Asymmetric Context-Dependent Mutation Patterns Revealed through Mutation–Accumulation Experiments

Way Sung,*1 Matthew S. Ackerman,1 Jean-François Gout,1 Samuel F. Miller,1 Emily Williams,1 Patricia L. Foster,1 and Michael Lynch1

1Department of Biology, Indiana University, Bloomington

*Corresponding author: E-mail: waysung@indiana.edu.

Abstract

Despite the general assumption that site-specific mutation rates are independent of the local sequence context, a growing body of evidence suggests otherwise. To further examine context-dependent patterns of mutation, we amassed 5,645 spontaneous mutations in wild-type (WT) and mismatch-repair deficient (MMR−) mutation–accumulation (MA) lines of the gram-positive model organism Bacillus subtilis. We then analyzed >7,500 spontaneous base-substitution mutations across B. subtilis, Escherichia coli, and Mesoplasma florum WT and MMR− MA lines, finding a context-dependent mutation pattern that is asymmetric around the origin of replication. Different neighboring nucleotides can alter site-specific mutation rates by as much as 75-fold, with sites neighboring G:C base pairs or dimers involving alternating pyrimidine–purine and purine–pyrimidine nucleotides having significantly elevated mutation rates. The influence of context-dependent mutation on genome architecture is strongest in M. florum, consistent with the reduced efficiency of selection in organisms with low effective population size. If not properly accounted for, the disparities arising from patterns of context-dependent mutation can significantly influence interpretations of positive and purifying selection.

Key words: context-dependent mutation, mutation rate, mismatch repair, Bacillus subtilis.

Introduction

Mutations, which are a primary source of genetic variation, have been thought to be determined by relatively generic sequence properties. Thus, classical evolutionary models do not account for the possibility that the mutation process can be influenced by local sequence context (Kimura 1980; Tajima 1996; Yang 1996). Although a growing body of evidence from in vivo reporter-construct studies (Koch 1971), phylogenetic comparisons in pseudogenes (Bulmer 1986; Bains 1992; Blake et al. 1992; Hess et al. 1994), and genome-wide scans (Hwang and Green 2004; Duret 2009; Baele et al. 2010; Lee et al. 2012; Schaibley et al. 2013; Zhu et al. 2014) have suggested that neighboring nucleotides (context) have an effect on site-specific substitution rates, the true influence that context has on the mutation process remains a relative mystery for three reasons. First, polymorphisms used in phylogenetic comparisons or genome-wide scans are influenced by selection (Kuo and Ochman 2010), making context-dependent patterns difficult to interpret. Second, context-dependent mutation patterns derived from reporter-construct studies can be heavily influenced by the location and sequence of the reporter (Hawk et al. 2005). Third, because the base-substitution mutation rate is on the order of 10−11 to 10−8 per site per generation (Drake et al. 1998; Sung, Ackerman et al. 2012), most experimental evolution studies have provided only a limited amount of empirical data in which to analyze context-dependent mutation patterns (Denver et al. 2009; Ossowski et al. 2010; Sung, Tucker et al. 2012). These issues have prevented the integration of context-dependent mutation patterns into evolutionary models, potentially leading to inaccurate estimates of rates of molecular evolution as well as incorrect inferences regarding the magnitude of positive or purifying selection (Hernandez et al. 2007).

To further our understanding of the context-dependent mutation process, we applied a mutation–accumulation (MA) strategy to 50 wild type (WT) and 19 mismatch-repair deficient (MMR−) mutS− knockout lines of the free-living gram-positive bacterium Bacillus subtilis subsp. subtilis str NCIB3610 (hereafter B. subtilis). The MA strategy uses repeated single-cell bottlenecks to minimize the efficiency of selection, which allows for the accumulation of all but the most deleterious mutations, thus providing an unbiased estimate of the rate and molecular spectrum of spontaneous mutations from the sequenced genomes (Haag-Liautard et al. 2008; Denver et al. 2009; Keightley et al. 2009; Kondrashov FA and Kondrashov AS 2010; Ossowski et al. 2010; Lee et al. 2012; Sung, Ackerman et al. 2012; Sung, Tucker et al. 2012; Lang et al. 2013; Schrider et al. 2013; Zhu et al. 2014). To accumulate a large number of mutations, the WT and MMR− MA lines were each propagated for ~5,080 and ~2,000 generations, respectively.

Results

Bacillus subtilis Mutation Rate

At the end of the MA process, the 50 WT and 19 MMR− B. subtilis MA lines were sequenced to ~100× coverage using high-throughput 100 bp paired-end Illumina sequencing. We identified 350 base-substitution mutations in the WT MA lines and 5,295 base-substitution mutations in the
MMR\(^{-}\) lines (fig. 1A and B), yielding a genome-wide base-substitution mutation rate of 3.28 (standard error of the mean \([\text{SEM}] = 0.22\) \( \times 10^{-10} \)) and 3.31 (\( \text{SEM} = 0.71\) \( \times 10^{-8} \)) per site per generation, respectively (fig. 1C, supplementary tables S1–S3, Supplementary Material online). Consistent with reporter-construct estimates in Bacillus species (Sasaki et al. 2000; Zeibell et al. 2007) and MMR\(^{-}\) MA experiments in Escherichia coli and Caenorhabditis elegans (Denver et al. 2005; Lee et al. 2012), the genome-wide base-substitution mutation rate of the MMR\(^{-}\) strain is \( \sim 100 \times \) greater than that of the WT strain (fig. 1C), an increase which is primarily due to a large elevation in the number of transition mutations.

