530.

Egan E. S., Fogel M. A., Waldor M. K., 2005 Divided genomes: Negotiating the cell cycle in prokaryotes with multiple chromosomes. Mol. Microbiol. **56**: 1129–1138.

Elena S. F., Ekunwe L., Hajela N., Oden S. A., Lenski R. E., 1998 Distribution of fitness effects caused by random insertion mutations in *Escherichia coli*. Genetica **102**: 349–358.

Estes S., Phillips P. C., Denver D. R., Thomas W. K., Lynch M., 2004 Mutation accumulation in populations of varying size: The distribution of mutational effects for fitness correlates in *Caenorhabditis elegans*. Genetics **166**: 1269–1279.

Eyre-Walker A., Keightley P. D., 2007 The distribution of fitness effects of new mutations. Nat. Rev. Genet. **8**: 610–8.

Field D., Magnasco M. O., Moxon E. R., Metzgar D., Tanaka M. M., Wills C., Thaler D. S., 1999 Contingency loci, mutator alleles, and their interactions: Synergistic strategies for microbial evolution and adpatation in pathogenesis. Mol. Strateg. Biol. Evol. **870**: 378–382.

Fijalkowska I. J., Schaaper R. M., Jonczyk P., 2012 DNA replication fidelity in *Escherichia coli*: a multi-DNA polymerase affair. Fems Microbiol. Rev. **36**: 1105–1121.

Flynn K. M., Vohr S. H., Hatcher P. J., Cooper V. S., 2010 Evolutionary rates and gene dispensability associate with replication timing in the archaeon *Sulfolobus islandicus*. Genome Biol. Evol. **2**: 859–869.

Foster P. L., Hanson A. J., Lee H., Popodi E. M., Tang H. X., 2013 On the mutational topology of the bacterial genome. G3-Genes Genomes Genet. **3**: 399–407.

Foster P. L., Lee H., Popodi E., Townes J. P., Tang H., 2015 Determinants of spontaneous mutation in the bacterium *Escherichia coli* as revealed by whole-genome sequencing. Proc. Natl. Acad. Sci. **112**: E5990–E5999.

Fry J. D., Keightley P. D., Heinsohn S. L., Nuzhdin S. V, 1999 New estimates of the rates and effects of mildly deleterious mutation in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U. S. A. **96**: 574–579.

Gao F., Zhang C.-T., 2008 Ori-Finder: a web-based system for finding *oriCs* in unannotated bacterial genomes. BMC Bioinformatics **9**: 79–85.

Gao F., Luo H., Zhang C. T., 2013 DoriC 5.0: an updated database of *oriC* regions in both bacterial and archaeal genomes. Nucleic Acids Res. **41**: D90–D93.

Ghosh R., Nair G. B., Tang L., Morris J. G., Sharma N. C., Ballal M., Garg P., Ramamurthy T., Stine O. C., 2008 Epidemiological study of *Vibrio cholerae* using variable number of tandem repeats. FEMS Microbiol. Lett. **288**: 196–201.

Goldberg S., Murphy J. R., 1983 Molecular epidemiological-studies of United-States gulf-coast *Vibrio cholerae* strains - integration site of mutator *Vibriophage* Vca-3.

Infect. Immun. **42**: 224–230.

Gout J.-F., Kelley Thomas W., Smith Z., Okamoto K., Lynch M., 2013 Large-scale detection of in vivo transcription errors. Proc. Natl. Acad. Sci. U. S. A. **110**: 18584–9.

Graf J., Dunlap P. V., Ruby E. G., 1994 Effect of transposon-induced motility mutations on colonization of the host light organ by *Vibrio fischeri*. J. Bacteriol. **176**: 6986–6991.

Graur D., Li W.-H., 2000 *Fundamentals of molecular evolution*. Sinauer Associates, Sunderland, Mass.

Hall L. M. C., Henderson-Begg S. K., 2006 Hypermutable bacteria isolated from humans - a critical analysis. Microbiology **152**: 2505–2514.

Hall D. W., Mahmoudizad R., Hurd A. W., Joseph S. B., 2008 Spontaneous mutations in diploid *Saccharomyces cerevisiae*: another thousand cell generations. Genet. Res. (Camb). **90**: 229–241.

Halligan D. L., Keightley P. D., 2009 Spontaneous mutation accumulation studies in evolutionary genetics. Annu. Rev. Ecol. Evol. Syst. **40**: 151–172.

Hawk J. D., Stefanovic L., Boyer J. C., Petes T. D., Farber R. A., 2005 Variation in efficiency of DNA mismatch repair at different sites in the yeast genome. Proc. Natl. Acad. Sci. U. S. A. **102**: 8639–8643.

Hazen T. H., Kennedy K. D., Chen S., Yi S. V, Sobecky P. a, 2009 Inactivation of mismatch repair increases the diversity of *Vibrio parahaemolyticus*. Environ. Microbiol. **11**: 1254–66.

Heckman K. L., Pease L. R., 2007 Gene splicing and mutagenesis by PCR-driven overlap extension. Nat Protoc **2**: 924–932.

Heilbron K., Toll-Riera M., Kojadinovic M., Maclean R. C., 2014 Fitness is strongly influenced by rare mutations of large effect in a microbial mutation accumulation experiment. Genetics **197**: 981–990.

Herrick J., 2011 Genetic variation and DNA replication timing, or why Is there late replicating DNA? Evolution (N. Y). **65**: 3031–3047.

Hershberg R., Petrov D. A., 2010 Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. **6**: e1001115.

Hildebrand F., Meyer A., Eyre-Walker A., 2010 Evidence of selection upon genomic GC-content in bacteria. PLoS Genet. **6**: e1001107.

Hudson R. E., Bergthorsson U., Roth J. R., Ochman H., 2002 Effect of chromosome location on bacterial mutation rates. Mol. Biol. Evol. **19**: 85–92.

Jasmin J., Lenormand T., 2015 Accelerating mutational load is not due to synergistic epistasis or mutator alleles in mutation accumulation lines of yeast. Genetics.

Jolley K. A., Maiden M. C. J., 2010 BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics **11**: 595–606.

Kahramanoglou C., Prieto A. I., Khedkar S., Haase B., Gupta A., Benes V., Fraser G.

M., Luscombe N. M., Seshasayee A. S. N., 2012 Genomics of DNA cytosine methylation in *Escherichia coli* reveals its role in stationary phase transcription. Nat. Commun. **3**: 886.

Katju V., Packard L. B., Bu L., Keightley P. D., Bergthorsson U., 2015 Fitness decline in spontaneous mutation accumulation lines of Caenorhabditis elegans with varying effective population sizes. Evolution (N. Y). **69**: 104–116.

Keightley P. D., 1994 The distribution of mutation effects on viability in *Drosophila melanogaster*. Genetics **138**: 1315–1322.

Keightley P. D., Caballero A., 1997 Genomic mutation rates for lifetime reproductive output and lifespan in *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. U. S. A. **94**: 3823–3827.

Keightley P. D., Lynch M., 2003 Toward a realistic model of mutations affecting fitness. Evolution (N. Y). **57**: 683–685.

Keightley P. D., Trivedi U., Thomson M., Oliver F., Kumar S., Blaxter M. L., 2009 Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. Genome Res. **19**: 1195–1201.

Kibota T. T., Lynch M., 1996 Estimate of the genomic mutation rate deleterious to overall fitness in *E. coli*. Nature **381**: 694–696.

Kim N., Jinks-Robertson S., 2012 Transcription as a source of genome instability. Nat. Rev. Genet. **13**: 204–214.

Kimura M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, New York.

Klapacz J., Bhagwat A. S., 2002 Transcription-dependent increase in multiple classes of base substitution mutations in *Escherichia coli*. J. Bacteriol. **184**: 6866–6872.

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

Koren S., Phillippy A. M., 2015 One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr. Opin. Microbiol. **23**: 110–120.

Kraemer S. A., Morgan A. D., Ness R. W., Keightley P. D., Colegrave N., 2015 Fitness effects of new mutations in *Chlamydomonas reinhardtii* across two stress gradients. J. Evol. Biol. **44**: n/a–n/a.

Kunkel T. A., 1992 Biological asymmetries and the fidelity of eukaryotic DNA replication. Bioessays **14**: 303–308.

Kunkel T. A., Erie D. A., 2005 DNA mismatch repair. Annu. Rev. Biochem. **74**: 681–710.

Kuo C.-H., Ochman H., 2009 Deletional bias across the three domains of life. Genome Biol. Evol. **1**: 145–152.

Lande R., 1994 Risk of population extinction from fixation of new deleterious mutations. Evolution (N. Y). **48**: 1460–1469.

Lang G. I., Murray A. W., 2011 Mutation rates across budding yeast chromosome VI are correlated with replication timing. Genome Biol. Evol. **3**: 799–811.

Lang G. I., Parsons L., Gammie A. E., 2013a Mutation rates, spectra, and genome-wide distribution of spontaneous mutations in mismatch repair deficient yeast. G3-Genes Genomes Genet. **3**: 1453–1465.

Lang G. I., Rice D. P., Hickman M. J., Sodergren E., Weinstock G. M., Botstein D., Desai M. M., 2013b Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature **500**: 571–574.

Lassalle F., Périan S., Bataillon T., Nesme X., Duret L., Daubin V., 2015 GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLOS Genet. **11**: e1004941.

Lawrence M. S., Stojanov P., Polak P., Kryukov G. V, Cibulskis K., Sivachenko A., Carter S. L., Stewart C., Mermel C. H., Roberts S. a, Kiezun A., Hammerman P. S., McKenna A., Drier Y., Zou L., Ramos A. H., Pugh T. J., Stransky N., Helman E., Kim J., Sougnez C., Ambrogio L., Nickerson E., Shefler E., Cortés M. L., Auclair D., Saksena G., Voet D., Noble M., DiCara D., Lin P., Lichtenstein L., Heiman D. I., Fennell T., Imielinski M., Hernandez B., Hodis E., Baca S., Dulak A. M., Lohr J., Landau D.-A., Wu C. J., Melendez-Zajgla J., Hidalgo-Miranda A., Koren A., McCarroll S. a, Mora J., Lee R. S., Crompton B., Onofrio R., Parkin M., Winckler W., Ardlie K., Gabriel S. B., Roberts C. W. M., Biegel J. a, Stegmaier K., Bass A. J., Garraway L. a, Meyerson M., Golub T. R., Gordenin D. a, Sunyaev S., Lander E. S., Getz G., 2013 Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature **499**: 214–8.

Lee H., Popodi E., Tang H. X., Foster P. L., 2012 Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. Proc. Natl. Acad. Sci. U. S. A. **109**: E2774–E2783.

Lenski R. E., Bouma J. E., 1987 Effects of segregation and selection on instability of plasmid pACYC184 in *Escherichia coli* B. J. Bacteriol. **169**: 5314–5316.

Lenski R. E., Rose M. R., Simpson S. C., Tadler S. C., 1991 Long-term experimental evolution in *Escherichia coli* .1. Adaptation and divergence during 2,000 generations. Am. Nat. **138**: 1315–1341.

Levy S. F., Blundell J. R., Venkataram S., Petrov D. A., Fisher D. S., Sherlock G., 2015 Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature **519**: 181–186.

Li H., Durbin R., 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**: 1754–1760.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 2009 The sequence alignment/map format and SAMtools. Bioinformatics **25**: 2078–2079.

Librado P., Rozas J., 2009 DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics **25**: 1451–1452.

Lind P. A., Andersson D. I., 2008 Whole-genome mutational biases in bacteria. Proc. Natl. Acad. Sci. U. S. A. **105**: 17878–17883.

LiPuma J. J., Spilker T., Coenye T., Gonzalez C. F., 2002 An epidemic *Burkholderia cepacia* complex strain identified in soil. Lancet **359**: 2002–2003.

Liu L., De S., Michor F., 2013 DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. Nat. Commun. **4**.

Long H., Sung W., Miller S. F., Ackerman M. S., Doak T. G., Lynch M., 2014 Mutation rate, spectrum, topology, and context-dependency in the DNA mismatch repair (MMR) deficient *Pseudomonas fluorescens* ATCC948. Genome Biol. Evol. **7**: 262–271.

