Supporting Information

The mutagenic cost of ribonucleotides in bacterial DNA

Jeremy W. Schroeder[1,4,5], Justin R. Randall[1,4], William G. Hirst[1], Michael E. O'Donnell[2,3], Lyle A. Simmons[1,3]

[1]*Department of Molecular, Cellular and Developmental Biology, University of Michigan, Ann Arbor, MI 48109, USA*

[2]*Howard Hughes Medical Institute and The Rockefeller University, New York, NY 10065, USA*

[3]To whom correspondence should be addressed: lasimm@umich.edu; odonnel@mail.rockefeller.edu

[4]These authors contributed equally to this work.

[5]Present address: Department *of Bacteriology, University of Wisconsin - Madison, Madison, WI 53706, USA*

SI Materials and Methods

Protein purification

Proteins were isolated as described (1). Briefly, cultures with each plasmid were grown in 3 L of terrific broth (1.2% tryptone, 2.4% yeast extract 72 mM K_2HPO_4, 17 mM KH_2PO_4, 0.4% glycerol) with constant shaking at 37°C to an OD of ≈ 0.75 and expression was induced by addition IPTG to a final concentration of 250 μM. Cultures were left for an additional 3 hours at 37°C with constant shaking followed by harvesting of cells with centrifugation. Cell pellets were frozen in liquid nitrogen before further processing. Pellets were thawed and resuspended in 50 mL of lysis buffer (50 mM Tris-HCl pH8, 300 mM NaCl, 10% sucrose, 10 mM imidizole) and lysed via sonication. Cell debris was pelleted via centrifugation and the supernatant was applied to a 2 mL Ni^{2+}-NTA agarose gravity-flow column. The column was washed (wash buffer: 50 mM Tris-HCl pH 8, 2M NaCl, 10 mM imidizole) and then eluted (elution buffer: 50 mM Tris-HCl pH 8, 50 mM NaCl, 500 mM imidizole). The SUMO protease was added to the eluate, which was dialyzed into SUMO Protease buffer (50 mM Tris-HCl pH 8, 150 mM NaCl, 1 mM DTT) 15-18 hours at 4°C to remove the affinity tag. The dialyzed protein solution was then applied to a 2 mL Ni^{2+}-NTA agarose gravity-flow column and fractions were collected and analyzed for purity via SDS-PAGE. Fractions containing the desired protein were pooled and each protein was applied to an anion exchange column (GE product number 17-5156-01) followed by elution with a 50-500 mM NaCl gradient. Fractions were analyzed via SDS-PAGE with the pure fractions pooled and concentrated into protein storage buffer (50 mM Tris-HCl pH 8, 150 mM NaCl, 25% glycerol) and frozen in liquid nitrogen. PolC and DnaE proteins were generous gifts from Dr. Charles McHenry at the University of Colorado (2). The Klenow fragment used in this work was purchased from New England Biolabs (product number M0210S).

Calculation of conditional mutation rate

Transitions rates for each base and replichore were calculated essentially as described (3):

$$\frac{M_{b,r}}{R_{b,r}xG} \quad (1)$$

where $M_{b,r}$ is the number of transitions from base b in the reference sequence of replichore r, $R_{b,r}$ is the total number of occurrences of base b in the reference sequence of replichore r, and G is the total number of generations surpassed during the mutation accumulation procedure. Transition rates in each dinucleotide sequence context were calculated essentially as described (3, 4):

$$\frac{M_d}{R_d xG} \quad (2)$$

where M_d is the number of transitions at the $5'$ position of a given dinucleotide in the lagging strand template, R_d is the total number of occurrences of the given dinucleotide in the lagging strand template, and G is the total number of generations the MA lines underwent during the mutation accumulation procedure.