The bottleneving process in MA experiments reduces the efficiency of selection to operate on new mutations, allowing for the accumulation of all but the most deleterious mutations. To test whether the efficiency of selection was minimized during MA propagation in this experiment, we determined whether the number of mutations at selectively constrained sites (nonsynonymous) and selectively unconstrained sites (synonymous) matched the random expectation. Given the codon usage in B. subtilis, we find that the ratio of nonsynonymous to synonymous mutations in the B. subtilis WT MA lines does not significantly differ from the random expectation of 3.17:1 (supplementary table S1, Sung et al. 2015).
Supplementary Material online, $\chi^2$ test, $P = 0.66$, df = 1), consistent with the idea that the efficiency of selection on the accumulated mutations was minimized during propagation of the B. subtilis WT MA lines. B. subtilis MMRR MA lines have a transition/transversion ratio that is 13.41-fold higher than that of WT lines (43.30:1 vs. 3.23:1), and because synonymous changes are mostly transitions, the expected ratio of nonsynonymous to synonymous mutations in the MMRR MA lines changes from 3.17:1 to 1.93:1. After taking this change into account, we find that the MA experiment generated an excess of nonsynonymous mutations (supplementary table S1, Supplementary Material online, $\chi^2$ test, $P = 2.82 \times 10^{-3}$, df = 1). Furthermore, across all coding sites, we find that there is an excess of mutations that have accumulated during the propagation of the MMRR MA lines (supplementary table S1, Supplementary Material online, $\chi^2$ test, $P = 5.62 \times 10^{-15}$, df = 1). Coding sites, in particular nonsynonymous sites, are generally under greater selective constraint, so an excess of mutations in these classes indicates that selection played a small role in eradicating mutations from the MMRR MA lines.

Context-Dependent Mutation Patterns

To study context-dependent mutation patterns, each base substitution found in the WT and MMRR MA lines was assigned to either the left or right replichore, defined as the left and right halves of the circular chromosome separated by the origin of replication (ORI) and terminus. We further partitioned each base substitution site into the 64 possible genomic triplets with respect to the 5’ and 3’ nucleotide of the premutated base on the leading-strand template (fig. 2A and B). Although we are unable to determine the strand from which the original base-substitution premutation arose, all base-substitution mutations are assigned to the leading-strand template so that identical triplets in different replichores are synthesized in the same manner. For example, in both replichores, figure 2A displays a G→X base-substitution mutation (where G is the leading-strand template and X is any other nucleotide) flanked by 5’A and 3’C on the leading strand. As a result, the base substitution is denoted in the “top” strand (the conventional nomenclature presentation found in the B. subtilis genome assembly) in the left replichore and in the “bottom” strand (the reverse complement of the top strand) in the right replichore. In both cases, this triplet is categorized as 5’A[G→X]C3’ with respect to the leading-strand template (fig. 2B), and the context-dependent mutation rate (defined as the mutations per occurrence of triplet in the genome) per site per generation of the center nucleotide of the triplet is given as $\mu_x$.

When categorized in this fashion, $\mu_x$ is bilaterally symmetrical around the ORI (asymmetric in the top and bottom strands), with reverse complementary triplets that are synthesized with the same 5’ and 3’ context in different replichores having a nearly identical $\mu_x$ in B. subtilis MMRR MA lines (fig. 2A and B, Pearson’s correlation, $r = 0.98$, $P = 2.20 \times 10^{-16}$, df = 62). The numbers of observed mutations in both B. subtilis MMRR and WT MA lines were not significantly different between the two replichores ($\chi^2$ test, MMRR $P = 0.27$, WT $P = 0.13$, df = 63). This finding is consistent with the large body of work showing an asymmetric strand-specific mutation process around the ORI (Lobry 1996; Frank and Lobry 1999; Lobry and Sueoka 2002), while going further in showing that the strand-specific replication errors are highly dependent on the local sequence context. An asymmetric context-dependent mutation process is also highly consistent with asymmetric nucleotide composition that have been observed in the leading and lagging strand (GC-skew) on either side of bacterial ORIs (McLean et al. 1998; Tillier and Collins 2000; Arakawa and Tomita 2007; Marin and Xia 2008), which are often used to determine the location of the ORI in different bacteria. When the data for each replichore are pooled, $\mu_x$ for the 64 genomic triplets are highly correlated between MMRR and WT lines (fig. 3A, Pearson’s correlation, $r = 0.75$, $P = 9.28 \times 10^{-11}$, df = 62). This result holds despite the fact that when comparing MMRR to WT MA lines, the mutation rate is 100× greater, the transition/transversion ratio is 13.4× greater, and there is variation in the proportion of mutations in different types of sites across the genome (intergenic, coding, synonymous, and nonsynonymous). This result suggests that, at least in B. subtilis, MMR operates in a generally impartial fashion, such that the effect of adjacent nucleotides on post-MMR mutations are similar to pre-MMR DNA replication errors.

If the mutation process were context-independent, $\mu_x$ would not be affected by the surrounding nucleotides. For example, we should observe no significant difference in $\mu_x$ for the following set of triplets (context) with the same 3’ nucleotide but a different 5’ nucleotide: 5’T[A→X]G3’, 5’G[A→X]G3’, 5’C[A→X]G3’, and 5’A[A→X]G3’. However, in B. subtilis MA MMRR lines, we find that $\mu_x$ for these four triplets with the identical substitution type (A→X) are significantly different from the null expectation based on the number of times each triplet occurs in the genome (fig. 2B, supplementary table S4, Supplementary Material online). Furthermore, we find that for all types of substitutions (A | C | G | T→X), $\mu_x$ is significantly different when the 5’ or 3’ neighboring nucleotide is altered and the other neighboring nucleotide remains unchanged ($\chi^2$ test, $P < 0.05$, df = 3, supplementary table S4, Supplementary Material online, 32/32 contexts). Separating the 350 B. subtilis WT MA mutations into the 64 possible triplets severely limits the resolution of contextual effects, yet we still find that $\mu_x$ is significantly different than expected given the genome-wide triplet content of B. subtilis for 11/32 contexts (Fisher’s exact test, $P < 0.05$, df = 3, supplementary table S4, Supplementary Material online). Thus, base-substitution mutations are context dependent, with different 5’ or 3’ nucleotides capable of elevating $\mu_x$ for the same mutation type by as much as 75-fold in MMRR B. subtilis and 10-fold in WT B. subtilis (fig. 2B, supplementary table S5, Supplementary Material online). Although there is a general correlation between the pattern of context-dependent mutation rates of MMRR and WT MA lines (fig. 3A), the triplets 5’C[A→X]A3’, 5’T[T→X]G3’, 5’G[G→X]C3’, and 5’A[T→X]G3’ appear overrepresented in the MMRR MA lines (supplementary table S4, Supplementary Material online). One possible explanation
for this difference is that MMR is known to interact with the replication fork in *B. subtilis* (Klocko et al. 2011), and inactivation of MMR enzymes may drive differential replication errors at certain contexts.