Long H., Kucukyildirim S., Sung W., Williams E., Lee H., Ackerman M., Doak T. G., Tang H., Lynch M., 2015 Background mutational features of the radiation-resistant bacterium *Deinococcus radiodurans*. Mol. Biol. Evol. **32**: 2383–2392.

Lopez De Saro F. J., Marinus M. G., Modrich P., O'Donnell M., 2006 The beta sliding clamp binds to multiple sites within MutL and MutS. J. Biol. Chem. **281**: 14340–14349.

Lyer R. R., Pluciennik A., Burdett V., Modrich P. L., 2006 DNA mismatch repair: Functions and mechanisms. Chem. Rev. **106**: 302–323.

Lynch M., Conery J., Burger R., 1995 Mutation accumulation and the extinction of small populations. Am. Nat. **146**: 489–518.

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

Lynch M., Blanchard J., Houle D., Kibota T., Schultz S., Vassilieva L., Willis J., 1999 Perspective: Spontaneous deleterious mutation. Evolution (N. Y). **53**: 645–663.

Lynch M., 2007 *The origins of genome architecture*. Sinauer Associates, Sunderland (MA).

Lynch M., Sung W., Morris K., Coffey N., Landry C. R., Dopman E. B., Dickinson W. J., Okamoto K., Kulkarni S., Hartl D. L., Thomas W. K., 2008 A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc. Natl. Acad. Sci. U. S. A. **105**: 9272–9277.

Lynch M., 2010a Rate, molecular spectrum, and consequences of human mutation. Proc. Natl. Acad. Sci. U. S. A. **107**: 961–968.

Lynch M., 2010b Evolution of the mutation rate. Trends Genet. **26**: 345–352.

Lynch M., 2011 The lower bound to the evolution of mutation rates. Genome Biol. Evol. **3**: 1107–1118.

Mahenthiralingam E., Urban T. A., Goldberg J. B., 2005 The multifarious, multireplicon *Burkholderia cepacia* complex. Nat. Rev. Microbiol. **3**: 144–156.

Mandel M. J., Wollenberg M. S., Stabb E. V, Visick K. L., Ruby E. G., 2009 A single regulatory gene is sufficient to alter bacterial host range. Nature **458**: 215–218.

Martin G., Lenormand T., 2006 The fitness effect of mutations across environments: a survey in light of fitness landscape models. Evolution **60**: 2413–2427.

Martincorena I., Seshasayee A. S. N., Luscombe N. M., 2012 Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature **485**: 95–98.

Marvig R. L., Johansen H. K., Molin S., Jelsbak L., 2013 Genome analysis of a transmissible lineage of *Pseudomonas aeruginosa* reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators. PLoS Genet. **9**: e1003741.

Mena A., Smith E. E., Burns J. L., Speert D. P., Moskowitz S. M., Perez J. L., Oliver A., 2008 Genetic adaptation of *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients is catalyzed by hypermutation. J. Bacteriol. **190**: 7910–7917.

Merrikh H., Zhang Y., Grossman A. D., Wang J. D., 2012 Replication-transcription conflicts in bacteria. Nat. Rev. Microbiol. **10**: 449–458.

Michaels M. L., Cruz C., Grollman A. P., Miller J. H., 1992 Evidence that *mutY* and *mutM* combine to prevent mutations by an oxidatively damaged form of guanine in DNA. Proc. Natl. Acad. Sci. U. S. A. **89**: 7022–7025.

Mira A., Ochman H., Moran N. A., 2001 Deletional bias and the evolution of bacterial genomes. Trends Genet. **17**: 589–596.

Mira A., Ochman H., 2002 Gene location and bacterial sequence divergence. Mol. Biol. Evol. **19**: 1350–1358.

Morrow J. D., Cooper V. S., 2012 Evolutionary effects of translocations in bacterial genomes. Genome Biol. Evol. **4**: 1256–1262.

Moxon E. R., Rainey P. B., Nowak M. A., Lenski R. E., 1994 Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr. Biol. **4**: 24–33.

Moxon R., Bayliss C., Hood D., 2006 Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. Annu. Rev. Genet. **40**: 307–33.

Mugal C. F., Wolf J. B. W., Grünberg H. H. Von, Ellegren H., 2010 Conservation of neutral substitution rate and substitutional asymmetries in mammalian genes. Genome Biol. Evol. **2**: 19–28.

Mukai T., 1964a Genetic Structure of Natural Populations of Drosophila Melanogaster .1. Spontaneous Mutation Rate of Polygenes Controlling Viability. Genetics **50**: 1–&.

Mukai T., 1964b The genetic structure of natural populations of *Drosophila melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability. Genetics **50**: 1–19.

Muller H. J., 1964 The relation of recombination to mutational advance. Mutat. Res. Mol. Mech. Mutagen. **1**: 2–9.

Ochman H., 2002 Bacterial evolution: Chromosome arithmetic and geometry. Curr. Biol. **12**: R427–R428.

Ochman H., 2003 Neutral Mutations and Neutral Substitutions in Bacterial Genomes. Mol. Biol. Evol. **20**: 2091–2096.

Oliver A., 2010 Mutators in cystic fibrosis chronic lung infection: Prevalence, mechanisms, and consequences for antimicrobial therapy. Int J Med Microbiol **300**: 563–572.

Ossowski S., Schneeberger K., Lucas-Lledo J. I., Warthmann N., Clark R. M., Shaw R. G., Weigel D., Lynch M., 2010 The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science **327**: 92–94.

Ostrowski E. A., Rozen D. E., Lenski R. E., 2005 Pleiotropic effects of beneficial mutations in *Escherichia coli*. Evolution (N. Y). **59**: 2343–2352.

Otto S. P., Lenormand T., 2002 Resolving the paradox of sex and recombination. Nat Rev Genet **3**: 252–261.

Pearl L. H., 2000 Structure and function in the uracil-DNA glycosylase superfamily. Mutat. Res. - DNA Repair **460**: 165–181.

Pearson T., Giffard P., Beckstrom-Sternberg S., Auerbach R., Hornstra H., Tuanyok A., Price E. P., Glass M. B., Leadem B., Beckstrom-Sternberg J. S., Allan G. J., Foster J. T., Wagner D. M., Okinaka R. T., Sim S. H., Pearson O., Wu Z., Chang J., Kaul R., Hoffmaster A. R., Brettin T. S., Robison R. A., Mayo M., Gee J. E., Tan P., Currie B. J., Keim P., 2009 Phylogeographic reconstruction of a bacterial species with high levels of lateral gene transfer. BMC Biol. **7**: 78–92.

Perfeito L., Sousa  a., Bataillon T., Gordo I., 2014 Rates of fitness decline and rebound suggest pervasive epistasis. Evolution (N. Y). **68**: 150–162.

R Development Core Team, 2013 R: A Language and Environment for Statistical Computing.

Raghavan R., Kelkar Y. D., Ochman H., 2012 A selective force favoring increased G plus C content in bacterial genes. Proc. Natl. Acad. Sci. U. S. A. **109**: 14504–14507.

Rasmussen T., Jensen R. B., Skovgaard O., 2007 The two chromosomes of *Vibrio cholerae* are initiated at different time points in the cell cycle. Embo J. **26**: 3124–3131.

Reyes G. X., Schmidt T. T., Kolodner R. D., Hombauer H., 2015 New insights into the mechanism of DNA mismatch repair. Chromosoma **124**: 443–462.

Roze D., Blanckaert A., 2014 Epistasis, pleiotropy, and the mutation load in sexual and asexual populations. Evolution (N. Y). **68**: 137–149.

Ruby E. G., Urbanowski M., Campbell J., Dunn A., Faini M., Gunsalus R., Lostroh P., Lupp C., McCann J., Millikan D., Schaefer A., Stabb E., Stevens A., Visick K., Whistler C., Greenberg E. P., 2005 Complete genome sequence of *Vibrio fischeri*: A symbiotic bacterium with pathogenic congeners. Proc. Natl. Acad. Sci. U. S. A. **102**: 3004–3009.

Sambrook J., Fritsch E. F., Maniatis T., 1989 *Molecular Cloning: A Laboratory Manual.* New York.

Sawabe T., Koizumi S., Fukui Y., Nakagawa S., Ivanova E. P., Kita-Tsukamoto K., Kogure K., Thompson F. L., 2009 Mutation is the main driving force in the diversification of the *Vibrio splendidus* clade. Microbes Environ. **24**: 281–285.

Schaack S., Allen D. E., Latta L. C., Morgan K. K., Lynch M., 2013 The effect of spontaneous mutations on competitive ability. J. Evol. Biol. **26**: 451–6.

Schmidt K. H., Reimers J. M., Wright B. E., 2006 The effect of promoter strength, supercoiling and secondary structure on mutation rates in *Escherichia coli*. Mol. Microbiol. **60**: 1251–1261.

Schrider D. R., Houle D., Lynch M., Hahn M. W., 2013 Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. Genetics **194**: 937–954.

Schultz S. T., Lynch M., Willis J. H., 1999 Spontaneous deleterious mutation in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. U. S. A. **96**: 11393–8.

Schuster-Böckler B., Lehner B., 2012 Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature **488**: 504–507.

Schwander T., Crespi B. J., 2009 Twigs on the tree of life? Neutral and selective models for integrating macroevolutionary patterns with microevolutionary processes in the analysis of asexuality. Mol. Ecol. **18**: 28–42.

Seemann T., 2014 Prokka: Rapid prokaryotic genome annotation. Bioinformatics **30**: 2068–2069.

Shaw R. G., Byers D. L., Darmo E., 2000 Spontaneous mutational effects on reproductive traits of *Arabidopsis thaliana*. Genetics **155**: 369–378.

Shaw F. H., Geyer C. J., Shaw R. G., 2002 A comprehensive model of mutations affecting fitness and inferences for *Arabidopsis thaliana*. Evolution **56**: 453–463.

Silander O. K., Tenaillon O., Chao L., 2007 Understanding the evolutionary fate of finite populations: the dynamics of mutational effects. PLoS Biol. **5**: e94.

Smith M. A., Bidochka M. J., 1998 Bacterial fitness and plasmid loss: the importance of culture conditions and plasmid size. Can. J. Microbiol. **44**: 351–355.

Sniegowski P. D., Lenski R. E., 1995 Mutation and adaptation - the directed mutation controversy in evolutionary perspective. Annu. Rev. Ecol. Syst. **26**: 553–578.

Sniegowski P. D., Gerrish P. J., Lenski R. E., 1997 Evolution of high mutation rates in experimental populations of *E. coli*. Nature **387**: 703–705.

Sniegowski P. D., Gerrish P. J., Johnson T., Shaver A., 2000 The evolution of mutation rates: separating causes from consequences. Bioessays **22**: 1057–1066.

Sobetzko P., Travers A., Muskhelishvili G., 2012 Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. Proc. Natl. Acad. Sci. U. S. A. **109**: E42–E50.

Stabb E. V., Ruby E. G., 2002 RP4-based plasmids for conjugation between *Escherichia coli* and members of the *vibrionaceae*. Methods Enzymol. **358**: 413–426.

Stamatoyannopoulos J. A., Adzhubei I., Thurman R. E., Kryukov G. V, Mirkin S. M., Sunyaev S. R., 2009 Human mutation rate associated with DNA replication timing.

Nat. Genet. **41**: 393–395.

Sung W., Ackerman M. S., Miller S. F., Doak T. G., Lynch M., 2012a Drift-barrier hypothesis and mutation-rate evolution. Proc. Natl. Acad. Sci. U. S. A. **109**: 18488–18492.

Sung W., Tucker A. E., Doak T. G., Choi E., Thomas W. K., Lynch M., 2012b Extraordinary genome stability in the ciliate *Paramecium tetraurelia*. Proc. Natl. Acad. Sci. U. S. A. **109**: 19339–19344.

Sung W., Ackerman M. S., Gout J.-F., Miller S. F., Williams E., Foster P. L., Lynch M., 2015 Asymmetric context-dependent mutation patterns revealed through mutation-accumulation experiments. Mol. Biol. Evol. **32**: 1672–1683.