Logistic regression of sequence context effect on transition rate

To determine the impact of local sequence context on guanosine to adenosine transitions in the lagging strand template, the logit of the probability of obtaining a guanosine to adenosine transition against nucleotide base identity up to five nucleotide positions in the 3′ direction of a guanosine in the lagging strand template:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta Base_{+1} + \beta\mathrm{Base}_{+2} + \beta\mathrm{Base}_{+3} + \beta\mathrm{Base}_{+4} + \beta\mathrm{Base}_{+5} \quad (3)$$

where p is the probability of guanosine in the lagging strand template having a transition occur and $Base_i$ represents either adenosine, cytidine, thymidine or guanosine at position i 3′ to the guanosine in question. Log-odds ratios presented in Figure 3B were calculated using adenosine as the baseline.

Motif identification at lagging strand template G to A transitions

A multifasta file containing the local sequence from all 162 sites where a G in the lagging strand template was transitioned to an A in Δ*rnhB* was prepared along with a multifasta file containing local sequence contexts at 1000 randomly selected G in the lagging strand template to control for random local sequence context. Specifically, the sequence contexts for motif detection extended one position 5′ through six positions 3′ to the transition site, from the perspective of the lagging strand template. MEME was applied in discriminative mode using the 162 sequence contexts surrounding G→A transitions in the lagging strand template as the primary sequences and the file containing 1000 negative control sequence contexts as the control sequences with the minimum possible motif width set to two (5). All other parameters were default values. The resulting MEME logo was trimmed to include only positions one through four nucleotides 3′ to the transition site.

RNase HII nicked substrate extension

A 5′ IR dye-labeled, 40-mer oligonucleotide containing one embedded rNMP (oJR297) was annealed to a complementary 40-mer oligonucleotide (oJR298) as described (see polymerase activity assay). RNase HII reactions were performed in a 200 µL reaction in the same buffer with 1 µM substrate and 200 nM RNase HII for 45 minutes at 37°C. Reactions were stopped by heating the sample to 95°C for 5 minutes. 45 µL of this reaction was then placed into three different microcentrifuge tubes containing 5 µL of either 1 µM Pol I, DnaE, or PolC and mixed by pipetting. Reactions were incubated at 25°C and 5 µL of the reaction was removed and quenched with formamide loading dye at 1, 5, 20, and 60 minutes and immediately placed on ice. After 60 minutes, NaOH was added to a final concentration of 300 mM and incubated at 55°C for 45 minutes to verify the rNMP was completely hydrolyzed. As a control the same extension reaction was performed without RNase HII cleavage. All reactions were then denatured at 100°C for two minutes and immediately snap cooled in an ice water bath. Products were resolved by electrophoresis through a 17% urea-polyacrylamide gel and visualized with a LI-COR Odyssey imager. Three independent reactions were performed. The percent full product extended was calculated by dividing the extended product in lane 8 by the cleaved substrate band in lane 4 and multiplying by 100. Plots show the mean and 95% confidence interval of all three reactions at each time point with each polymerase.

Sequence context-dependent mutagenic resynthesis

A 5′ IR dye-labeled primer was annealed as described above (see polymerase activity assay) to a complementary template strand intended to modeling an NER-processed rNMP incorporation at a 5′-GCCTT-3′ sequence (oJR895 and oJR283). Extension reactions were performed in 10 μL reactions in the same buffer. Each reaction contained 1 μM substrate without polymerase or with Pol I, DnaE, or PolC. These reactions were performed with all dNTPs at 200 μM, or in the absence of dCTP. Reactions were stopped with 10 μL of formamide loading dye at 5 and 45 minutes. Reactions were denatured at 100°C for two minutes and immediately snap cooled in an ice water bath. Products were resolved via electrophoresis through 17% urea-polyacrylamide gels and visualized with a LI-COR Odyssey imager. Two independent reactions were performed and the relative primer extension was calculated by dividing the intensity of the band from 5 minute extension with no dCTP by the intensity of the band from the 5 minute extension with all dNTPs and multiplying by 100. Bar height and error bars represent the mean and range, respectively, of two independent experiments.