Given the general similarities between context-dependent mutation patterns in MMR− and WT lines (fig. 3A) and the statistical power provided by the number of mutations in the MMR− lines, we used the MMR− lines to initially identify two nucleotide motifs specifically associated with the elevation of \( \mu_x \) in *B. subtilis*. First, triplets with a strong base (S: G or C) on the 5′ side exhibit a significantly elevated \( \mu_x \) compared with triplets with weak base (W: A or T) on the 5′ side, and again when the triplet involves a strong base on the 3′ side (fig. 1D, supplementary table S7, Supplementary Material online).

Second, triplets containing pyr-pur (pyrimidine–purine) or pur-pyr (purine–pyrimidine) dimers (where the mutated base is within the dimer) have a significantly elevated \( \mu_x \) when compared with mutated bases involving pyr-pyr or pur-pur dimers (fig. 1D, supplementary table S7, Supplementary Material online). The same triplet combinations are significantly elevated in \( \mu_x \) in *B. subtilis* WT MA lines (supplementary table S7, Supplementary Material online). The strong flanking and pur-pyr/pyr-pur motifs contributing to an elevation of \( \mu_x \) in *B. subtilis* are consistent with the observation of elevated (C→X) mutations in both AC/TG dimers when the T is templating the leading strand, and GC/CG dimers when the G is templating the leading strand in *E. coli* (Lee et al. 2012). This observation suggests that the

**Fig. 2.** Bilaterally symmetrical context-dependent mutation patterns in *Bacillus subtilis* MMR− MA lines. A. Bidirectional fork at ORI displaying a 5′A-[G→X]-C3′ triplet in the leading-strand template. B. Heatmap of conditional base-substitution mutation rate (\( \mu_x \)) of 64 possible triplet combinations in the left and right replichores, corresponding to the 5′ nucleotide, original nucleotide, and 3′ nucleotide with respect to the leading-strand template (Pearson’s correlation of the same triplet in each replichore, \( r = 0.98, P = 2.20 \times 10^{-16}, df = 62 \)).
Context-dependent mutation patterns may arise from elevated mutation rates at certain motifs. For example, deoxyadenosine methylase (Dam) and DNA cytosine methylase (Dcm) are well-studied enzymes that methylate G[A]TC and C[C]GG motifs for strand identification and gene regulation. Methylated adenines and cytosines are biochemically prone to depurination and deamination and can result in G(\(\rightarrow\)T]TC transversions and C[C\(\rightarrow\)T]GG transitions at these motifs. Although no known Dam and Dcm genes exist in \textit{B. subtilis} (Dreiseikelmann and Wackernagel 1981), we searched for an excess of methylation-related mutations at canonical Dam and Dcm motifs in both WT and MMR
\textsuperscript{-} MA lines. In the WT MA lines, no (0/28) A:T \(\textless\) T:A transversions or (0/27) A:T \(\textgreater\) C:G transversions were associated with GATC motifs, and no (0/132) G:C \(\textgreater\) A:T transitions were associated with CCA/TGG motifs. In the MMR
\textsuperscript{-} MA lines, 1.6\% (1/62) of all A:T \(\textless\) T:A transversions and no (0/43) A:T \(\textless\) C:G transversions were associated with GATC motifs, and 0.3\% (9/2,614) of all G:C \(\textgreater\) A:T transitions were associated with CCA/TGG motifs. GATC motifs incorporate 1.5\% (1,800/1,187,744) of all adenines in the genome, and CCA/TGG motifs incorporate 0.3\% (3,256/918,965) of all cytosines in the genome, so the expected number of mutations at these sites do not significantly differ from random expectation ($\chi^2$ test, $P = 1$, df = 1). Taken together, Dam and Dcm methylation do not appear to be driving context-dependent mutation patterns in \textit{B. subtilis}.

Organism-Specific Similarities and Differences in Context-Dependent Mutation Patterns

The mechanisms involved in DNA synthesis and repair are highly conserved across life, and we observe some consistent context-dependent mutation patterns across multiple organisms.

Analysis of context-dependent mutation patterns in \textit{Mesoplasma florum} (Sung, Ackerman et al. 2012), a bacterium

![Fig. 3](http://mbe.oxfordjournals.org/)

**Fig. 3.** \textit{Bacillus subtilis} MMR
\textsuperscript{-} context-dependent mutation patterns compared with WT and other organisms. A. Log–log plot showing the correlation between context-dependent mutation rates for identical triplets in MMR
\textsuperscript{-} and WT \textit{B. subtilis} MA lines (Pearson's correlation, $r = 0.75$, $P = 9.28 \times 10^{-11}$, df = 62). B. Log–log plot displaying the relationship between the context-dependent mutation rates of identically synthesized triplets in the left and right replications of MMR
\textsuperscript{-} MA lines of \textit{Mesoplasma florum}, \textit{B. subtilis}, and \textit{Escherichia coli}. The joint linear regression with equation log$_{10}y = 0.87-0.99\log_{10}x$ includes all points ($r^2 = 0.77$, $P < 1 \times 10^{-6}$, df = 162).
that lacks MMR, and *E. coli* mutT*−* MMR− MA lines (Lee et al. 2012) reveals a consistent symmetric context-dependent mutation pattern around the ORI (fig. 3B, supplementary table S6, Supplementary Material online). Pur-pyr or pyr-pur and pur-pyr-pur or pyr-pur-pyr motifs dimers significantly elevate $\mu_\text{X}$ across all analyzed organisms (supplementary table S7, Supplementary Material online), with the exception of *E. coli* WT lines which show a nonsignificant elevation of $\mu_\text{X}$ (significance limited by the mutation sample size). On the other hand, we do observe species-specific differences in context-dependent mutation patterns. Strong base pairing of adjacent nucleotides which significantly elevates $\mu_\text{X}$ in *B. subtilis*, also significantly elevates $\mu_\text{X}$ in *E. coli* MA MMR− lines (slight elevation *E. coli* MA WT lines), but not in *M. florum* (supplementary table S7, Supplementary Material online). We find that the influence of adjacent strong base pairing depends on the directionality (5′ or 3′) of the adjacent nucleotide. For example, we find a significant elevation in $\mu_\text{X}$ when a G is 3′ to a mutation at a T site (5′N[T→X]G3′ > 5′G[T→X]N3′ where N is any leading-strand template nucleotide, supplementary table S7, Supplementary Material online) in *B. subtilis*, but find that the elevation occurs when the strong base pairing occurs on the opposite side in *E. coli* (5′G[T→X]N3′ > 5′N[T→X]G3′, supplementary table S7, Supplementary Material online). Although further investigation is necessary, the observed strand-specific differences in context-dependent mutation patterns may be driven by strand-specific differences in synthesis and repair across organisms. For example, DnaE is responsible for both leading and lagging-strand synthesis in *E. coli*, whereas in *B. subtilis*, PolC is solely responsible for leading-strand synthesis, whereas DnaE (which lacks proofreading capability in *B. subtilis*) participates in lagging-strand synthesis (McHenry 2011; Timinskas et al. 2014). Alternatively, different repair enzymes are absent in each organism (*mutS* − *B. subtilis*, mutT*−* *E. coli*, and *M. florum* which lacks both enzymes), and knockout of a single part of the MMR pathway may leave MMR with some minor functionality that may lead to differences in context-dependent mutation patterns.

