LETTER

doi:10.1038/nature11989

Accelerated gene evolution through replication–transcription conflicts

Sandip Paul[1], Samuel Million-Weaver[1], Sujay Chattopadhyay[1], Evgeni Sokurenko[1] & Houra Merrikh[1]

Several mechanisms that increase the rate of mutagenesis across the entire genome have been identified; however, how the rate of evolution might be promoted in individual genes is unclear. Most genes in bacteria are encoded on the leading strand of replication[1–4]. This presumably avoids the potentially detrimental head-on collisions that occur between the replication and transcription machineries when genes are encoded on the lagging strand[1–4]. Here we identify the ubiquitous (core) genes in *Bacillus subtilis* and determine that 17% of them are on the lagging strand. We find a higher rate of point mutations in the core genes on the lagging strand compared with those on the leading strand, with this difference being primarily in the amino-acid-changing (nonsynonymous) mutations. We determine that, overall, the genes under strong negative selection against amino-acid-changing mutations tend to be on the leading strand, co-oriented with replication. In contrast, on the basis of the rate of convergent mutations, genes under positive selection for amino-acid-changing mutations are more commonly found on the lagging strand, indicating faster adaptive evolution in many genes in the head-on orientation. Increased gene length and gene expression amounts are positively correlated with the rate of accumulation of nonsynonymous mutations in the head-on genes, suggesting that the conflict between replication and transcription could be a driving force behind these mutations. Indeed, using reversion assays, we show that the difference in the rate of mutagenesis of genes in the two orientations is transcription dependent. Altogether, our findings indicate that head-on replication–transcription conflicts are more mutagenic than co-directional conflicts and that these encounters can significantly increase adaptive structural variation in the coded proteins. We propose that bacteria, and potentially other organisms, promote faster evolution of specific genes through orientation-dependent encounters between DNA replication and transcription.

Concurrent DNA replication and transcription leads to conflicts that stall replication, especially on the lagging strand, where the two machineries meet head-on (Supplementary Fig. 1)[5–8]. Presumably, to avoid these encounters that can delay replication, bacteria have co-oriented most of their genes with replication by encoding them on the leading strand[1,2]. However, it is unclear why, even in species with a strong orientation bias such as *B. subtilis*, 25% of all genes and 6% of all essential genes remain on the lagging strand[1,2].

We investigated the relation between gene orientation and mutagenesis, by analysing the rate of mutations of ubiquitous (core) genes from five clonally divergent strains of *B. subtilis* (Supplementary Table 1). To be defined as a core gene, the gene had to be present in all five strains, have at least 95% nucleotide identity and at least 95% of the length coverage, to exclude the highly diverse genes affected by non-homologous gene shuffling[9]. Among the 759 core genes, 132 (17%) were encoded on the lagging strand (Supplementary Table 2). Of the 148 core genes that were determined to be essential[10], only six (4%) were on the lagging strand, consistent with the previously described strong orientation bias in the essential genes[1,2].

We compared the rates of silent or synonymous (dS), and amino-acid-changing or nonsynonymous (dN) mutations, of the core genes on the leading and lagging strands. The rate of synonymous mutation on the lagging was marginally (2%) higher than on the leading strand (Fig. 1a). In contrast, the rate of nonsynonymous mutations of the lagging strand genes was 42% higher than the genes on the leading strand (Fig. 1b, $P < 0.0001$), without a significant number of outliers (Supplementary Fig. 2). The relative increase of dN in the genes on the lagging compared with the leading strand was irrespective of the dS amount of the corresponding genes (Supplementary Fig. 3). We found a similar trend when we analysed the mutagenesis rate in all leading- or lagging-strand genes: that is, besides the core genes, on the two strands (Supplementary Table 3). However, among the core genes that were previously identified as essential, we did not detect a difference in structural variability between the two strands (Supplementary Fig. 4).