Sung W., Ackerman M. S., Dillon M. M., Platt T., Fuqua C., Cooper V. S., Lynch M., 2016 Genetic drift limits insertion-deletion mutation rate evolution. Rev.

Taxis C., Keller P., Kavagiou Z., Jensen L. J., Colombelli J., Bork P., Stelzer E. H. K., Knop M., 2005 Spore number control and breeding in *Saccharomyces cerevisiae*: A key role for a self-organizing system. J. Cell Biol. **171**: 627–640.

Tenaillon O., Toupance B., Nagard H. Le, Taddei F., Godelle B., 1999 Mutators, population size, adaptive landscape and the adaptation of asexual populations of bacteria. Genetics **152**: 485–493.

Thompson F. L., Iida T., Swings J., 2004 Biodiversity of *Vibrios*. Microbiol. Mol. Biol. Rev. **68**.

Traverse C. C., Mayo-Smith L. M., Poltak S. R., Cooper V. S., 2013 Tangled bank of experimentally evolved *Burkholderia* biofilms reflects selection during chronic infections. Proc. Natl. Acad. Sci. U. S. A. **110**: E250–E259.

Trindade S., Perfeito L., Gordo I., 2010 Rate and effects of spontaneous mutations that affect fitness in mutator *Escherichia coli*. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **365**: 1177–1186.

Val M. E., Skovgaard O., Ducos-Galand M., Bland M. J., Mazel D., 2012 Genome engineering in *Vibrio cholerae*: A feasible approach to address biological issues. Plos Genet. **8**: e1002472.

Val M.-E., Soler-Bistué A., Bland M. J., Mazel D., 2014 Management of multipartite genomes: the *Vibrio cholerae* model. Curr. Opin. Microbiol. **22**: 120–126.

Vassilieva L. L., Hook  a M., Lynch M., 2000 The fitness effects of spontaneous mutations in *Caenorhabditis elegans*. Evolution (N. Y). **54**: 1234–46.

Vos M., Didelot X., 2009 A comparison of homologous recombination rates in bacteria and archaea. ISME J. **3**: 199–208.

Warnecke T., Supek F., Lehner B., 2012 Nucleoid-associated proteins affect mutation dynamics in *E. coli* in a growth phase-specific manner. PLoS Comput. Biol. **8**: e1002846.

Watterson G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7**: 256–276.

Wei W., Ning L.-W., Ye Y.-N., Li S.-J., Zhou H.-Q., Huang J., Guo F.-B., 2014 SMAL: A Resource of Spontaneous Mutation Accumulation Lines. Mol. Biol. Evol. **31**: 1302–8.

Wielgoss S., Barrick J. E., Tenaillon O., Cruveiller S., Chane-Woon-Ming B., Medigue C., Lenski R. E., Schneider D., 2011 Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. G3-Genes Genomes Genet. **1**: 183–186.

Wloch D. M., Szafraniec K., Borts R. H., Korona R., 2001 Direct estimate of the mutation rate and the distribution of fitness effects in the yeast *Saccharomyces cerevisiae*. Genetics **159**: 441–452.

Wollenberg M. S., Ruby E. G., 2012 Phylogeny and fitness of Vibrio fischeri from the light organs of *Euprymna scolopes* in two Oahu, Hawaii populations. ISME J. **6**: 352–362.

Ye K., Schulz M. H., Long Q., Apweiler R., Ning Z. M., 2009 Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics **25**: 2865–2871.

Zeyl C., Visser J. A. de, 2001 Estimates of the rate and distribution of fitness effects of spontaneous mutation in Saccharomyces cerevisiae. Genetics **157**: 53–61.

Zhang X. L., Mathews C. K., 1995 Natural DNA precursor pool asymmetry and base sequence context as determinants of replication fidelity. J. Biol. Chem. **270**: 8401–8404.

Zhu Y. O., Siegal M. L., Hall D. W., Petrov D. A., 2014 Precise estimates of mutation rate and spectrum in yeast. Proc. Natl. Acad. Sci. U. S. A. **111**: E2310–E2318.

# APPENDICES

# Appendix A Chapter I Supplemental Material

Figures A.1-6 and Tables A.1-2

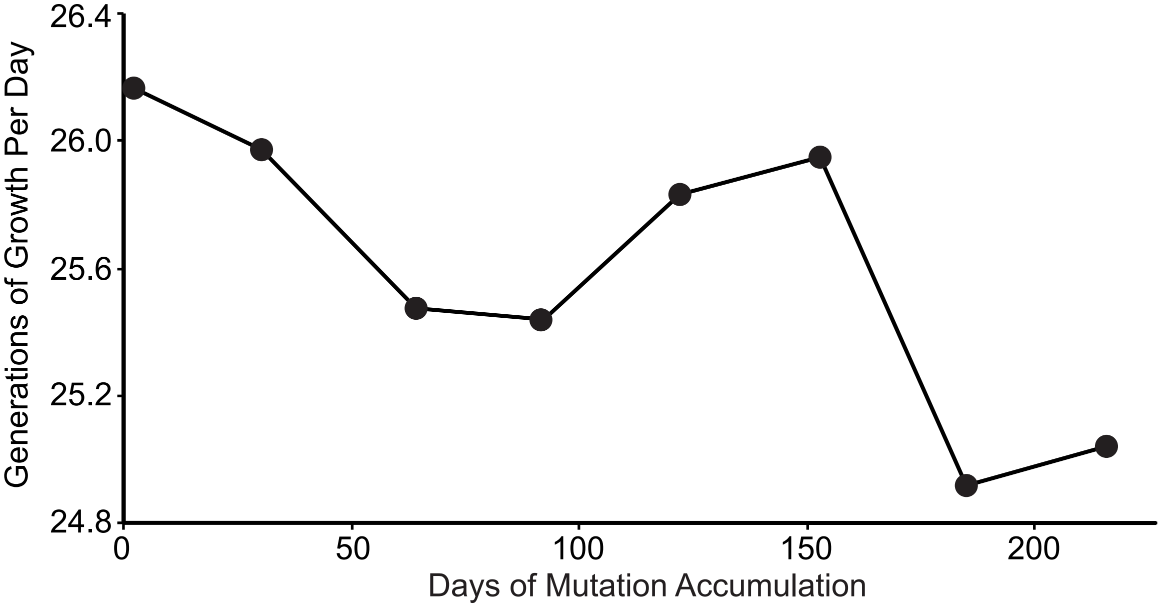

**Figure A.1. Estimates of the average number of generations per day experienced by the *Burkholderia cenocepacia* mutation accumulation lineages.** Each measurement was taken using the average of ten representative lineages per MA experiment and measurement error is such that error bars representing 95% confidence intervals are not visible outside the markers.

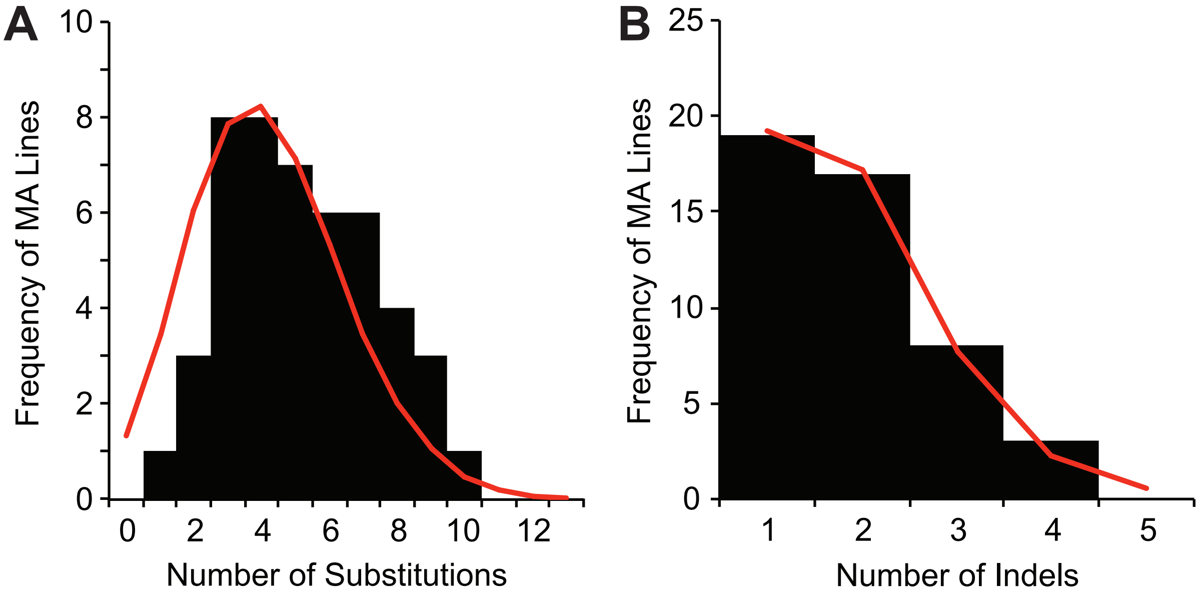

**Figure A.2. Frequency distributions of the number of base-substitution (bpsm) and insertion-deletion (indel) mutations per lineage in the *Burkholderia cenocepacia* mutation accumulation (MA) experiment.** Neither the distribution for bpsm (A) or indels (B) differs significantly from a Poisson distribution (bps: $\chi^2$ = 1.81, p = 0.99; indels: $\chi^2$ = 0.48, p = 0.92).

.

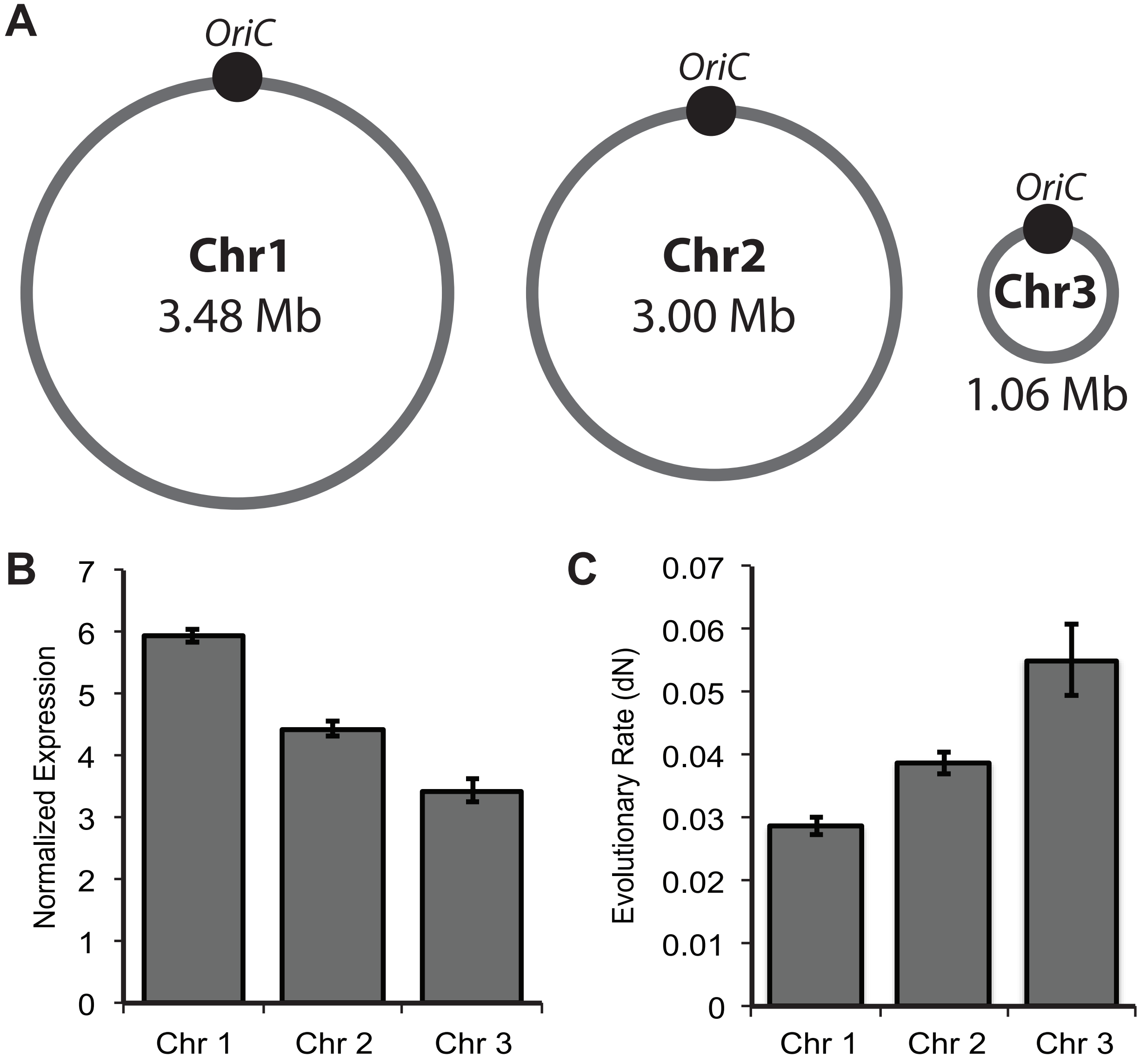

**Figure A.3. Chromosome size (A), expression (B), and evolutionary rate (C) differences between the three chromosomes of *Burkholderia cenocepacia*.** Chromosome sizes were obtained from the complete *B. cenocepacia* H!2424 genome (NCBI), expression data was derived using RNAseq as described in (Gout *et al.* 2013), and evolutionary rates were obtained from (Morrow and Cooper 2012).