Even considering the enzymatic differences in synthesis and repair, one consistent pattern can be observed across all three bacteria. The conditional mutation rate of the central nucleotide in most pur-pyr-pur and pyr-pur-pyr triplets is elevated compared with pur-pyr-pur and pyr-pyr-pyr triplets (supplementary table S7, Supplementary Material online). Phylogenetic comparisons have also identified pyr-pur-pyr patterns may be driven by strand-specific differences in context-dependent mutation patterns. To evaluate how context-dependent mutation effects can influence the evolution of genome-wide nucleotide composition in organisms with small $N_e$ we determined the expected genome composition of WT lines of *B. subtilis*, *E. coli*, and *M. florum* at context-dependent mutation equilibrium by computer simulations using the context-dependent mutation spectrum derived from the corresponding MA experiment (supplementary tables S5 and S6, Supplementary Material online). A plot of the current genome-wide count of each triplet against the expected genome-wide count of each triplet at context-dependent mutation equilibrium shows that *M. florum* has a current triplet distribution highly correlated with its equilibrium neutral expectation (fig. 4, supplementary table S8, Supplementary Material online, $r^2 = 0.90$, $P < 1 \times 10^{-6}$, df = 62), whereas *B. subtilis* and *E. coli* do not (supplementary table S8, Supplementary Material online, *B. subtilis*: $r^2 = 0.1$, $P = 0.2$, df = 62, *E. coli*: $r^2 = 0$, $P = 0.86$, df = 62). The simulation result using context-dependent mutation patterns from a randomly selected sample of 350 mutations from the *M. florum* dataset (same number of mutations as the *B. subtilis* WT MA line) also yield high correlation (supplementary table S8, Supplementary Material online, *M. florum*: 350, $r^2 = 0.91$, $P < 1 \times 10^{-6}$, df = 62). Using direct estimates of the mutation rate ($u$) derived from MA experiments (Lee et al. 2012; Sung, Ackerman

![Image](https://example.com/image.png)

**FIG. 4.** Observed correlation between existing genome-wide triplet count and expected genome-wide triplet usage at context-dependent mutation equilibrium in *Mesoplasma florum*. Log–log plot showing the relationship of the current genome-wide count of the 64 possible nucleotide triplets against the genome-wide count of the 64 possible nucleotide triplets at context-dependent mutation equilibrium in *M. florum*. Linear regression with equation $\log_{10} y = 2.80 + 0.37 \log_{10} x$ ($r^2 = 0.90$, $P < 1 \times 10^{-6}$, df = 62).

**Influence of Context-Dependent Mutation Pressure on Genome Nucleotide Composition**

Because the efficiency of selection is minimized in organisms with a small effective population size ($N_e$), the evolution of their genome nucleotide composition should reflect neutral processes more than in high-$N_e$ species (Lynch and Conery 2003). Although most bacterial genomes have been suggested to be compositionally different from equilibrium expectations based on mutation alone (Rocha et al. 2006), prior studies have not taken into account context-dependent mutation patterns. To evaluate how context-dependent mutation effects can influence the evolution of genome-wide nucleotide composition in organisms with small $N_e$ we determined the expected genome composition of WT lines of *B. subtilis*, *E. coli*, and *M. florum* at context-dependent mutation equilibrium by computer simulations using the context-dependent mutation spectrum derived from the corresponding MA experiment (supplementary tables S5 and S6, Supplementary Material online). A plot of the current genome-wide count of each triplet against the expected genome-wide count of each triplet at context-dependent mutation equilibrium shows that *M. florum* has a current triplet distribution highly correlated with its equilibrium neutral expectation (fig. 4, supplementary table S8, Supplementary Material online, $r^2 = 0.90$, $P < 1 \times 10^{-6}$, df = 62), whereas *B. subtilis* and *E. coli* do not (supplementary table S8, Supplementary Material online, *B. subtilis*: $r^2 = 0.1$, $P = 0.2$, df = 62, *E. coli*: $r^2 = 0$, $P = 0.86$, df = 62). The simulation result using context-dependent mutation patterns from a randomly selected sample of 350 mutations from the *M. florum* dataset (same number of mutations as the *B. subtilis* WT MA line) also yield high correlation (supplementary table S8, Supplementary Material online, *M. florum*: 350, $r^2 = 0.91$, $P < 1 \times 10^{-6}$, df = 62). Using direct estimates of the mutation rate ($u$) derived from MA experiments (Lee et al. 2012; Sung, Ackerman