Gene reporter sequencing:

A total of 48 independent 100 μL cultures of JWS328, JWS329, JWS330 or JWS331 were grown in 96-well plates at 37°C with constant shaking in $S7_{50}$ minimal medium supplemented with 0.1 μM final concentration each of tryptophan and phenylalanine and 200 μM final concentration thymidine. The entirety of each culture was plated on $S7_{50}$ minimal medium agar plates supplemented with 0.1 μM final concentration each of tryptophan and phenylalanine, 200 μM final concentration thymidine, 34 μM final concentration trimethoprim and 0.2% (w/v) final concentration casaminoacids. Plates were incubated at 37°C for two days, at which time fifteen colonies from each plate were inoculated into 96-well plates containing LB supplemented with 200 μM thymidine and 34 μM trimethoprim, for a total of 720 trimethoprim-resistant colonies of each genotype. 96-well plates were incubated with constant shaking for 8 hours at 37°C. Plates were stored at 4°C overnight. Cultures were diluted 16-fold in water. A Phusion Hot Start II DNA polymerase PCR master mix was prepared for amplification of the *thyA* locus using primers oJS858 and oJS859. For amplification 19 μL of the PCR master mix was deposited into each well of 96-well plates. For template, 1 μL of each diluted culture was added to each PCR reaction and mixed by pipetting. PCR amplification was performed and products were purified using 1x volume of MagNA bead slurry consisting of 2% (v/v) Sera-Mag SpeedBeads, 18% (w/v) PEG-8000, 1 M NaCl, 10 mM Tris HCl, 1 mM EDTA, 0.05% (v/v) Tween 20 (6-8). The concentration of each purified PCR product was determined using the QuantIT – High sensitivity kit (Fisher Scientific, Q33120). The fifteen PCR products that originated from colonies from a common plate were pooled at an equimolar concentration to achieve a combined final DNA concentration of 0.5 ng/μL. Sequencing libraries were prepared from the pooled PCR products according to the method of Baym *et al*. (8). DNA was sequenced at the University of Michigan DNA Sequencing Core using an Illumina Hi-seq 4000.

Paired-ended 150 base reads were trimmed to remove Illumina sequencing adapters. Alignment to the *thyA* open reading frame and the 500 bases flanking each side was performed using bwa (9) and alignment files prepared using samtools (10). Variants present in greater than 2% of the aligned reads at a given location were detected using Freebayes (11). Potential variants were then stringently filtered, retaining only variants with a quality score greater than 500 for further analysis. The final list of variants included 181 from JWS328, 266 from JWS329, 185 from JWS330, and 193 from JWS331. Testing for differences in mutation spectra was performed using a Chi-square test.

Table S1. Overview of the variants detected in mutation accumulation lines

	Wild type [a]	Δ*rnhB*
Number of lines	75	81
Total mutations	295	462
BPSs	254	420
Indels	41	42
Transitions	198	363
AT to GC	74	90
GC to AT	124	273
Transversions	56	57
AT to CG	25	21
AT to TA	10	13
GC to CG	4	3
GC to TA	17	20
Generations per line	3636	3610
Mutation rate per genome replication ($x10^3$) $[95\% CI]_b$	1.1 [1.0-1.2]	1.6 [1.4-1.7]
Mutation rate per nucleotide ($x10^{10}$) $[95\% CI]_b$	2.7 [2.4-3.0]	4.0 [3.6-4.3]

[a] Wild type data were published in (3). All variants are presented in Dataset 1.

Table S2. Mutation rate of cells with *rnhB* and *uvrA* deletions

Strain	No. of cultures	Mutations per culture [+95%CI]	Mutation rate (Mutations per generation 10^{-8} ± [95%CI]	Relative mutation rate
Wild Type	30	0.5 [0.3-0.5]	1.15 [0.7-1.2]	1
Δ*rnhB*	30	0.9 [0.5-1.2]	1.9 [1.0-2.5]	1.7
ΔuvrA	27	1.0 [0.5-1.5]	1.8 [0.8-2.6]	1.6
Δ*rnhB, ΔuvrA*	30	1.6 [0.7-2.4]	4.45 [1.9-6.7]	3.9

Mutagenesis assays were done as described using rif^R as an indicator. Mutation rate and mutations per culture were calculated using the Ma-Sandri-Sarkar Maximum Likelihood Estimator with the web-based tool FALCOR (12).

