The Accelerated Evolution of Lagging Strand Genes Is Independent of Sequence Context

Christopher N. Merrikh, Eli Weiss, and Houra Merrikh*

Microbiology Department, University of Washington, Seattle

*Corresponding author: E-mail: merrikh@uw.edu.

Accepted: November 6, 2016

Data deposition: The genomic sequences of the B. subtilis strains analyzed here are available at Genbank under the accession numbers: AL009126, CP003329, CP002468, CP003695, CP002905, CP002183, CP002906, and CP004019.

Abstract

We previously discovered that lagging strand genes evolve faster in *Bacillus subtilis* (and potentially other bacteria). Lagging strand genes are transcribed in the head-on orientation with respect to DNA replication, leading to collisions between the two machineries that stall replication and can destabilize genomes. Our previous work indicated that the increased mutagenesis of head-on genes depends on transcription-coupled repair and the activity of an error prone polymerase which is likely activated in response to these collisions. Recently, it was proposed that sequence context is a major contributor to the increased mutagenesis and evolution of head-on genes. These models are based on laboratory-based evolution experiments performed in *B. subtilis*. However, critical evolutionary analyses of naturally occurring single nucleotide polymorphisms (SNPs) in wild strains were not performed. Using the genomic sequences from nine closely related wild *B. subtilis* strains, we analyzed over 200,000 naturally occurring SNPs as a proxy for natural mutation patterns for all genes and in particular, head-on genes. Our analysis suggests that (frame-independent) triplet sequence context can impact mutation rates: certain triplet sequences (TAG, CCC, CTA, and ACC) accumulate SNPs at a higher rate and are depleted from the genome. However, the triplet sequences previously identified as mutagenic in laboratory experiments (CCG, GCG, and CAC) do not have an elevated rate of SNP accumulation and are not depleted from the genome. Importantly, dN/dS analyses indicate that the accelerated evolution of head-on genes is not dependent on any particular triplet sequence. Thus, in agreement with our previous results, mutagenic transcription-coupled repair, rather than sequence context, is sufficient to explain the accelerated evolution of head-on genes.

Key words: replication-transcription conflicts, sequence context, accelerated evolution, transcription-coupled repair.

Introduction

Genes encoded on the lagging strand of DNA replication forks are transcribed in the opposite orientation (head-on) with respect to the movement of replication machinery, generating conflicts. As a consequence of these encounters, head-on genes mutate at a higher rate, and evolve faster than genes co-oriented with replication (Paul et al. 2013; Million-Weaver et al. 2015). The functional enrichment and increased prevalence of convergent mutations in head-on genes suggests that this mechanism is adaptive, and therefore has major implications for both genomic organization and the process of evolution (Paul et al. 2013).

We previously demonstrated that, when transcribed, reporter genes oriented head-on to replication acquire phenotype-reverting mutations at a higher rate than the same gene oriented co-directionally. The mechanistic basis for this increase in spontaneous mutagenesis involves the activity of the transcription-coupled nucleotide excision repair (TC-NER) pathway and the mutagenic, constitutively expressed, Y-family polymerase PolY1 (Million-Weaver et al. 2015). Though it is possible that PolY1 promotes mutations in head-on genes via an association with the replication fork, our epistasis analyses suggested that it functions as part of TC-NER (Million-Weaver et al. 2015).

The increased frequency of spontaneous mutations in head-on genes raised important implications for the evolution of bacteria. Therefore, as part of previous studies, we also examined single nucleotide polymorphisms (SNPs) from wild *Bacillus subtilis* strains. Our analyses indicated that head-on genes have a higher rate of non-synonymous mutations

(dN) than co-directional genes, suggesting that the increase in spontaneous mutations leads to a higher rate of evolution (Paul et al. 2013; Million-Weaver et al. 2015). These experiments also revealed a mutational footprint consistent with TC-NER and PolY1 driving the accelerated evolution of head-on genes (Million-Weaver et al. 2015).