**Influence of Context-Dependent Mutation Pressure on Genome Nucleotide Composition**

Because the efficiency of selection is minimized in organisms with a small effective population size ($N_e$), the evolution of their genome nucleotide composition should reflect neutral processes more than in high-$N_e$ species (Lynch and Conery 2003). Although most bacterial genomes have been suggested to be compositionally different from equilibrium expectations based on mutation alone (Rocha et al. 2006), prior studies have not taken into account context-dependent mutation patterns. To evaluate how context-dependent mutation effects can influence the evolution of genome-wide nucleotide composition in organisms with small $N_e$ we determined the expected genome composition of WT lines of *B. subtilis*, *E. coli*, and *M. florum* at context-dependent mutation equilibrium by computer simulations using the context-dependent mutation spectrum derived from the corresponding MA experiment (supplementary tables S5 and S6, Supplementary Material online). A plot of the current genome-wide count of each triplet against the expected genome-wide count of each triplet at context-dependent mutation equilibrium shows that *M. florum* has a current triplet distribution highly correlated with its equilibrium neutral expectation (fig. 4, supplementary table S8, Supplementary Material online, $r^2 = 0.90$, $P < 1 \times 10^{-6}$, df = 62), whereas *B. subtilis* and *E. coli* do not (supplementary table S8, Supplementary Material online, *B. subtilis*: $r^2 = 0.1$, $P = 0.2$, df = 62, *E. coli*: $r^2 = 0$, $P = 0.86$, df = 62). The simulation result using context-dependent mutation patterns from a randomly selected sample of 350 mutations from the *M. florum* dataset (same number of mutations as the *B. subtilis* WT MA line) also yield high correlation (supplementary table S8, Supplementary Material online, *M. florum*: 350, $r^2 = 0.91$, $P < 1 \times 10^{-6}$, df = 62). Using direct estimates of the mutation rate ($u$) derived from MA experiments (Lee et al. 2012; Sung, Ackerman

![Image](https://example.com/image.png)

**FIG. 4.** Observed correlation between existing genome-wide triplet count and expected genome-wide triplet usage at context-dependent mutation equilibrium in *Mesoplasma florum*. Log–log plot showing the relationship of the current genome-wide count of the 64 possible nucleotide triplets against the genome-wide count of the 64 possible nucleotide triplets at context-dependent mutation equilibrium in *M. florum*. Linear regression with equation $\log_{10} y = 2.80 + 0.37 \log_{10} x$ ($r^2 = 0.90$, $P < 1 \times 10^{-6}$, df = 62).
et al. 2012), and silent-site diversity ($\pi_s$) from population data (Sung, Ackerman et al. 2012), we have previously estimated the $N_e$ of *M. florum* to be $\sim 10^6$, an order of magnitude less than *B. subtilis*, and two orders of magnitude less than *E. coli*. Thus, assuming that these three organisms are at mutation-drift-selection equilibrium with respect to nucleotide usage, the reduced efficiency of selection on nucleotide composition in *M. florum* is consistent with the expectation for a species with depressed $N_e$ (Kimura 1983; Lynch and Conery 2003; Sung, Ackerman et al. 2012). It remains possible that if selection and mutation biases are identical, the same result can be generated. However, given the depressed $N_e$ in *M. florum*, and that only *M. florum* has a genome that is in context-dependent mutation equilibrium, we propose that organisms with lower $N_e$ (e.g., eukaryotes and endosymbionts [Moran et al. 2009; Sung, Ackerman et al. 2012]) will have genome-wide nucleotide compositions more strongly driven by context-dependent mutation patterns than those with high $N_e$.

**Discussion**

Two major results can be derived from this *B. subtilis* MA study. First, we provide a direct estimate of the rate and spectrum of genome-wide base substitutions for *B. subtilis* under both WT and MMR– conditions. This *B. subtilis* study allows for a direct comparison with parallel WT and MMR– data in the gram-negative bacterium *E. coli* (Lee et al. 2012). Consistent with *E. coli*, we find that MMR deficiency results in a significant elevation in mutation rate, in particular, for transitions. However, *E. coli* mutL– shows a 465-fold elevation in A:T $\gg$ G:C transitions when compared with WT lines (Lee et al. 2012), whereas *B. subtilis* mutS– displays an elevation of $\sim$100-fold in both A:T $\gg$ G:C and G:C $\gg$ AT transitions when compared with WT lines (fig. 1C). One possible explanation for this difference is that the MMR pathway differentially repairs transition errors in the two organisms. An alternative explanation is that a knockout of a single enzyme in the MMR pathway does not entirely preclude the remaining enzymes from participating in DNA repair, which may be the case, as different components of MMR was knocked out in each study (mutL– vs. mutS–). In the future, a double-knockout of both enzymes can provide details on whether the differential elevation of transition mutation types is due to incomplete knockout of the MMR system, or unique properties of repair in each organism. Unlike *E. coli* (Lee et al. 2012), we find that Dam or Dcm methylation does not appear to influence the mutation spectrum *B. subtilis*. Thus, organisms that directly methylate DNA for strand identification or gene expression will have elevated mutation rates at those motifs and may require evolution of additional DNA repair mechanisms to compensate (Sedgwick 2004).

The second major result of this study is to provide a detailed analysis on the influence of adjacent nucleotides on site-specific replication errors. Prior studies have observed asymmetrical substitution patterns in the two DNA strands of bacteria, and patterns of GC-skew surrounding the ORI and terminus (Lobry 1996; McLean et al. 1998; Tillier and Collins 2000; Lobry and Sueoka 2002; Arakawa and Tomita 2007; Marin and Xia 2008). Larger asymmetry in intergenic and synonymous sites, which are subject to a lesser degree of selective constraint, suggest that a mutational bias is responsible for the pattern. Furthermore, widely-variable substitution patterns across different organisms suggest that the mutation pattern is multifactorial (Rocha et al. 2006). Using MA lines, we show that not only are the mutations symmetric around the ORI, but that the fidelity of replication of a nucleotide is highly dependent on its local sequence context—the two nucleotides that are directly adjacent to it. Context-dependent mutation patterns have been previously observed at GC dinucleotides (CpG islands) and have been attributed to deamination of actively methylated cytosines (Beletskii and Bhagwat 1996; Lobry and Sueoka 2002). Furthermore, canonical motifs involved with Dcm methylation have been shown to influence mutation rates (Lee et al. 2012) and can generate context-dependent mutation patterns. Yet, *B. subtilis*, an organism that lacks the enzymes responsible for active methylation of cytosine (Dcm), also exhibits elevated mutation rates at cytosines flanked by a strong base pair (fig. 1D). Thus, although context-dependent mutation patterns are partly attributable to methyl-induced mutagenesis at specific motifs, the finding of context-dependent mutation patterns in an organism lacking these enzymes suggest that base-pairing and dimer-stacking interactions during replication also play a large role in replication errors and mutagenesis.