Table S3. Oligonucleotides used in this study

Name	Sequence
oJR46	TCGAGCACCACCACCACCACCACTGAG
oJR47	ACCTCCAATCTGTTCGCGGTGAGCCTCAATAATATCG
oJR88	CCGCGAACAGATTGGAGGTGTGAATACATTAACCGTAAAGGACATTAAAGACC
oJR89	TGGTGGTGGTGGTGCTCGATTATCTGAAAGATTGAACAGGAGCG
oJR90	CCGCGAACAGATTGGAGGTGTGTCCCATTCAGTGATAAAAGTATCG
oJR91	TGGTGGTGGTGGTGCTCGACTATGAACGTTTTTTATCAGCAAGGCG
oJR94	GTTGcCGcGGTCGGCCG
oJR95	GACCgCGgCAACACCTGCAATC
oJR96	TCTGcCGcAGTCGGAACCGG
oJR97	GACTgCGgCAGAACCGATAACAGACATTCC
oJR102	CCGCGAACAGATTGGAGGTATGACGGAACGAAAAAAATTAGTGC
oJR103	TGGTGGTGGTGGTGCTCGATTATTTCGCATCGTACCAAGATGG
oJR145	GCAGAGCTAGCTTACGATCG
oJR209	/5IRD800CWN/CGATCGTAArGCTAGCTCTGC
oJR210	/5IRD800CWN/CGATCGTArArGrCrUAGCTCTGC
oJR283	AACTTAGCCTTGCGCATCAGCTGCAG
oJS895	/5IRD800CWN/CTGCAGCTGATGCGCA
oJR297	/5IRD800CWN/GTACTCGGTGATCCGTACATrGGCGCATCAGCTGCAGTAG
oJR298	CTACTGCAGCTGATGCGCCATGTACGGATCACCGAGTAC
oJS858	ACCGCAATATCAAACCATTTCGT
oJS859	TGGTCATAGTTCGTTTCTTCAGC

Lowercase letters in oligonucleotide sequences indicate sites of mutagenesis for catalytic mutants. /5IRD800CWN/ preceding the oligonucleotide sequence indicates the oligonucleotide was 5′ end labeled by Integrated DNA Technologies with IRDye CW800 (NHS ester) for imaging on a Li-COR imaging system.

Table S4. Plasmids used in this study

Name	Vector	Insert
pJR17	pE-SUMO	*rnhB*
pJR18	pE-SUMO	*rnhB (D78A, E79A)*
pJR19	pE-SUMO	*rnhC*
pJR20	pE-SUMO	*rnhC (D100A, E101A)*
pJR22	pE-SUMO	*polA*

Table S5. Strains used in this study

Name	Relevant Genotype	Source
PY79	Wild type SPβ^{o}	(13)
JWS105	Δ*rnhB*	(14)
PEB307	Δ*uvrA*	
JWS305	Δ*rnhB,* Δ*uvrA*	This work
JWS328	*thyB::erm*	This work
JWS329	Δ*rnhB*, *thyB::erm*	This work
JWS330	Δ*uvrA*, *thyB::erm*	This work
JWS331	Δ*rnhB*, Δ*uvrA*, *thyB::erm*	This work

All strains are derived from PY79.

Dataset 1. Spreadsheet with information on all variants detected in MA lines.

Dataset 2. Spreadsheet with information on all variants detected in *thyA* resequencing located.