The observed difference between the rates at which leading- and lagging-strand genes vary may be due to orientation-dependent encounters between replication and transcription. It was previously suggested that transcript length contributes to severity of conflicts on the basis of the observation that gene operons oriented head-on tend to be of a relatively small size[4]. Interestingly, we found that most (82%) core genes on the lagging strand are not grouped together, and are organized as single genes rather than operons (Supplementary Fig. 5). Moreover, there is a significant bias for genes on the lagging strand to be significantly shorter on average, with the average gene size being 48% larger on the leading strand (681 compared with 459 base pairs). In particular, genes coding for proteins with a length longer than 200 amino acids are underrepresented on the lagging strand, with only 26% of genes (34 of 132 genes) being of this size category, whereas 48% of genes exceed 200 amino acids on the leading strand (Fig. 2a, $P < 0.0001$). These data are consistent with selection for genes on the lagging strand to be shorter, potentially to decrease the frequency of head-on conflicts.

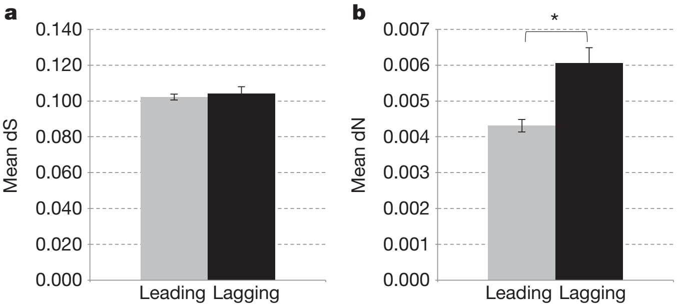

Figure 1 | Highly conserved 'core' genes show a higher rate of nonsynonymous mutations on the lagging compared with the leading strand. The mean values for the core genes in relation to diversity are presented. The dS (**a**) and dN (**b**) values are plotted for the leading (grey) and lagging (black) strands. Error bars, s.e.m. A statistically significant difference in dN between the two strands is detected with $*P < 0.0001$. Analysis of statistical significance was performed using the Z-test for dS and dN values (see Methods).

[1] Department of Microbiology, Health Sciences Building—J-wing, University of Washington, Seattle, Washington 98195, USA.