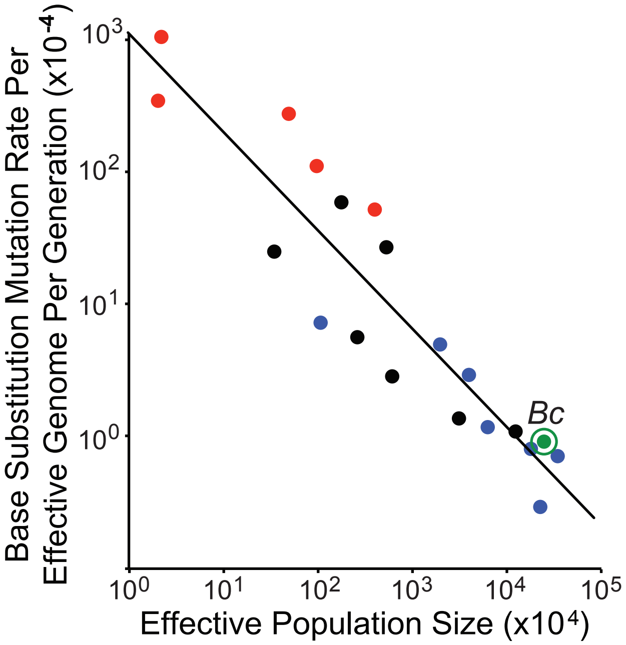

**Figure A.4. Relationship between base-substitution mutation (bpsm) rate per effective genome size per generation with effective population size ($N_E$).** Base-substitution mutation rates were measured in five multicellular eukaryotes (red), seven unicellular eukaryotes (black), and eight prokaryotes (blue; *B. cenocepacia* – green) (Sung *et al.* 2012a). The log-linear regression is highly significant ($r^2$=0.85, p<0.0001, df=19).

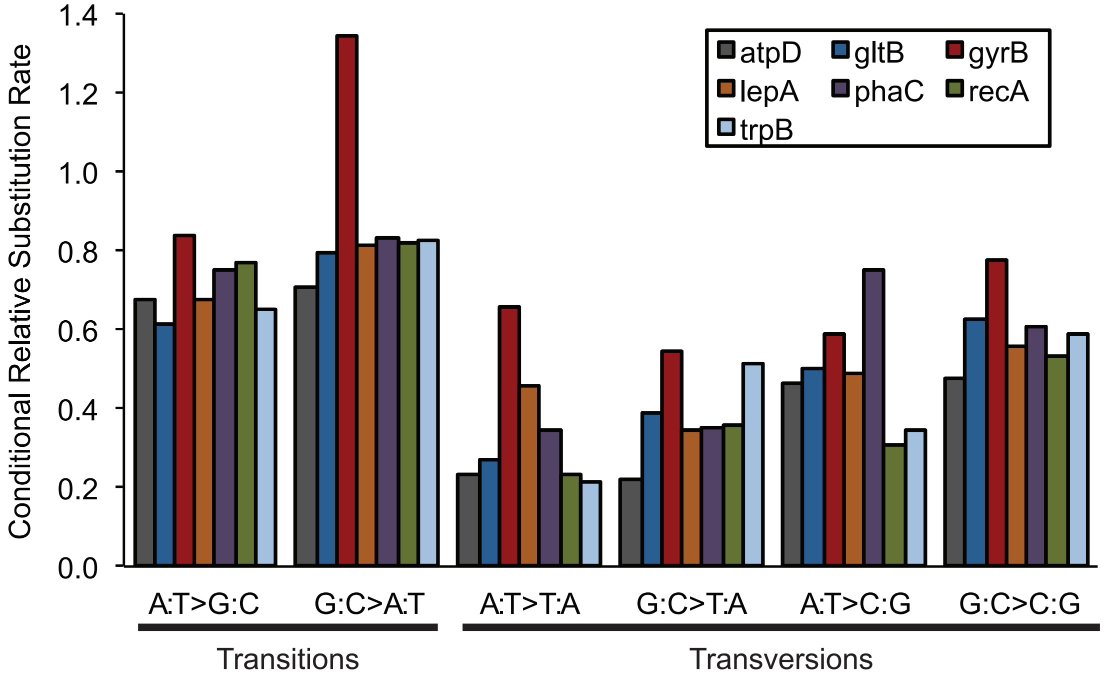

**Figure A.5. Conditional relative substitution rates at seven *Burkholderia cenocepacia* loci, including *atpD*, *gltB*, *gyrB*, *lepA*, *phaC*, *recA*, and *trpB*.** Relative conditional substitution rates are estimated by assuming that the most common nucleotide at each site is ancestral and any deviation from that nucleotide is caused by a single mutation. Substitution rates were calibrated to the nucleotide content at polymorphic sites for each gene, whereby only covered sites capable of producing a given substitution are used in the denominator of each calculation.

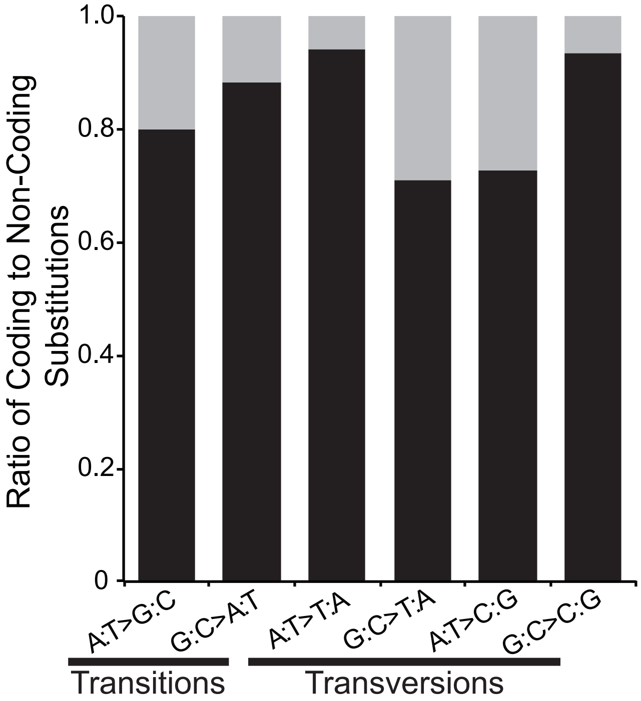

**Figure A.6. Ratio of coding (black) to non-coding (grey) base-substitution mutations (bpsms) for each *Burkholderia cenocepacia* bpsm type.**

**Table A.1. All base-substitution mutations (bpsms) identified in 47 independent MA lineages when aligned to *Burkholderia cenocepacia* HI2424 (NC_008542; NC_008543; NC_008544; NC_008545) with BWA and Novoalign.**

https://drive.google.com/drive/folders/0Bz0jXLefrfhaYWlnZUtkSXR4X00

**Table A.2. All insertion-deletion (indel) mutations identified in 47 independent MA lineages when aligned to *Burkholderia cenocepacia* HI2424 (NC_008542; NC_008543; NC_008544; NC_008545).**

https://drive.google.com/drive/folders/0Bz0jXLefrfhaYWlnZUtkSXR4X00

# Appendix B Chapter II Supplemental Material

Figures B.1-2 and Tables B.1-4

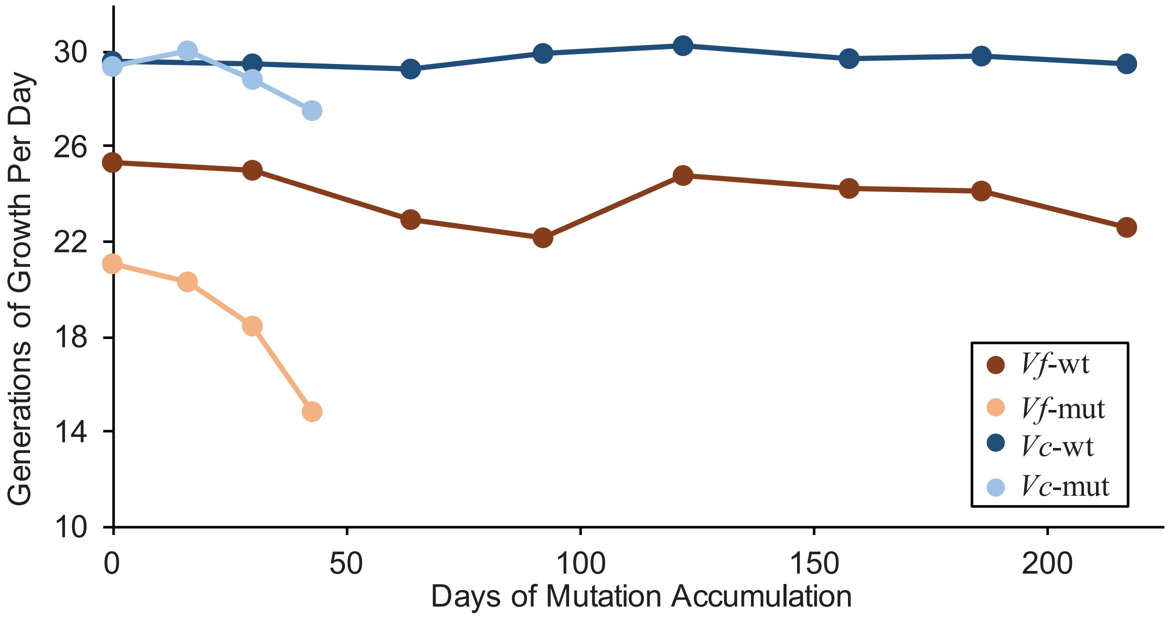

**Figure B.1. Estimates of the average number of generations per day experienced by the *Vibrio fischeri* wild-type, *Vibrio cholerae* wild-type, *Vibrio fischeri ΔmutS*, and *Vibrio cholerae ΔmutS* mutation accumulation lineages.** Each measurement was taken using the average of ten representative lineages per MA experiment and measurement error is such that error bars representing 95% confidence intervals are not visible outside of the markers.