Recently, other groups proposed that differential sequence context in head-on genes may lead to an inherently higher spontaneous mutation rate, thereby promoting their accelerated evolution (Chen and Zhang 2013; Schroeder et al. 2016). These studies demonstrated that the fidelity of nucleotide selection by the replication fork is influenced by the adjacent upstream and downstream nucleotides, that is, its "triplet context" (Sung et al. 2015; Schroeder et al. 2016). Both studies identified this effect in mismatch repair (MMR) mutants of *B. subtilis*, which have elevated mutation rates due to their inability to repair replication fork derived mismatches. They also observed that certain triplet DNA sequences are particularly mutagenic, and that a similar pattern appears in their corresponding isogenic MMR+ (i.e. wild-type) cells. These data suggest that triplet context could influence the long-term evolution of wild bacteria. One group also suggested that the accelerated evolution of head-on genes may be explained by a slight overabundance of the highly mutagenic triplet CCG in head-on genes (Schroeder et al. 2016). However, these models have not been tested in the context of natural evolution.

Here we analyze the effects of triplet sequence context on the evolution of nine closely related wild *B. subtilis* strains using 218,182 SNPs as a proxy for mutations. We also re-examine the potential impact of triplet sequences on our previously published reporter-based phenotypic reversion assays. Together, these studies address both the spontaneous mutation rate and retained mutation patterns of coding genes. Our data suggest that triplet sequence context can affect mutation rates in wild cells over evolutionary time periods. However, we identify a set of potentially mutagenic triplets that are distinct from those previously identified by others in laboratory experiments. Using dN/dS analysis, we also show that SNPs in mutagenic triplets are not required for the accelerated evolution of head-on genes. In parallel, we analyze the sequences of our reporter genes and find that mutagenic triplets are not more abundant in the leading strand sequences of the head-on genes. As such, differential triplet context is insufficient to explain the increase in spontaneous mutations in head-on transcription units. Overall, we find that triplet context, and indeed sequence context in general, is not responsible for the accelerated evolution of head-on genes.

Triplets with High SNP Frequency Are Depleted from the Genome

Recent studies indicated that triplet sequence context influences spontaneous mutation rates by altering the fidelity of the replication fork (Sung et al. 2015; Schroeder et al. 2016). In particular, the sequences CCG, GCG, and CAC were shown to be highly mutagenic. Therefore, we anticipated that SNPs in wild strains might be more abundant in these sequences. To test this, we identified SNPs in nine wild *B. subtilis* strains using the Prokaryotic Genome Analysis Tool kit (Brittnacher et al. 2011). We then analyzed the distribution of these SNPs with regard to triplet context on the leading strand of the left and right replichores as performed previously (Sung et al. 2015; Schroeder et al. 2016). Because the abundance of triplet sites available for mutation varies, raw SNP abundance values will be skewed. To account for this, we counted the number of each triplet present in the *B. subtilis 168* chromosome. We then normalized the relative SNP abundance at each triplet to the relative abundance of each triplet site. For example, TCG represents 1.27% of total triplets in the *B. subtilis* genome, whereas 3.27% of total SNPs were found at the central C of TCG. Therefore, we divided 3.27% by 1.27%, yielding a SNP frequency of 2.56. The SNP frequency values should serve as a proxy for the relative mutation rate of each triplet. A value of >1.0 suggests that a triplet is more mutagenic than predicted by the null hypothesis, and vice versa.

We conducted our analyses using all genes (218,182 SNPs) or core genes only (25,477 SNPs) as previously defined (Paul et al. 2013). We also examined all core gene mutations together or transition mutations only. Each analysis produced similar results, and we present data for all SNPs found in either all genes or core genes (fig. 1*A* and *B*, supplementary table S1, Supplementary Material online). Our data show that among the 64 possible triplets, SNP frequency ranges from 0.39 to 2.56%, consistent with the hypothesis that triplet context can influence the rate of spontaneous mutagenesis in nature (fig. 1, gray columns).