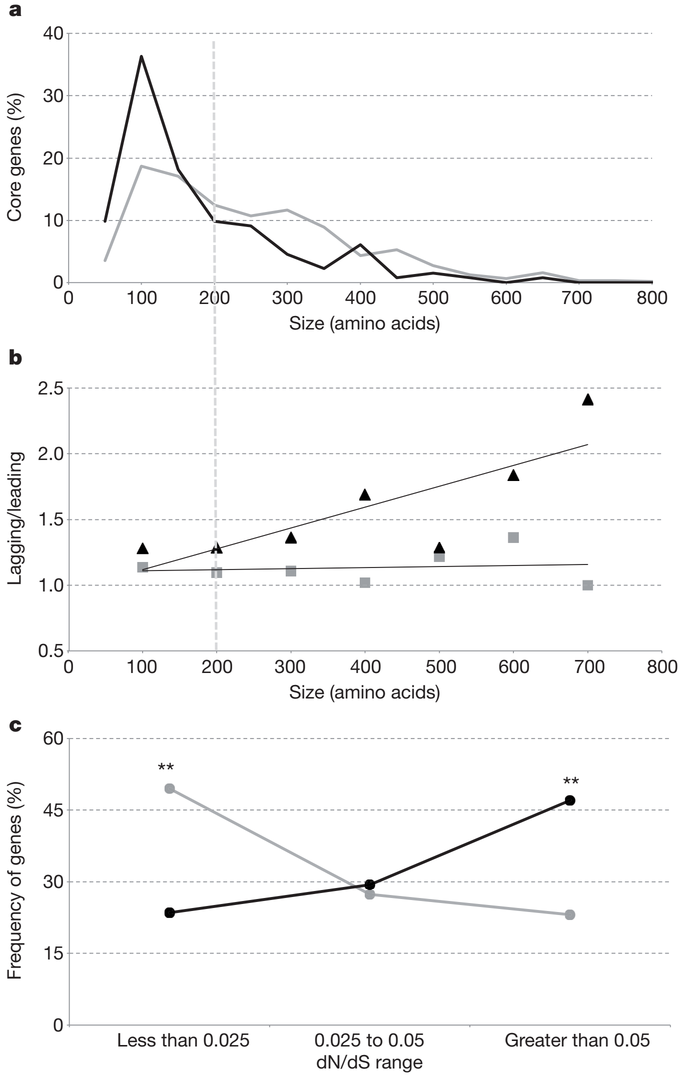

Figure 2 | For head-on genes, a significant positive correlation exists between increased length, mutagenesis and positive selection. a, The percentage of genes in each size category for lagging (black) and leading (grey) strands in increasing windows of 50 amino acids. **b**, The fold difference between either dN (black) or dS (grey) for lagging compared with the leading-strand genes in seven size categories ($R^2 = 0.65$ for dN and 0.02 for dS, over length). **c**, dN/dS ratios for genes longer than 200 amino acids on the leading (grey) and lagging (black) strands (**$P < 0.005$). P values were determined using a 2 × 2 χ^2 test.

Because gene size probably correlates with conflict level, we analysed the relation between nonsynonymous mutation rates and gene length. Indeed, there was a significant positive correlation between increased length and increased dN on the lagging compared with the leading strand (Fig. 2b), with the relative difference in dN being most apparent among the genes coding for proteins longer than 200 amino acids. In this size category (303 genes on the leading strand and 34 genes on the lagging strand), we also find a positive correlation between transcript abundance and dN, with the mean expression values being threefold higher in the group with higher dN amounts compared with those with the lower dN amounts, only in the lagging-strand genes (Supplementary Fig. 6 and Supplementary Table 5). Together, the positive correlation between increased mutation rates and gene length, as well as expression levels, suggest that replication–transcription conflicts are probably responsible for the increased mutation rates in the lagging-strand genes.

We examined the ratio of dN to dS, to evaluate the relative strength of negative selection against structural changes in the genes coding proteins more than 200 amino acids in length. A significantly higher proportion of the genes on the leading strand (about half) are in the very low dN/dS range of less than 0.025 compared with those on the lagging strand ($P = 0.004$, Fig. 2c). Furthermore, there are 25 genes on the leading strand but none on the lagging strand that lack any structural variation at all. Thus, genes that are under especially strong selection against nonsynonymous changes are heavily biased to be co-oriented with replication. In contrast, almost half of the genes on the lagging strand had a relatively higher dN/dS range of more than 0.05, significantly more than that on the leading strand ($P = 0.002$). To determine whether the increased dN/dS on the lagging-strand genes could be in part due to selection for nonsynonymous mutations, we looked for the presence of convergent amino-acid mutations, which is strong evidence for positive selection and adaptive evolution[11–13]. Among the genes encoding proteins longer than 200 amino acids, convergent mutations were detected in 24% of the core genes on the lagging strand in contrast to only 11% of the genes on the leading strand ($P < 0.04$) (Table 1 and Supplementary Table 4), indicating faster adaptive evolution of the corresponding genes.

Table 1 | Distribution of genes with convergent amino-acid mutations in lagging and leading strands

Strand	Number of genes*	Genes with convergent amino-acid mutations	2 × 2 χ^2 *P* value
Lagging	34	8	0.04
Leading	303	34	

* Genes that encode proteins with length of 200 or more amino acids.