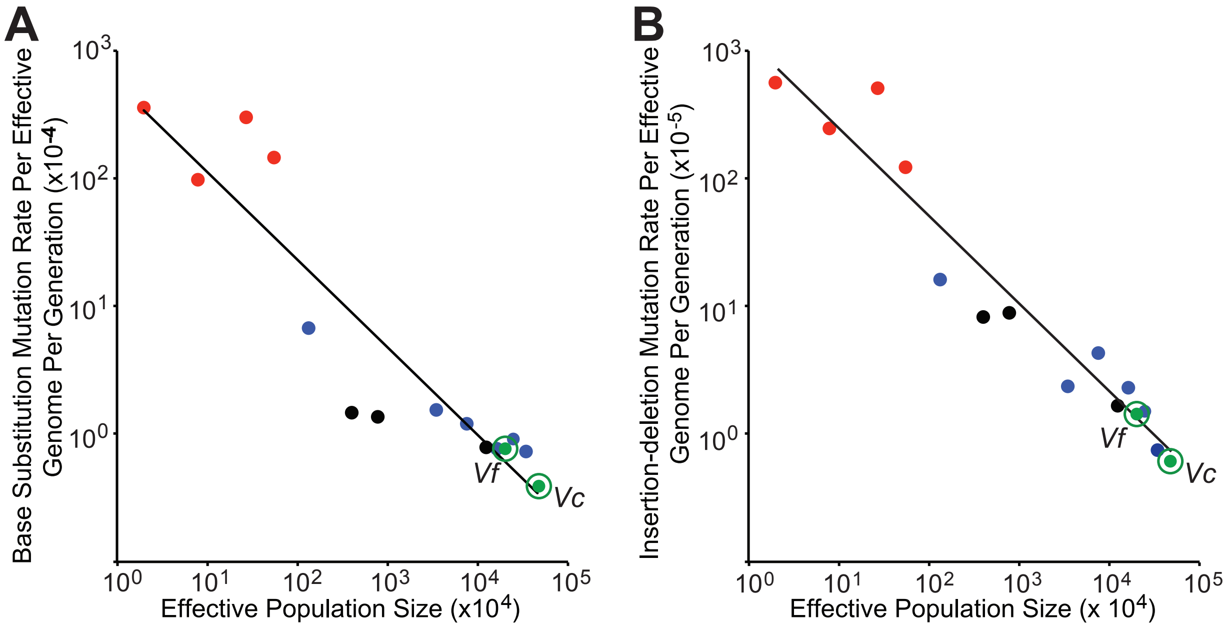

**Figure B.2. Relationship between base-substitution mutation (bpsm) rate and insertion-deletion (indel) rate per effective genome size per generation with effective population size ($N_E$).** Four multicellular eukaryotes are shown in red, three unicellular eukaryotes are shown in black, and eight prokaryotes are shown in blue. *Vibrio fischeri* and *Vibrio cholerae* wild-type bpsm and indel rates rates estimated in this study are highlighted in green. The log-linear regressions are highly significant for both bpsm rate ($r^2 = 0.86$, $p < 0.0001$, df = 14) and indel rate ($r^2 = 0.94$, $p < 0.0001$, df = 14).

**Table B.1. All base-substitution mutations (bpsms) identified in the *V. fischeri* wild-type, *V. cholerae* wild-type, *V. fischeri ΔmutS*, and *V. cholerae ΔmutS* MA lineages when aligned to their respective reference genomes.**

https://drive.google.com/drive/folders/0Bz0jXLefrfhaYWlnZUtkSXR4X00

**Table B.2. All insertion-deletion mutations (indel) identified in the *V. fischeri* wild-type, *V. cholerae* wild-type, *V. fischeri ΔmutS*, and *V. cholerae ΔmutS* MA lineages when aligned to their respective reference genomes**.

https://drive.google.com/drive/folders/0Bz0jXLefrfhaYWlnZUtkSXR4X00

**Table B.3. Conditional base-substitution mutation (bpsm) rates for wild-type *V. fischeri* and *V. cholerae* in different replication timing regions.** Early chr1 regions are regions on chr1 replicated prior to the initiation of chr2 replication, late chr1 regions are regions on chr1 replicated concurrently with chr2, and chr2 regions are the bpsm rates on chr2 itself.

| Species | Bps Type | Early chr1 | | Late chr1 | | Chr2 | |
|---|---|---|---|---|---|---|---|
| | | Avg | SEM | Avg | SEM | Avg | SEM |
| *V. fischeri* | A:T>G:C | $4.25\cdot10^{-11}$ | $1.37\cdot10^{-11}$ | $3.57\cdot10^{-11}$ | $1.28\cdot10^{-11}$ | $5.43\cdot10^{-11}$ | $1.47\cdot10^{-11}$ |
| | G:C>A:T | $1.69\cdot10^{-10}$ | $3.24\cdot10^{-11}$ | $1.43\cdot10^{-10}$ | $3.74\cdot10^{-11}$ | $2.24\cdot10^{-10}$ | $4.43\cdot10^{-11}$ |
| | A:T>T:A | $4.25\cdot10^{-12}$ | $4.30\cdot10^{-12}$ | $9.78\cdot10^{-12}$ | $6.91\cdot10^{-12}$ | $4.80\cdot10^{-12}$ | $4.85\cdot10^{-12}$ |
| | G:C>T:A | $1.10\cdot10^{-10}$ | $2.73\cdot10^{-11}$ | $1.98\cdot10^{-10}$ | $4.12\cdot10^{-11}$ | $2.78\cdot10^{-10}$ | $4.86\cdot10^{-11}$ |
| | A:T>C:G | $3.83\cdot10^{-11}$ | $1.18\cdot10^{-11}$ | $3.91\cdot10^{-11}$ | $1.47\cdot10^{-11}$ | $3.84\cdot10^{-11}$ | $1.44\cdot10^{-11}$ |
| | G:C>C:G | $4.54\cdot10^{-11}$ | $1.87\cdot10^{-11}$ | $1.58\cdot10^{-11}$ | $1.12\cdot10^{-11}$ | $8.17\cdot10^{-12}$ | $8.26\cdot10^{-12}$ |
| *V. cholerae* | A:T>G:C | $3.89\cdot10^{-11}$ | $1.10\cdot10^{-11}$ | $2.74\cdot10^{-11}$ | $1.18\cdot10^{-11}$ | $1.08\cdot10^{-11}$ | $7.67\cdot10^{-12}$ |
| | G:C>A:T | $1.12\cdot10^{-10}$ | $1.85\cdot10^{-11}$ | $7.92\cdot10^{-11}$ | $2.11\cdot10^{-11}$ | $1.23\cdot10^{-10}$ | $2.50\cdot10^{-11}$ |
| | A:T>T:A | $9.72\cdot10^{-12}$ | $5.55\cdot10^{-12}$ | $5.47\cdot10^{-12}$ | $5.52\cdot10^{-12}$ | $1.08\cdot10^{-11}$ | $7.66\cdot10^{-12}$ |
| | G:C>T:A | $4.18\cdot10^{-11}$ | $1.18\cdot10^{-11}$ | $3.05\cdot10^{-11}$ | $1.59\cdot10^{-11}$ | $4.93\cdot10^{-11}$ | $1.63\cdot10^{-11}$ |
| | A:T>C:G | $1.94\cdot10^{-11}$ | $8.91\cdot10^{-12}$ | $2.74\cdot10^{-11}$ | $1.18\cdot10^{-11}$ | $3.25\cdot10^{-11}$ | $1.68\cdot10^{-11}$ |
| | G:C>C:G | $6.97\cdot10^{-12}$ | $4.93\cdot10^{-12}$ | $6.10\cdot10^{-12}$ | $6.16\cdot10^{-12}$ | $1.85\cdot10^{-11}$ | $1.05\cdot10^{-11}$ |

**Table B.4. Relative frequencies of insertion-deletion mutations observed in the wild-type and mutator mutation accumulation experiments with *V. fischeri* and *V. cholerae*.** Chi-square tests were conducted to test whether indels in each size category were significantly over-represented in the mutator lineages after correcting for differences in total number of sites analyzed across all lineages and the number of generations in the wild-type and mutator experiments.

| Species | Indel | Indels Observed | | Expected Frequencies | | $\chi^2$ | df | p |
|---|---|---|---|---|---|---|---|---|
| | Length | Wt | Mut | Wt | Mut | | | |
| *V. fischeri* | 1 | 12 | 352 | 0.942 | 0.058 | 5460.000 | 1 | <0.0001 |
| | 2 | 0 | 11 | 0.942 | 0.058 | 177.200 | 1 | <0.0001 |
| | 3 | 0 | 5 | 0.942 | 0.058 | 80.542 | 1 | <0.0001 |
| | 4 | 1 | 0 | 0.942 | 0.058 | 0.062 | 1 | 0.803 |
| | 5 | 2 | 0 | 0.942 | 0.058 | 0.124 | 1 | 0.725 |
| | 6 | 4 | 2 | 0.942 | 0.058 | 8.238 | 1 | 0.004 |
| | 7 | 19 | 4 | 0.942 | 0.058 | 5.572 | 1 | 0.018 |
| | 8 | 3 | 0 | 0.942 | 0.058 | 0.186 | 1 | 0.666 |
| | 9 | 0 | 0 | 0.942 | 0.058 | - | - | - |
| | 10 | 0 | 2 | 0.942 | 0.058 | 32.217 | 1 | <0.0001 |
| *V. cholerae* | 1 | 8 | 233 | 0.920 | 0.080 | 2567.000 | 1 | <0.0001 |
| | 2 | 0 | 30 | 0.920 | 0.080 | 343.900 | 1 | <0.0001 |
| | 3 | 2 | 7 | 0.920 | 0.080 | 59.332 | 1 | <0.0001 |
| | 4 | 2 | 0 | 0.920 | 0.080 | 0.174 | 1 | 0.676 |
| | 5 | 0 | 0 | 0.920 | 0.080 | - | - | - |
| | 6 | 0 | 1 | 0.920 | 0.080 | 11.462 | 1 | 0.001 |
| | 7 | 0 | 0 | 0.920 | 0.080 | - | - | - |
| | 8 | 0 | 0 | 0.920 | 0.080 | - | - | - |
| | 9 | 1 | 1 | 0.920 | 0.080 | 4.775 | 1 | 0.029 |
| | 10 | 2 | 0 | 0.920 | 0.080 | 0.174 | 1 | 0.676 |

# Appendix C Chapter III Supplemental Material

Tables C.1-4

**Table C.1. All base-substitution mutations (bpsms) identified in the *V. fischeri ΔmutS*, *V. cholerae ΔmutS, V. fischeri* wild-type, *V. cholerae* wild-type, and *B. cenocepacia* wild-type MA lineages when aligned to their respective reference genomes.**

https://drive.google.com/drive/folders/0Bz0jXLefrfhaYWlnZUtkSXR4X00

**Table C.2. Linear regression statistics for correlations between the base-substitution mutation (bpsm) rates in concurrently replicated regions of opposing replichores in *Vibrio fischeri ΔmutS* at various interval lengths.** Bpsm rates are calculated as the number of mutations observed in each interval, divided by product of the total number of sites analyzed in that interval across all lines and the number of generations of mutation accumulation.

| Chromosome | Interval Length | + or - | F | df | p | $r^2$ |
|---|---|---|---|---|---|---|
| Chr1[a] | 500 Kb | + | 4.365 | 1, 1 | 0.284 | 0.814 |
| | 250 Kb | + | 13.680 | 1, 4 | 0.021 | 0.774 |
| | 100 Kb | + | 10.980 | 1, 13 | 0.006 | 0.458 |
| | 50 Kb | + | 12.810 | 1, 27 | 0.001 | 0.322 |
| | 25 Kb | + | 9.355 | 1, 56 | 0.003 | 0.143 |
| | 10 Kb | + | 8.267 | 1, 143 | 0.005 | 0.055 |
| Chr2[b] | 500 Kb | NA | NA | NA | NA | NA |
| | 250 Kb | + | 1.205 | 1, 1 | 0.470 | 0.546 |
| | 100 Kb | - | 0.021 | 1, 6 | 0.891 | 0.003 |
| | 50 Kb | + | 0.752 | 1, 12 | 0.403 | 0.059 |
| | 25 Kb | + | 2.634 | 1, 25 | 0.117 | 0.095 |
| | 10 Kb | + | 1.416 | 1, 65 | 0.238 | 0.021 |

[a] The final intervals on each of the replichores of chr1 are of equal length, but shorter than the specified interval length to account for the fact that the size of the chromosome is not exactly divisible by the interval length.

[b] The chr2 intervals are calibrated to reflect their concurrent replication with chromosome 1 intervals. As such, the final intervals on each replichore are shorter than the specified interval length but equal to each other and the final intervals of chr1, while the first intervals on each replichore are equal to each other but shorter than the specified interval length.