We then hypothesized that if triplet context truly alters spontaneous mutation rates, highly mutagenic triplets should be depleted from the genome. In contrast, less mutagenic triplets should be over-represented. Given an equal rate of mutation (and barring other constraints), each of the 64 triplets would be expected to represent, on average, 1.56% of the genome (64 triplets divided by 100%). However, the relative abundance of A, C, G, and T is likely constrained. This should cause the relative abundance of each triplet to differ from the mathematical average. To account for this, we developed a prediction for the relative abundance of each triplet based upon the ratio of A, C, G, and T nucleotides in the leading strand of all genes or only the core genes of *B. subtilis 168* (supplementary table S1, Supplementary Material online). This prediction assumes that no other factors are influencing triplet abundance. We then divided the actual triplet abundance by the theoretical abundance and used the arbitrary cutoff of 50% difference between the two as an indicator of significant depletion or over-representation. Using this criteria, we observed that eight triplets are underrepresented,

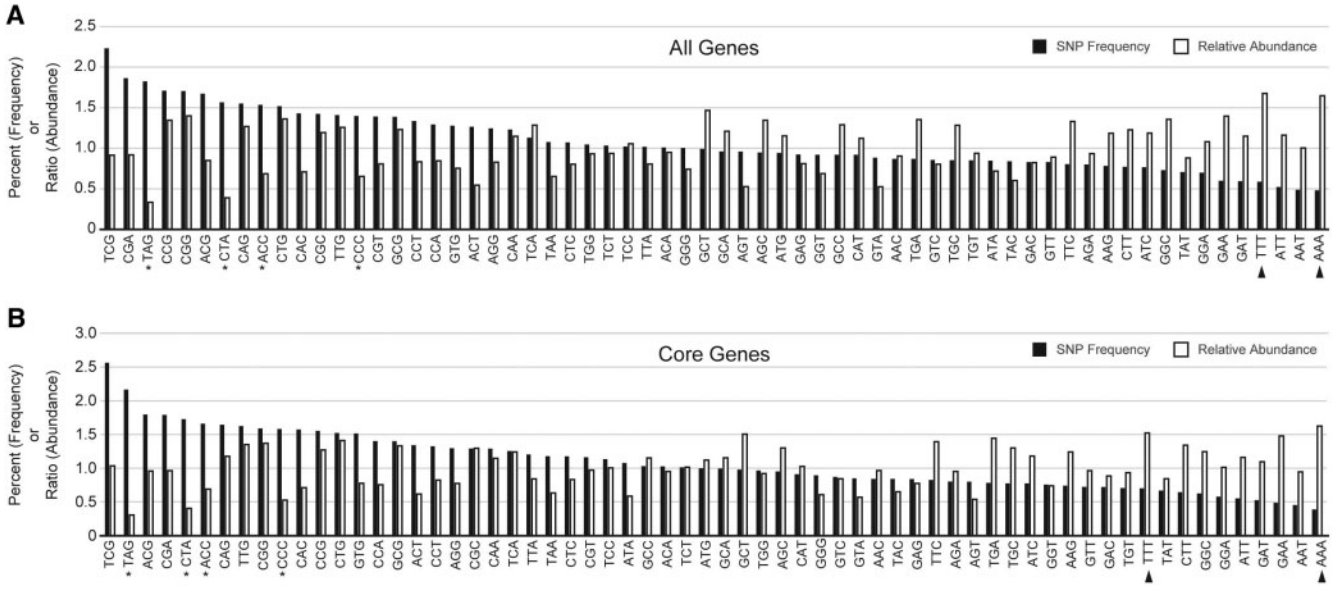

Fig. 1.—SNP frequency at codon-independent triplet DNA sequences, and the relative abundance of all triplets in the *B. subtilis* genome. The SNP frequency (% of total SNPs/% of all triplets in the *B. subtilis 168* genome) at all triplet sequences is shown, sorted from highest to lowest (black columns). The relative abundance (actual abundance/theoretical abundance) of each triplet in the *B. subtilis* genome (white columns) is also shown. Theoretical abundance values are based upon the A/C/G/T content of the indicated group of genes. (***A***) Data for all genes. (***B***) Data for core genes. Asterisks indicate triplets with both a significantly elevated SNP frequency and a significantly lower than expected abundance. Triangles indicate triplets with both a significantly lower than expected SNP frequency and a higher than expected abundance.

and two are overrepresented in core genes. (fig. 1*A* white columns, supplementary table S1, Supplementary Material online, left side). Among all genes, we observe 10 underrepresented and 3 overrepresented triplets (fig. 1*B*, white columns, supplementary table S1, Supplementary Material online, right side). The over/underrepresented triplets in these two groups are nearly identical. Together, these data suggest that some triplets may in fact be depleted or overrepresented in the genome, supporting the hypothesis that triplet context can affect the rate of spontaneous mutagenesis.