We next determined whether replication–transcription conflicts are responsible for the increased mutation rates in the lagging-strand genes experimentally, using classic reversion assays. The positive correlation between increased gene length, as well as expression levels with increased dN in the lagging-strand genes, strongly suggest that conflicts are driving the mutagenesis of the head-on genes (Fig. 2b and Supplementary Fig. 6). We engineered strains auxotrophic for histidine biosynthesis, and integrated an ectopic copy of the histidine synthetase (*hisC*) gene with a premature stop codon at the 318th position, under the control of the isopropyl-D-thiogalactopyranoside (IPTG inducible promoter $P_{spank(hy)}$ onto the chromosome, in either the head-on or co-directional orientation (Fig. 3). Reversion assays and

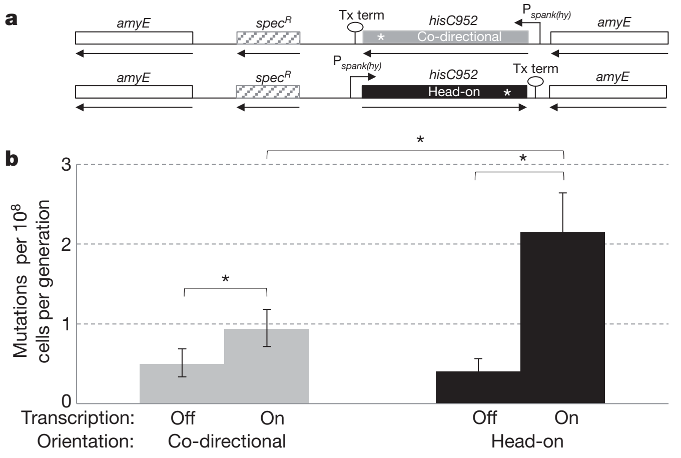

Figure 3 | Increased rates of mutagenesis in the head-on orientation are transcription dependent. a, The *amyE* locus containing reporter co-directionally (HM420) or head-on (HM419) to replication. **b**, Mutation rates on the basis of reversion assays for *hisC952* when the gene was oriented head on or co-directionally to replication, in the presence (transcription on) or absence (transcription off) of IPTG (*$P < 0.05$). Error bars, 95% confidence intervals. We confirmed by sequencing that the transcription-dependent mutations were indeed occurring in the *hisC* allele at *amyE* (Supplementary Table 9). YB955 did not show an increase in *hisC952* reversions with IPTG treatment (data not shown).