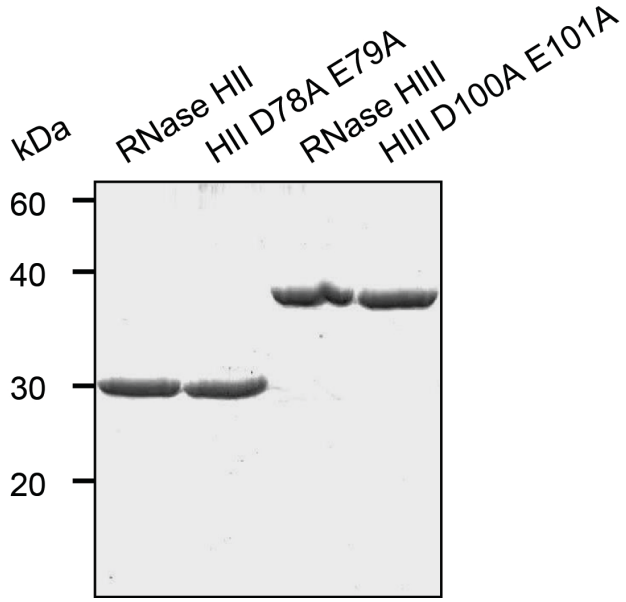

Figure S1. SDS-PAGE of RNase HII, HIII and catalytically inactive variants. Shown is an SDS-PAGE with 2 µg of purified RNase HII, RNase HIII, RNase HII D78A E79A, and RNase HIII D100A E101A followed by coomassie blue staining.

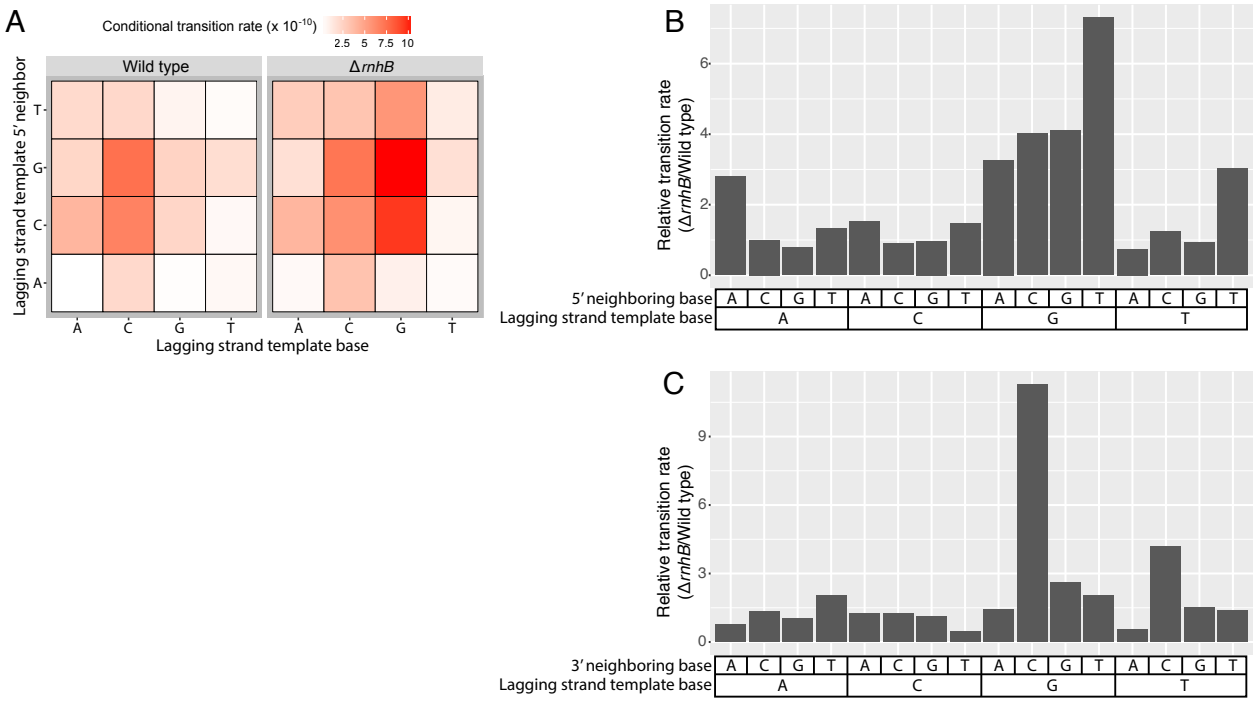

Figure S2. Context-dependence of transition rate in the lagging strand template is specific to 3′ neighboring bases. (A) A heatmap displaying the transition rate for a nucleotide with the base given on the x-axis with the 5′ neighboring base indicated in the y-axis. **(B)** A barplot displaying the relative transition rate in Δ*rnhB*/wild type from the indicated base with the given 5′ neighbor showing the increased mutation rate at guanosine nucleotides is largely independent of 5′ neighboring nucleotide identity. **(C)** A barplot showing the relative transition rate in Δ*rnhB*/wild type from the indicated base with the given 3′ neighbor shows the increase in transitions at guanosine nucleotides is highly dependent on 3′ sequence context.