As anticipated, we note that triplet abundance is negatively correlated with SNP frequency (Pearson's correlation of the two data sets, $r = -0.499$) (fig. 1, gray and blue columns). This is consistent with the idea that sequences which mutate at a higher rate should be present at a lower frequency in the genomes. Qualitatively this relationship appears more pronounced for triplets with the highest and lowest normalized SNP abundances. We therefore identified the triplets in the top and bottom 10% (six triplets each) in each data set. Given two randomly chosen sets of 12 triplets, 2.25 would be expected to be present in both data sets ((12 triplets/64 possible)^2 $\times$ 64). Here, six triplets are common to both data sets: among the six triplets with highest SNP frequency, three are also in the lowest abundance in the genome. For the six triplets with the lowest SNP frequency, three are also in the highest 10% in terms of abundance. These data further suggest that triplet context can in fact influence the mutation rate of certain triplets. Of particular note, the triplets TAG, CCC, CTA, and ACC have both an elevated apparent mutation rate and a lower than expected abundance in the genome (fig. 1, asterisks). Therefore, these particular triplets may be depleted from the genome due to an elevated mutation rate in the context of natural evolution. Similarly, the number of SNPs in the triplets AAA and TTT is lower than anticipated, and these triplets are overrepresented in the genome (fig. 1, triangles). This suggests that AAA and TTT triplets have a decreased mutation rate in the wild.

As an additional consideration, we note that the three triplets previously identified as mutagenic in mutation accumulation lines (CAC, CCG, and GCG) also have SNP frequency values >1. However, among the three, only CAC is underrepresented in the genome (its abundance is 0.86%. The average abundance of all triplets is 1.56 with a lower limit of 0.89% for a single SD). Therefore, with the possible exception of CAC, the data suggest that during natural evolution the reported mutagenic triplets are not a significant determinant of the overall mutational profile of the genome (Sung et al. 2015; Schroeder et al. 2016).

The Accelerated Evolution of Head-on Genes Does not Depend on Any Single Triplet Sequence.

The leading strand sequences of head-on and co-directional gene regions have slightly different triplet sequence compositions (fig. 2, and Schroeder et al. 2016). Because triplets