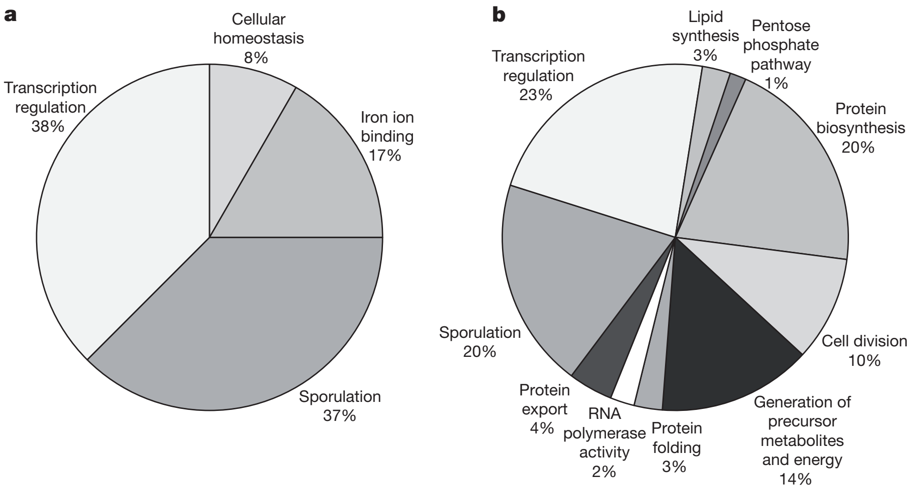

Figure 4 | Functional categories of leading- and lagging-strand genes. Pie charts presenting significantly ($P < 0.05$) overrepresented functional categories (as determined by medium stringency using DAVID[23]) of protein products from lagging (**a**) and leading (**b**) strands of replication for *B. subtilis* strain 168. The genes categorized for the lagging strand were *abrB*, *acuA*, *acuB*, *ahpC*, *ahpF*, *antE*, *cotM*, *cspD*, *cueR*, *def*, *fer*, *gerPF*, *hpr*, *ispG*, *katA*, *lexA*, *mrgA*, *nasE*, *sacY*, *sda*, *sigM*, *sigV*, *sinI*, *sinR*, *spoIISB*, *sspB*, *sspD*, *sspI*, *sspK*, *sspL*, *sspM*, *sspO*, *tnrA*, *ycnC*, *ydbP*, *ydgJ*, *yhdK*, *yjbI*, *ynzD*, *ypoP*, *ytzE*, *yutI*, *yuxN* and *ywoH*.

fluctuation analyses were performed in the presence and absence of transcription, for both orientations (see Methods). In the absence of transcription, the orientation of the gene did not affect mutation rates (Fig. 3 and Supplementary Table 6). When activated, transcription led to an increase in the mutation rate in both orientations, but the increase was much more prominent for the head-on orientation (Fig. 3 and Supplementary Table 6). The dependence of increased mutation rates on both transcription and orientation indicates that head-on replication–transcription conflicts lead to increased mutagenesis. These data demonstrate that, in the absence of transcription, mutagenesis rates of the two strands are not significantly different.

We analysed the functional categories of the core genes on the lagging strand and found enrichment in genes for (1) sporulation, (2) iron binding, (3) transcription regulation and (4) cellular homeostasis (Fig. 4). In general, most of these genes represent a variety of stress responses. For example, the transcription regulation genes include extracytoplasmic function (ECF)-type sigma factors (SigV and SigM), which are known to be the most variable of the sigma factors and are involved in responding to extracytoplasmic functions, as well as master regulators of biofilm formation (SinI and SinR), which control the transition from a free-swimming lifestyle to a stress-tolerant, biofilm lifestyle. We propose that these particular functions are kept relatively variable, possibly to produce population heterogeneity that can more easily respond to rapid changes in the environment.

Taken together, our data indicate that genes positioned on the lagging strand evolve at a significantly higher rate than those encoded on the leading strand, and that the mutagenic nature of the lagging strand is due, at least in part, to the conflicts between replication and transcription. The increased rate of mutagenesis in the genes on the lagging strand is probably detrimental for many, especially essential genes, and is probably responsible for the strong bias for these genes to be co-oriented with replication. Two models were proposed previously to explain the underlying cause of the bias against head-on orientation of genes: (1) the detrimental effect of increased mutagenesis, by Mirkin and Mirkin[14]; (2) increased rate of gene transcript truncations, by Rocha and Danchin[1]. Experimental work from *B. subtilis*, *Escherichia coli* and *Saccharomyces cerevisiae* has shown that genomic instability (which may increase the rate of mutations) is a potential consequence of head-on conflicts[8,15,16]. Indeed, our findings show that the bias for the co-directional orientation of genes in bacteria is at least partly determined by negative selection against increased rate of nonsynonymous mutations, consistent with the Mirkin and Mirkin model.

The other side of the coin, however, in contrast to the essential genes, is that the increased mutation rate could also be the reason for the head-on orientation of certain core genes: that is, head-on orientation could be promoted by positive selection. The action of positive selection is apparent from the high rate of convergent evolution in the proteins coded by genes on the lagging strand—a strong indication of their adaptive significance. In contrast to the nonsynonymous changes, synonymous mutations are generally (but not always) considered to be fitness-neutral in nature and their accumulation over time is expected to be driven largely by genetic drift. Thus accumulation of synonymous mutations could be a relatively slow process compared with positive selection for the adaptive changes, possibly explaining the low level of accumulation of the synonymous mutations in the lagging-strand genes compared with the nonsynonymous mutations. Although it is possible that there is a stronger negative selection against synonymous mutations in the lagging-strand genes owing to codon usage bias in highly expressed genes, we did not see a large difference between codon usage in the two strands' genes (Supplementary Fig. 7).