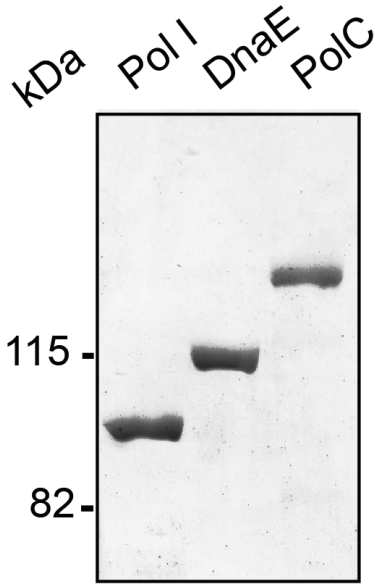

Figure S3. Coomassie blue-stained 8% SDS-polyacrylamide gel with 2 μg each of Pol I, DnaE and PolC.

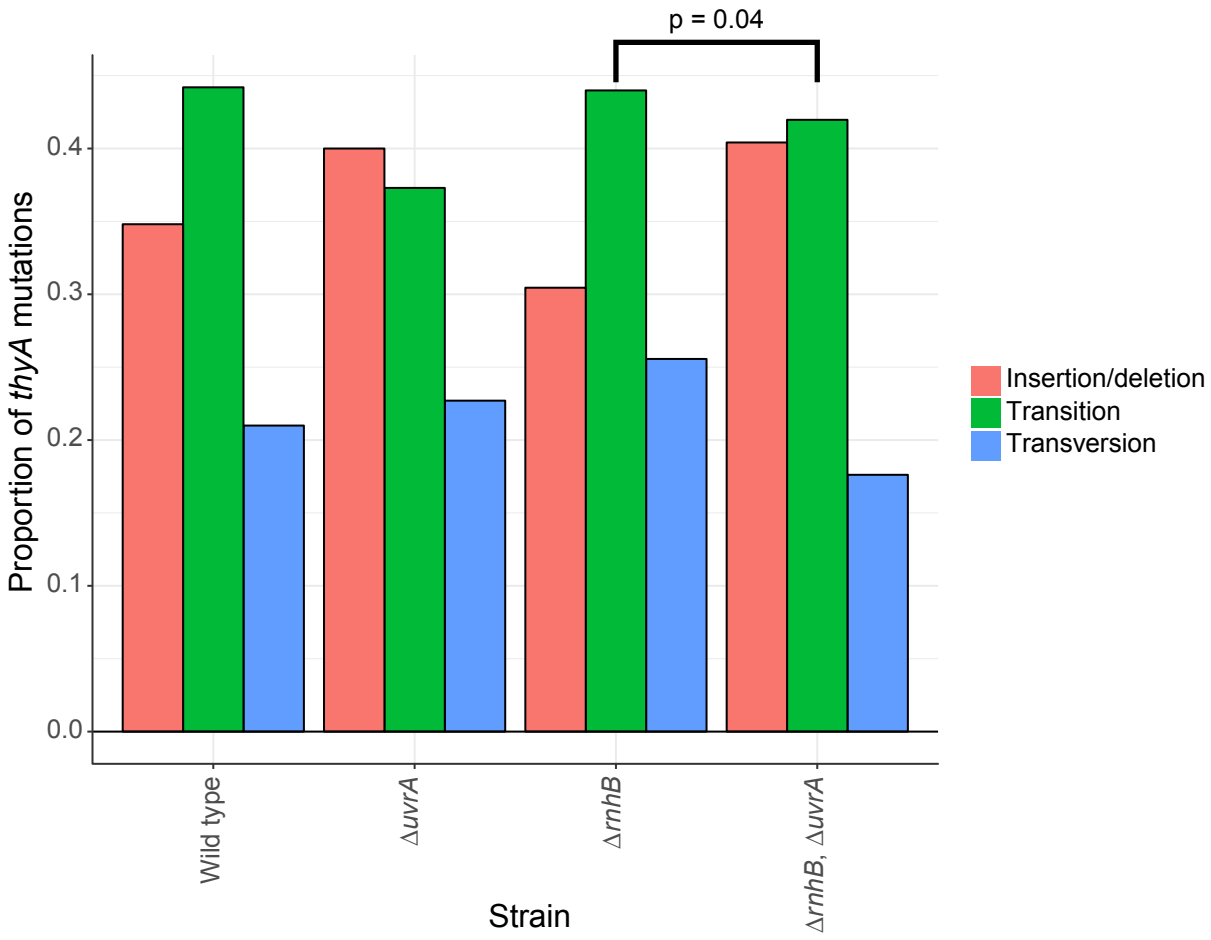

Figure S4. Mutation spectra of cells with *rnhB* and *uvrA* deletions. Shown is a barplot with the mutation spectra for the *thyA* gene in trimethoprim-resistant mutants of the indicated strains in a *thyB::erm* genetic background. The spectrum of **Δ***rnhB* is significantly different from that of the **Δ***rnhB*, **Δ***uvrA* strain. The *thyA* reporter represents a selection for mutations at a specific locus with a limited sequence context (15, 16). In contrast, mutation accumulation lines measure mutations that occur in the near absence of selection sampling sequence context on a genome-wide scale (3). Therefore, the spectra observed at the *thyA* locus differ from what we observed genome-wide in the mutation accumulation lines because spectra are sequence context dependent (3, 4, 17). The p-value is the result of a Chi-square test. All variants are listed in Dataset 2.

References

1. Liao Y, Schroeder JW, Gao B, Simmons LA, & Biteen JS (2015) Single-molecule motions and interactions in live cells reveal target search dynamics in mismatch repair. *Proc Natl Acad Sci U S A* 112(50):E6898-6906.
2. Sanders GM, Dallmann HG, & McHenry CS (2010) Reconstitution of the B. subtilis replisome with 13 proteins including two distinct replicases. *Mol Cell* 37(2):273-281.
3. 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(5):692-697.
4. Sung W*, et al.* (2015) Asymmetric Context-Dependent Mutation Patterns Revealed through Mutation-Accumulation Experiments. *Mol Biol Evol* 32(7):1672-1683.