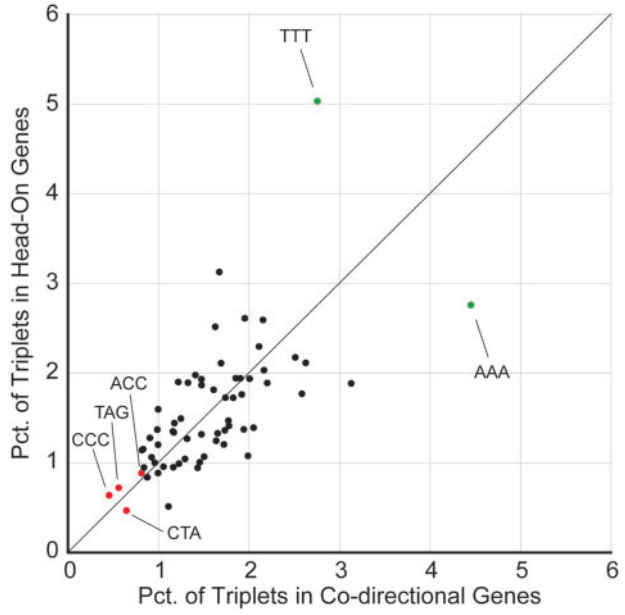

Fig. 2.—Relative Triplet abundance in the *B. subtilis 168* genome. The abundance (%) of each triplet in core head-on or core co-directional genes. Triplets with an abundance more than 1 standard deviation lower (red) or higher (green) than the mean are indicated.

appear to influence mutation rates, differential triplet content could conceivably promote the higher rate of evolution in head-on genes we previously reported (Paul et al. 2013). To test this possibility, we re-calculated the dN and dS of head-on and co-directional genes after removing SNPs present at the central nucleotide position of individual triplets. To account for the possibility that a given SNP actually occurred in the reverse direction relative to our reference genome (*B. subtilis 168*), we collectively removed all SNPs in a given group of triplets. For example, we removed SNPs found at the central position in CCG, CTG, and CAG, and CGG triplets (labeled collectively as "CNG"). Using the modified data, we then re-analyzed the mean dN and dS for head-on and co-directional core genes (fig. 3*A* and *B*). The removal of any group of SNPs reduced the average dN and dS, as expected (fig. 3*A* and *B*). However, in each case, the average dN in head-on genes remained significantly higher than the dN for co-directional genes (ratios of head-on dN/co-directional dN ranged from 1.27 to 1.39-fold vs. the original value of 1.315-fold). We also tested the effect of removing SNPs at the specific triplet CCG, which is more abundant in head-on genes and had a high rate of mutation in laboratory experiments (Sung et al. 2015; Schroeder et al. 2016). Based upon these observations, it was hypothesized that mutations in CCG might promote the accelerated evolution of head-on genes. We observed that the removal of SNPs in CCG had no significant effect on the dN or dS levels of genes of either orientation (the ratio of the head-on dN/co-directional dN is 1.30). These data indicate that triplet context is not a significant driver of accelerated evolution of head-on genes.

Head-on Gene Orientation Leads to a Higher Mutation Rate, Regardless of Sequence Content

Previously, we identified a higher rate of spontaneous mutagenesis in three distinct reporter genes when they are oriented head-on to replication (Million-Weaver et al. 2015). For these experiments we quantified the phenotypic reversion rate of strains harboring the *hisC952*, *metB5*, and *leuC427* genes when they are integrated onto the chromosome in either orientation. Each reporter yielded a similar result despite their entirely different coding sequences and the use of two different chromosomal integration loci (Million-Weaver et al. 2015). To assess the potential contribution of triplet sequences, we analyzed the leading strand triplet content of our reporter genes in both orientations. We found that none of the inactivating mutation sites contain any of the previously identified mutagenic triplets (CCG, GCG, and CAC) in any frame, regardless of their orientation. Among the potentially mutagenic triplets identified in this study (CCC, CTA, ACC, and TAG), only CTA and TAG are present at inactivating mutation sites. Specifically, CTA is found in the inactivating mutation site on the leading strand of head-on *hisC952* and *leuC427*. Conversely, TAG is found only in the inactivating mutation sites of the co-directionally oriented *hisC952* and *leuC427* genes. Yet, for all reporters, we observed a higher rate of mutation when they were oriented head-on to replication (Million-Weaver et al. 2015). Consistent with the results of the evolutionary analyses presented earlier, the reporter gene data clearly indicate that the 2–2.5-fold increase in spontaneous mutation rate we observed in head-on genes is not attributable to triplet sequences. Additionally, the diverse coding sequences of the three reporters strongly suggest that this process is independent of sequence context in general.