The existence of context-dependent mutation biases suggests a need to revisit models of molecular evolution used to estimate the rate at which one nucleotide site is replaced by another. Current evolutionary models correct for transition and transversion mutation biases when calculating evolutionary distance (King and Jukes 1969), but they do not incorporate context-dependent mutation processes, which can confer up to a $\sim$75-fold variation in site-specific mutation rates for similar substitution types (supplementary table S5, Supplementary Material online, $5'T[T\rightarrow X]T3'$ vs. $5'G[T\rightarrow X]G3'$). This oversimplification of the mutation models underlying the calculations of widely used measures, like synonymous (dS) and nonsynonymous (dN) substitution rates, might lead to erroneous estimations of these important evolutionary measurements. For example, synonymous sites that are adjacent to G:C base pairs or involving pyr-pur dimers have an elevated rate of mutation, and without taking context-dependent patterns into account, the interpretation of evolutionary distance at these sites will be inflated. Thus, understanding context-dependent mutation patterns are essential in studies which incorporate measures of evolutionary distance.

Context-dependent mutation bias can also impact how we interpret signatures of positive and purifying selection. Given the codon usage of *B. subtilis, E. coli,* and *M. florum,* and the context-dependent mutation patterns in these organisms, the expected ratios of nonsynonymous substitutions per nonsynonymous site to the synonymous substitutions per synonymous site (dN/dS) at neutrality in *B. subtilis, E. coli,* and *M. florum* are 0.63, 0.60, and 1.51, respectively. The measurement of dN/dS is often used as a proxy for signatures of selection, with dN/dS values $> 1.0$ being taken to indicate positive selection, and dN/dS values $< 1.0$ to be
indicative of purifying selection (Kimura 1977; Yang and Bielawski 2000). These results for B. subtilis and E. coli imply that dN/dS in the range 0.60 and 1.0 is actually consistent with a signature of positive selection, and not the usual interpretation of purifying selection (for all dN/dS < 1.0). Conversely, for M. florum, any observed dN/dS within the range of 1.00–1.51 is consistent with purifying selection, and not the usual interpretation of positive selection. Some evolutionary models attempt to correct for differences in mutational biases (e.g., transition/transversion ratio) when calculating dN/dS (Yang 2007), but generally ignore the possibility that mutations are strand-specific and context-dependent, potentially misconstruing signatures of both positive and purifying selection.

Patterns of context-dependent mutations can also alter the expected probability that a site will either generate the same mutation in multiple individuals (parallel mutations) or remain the same nucleotide over time. For instance, in B. subtilis WT MA lines, 5’T[T→X]T3’ sites on the leading-strand template have a mutation rate of 0.66 × 10^{-11} per site per generation, 5.97 times lower than the average T site (3.94 × 10^{-10} per site per generation, supplementary table S7, Supplementary Material online). On the other hand, 5’G[T→X]G3’ sites on the leading-strand template have a mutation rate of 9.84 × 10^{-10} per site per generation, 2.50 times greater than the average T site. Without taking context-dependent mutation patterns into account, the probability of observing parallel 5’G[T→X]G3’ mutations or no mutation at a 5’T[T→X]T3’ site is then 2.50× and 5.97× greater than expected, respectively. When comparing the two motifs, 5’G[T→X]G3’ mutations arise at 14.91 times more often than 5’T[T→X]T3’ mutations. Thus, our results show that context-dependent mutation patterns can have a large impact on the null expectation that a parallel mutation will or will not arise at a particular site.

An understanding of the context-dependent mutation process is essential to understanding patterns of organismal codon usage. The forces responsible for determining codon usage are assumed to be a balance between the selective advantage of translational efficiency and the mutational bias (Bulmer 1991; Sharp et al. 1993; McVean and Charlesworth 1999; Hershberg and Petrov 2008; Powdel et al. 2010; Sharp et al. 2010; Cusack et al. 2011; Liu 2012). However, deviations from the neutral expectation of codon usage will arise by not taking into account context-dependent mutation processes, which have been implicated to be a major factor in influencing codon usage in mitochondrial genomes (Jia and Higgs 2008) and weakly expressed genes (Powdel et al. 2010). As an example, GTG and GTT both code for the amino acid valine, and in B. subtilis, μv for 5’G[T→X]G3’ is 11.54-fold higher than 5’G[T→X]T3’. Because changes at the second position are nonsynonymous, GTG codons will naturally evolve faster than GTT codons simply because they are more mutable, and encoding valine with GTT will minimize the chance that a mutation will change the amino acid (mutational hazard). If codon usage is adapted to minimize mutational hazard, we would observe that the most frequently used codons mutate more often to an amino acid of similar charge and polarity. For example, the polar amino acid Threonine is encoded by ACT, ACC, ACA, and ACG. At the center nucleotide of ACT and ACC, C→A and C→G mutations result in an amino acid that remains polar, and a C→T mutation yields an amino acid that is nonpolar. C→A and C→G mutations at the center nucleotide in ACA and ACG yield an amino acid that is basic, and a C→T mutation yields an amino acid that is nonpolar. In an effort to minimize mutational protein hazard, it might be expected that B. subtilis may evolve low mutation rates for the contexts which change the property of the encoded amino acid, or minimize the use of amino acids which contexts have high mutation rates.