5. Bailey TL & Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2:28-36.
6. Rohland N & Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res* 22(5):939-946.
7. DeAngelis MM, Wang DG, & Hawkins TL (1995) Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic acids research* 23(22):4742-4743.
8. Baym M*, et al.* (2015) Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One* 10(5):e0128036.
9. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-1760.
10. Li H*, et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
11. Garrison E & Marth G (2012) Haplotype-based variant detection from short-read sequencing. *ArXiv e-prints.* 1207(3907).
12. Hall BM, Ma CX, Liang P, & Singh KK (2009) Fluctuation analysis CalculatOR: a web tool for the determination of mutation rate using Luria-Delbruck fluctuation analysis. *Bioinformatics* 25(12):1564-1565.
13. Youngman P, Perkins JB, & Losick R (1984) Construction of a cloning site near one end of Tn917 into which foreign DNA may be inserted without affecting transposition in Bacillus subtilis or expression of the transposon-borne erm gene. *Plasmid* 12(1):1-9.
14. Yao NY, Schroeder JW, Yurieva O, Simmons LA, & O'Donnell ME (2013) Cost of rNTP/dNTP pool imbalance at the replication fork. *Proceedings of the National Academy of Sciences of the United States of America* 110(32):12942-12947.
15. Walsh BW*, et al.* (2014) RecD2 helicase limits replication fork stress in Bacillus subtilis. *J Bacteriol* 196(7):1359-1368.
16. Sankar TS, Wastuwidyaningtyas BD, Dong Y, Lewis SA, & Wang JD (2016) The nature of mutations induced by replication-transcription collisions. *Nature* 535(7610):178-181.
17. 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):E2774-2783.