Discussion

An increased mutation rate has been demonstrated at certain triplet sequences and in head-on genes. The highly consistent data from our reporter studies and our analysis of their triplet sequence content indicate that head-on genes experience this higher mutation rate regardless of their sequence. Therefore, we conclude that our initial model is accurate: though transcription is inherently mutagenic, the orientation of transcription further alters the rate of spontaneous mutation within the transcribed region by ~2-fold. On an evolutionary time scale, this difference precipitates a significantly higher rate of evolution in head-on genes.

Our analyses of SNPs from wild cells yielded results that are consistent with previous publications showing that sequence context can affect local spontaneous mutation rates (Sung et al. 2015; Schroeder et al. 2016). Interestingly however, the triplets with the highest SNP frequency in our evolutionary

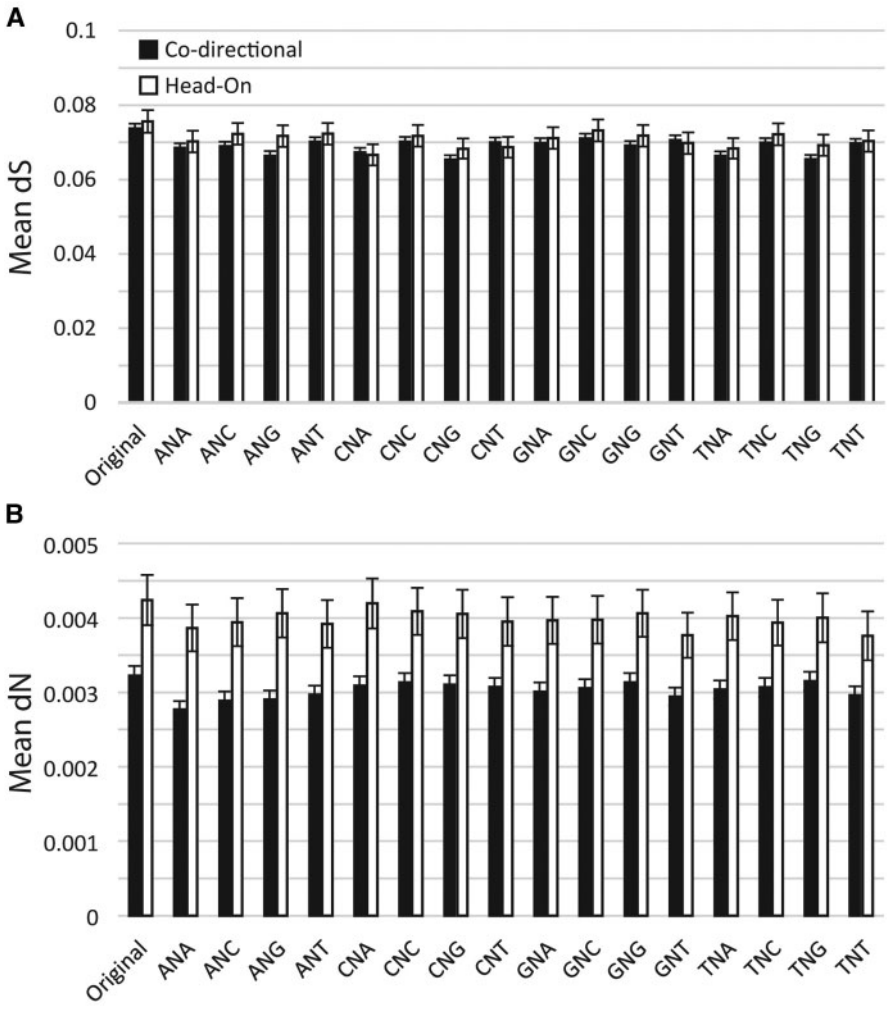

Fig. 3.—Mean dN and dS values of core genes before (original) or after the removal of SNPs found at the specified group of triplets. For each data point, SNPs in the original data set were reverted to the *B. subtilis 168* wild type sequence if they occurred at the central nucleotide of the specified triplet group. Data for head-on genes (white columns) and co-directional genes (black columns) are shown. Using the Z-test, we identified a statistically significant difference between the head-on and co-directional dN for each triplet group ($P \leq 0.01$).

analyses are distinct from those identified as mutagenic in laboratory experiments. There are at least two potential explanations for this: 1) In the wild, cells should generally have a functional MMR system that reduces the incidence of replication-fork derived mutations to a low level. DNA may also be directly damaged due to environmentally derived sources of genotoxic chemicals or ultra-violet light. Together, these naturally occurring conditions could significantly alter the mutational profile of cells living the the natural world relative to the same cells grown in the laboratory without stress (i.e. the conditions used during mutation accumulation line experiments). 2) The triplets identified as mutagenic in the laboratory may also be particularly mutagenic in the wild, but due to negative or purifying selection, the prevalence of these mutations are reduced. These possibilities are not mutually exclusive. Regardless of the reason, the SNP record suggests that sequence context and replication fork derived mutations are not the predominant mechanism driving the evolution of *B. subtilis*.