**Table C.3. Linear regression statistics for correlations between the base-substitution mutation (bpsm) rates in concurrently replicated regions of opposing replichores in *Vibrio cholerae* Δ*mutS* at various interval lengths.** Bpsm rates are calculated as the number of mutations observed in each interval, divided by product of the total number of sites analyzed in that interval across all lines and the number of generations of mutation accumulation.

| Chromosome | Interval Length | + or - | F | df | p | $r^2$ |
|---|---|---|---|---|---|---|
| Chr1 | 500 Kb | + | 61.550 | 1, 1 | 0.081 | 0.984 |
| | 250 Kb | + | 4.445 | 1, 4 | 0.103 | 0.526 |
| | 100 Kb | + | 6.759 | 1, 13 | 0.022 | 0.342 |
| | 50 Kb | + | 4.441 | 1, 28 | 0.044 | 0.137 |
| | 25 Kb | + | 5.083 | 1, 58 | 0.028 | 0.081 |
| | 10 Kb | + | 0.722 | 1, 148 | 0.397 | 0.005 |
| Chr2 | 500 Kb | NA | NA | NA | NA | NA |
| | 250 Kb | + | 3.698 | 1, 1 | 0.305 | 0.787 |
| | 100 Kb | - | 0.063 | 1, 4 | 0.814 | 0.016 |
| | 50 Kb | - | 4.267 | 1, 10 | 0.066 | 0.299 |
| | 25 Kb | - | 1.948 | 1, 21 | 0.177 | 0.085 |
| | 10 Kb | + | 0.091 | 1, 54 | 0.769 | 0.002 |

[a] The final intervals on each of the replichores of chr1 are of equal length, but shorter than the specified interval length to account for the fact that the size of the chromosome is not exactly divisible by the interval length.

[b] The chr2 intervals are calibrated to reflect their concurrent replication with chromosome 1 intervals. As such, the final intervals on each replichore are shorter than the specified interval length but equal to each other and the final intervals of chr1, while the first intervals on each replichore are equal to each other but shorter than the specified interval length.

**Table C.4. Sum of the residuals between the bpsm rates on chromosome 2 (chr2) and the bpsm rates on chromosome 1 (chr1), when the 100 Kb intervals on each replichore of chr2 are mapped to every possible replication timing location on chr1 for all MA experiments.** The lowest sum of the residuals, corresponding to the best fit for the chr2 intervals on chr1 for each analysis is bolded and underlined.

| Ma Lines | First chr1 Interval[a] | Last chr1 Interval[a] | Sum of Residuals |
|---|---|---|---|
| *Vf*-mut | 1 | 8 | 19.69 |
| | 2 | 9 | 18.14 |
| | 3 | 10 | 19.45 |
| | 4 | 11 | 22.14 |
| | 5 | 12 | 26.70 |
| | 6 | 13 | 25.25 |
| | 7 | 14 | 18.13 |
| | **8** | **15** | **14.01** |
| *Vc*-mut | 1 | 6 | 3.84 |
| | 2 | 7 | 3.88 |
| | 3 | 8 | 5.21 |
| | 4 | 9 | 5.76 |
| | 5 | 10 | 5.33 |
| | 6 | 11 | 5.46 |
| | 7 | 12 | 5.06 |
| | 8 | 13 | 4.86 |
| | 9 | 14 | 3.45 |
| | **10** | **15** | **2.53** |
| *Vf*-wt | 1 | 8 | 19.34 |
| | 2 | 9 | 18.98 |
| | 3 | 10 | 19.86 |
| | **4** | **11** | **18.09** |
| | 5 | 12 | 21.47 |
| | 6 | 13 | 19.45 |
| | 7 | 14 | 18.14 |
| | 8 | 15 | 19.69 |
| *Vc*-wt | **1** | **6** | **3.84** |
| | 2 | 7 | 3.88 |
| | 3 | 8 | 5.21 |
| | 4 | 9 | 5.76 |
| | 5 | 10 | 5.33 |
| | 6 | 11 | 5.46 |
| | 7 | 12 | 5.06 |
| | 8 | 13 | 9.95 |
| | 9 | 14 | 9.00 |
| | 10 | 15 | 6.74 |
| *Bc*-wt | 1 | 16 | 29.45 |
| | 2 | 17 | 28.68 |
| | **3** | **18** | **26.97** |

[a] The first chr1 interval for each analysis is the first bpsm interval on each replichore of chr1 that the chr2 bpsm intervals are mapped to, while the last chr1 interval is where the the final chr2 interval in the analysis is mapped.

# Appendix D Chapter IV Supplemental Material

Figure D.1-3 and Tables D.1-3

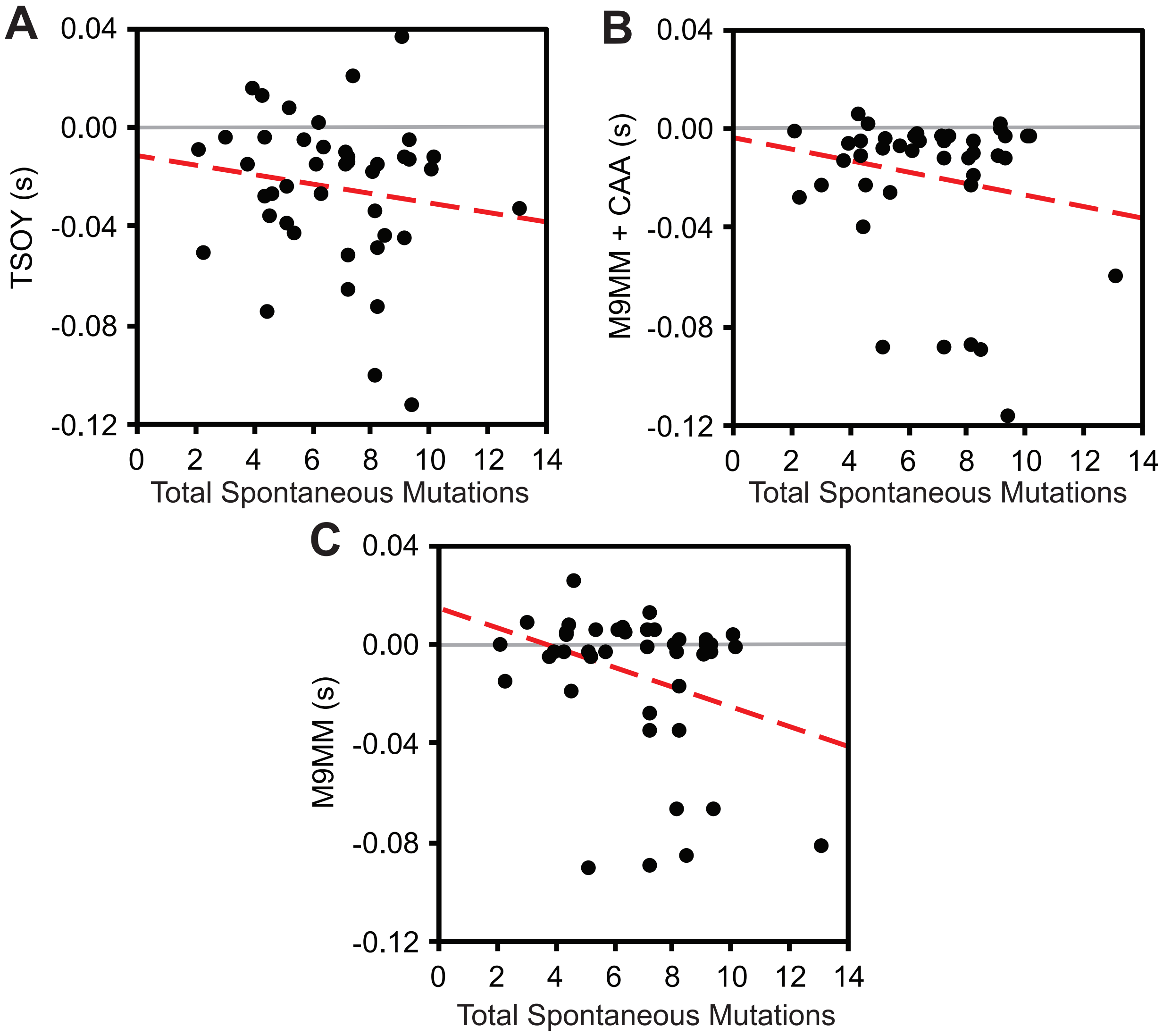

**Figure D.1. Relationship between the selection coefficients in *Burkholderia cenocepacia* MA lineages and the number of spontaneous mutations that they harbor for all three environments**. All linear regressions are negative, but none are statistically significant (A: F = 1.401, df = 41, p = 0.2434, $r^2$ = 0.0330; B: F = 1.354, df = 41, p = 0.2513, $r^2$ = 0.0320; C: F = 2.957, df = 41, p = 0.0930, $r^2$ = 0.0673).

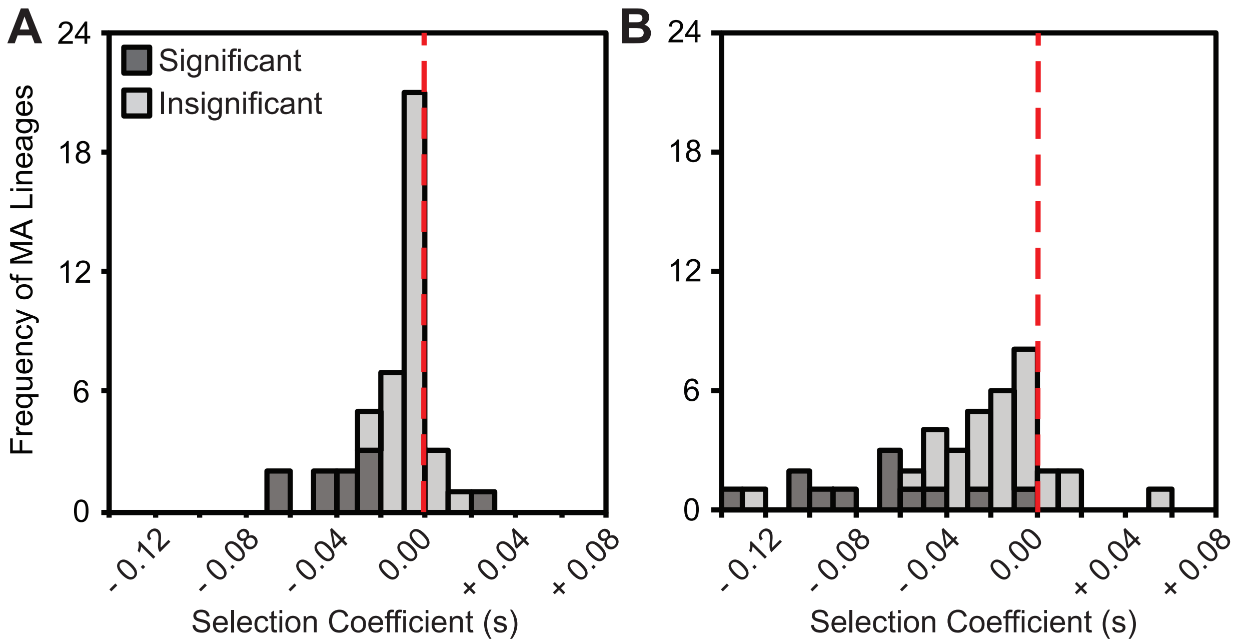

**Figure D.2. Preliminary analysis of the selection coefficients of forty-five of the *Vibrio fischeri* MA lineages from Chapter 2 in tryptic soy broth supplemented with NaCl (A), and HEPES minimal medium supplemented with casamino acids (B).** Fitness assays were carried out as described in Chapter 4, except that competitions were incubated at 28° and relative frequencies were measured with flow cytometry using a fluorescent reporter plasmid in the the *V. fischeri* ES114 ancestor. Significance was determined from independent two-tailed t-tests on four replicate fitness assays for each lineage. P-values were corrected for multiple comparisons using a Benjamini-Hochberg correction, and corrected p-values that remained below 0.05 were considered significant. One lineage with significantly reduced fitness on panel A (*s* = - 0.23) and two lineages with significantly reduced fitness on panel B (*s* = - 0.38 and *s* = - 0.44) are not shown on the plot.