We analyzed whether the codon usage patterns minimized mutational protein hazard in B. subtilis, using context-dependent mutation rates from both WT and MMR MA lines. For each codon, we weigh the frequency of that codon in the genome with the context-dependent mutation rate to conservative (polarity and charge maintained) and nonconservative (polarity or charge differs) amino acid changes. For simplicity, we focus on the first frame of coding triplets, whereby a mutation at the center nucleotide (second position of the coding triplet) always results in an amino acid change. We find that the average context-dependent mutation rate per codon is significantly reduced at codons which yield nonconservative amino acid changes using context-dependent mutation patterns from both B. subtilis WT (t-test, P < 0.04, df = 109.60, supplementary table S10, Supplementary Material online) and MMR MA lines (t-test, P < 0.01, df = 99.52, supplementary table S10, Supplementary Material online). This result suggests that context-dependent mutation patterns are tuned to minimize nonconservative amino acid changes and that codon usage and context-dependent mutation patterns may be adapted to each other. Given that the genetic code and properties of amino acids are maintained across organisms, this further suggests that context-dependent mutation patterns may be universal. A larger study that integrates context-dependent mutation patterns with expression data may help decipher the individual contribution that mutation pressure and translational efficiency have on codon usage bias.

In summary, our finding of context-dependent mutation patterns has significant implications for the field of evolutionary genomics (Lee et al. 2012; Zhu et al. 2014). The effect that local sequence contexts can have on site-specific mutation rates suggests a need to reevaluate the current rate models of molecular evolution used to estimate signatures of positive and purifying selection, selective constraints on codon usage, likelihood of parallel mutations, and measurements of evolutionary distance. Further study of long-term MA lines is necessary to determine the degree to which closely related species share context-dependent mutation patterns, which would allow us to infer broader, more general, context-dependent mutation parameters for use in evolutionary models. Context-dependent mutation patterns appear to reflect the thermodynamic stability of both base pairing and dimer interactions, such that an increased local thermodynamic stability is capable of stabilizing mispaired nucleotides, which facilitates inaccurate proofreading during DNA
synthesis and repair. It is interesting to note that changes in salt concentration and temperature can alter the thermodynamic interactions associated with dimer stability (Yakovchuk et al. 2006), which suggests that organisms living in extreme environments may exhibit entirely different context-dependent mutation patterns. As context-dependent mutation bias appears to be strand-specific, understanding context-dependent mutation patterns will become even more complicated in eukaryotic organisms with multiple ORIs at unknown locations. Our work provides evidence that reverse complementary triplets in different replicores have predictable context-dependent mutation patterns, which may prove useful in developing future theoretical and empirical studies to identify replication origins and determine patterns of context-dependent mutations.

Methods
MA Process
Seventy-five independent B. subtilis sp. 3,610 WT (undomesticated strain 3,610) and MMR<sup>−</sup> (mutS:Tn10 strain DSS527) MA lines were each initiated from a single colony. Lines were grown at 37 °C on 100 × 15 mm petri dishes containing LB agar (tryptone/yeast extract/sodium chloride/agar). Every day, a single isolated B. subtilis colony from each MA line was bottlenecks by transferring to a fresh plate containing LB agar. The bottlenecking process ensures that mutations accumulate in an effectively neutral fashion, as previously demonstrated in prior MA studies (Kibota and Lynch 1996; Lee et al. 2012).

To estimate B. subtilis generation times, every month, an entire colony from five randomly selected WT MA lines were transferred to 1× PBS saline buffer. These suspensions were vortexed, serially diluted, and replated. Cell densities were calculated from viable cell counts, yielding an average generation time at 32 °C of 27.53 divisions every day over the entire experiment. After each transfer, the original plate was retained as a backup plate at 4 °C. If the destination plate was contaminated or we were unable to pick a single colony, we picked a single colony from the most recent backup plate available. Using the procedure described earlier, we also estimated an average generation time of 13.45/divisions/day for 2 days and 8.96/divisions/day for 4 days at 4 °C. We reached the final calculation for each MA line by the sum of the generations per divisions per day and the weighted sum of the generations per division per day if backups were used (supplementary table S2, Supplementary Material online). On average, the WT and MMR<sup>−</sup> MA lines were propagated for ~5,080 generations and ~2,000 generations, respectively.

DNA extraction from WT and MMR<sup>−</sup> B. subtilis sp. 3,610 MA lines using the wizard DNA extraction kit (Promega) were followed by phenol/chloroform extractions to Illumina library standards.

Sequencing and Alignment
We applied 101-bp paired-end Illumina (Illumina Hi-Seq platform) sequencing to 50 randomly selected B. subtilis sp. 3,610 WT MA lines and 19 B. subtilis sp. 3,610 MMR<sup>−</sup> MA lines. Each B. subtilis MA line was sequenced to a coverage depth of ~100× with an average library fragment size (distance between paired end reads) of ~175 bp. The paired-end reads for each B. subtilis MA line were individually mapped against the B. subtilis sp. 3,610 reference genome (assembly and annotation available from the National Center for Biotechnology Information, https://www.ncbi.nlm.nih.gov) using two separate alignment algorithms: BWA (Li and Durbin 2009) and NOVOALIGN (available at www.novocraft.com). The resulting pileup files were converted to SAM format using SAMTOOLS (Li et al. 2009). Using in-house Perl scripts, the alignment information was further parsed to generate forward and reverse mapping information at each site, resulting in a configuration of eight numbers for each line (A, a, C, c, G, g, T, t), corresponding to the number of reads mapped at each genomic position in the reference sequence.

Consensus Method
To identify putative mutations, each individual line (focal line) was compared with the consensus of all the remaining lines. This consensus approach is ideal with a large number of samples and low variance in coverage and is robust against sequencing or alignment errors in the reference genome. Previous application of the consensus method provided very low false-positive rates (Lynch et al. 2008; Denver et al. 2009; Ossowski et al. 2010).

The consensus approach employs three steps in mutation identification: 1) At each nucleotide position, the consensus is identified for each individual line, requiring 80% of the reads in a line to indicate the same nucleotide (A | C | T | G), with at least two forward and two reverse reads. 2) The overall consensus base call is identified, requiring 50% of the reads across all lines to indicate the same nucleotide (A | C | T | G). 3) The individual consensus for each line is compared against the overall consensus. If the line-specific consensus has a base call that differs from the overall consensus, and at least two other lines contained enough reads to be used in the comparison, the site was designated as a putative mutation for the discordant line.