Further investigation of accelerated head-on gene evolution is important given that in certain studies where genome-wide comparisons of either transcription level or head-on and co-directional genes were performed, no differences in spontaneous mutation rate were seen (Lee et al. 2012; Foster et al. 2015; Sung et al. 2015; Schroeder et al. 2016). For example, Lee et al. detected no effect of transcription level (regardless of its orientation) in *Escherichia coli* mutation accumulation lines. The disparity between reporter-based and genome-wide data sets is potentially attributable to a number of sources: Genome-wide studies require the comparison of a group of co-directional genes to an entirely different set of genes oriented head-on to replication. Inherent differences in the genes

making up each group can include their length, transcription level, sequence, function, essentiality, and orientation. This apples-to-oranges style comparison may make it difficult to parse out the specific effect of orientation. Hence, we propose that the use of otherwise identical inducible reporter genes may provide a clearer view of the effects of conflicts on orientation and transcription level-dependent mutation rates. It is also important to note that few head-on oriented genes are transcribed during low stress growth in rich media, at least in *B. subtilis*. This is likely due to the enrichment of stress response genes in the head-on orientation. In mutation accumulation line based experiments, which specifically avoid selection by minimizing stress, conflicts at head-on genes should be minimized. In contrast, in the wild, cells should occasionally experience stress, leading to the induction of various head-on genes, and subsequently inducing mutagenesis of these regions via PolY1 activity. Thus, the lack of increased mutagenesis of head-on genes in the mutation accumulation line experiments is not entirely surprising. For these reasons, mutation accumulation line experiments may not be an ideal method for studying gene orientation-based mutagenesis. Future studies are needed to clarify these possibilities.

In conclusion, regardless of the underlying reason for the different results, our reporter data and SNP analyses, presented here and previously, demonstrate that head-on genes undergo accelerated evolution independently of sequence context-based effects (Million-Weaver et al. 2015). Therefore, the established activities of TC-NER and PolY1 appear sufficient to explain this process.