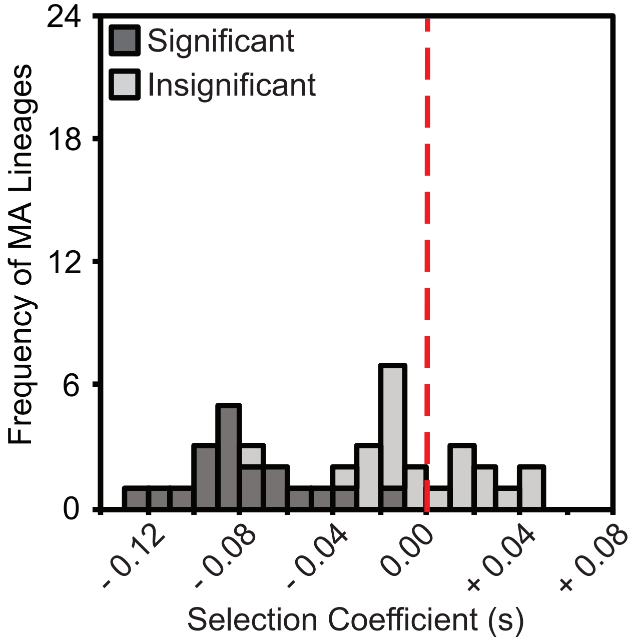

**Figure D.3. Preliminary analysis of the selection coefficients of forty-two of the *Vibrio cholerae* MA lineages from Chapter 2 in HEPES minimal medium.** Fitness assays were carried out as described in Chapter 4, except that relative frequencies were measured with flow cytometry using a fluorescent reporter inserted into the genome of *V. cholerae* 2740-80 ancestor. Significance was determined from independent two-tailed t-tests on four replicate fitness assays for each lineage. P-values were corrected for multiple comparisons using a Benjamini-Hochberg correction, and corrected p-values that remained below 0.05 were considered significant. One lineage with significantly reduced fitness (*s* = - 0.27) is not shown on the plot.

**Table D.1. All base-substitution (bpsms) and insertion-deletion (indels) mutations identified in forty-three independent MA lineages when aligned to *Burkholderia cenocepacia* HI2424 (NC_008542; NC_008543; NC_008544; NC_008545) with BWA and Novoalign.**

https://drive.google.com/drive/folders/0Bz0jXLefrfhaYWlnZUtkSXR4X00

**Table D.2. Estimates of the number of false negative base-substitution (bpsm) and insertion-deletion (indel) mutations from each of the *Burkholderia cenocepacia* mutation accumulation lineages.** The number of missed bpsms and indels are calculated as the number of unanalyzed sites in each lineage, multiplied by the product of the number of generations experienced per lineage and the experiment-wide estimates of bpsm and indel rates, respectively.

| | Identified Mutations | | | False Negative Mutations | | | |
|---|---|---|---|---|---|---|---|
| Lineage | Analyzed Sites | Verified Bpsms | Verified Indels | Unanalyzed Sites | Missed Bpsms | Missed Indels | Total Mutations |
| 1 | 7540220 | 8 | 2 | 162620 | 0.12 | 0.02 | 10.14 |
| 2 | 7545560 | 9 | 0 | 157280 | 0.11 | 0.02 | 9.13 |
| 4 | 7414821 | 3 | 1 | 288019 | 0.21 | 0.04 | 4.25 |
| 6 | 7205001 | 3 | 1 | 497839 | 0.36 | 0.07 | 4.43 |
| 8 | 6974428 | 3 | 1 | 728412 | 0.53 | 0.10 | 4.63 |
| 10 | 7160177 | 7 | 1 | 542663 | 0.39 | 0.07 | 8.46 |
| 11 | 7261465 | 5 | 2 | 441375 | 0.32 | 0.06 | 7.38 |
| 12 | 7305309 | 3 | 1 | 397531 | 0.29 | 0.05 | 4.34 |
| 13 | 6587366 | 2 | 1 | 1115474 | 0.81 | 0.15 | 3.96 |
| 14 | 7545418 | 7 | 1 | 157422 | 0.11 | 0.02 | 8.13 |
| 15 | 7443040 | 4 | 2 | 259800 | 0.19 | 0.03 | 6.22 |
| 20 | 7305503 | 4 | 0 | 397337 | 0.29 | 0.05 | 4.34 |
| 23 | 7375775 | 9 | 0 | 327065 | 0.24 | 0.04 | 9.28 |
| 24 | 7277839 | 5 | 1 | 425001 | 0.31 | 0.06 | 6.37 |
| 26 | 6340689 | 3 | 2 | 1362151 | 0.99 | 0.18 | 6.17 |
| 27 | 7372095 | 5 | 1 | 330745 | 0.24 | 0.04 | 6.28 |
| 28 | 7250123 | 9 | 0 | 452717 | 0.33 | 0.06 | 9.39 |
| 29 | 6790602 | 3 | 0 | 912238 | 0.66 | 0.12 | 3.78 |
| 30 | 7519782 | 7 | 0 | 183058 | 0.13 | 0.02 | 7.15 |
| 31 | 7657324 | 3 | 0 | 45516 | 0.03 | 0.01 | 3.04 |
| 32 | 7097422 | 4 | 0 | 605418 | 0.44 | 0.08 | 4.52 |
| 33 | 7591755 | 10 | 3 | 111085 | 0.08 | 0.01 | 13.09 |
| 34 | 7507586 | 4 | 3 | 195254 | 0.14 | 0.03 | 7.17 |
| 35 | 7374130 | 2 | 0 | 328710 | 0.24 | 0.04 | 2.28 |
| 36 | 7422439 | 6 | 1 | 280401 | 0.20 | 0.04 | 7.24 |
| 37 | 7376233 | 8 | 1 | 326607 | 0.24 | 0.04 | 9.28 |
| 41 | 6859606 | 4 | 1 | 843234 | 0.61 | 0.11 | 5.72 |
| 43 | 7526671 | 4 | 2 | 176169 | 0.13 | 0.02 | 6.15 |
| 45 | 7299446 | 4 | 1 | 403394 | 0.29 | 0.05 | 5.34 |
| 47 | 7578997 | 5 | 0 | 123843 | 0.09 | 0.02 | 5.11 |
| 49 | 7474089 | 7 | 0 | 228751 | 0.17 | 0.03 | 7.2 |
| 51 | 7657605 | 6 | 2 | 45235 | 0.03 | 0.01 | 8.04 |
| 53 | 7588758 | 2 | 0 | 114082 | 0.08 | 0.02 | 2.1 |
| 56 | 7532631 | 7 | 0 | 170209 | 0.12 | 0.02 | 7.14 |
| 57 | 7504288 | 8 | 1 | 198552 | 0.14 | 0.03 | 9.17 |
| 59 | 7455873 | 5 | 0 | 246967 | 0.18 | 0.03 | 5.21 |
| 62 | 7421688 | 5 | 1 | 281152 | 0.20 | 0.04 | 6.24 |
| 63 | 7554439 | 6 | 2 | 148401 | 0.11 | 0.02 | 8.13 |
| 65 | 7632258 | 8 | 2 | 70582 | 0.05 | 0.01 | 10.06 |
| 68 | 7459520 | 6 | 1 | 243320 | 0.18 | 0.03 | 7.21 |
| 69 | 7496022 | 7 | 1 | 206818 | 0.15 | 0.03 | 8.18 |
| 71 | 7599340 | 5 | 0 | 103500 | 0.08 | 0.01 | 5.09 |
| 74 | 7663293 | 6 | 3 | 39547 | 0.03 | 0.01 | 9.04 |

**Table D.3. Selection coefficients (s) and two-tailed t-statistics for each of the forty-three *Burkholderia cenocepacia* mutation accumulation lineages in each of the environments studied.** Uncorrected p-values and Benjamini-Hochberg corrected p-values are both provided.