Data Processing
In the WT lines, the consensus approach identified 365 base substitutions when applied to the BWA mapping output and 352 base substitutions when applied to the NOVOALIGN mapping output. Three hundred and fifty of the base substitutions overlapped between the two algorithms (supplementary table S2, Supplementary Material online). When closely examined, the remaining 17 base substitutions were either shared across all lines (not an MA-derived mutation) or directly adjacent to an indel, resulting from misalignment at the site. These 17 base substitutions were discarded. The same method was used to identify the 5,295 base substitutions in the 19 MMR<sup>−</sup> lines.

Mutation Verification
We designed primer sets to PCR amplify 300–500 bp regions surrounding randomly selected base substitutions...
One hundred percentage (69/69) of the base substitutions which could be amplified were directly confirmed using standard Sanger sequencing technology at the Indiana Molecular Biology Institute at Indiana University. For all cases, the WT nucleotide was also confirmed at the mutation site in at least one other line without the mutation.

**Mutation Rate Calculations**

To calculate the base-substitution mutation rate per cell division for each line, we used the following equation.

$$u_{bs} = \frac{m}{nT}$$

where $u_{bs}$ is the base-substitution mutation rate (per nucleotide site per generation), $m$ is the number of observed base substitutions, $n$ is the number of nucleotide sites analyzed, $N$ is the total number of lines, and $T$ is the number of generations that occurred in the MA line studied. The pooled standard error across all lines is given by

$$SE_{u_{bs}} = \sqrt{\frac{u_{bs}}{NnT}}$$

The context-dependent mutation rate ($u_x$) is given by the base-substitution mutation rate for the center nucleotide of a triplet divided by the observed number of triplets in the genome.

$$u_x = \frac{u_{bs} \text{ of center nucleotide of triplet}}{\# \text{triplet}}$$

The 95% confidence interval (CI) for a class of triplets (fig. 1D) is weighted by the number of observations for each triplet in the class. The equation for the 95% CI is given by the following equation, where $u_x$ is the mutation rate per site per generation at the center nucleotide of a triplet and $p_x$ is the count of that triplet in the genome.

$$95\% \text{ CI} = \pm 1.96 \times \sqrt{\frac{u_1 + u_2 + \ldots + u_x}{(p_1 + p_1 + \ldots p_x)}^2}$$

**Simulation of Context-Dependent Mutation Equilibrium**

To determine the context-dependent mutation equilibrium for each organism, we started with the current genome and saturated each of the three genomes with context-dependent mutations, effectively simulating the evolution of these genomes in the absence of natural selection. Each nucleotide site in the genome has a probability of mutating to one of the other three bases given the genome-wide context-dependent mutation rate measured in the MA experiment ($5'N[N\rightarrow X]N3'$, supplementary tables S5 and S6, Supplementary Material online). If no mutation rate is available for a particular context, a probability 10-fold lower than the lowest context-dependent mutation rate in that organism was assigned to that context. Starting from the organism’s current genome (reference genome assembly), we iterated across the entire genome, determining whether each site mutates to any of the other three bases. New mutations can influence the context of its surrounding nucleotides, so that mutation is immediately integrated into the new genome. Mutations were distributed recursively until the number of mutations per site exceeded three, ensuring saturation of the genome (supplementary table S8, Supplementary Material online).

**Data Access**

Illumina DNA sequences for the WT and MMR− B. subtilis MA lines used in this study are deposited under the Bioproject PRJNA256312 at the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA) http://www.ncbi.nlm.nih.gov/sra (last accessed March 26, 2015).

**Supplementary Material**

Supplementary tables S1–S10 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

The authors thank all members of the Lynch Lab for helpful discussion, Daniel Kearns for providing the B. subtilis stock strains, and Elizabeth Housworth for statistical advice. This work was supported by the Multidisciplinary University Research Initiative Award W911NF-09-1-0444 from the US Army Research Office to M.L., P.L.F., H. Tang, and S. Finkel and National Institutes of Health Award F32GM103164 to W.S. and R01 GM036827 to M.L. and W.K. Thomas. This material is based upon work supported by the National Science Foundation under Grant No. CNS-0521433, CNS-0723054, and ABI-1062432.

**References**


Bulmer M. 1986. Neighboring base effects on substitution rates in *B. subtilis* stock strains, and Elizabeth Housworth for statistical advice. This work was supported by the Multidisciplinary University Research Initiative Award W911NF-09-1-0444 from the US Army Research Office to M.L., P.L.F., H. Tang, and S. Finkel and National Institutes of Health Award F32GM103164 to W.S. and R01 GM036827 to M.L. and W.K. Thomas. This material is based upon work supported by the National Science Foundation under Grant No. CNS-0521433, CNS-0723054, and ABI-1062432.

**References**


Bulmer M. 1986. Neighboring base effects on substitution rates in *B. subtilis* stock strains, and Elizabeth Housworth for statistical advice. This work was supported by the Multidisciplinary University Research Initiative Award W911NF-09-1-0444 from the US Army Research Office to M.L., P.L.F., H. Tang, and S. Finkel and National Institutes of Health Award F32GM103164 to W.S. and R01 GM036827 to M.L. and W.K. Thomas. This material is based upon work supported by the National Science Foundation under Grant No. CNS-0521433, CNS-0723054, and ABI-1062432.

**References**


Bulmer M. 1986. Neighboring base effects on substitution rates in *B. subtilis* stock strains, and Elizabeth Housworth for statistical advice. This work was supported by the Multidisciplinary University Research Initiative Award W911NF-09-1-0444 from the US Army Research Office to M.L., P.L.F., H. Tang, and S. Finkel and National Institutes of Health Award F32GM103164 to W.S. and R01 GM036827 to M.L. and W.K. Thomas. This material is based upon work supported by the National Science Foundation under Grant No. CNS-0521433, CNS-0723054, and ABI-1062432.