Methods

SNP Analysis: The genomic sequences of nine closely related *B. subtilis* subspecies (NCBI INSDC numbers AL009126, CP003329, CP002468, CP003695, CP002905, CP002183, CP002906, CP004019) were analyzed using the Prokaryotic Genome Analysis Tool (http://tools.uwgenomics.org/pgat/), yielding 218,182 SNPs. Among core genes (previously defined by Paul et al. 2013), we identified 25,477 SNPs. We used python scripts to identify the triplet context of each SNP and to tally the total number of each codon-independent triplet in the leading strand of the left and right replication forks, within protein coding regions in the *B. subtilis 168* genome. Raw SNP abundance at each triplet was converted to the percent of total SNPs. SNP abundance as percent was then normalized to account for the relative abundance of each triplet present in the genome: For each triplet, the SNP abundance was divided by the triplet abundance (as the percent of total triplets) in the genome, yielding the SNP frequency.

Identification of depleted/overrepresented triplets: Our null hypothesis is that all triplet sequences are equally mutagenic. To model the expected relative abundance of triplets this hypothesis suggests, we tallied the A/C/G and T abundance within either all genes, or core genes only. (Data for core genes—A:0.302, C: 0.200, G: 0.236, T: 0.261, and all genes—A:0.291, C: 0.204, G: 0.237, T: 0.267). Abundance values were used in place of probability to calculate the likelihood that a given triplet would appear in the genome. E.g. the probability of finding AGT within core genes is $0.302 \times 0.236 \times 0.261$. Triplets with a SNP frequency value that differs by more than 50% from the null hypothesis value (i.e., values of $\geq$1.5- or $\leq$0.66-fold) were considered significant.dN/dS Analysis: To calculate mean dN and dS, the program PAML was used (Xu and Yang 2013). After producing aligned nucleotide sequences for each gene, we used python scripts to revert specific sets of SNPs (e.g., for the group "ANT", the SNPs in AAT, ACT, AGT, and ATT were reverted to the *B. subtilis* consensus sequence) to the *B. subtilis* wild type coding sequence, thereby removing that group of SNPs from our data files. The modified nucleotide alignment files were then used as input for dN and dS analysis by PAML. Analysis of statistical significance was performed using the Z-test for dN and dS values.

Supplementary Material

Supplementary materials are available at *Genome Biology and Evolution* online.

Acknowledgments

We would like to thank Hillary Hayden for helpful discussions. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number DP2GM110773 (to H.M.). We also thank the Genomics Core which is funded by the Cystic Fibrosis Research Center of the National Institutes of Health under award number NIDDK KD089507, and the Cystic Fibrosis Foundation under award SINGH15R0.

Literature Cited

Brittnacher MJ, et al. 2011. PGAT: a multistrain analysis resource for microbial genomes. Bioinformatics 27(17):2429–2430.

Chen X, Zhang J. 2013. Why are genes encoded on the lagging strand of the bacterial genome? Genome Biol Evol. 5(12):2436–2439.

Foster PL, Lee H, Popodi E, Townes JP, Tang H. 2015. Determinants of spontaneous mutation in the bacterium *Escherichia coli* as revealed by whole-genome sequencing. Proc Natl Acad Sci U S A. 112(44):E5990–E5999.

Lee H, Popodi E, Tang H, Foster PL. 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(41):E2774–E2783.

Million-Weaver S, et al. 2015. An underlying mechanism for the increased mutagenesis of lagging-strand genes in *Bacillus subtilis*. Proc Natl Acad Sci U S A. 112:E1096–E1105.

Paul S, Million-Weaver S, Chattopadhyay S, Sokurenko E, Merrikh H. 2013. Accelerated gene evolution through replication-transcription conflicts. Nature 495:512–515.

Schroeder JW, Hirst WG, Szewczyk GA, Simmons LA. 2016. The effect of local sequence context on mutational bias of genes encoded on the leading and lagging strands. Curr Biol. 26:692–697.

Sung W, et al. 2015. Asymmetric context-dependent mutation patterns revealed through mutation-accumulation experiments. Mol Biol Evol. 32(7):1672–1683.

Xu B, Yang Z. 2013. PAMLX: a graphical user interface for PAML. Mol Biol Evol. 30(12):2723–2724.

Associate editor: Tal Dagan