| Environment | Lineage | s | SEM | T | DF | P | P (BH) |
|---|---|---|---|---|---|---|---|
| T-Soy | 1 | -0.0117 | 0.0039 | -3.4621 | 3 | 0.0406 | 0.0831 |
| T-Soy | 2 | -0.0445 | 0.0210 | -2.4470 | 3 | 0.0919 | 0.1464 |
| T-Soy | 4 | 0.0131 | 0.0111 | 1.3553 | 3 | 0.2683 | 0.3205 |
| T-Soy | 6 | -0.0739 | 0.0084 | -10.2146 | 3 | 0.0020 | 0.0143 |
| T-Soy | 8 | -0.0265 | 0.0091 | -3.3814 | 3 | 0.0430 | 0.0841 |
| T-Soy | 10 | -0.0432 | 0.0044 | -11.3750 | 3 | 0.0015 | 0.0157 |
| T-Soy | 11 | 0.0212 | 0.0103 | 2.3666 | 3 | 0.0988 | 0.1465 |
| T-Soy | 12 | -0.0278 | 0.0077 | -4.1930 | 3 | 0.0247 | 0.0560 |
| T-Soy | 13 | 0.0155 | 0.0089 | 2.0107 | 3 | 0.1379 | 0.1853 |
| T-Soy | 14 | -0.1001 | 0.0103 | -11.2292 | 3 | 0.0015 | 0.0130 |
| T-Soy | 15 | -0.0519 | 0.0107 | -5.5779 | 3 | 0.0114 | 0.0288 |
| T-Soy | 20 | -0.0039 | 0.0065 | -0.7025 | 3 | 0.5330 | 0.5590 |
| T-Soy | 23 | -0.0049 | 0.0075 | -0.7507 | 3 | 0.5073 | 0.5454 |
| T-Soy | 24 | -0.0079 | 0.0102 | -0.8952 | 3 | 0.4366 | 0.4941 |
| T-Soy | 26 | -0.0648 | 0.0017 | -43.7280 | 3 | 0.0000 | 0.0011 |
| T-Soy | 27 | -0.0270 | 0.0036 | -8.6163 | 3 | 0.0033 | 0.0157 |
| T-Soy | 28 | -0.1115 | 0.0131 | -9.8543 | 3 | 0.0022 | 0.0119 |
| T-Soy | 29 | -0.0144 | 0.0071 | -2.3350 | 3 | 0.1017 | 0.1458 |
| T-Soy | 30 | -0.0097 | 0.0040 | -2.7738 | 3 | 0.0693 | 0.1193 |
| T-Soy | 31 | -0.0037 | 0.0049 | -0.8825 | 3 | 0.4425 | 0.4879 |
| T-Soy | 32 | -0.0354 | 0.0073 | -5.5848 | 3 | 0.0113 | 0.0305 |
| T-Soy | 33 | -0.0323 | 0.0059 | -6.2940 | 3 | 0.0081 | 0.0232 |
| T-Soy | 34 | -0.0143 | 0.0017 | -9.9464 | 3 | 0.0022 | 0.0133 |
| T-Soy | 35 | -0.0503 | 0.0076 | -7.6021 | 3 | 0.0047 | 0.0185 |
| T-Soy | 36 | -0.0113 | 0.0058 | -2.2572 | 3 | 0.1092 | 0.1515 |
| T-Soy | 37 | -0.0125 | 0.0095 | -1.5233 | 3 | 0.2251 | 0.2846 |
| T-Soy | 41 | -0.0046 | 0.0130 | -0.4120 | 3 | 0.7080 | 0.7248 |
| T-Soy | 43 | -0.0147 | 0.0020 | -8.2935 | 3 | 0.0037 | 0.0158 |
| T-Soy | 45 | -0.0426 | 0.0156 | -3.1610 | 3 | 0.0508 | 0.0950 |
| T-Soy | 47 | -0.0381 | 0.0036 | -12.2712 | 3 | 0.0012 | 0.0167 |
| T-Soy | 49 | -0.0726 | 0.0033 | -25.5106 | 3 | 0.0001 | 0.0028 |
| T-Soy | 51 | -0.0178 | 0.0028 | -7.2197 | 3 | 0.0055 | 0.0196 |
| T-Soy | 53 | -0.0092 | 0.0076 | -1.3940 | 3 | 0.2576 | 0.3165 |
| T-Soy | 56 | -0.0152 | 0.0091 | -1.9428 | 3 | 0.1473 | 0.1919 |
| T-Soy | 57 | -0.0114 | 0.0055 | -2.3800 | 3 | 0.0976 | 0.1499 |
| T-Soy | 59 | 0.0076 | 0.0096 | 0.9167 | 3 | 0.4269 | 0.4961 |
| T-Soy | 62 | 0.0020 | 0.0094 | 0.2488 | 3 | 0.8196 | 0.8196 |
| T-Soy | 63 | -0.0338 | 0.0140 | -2.7804 | 3 | 0.0690 | 0.1236 |
| T-Soy | 65 | -0.0171 | 0.0076 | -2.5902 | 3 | 0.0811 | 0.1341 |
| T-Soy | 68 | -0.0486 | 0.0082 | -6.8805 | 3 | 0.0063 | 0.0193 |
| T-Soy | 69 | -0.0148 | 0.0024 | -7.1234 | 3 | 0.0057 | 0.0188 |
| T-Soy | 71 | -0.0233 | 0.0063 | -4.2732 | 3 | 0.0235 | 0.0562 |
| T-Soy | 74 | 0.0371 | 0.0104 | 4.1282 | 3 | 0.0258 | 0.0554 |
| M9-MM+CAA | 1 | -0.0026 | 0.0039 | -0.7748 | 3 | 0.4949 | 0.5457 |
| M9-MM+CAA | 2 | 0.0023 | 0.0023 | 1.1499 | 3 | 0.3335 | 0.5122 |
| M9-MM+CAA | 4 | 0.0060 | 0.0061 | 1.1463 | 3 | 0.3348 | 0.4965 |
| M9-MM+CAA | 6 | -0.0397 | 0.0072 | -6.3789 | 3 | 0.0078 | 0.0305 |
| M9-MM+CAA | 8 | 0.0021 | 0.0025 | 0.9685 | 3 | 0.4042 | 0.5112 |
| M9-MM+CAA | 10 | -0.0888 | 0.0072 | -14.1932 | 3 | 0.0008 | 0.0065 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| M9-MM+CAA | 11 | -0.0033 | 0.0046 | -0.8372 | 3 | 0.4639 | 0.5392 |
| M9-MM+CAA | 12 | -0.0108 | 0.0059 | -2.0927 | 3 | 0.1275 | 0.2610 |
| M9-MM+CAA | 13 | -0.0055 | 0.0058 | -1.0929 | 3 | 0.3543 | 0.4762 |
| M9-MM+CAA | 14 | -0.0223 | 0.0043 | -5.9552 | 3 | 0.0095 | 0.0339 |
| M9-MM+CAA | 15 | -0.0884 | 0.0078 | -13.1076 | 3 | 0.0010 | 0.0059 |
| M9-MM+CAA | 20 | -0.0054 | 0.0038 | -1.6581 | 3 | 0.1959 | 0.3509 |
| M9-MM+CAA | 23 | -0.0115 | 0.0020 | -6.5474 | 3 | 0.0072 | 0.0311 |
| M9-MM+CAA | 24 | -0.0049 | 0.0051 | -1.1067 | 3 | 0.3492 | 0.5005 |
| M9-MM+CAA | 26 | -0.0121 | 0.0029 | -4.8932 | 3 | 0.0163 | 0.0502 |
| M9-MM+CAA | 27 | -0.0022 | 0.0030 | -0.8359 | 3 | 0.4645 | 0.5257 |
| M9-MM+CAA | 28 | -0.1161 | 0.0072 | -18.6177 | 3 | 0.0003 | 0.0145 |
| M9-MM+CAA | 29 | -0.0126 | 0.0069 | -2.0962 | 3 | 0.1270 | 0.2731 |
| M9-MM+CAA | 30 | -0.0031 | 0.0059 | -0.6168 | 3 | 0.5810 | 0.6093 |
| M9-MM+CAA | 31 | -0.0223 | 0.0055 | -4.7190 | 3 | 0.0180 | 0.0517 |
| M9-MM+CAA | 32 | -0.0230 | 0.0022 | -12.0361 | 3 | 0.0012 | 0.0066 |
| M9-MM+CAA | 33 | -0.0595 | 0.0042 | -16.3089 | 3 | 0.0005 | 0.0054 |
| M9-MM+CAA | 34 | -0.0052 | 0.0065 | -0.9169 | 3 | 0.4268 | 0.5098 |
| M9-MM+CAA | 35 | -0.0279 | 0.0041 | -7.7713 | 3 | 0.0044 | 0.0212 |
| M9-MM+CAA | 36 | -0.0040 | 0.0034 | -1.3567 | 3 | 0.2679 | 0.4267 |
| M9-MM+CAA | 37 | -0.0030 | 0.0034 | -1.0005 | 3 | 0.3908 | 0.5092 |
| M9-MM+CAA | 41 | -0.0067 | 0.0038 | -2.0392 | 3 | 0.1342 | 0.2622 |
| M9-MM+CAA | 43 | -0.0090 | 0.0029 | -3.5854 | 3 | 0.0371 | 0.0841 |
| M9-MM+CAA | 45 | -0.0261 | 0.0023 | -13.2371 | 3 | 0.0009 | 0.0067 |
| M9-MM+CAA | 47 | -0.0876 | 0.0059 | -17.1446 | 3 | 0.0004 | 0.0062 |
| M9-MM+CAA | 49 | -0.0050 | 0.0038 | -1.5533 | 3 | 0.2182 | 0.3752 |
| M9-MM+CAA | 51 | -0.0120 | 0.0075 | -1.8478 | 3 | 0.1618 | 0.3025 |
| M9-MM+CAA | 53 | -0.0014 | 0.0021 | -0.7709 | 3 | 0.4969 | 0.5342 |
| M9-MM+CAA | 56 | -0.0029 | 0.0009 | -3.7905 | 3 | 0.0322 | 0.0770 |
| M9-MM+CAA | 57 | -0.0002 | 0.0041 | -0.0587 | 3 | 0.9569 | 0.9569 |
| M9-MM+CAA | 59 | -0.0042 | 0.0043 | -1.1056 | 3 | 0.3496 | 0.4849 |
| M9-MM+CAA | 62 | -0.0029 | 0.0034 | -0.9682 | 3 | 0.4044 | 0.4968 |
| M9-MM+CAA | 63 | -0.0869 | 0.0056 | -18.0102 | 3 | 0.0004 | 0.0080 |
| M9-MM+CAA | 65 | -0.0025 | 0.0053 | -0.5290 | 3 | 0.6335 | 0.6485 |
| M9-MM+CAA | 68 | -0.0190 | 0.0053 | -4.1269 | 3 | 0.0258 | 0.0653 |
| M9-MM+CAA | 69 | -0.0094 | 0.0020 | -5.3254 | 3 | 0.0129 | 0.0428 |
| M9-MM+CAA | 71 | -0.0076 | 0.0021 | -4.1397 | 3 | 0.0256 | 0.0688 |
| M9-MM+CAA | 74 | -0.0107 | 0.0082 | -1.5130 | 3 | 0.2275 | 0.3762 |
| M9-MM | 1 | -0.0007 | 0.0051 | -0.1551 | 3 | 0.8866 | 0.9775 |
| M9-MM | 2 | -0.0003 | 0.0070 | -0.0512 | 3 | 0.9624 | 1.0346 |
| M9-MM | 4 | -0.0026 | 0.0024 | -1.2488 | 3 | 0.3003 | 0.5166 |
| M9-MM | 6 | 0.0079 | 0.0107 | 0.8525 | 3 | 0.4566 | 0.6770 |
| M9-MM | 8 | 0.0259 | 0.0033 | 9.1263 | 3 | 0.0028 | 0.0133 |
| M9-MM | 10 | -0.0849 | 0.0051 | -19.1123 | 3 | 0.0003 | 0.0045 |
| M9-MM | 11 | 0.0063 | 0.0037 | 1.9570 | 3 | 0.1453 | 0.3123 |
| M9-MM | 12 | 0.0036 | 0.0037 | 1.1119 | 3 | 0.3473 | 0.5531 |
| M9-MM | 13 | -0.0030 | 0.0073 | -0.4674 | 3 | 0.6720 | 0.8757 |
| M9-MM | 14 | -0.0027 | 0.0074 | -0.4144 | 3 | 0.7064 | 0.8934 |
| M9-MM | 15 | -0.0891 | 0.0104 | -9.8618 | 3 | 0.0022 | 0.0119 |
| M9-MM | 20 | 0.0048 | 0.0070 | 0.7841 | 3 | 0.4902 | 0.6800 |
| M9-MM | 23 | -0.0024 | 0.0076 | -0.3674 | 3 | 0.7377 | 0.9063 |
| M9-MM | 24 | 0.0049 | 0.0039 | 1.4337 | 3 | 0.2471 | 0.4427 |
| M9-MM | 26 | -0.0341 | 0.0064 | -6.1857 | 3 | 0.0085 | 0.0305 |
| M9-MM | 27 | 0.0068 | 0.0051 | 1.5296 | 3 | 0.2236 | 0.4180 |
| M9-MM | 28 | -0.0662 | 0.0063 | -12.1133 | 3 | 0.0012 | 0.0104 |
| M9-MM | 29 | -0.0051 | 0.0011 | -5.3954 | 3 | 0.0125 | 0.0358 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| M9-MM | 30 | 0.0061 | 0.0024 | 2.9657 | 3 | 0.0593 | 0.1341 |
| M9-MM | 31 | 0.0092 | 0.0109 | 0.9740 | 3 | 0.4019 | 0.6172 |
| M9-MM | 32 | -0.0190 | 0.0045 | -4.8507 | 3 | 0.0167 | 0.0423 |
| M9-MM | 33 | -0.0812 | 0.0033 | -28.5657 | 3 | 0.0001 | 0.0041 |
| M9-MM | 34 | 0.0129 | 0.0028 | 5.3396 | 3 | 0.0128 | 0.0345 |
| M9-MM | 35 | -0.0146 | 0.0031 | -5.3994 | 3 | 0.0125 | 0.0382 |
| M9-MM | 36 | -0.0276 | 0.0053 | -5.9960 | 3 | 0.0093 | 0.0307 |
| M9-MM | 37 | -0.0002 | 0.0091 | -0.0251 | 3 | 0.9816 | 1.0049 |
| M9-MM | 41 | -0.0031 | 0.0043 | -0.8410 | 3 | 0.4621 | 0.6623 |
| M9-MM | 43 | 0.0060 | 0.0010 | 6.7421 | 3 | 0.0067 | 0.0260 |
| M9-MM | 45 | 0.0062 | 0.0020 | 3.6410 | 3 | 0.0357 | 0.0853 |
| M9-MM | 47 | -0.0903 | 0.0091 | -11.4516 | 3 | 0.0014 | 0.0088 |
| M9-MM | 49 | -0.0163 | 0.0011 | -16.6690 | 3 | 0.0005 | 0.0051 |
| M9-MM | 51 | 0.0001 | 0.0035 | 0.0392 | 3 | 0.9712 | 1.0186 |
| M9-MM | 53 | 0.0000 | 0.0060 | -0.0083 | 3 | 0.9939 | 0.9939 |
| M9-MM | 56 | -0.0009 | 0.0056 | -0.1963 | 3 | 0.8569 | 0.9697 |
| M9-MM | 57 | 0.0024 | 0.0090 | 0.3065 | 3 | 0.7793 | 0.9308 |
| M9-MM | 59 | -0.0045 | 0.0029 | -1.8087 | 3 | 0.1682 | 0.3444 |
| M9-MM | 62 | 0.0064 | 0.0010 | 7.2277 | 3 | 0.0055 | 0.0235 |
| M9-MM | 63 | -0.0659 | 0.0063 | -12.1051 | 3 | 0.0012 | 0.0087 |
| M9-MM | 65 | 0.0038 | 0.0071 | 0.6175 | 3 | 0.5806 | 0.7802 |
| M9-MM | 68 | -0.0342 | 0.0016 | -24.1572 | 3 | 0.0002 | 0.0033 |
| M9-MM | 69 | 0.0016 | 0.0060 | 0.3034 | 3 | 0.7814 | 0.9081 |
| M9-MM | 71 | -0.0031 | 0.0023 | -1.5408 | 3 | 0.2210 | 0.4320 |
| M9-MM | 74 | -0.0035 | 0.0035 | -1.1631 | 3 | 0.3289 | 0.5439 |