It is unclear how the phenomenon, described here, extends to other Gram-positive bacteria or to Gram-negative organisms such as *E. coli* and *Salmonella*. Orientation, transcription and rates of evolution in core genes have not been systematically investigated in these organisms, and previous studies looking at some aspects of these questions have produced contradictory results depending on methodology[17–22].

Replication–transcription conflict-mediated mutagenesis could be used by many organisms, including eukaryotes, as a universal strategy to link gene expression and evolution rate under selection. Genes on the lagging strand can evolve at a relatively fast rate because of head-on conflicts. Thus a simple switch in orientation, using short sequence homologies and recombination-dependent mechanisms, could facilitate evolution in specific genes in a targeted way. Investigating the main targets of conflict-mediated mutagenesis is likely to show far-reaching biological insights into adaptation and evolution of organisms.

METHODS SUMMARY

Strains are listed in Supplementary Tables 1 and 7. Plasmids are listed in Supplementary Table 8. Relevant properties of strains are described in the text. Strain constructions, growth conditions and bioinformatics analyses are described both in Methods and Supplementary Methods. Standard procedures were used for the reversion assays and calculation of mutation rates and are described (Methods). The rate of mutations (dS and dN) were determined on the basis of the possible number of changes compared with the actual number of changes observed, either for synonymous or nonsynonymous mutations (see Methods for details).

Full Methods and any associated references are available in the online version of the paper.

Received 8 November 2012; accepted 7 February 2013.

1. Rocha, E. P. & Danchin, A. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res.* **31,** 6570–6577 (2003).
2. Rocha, E. P. & Danchin, A. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nature Genet.* **34,** 377–378 (2003).
3. Rocha, E. P. The replication-related organization of bacterial genomes. *Microbiology* **150,** 1609–1627 (2004).
4. Price, M. N., Alm, E. J. & Arkin, A. P. Interruptions in gene expression drive highly expressed operons to the leading strand of DNA replication. *Nucleic Acids Res.* **33,** 3224–3234 (2005).
5. Merrikh, H., Zhang, Y., Grossman, A. D. & Wang, J. D. Replication–transcription conflicts in bacteria. *Nature Rev. Microbiol.* **10,** 449–458 (2012).
6. Mirkin, E. V. & Mirkin, S. M. Replication fork stalling at natural impediments. *Microbiol. Mol. Biol. Rev.* **71,** 13–35 (2007).
7. Pomerantz, R. T. & O'Donnell, M. Direct restart of a replication fork stalled by a head-on RNA polymerase. *Science* **327,** 590–592 (2010).
8. De Septenville, A. L., Duigou, S., Boubakri, H. & Michel, B. Replication fork reversal after replication-transcription collision. *PLoS Genet.* **8,** e1002622 (2012).
9. Chattopadhyay, S. *et al.* High frequency of hotspot mutations in core genes of *Escherichia coli* due to short-term positive selection. *Proc. Natl Acad. Sci. USA* **106,** 12412–12417 (2009).
10. Kobayashi, K. *et al.* Essential *Bacillus subtilis* genes. *Proc. Natl Acad. Sci. USA* **100,** 4678–4683 (2003).
11. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335,** 167–170 (1988).
12. Christin, P. A., Weinreich, D. M. & Besnard, G. Causes and evolutionary significance of genetic convergence. *Trends Genet.* **26,** 400–405 (2010).
13. Tenaillon, O. *et al.* The molecular diversity of adaptive convergence. *Science* **335,** 457–461 (2012).
14. Mirkin, E. V. & Mirkin, S. M. Mechanisms of transcription-replication collisions in bacteria. *Mol. Cell. Biol.* **25,** 888–895 (2005).
15. Srivatsan, A., Tehranchi, A., MacAlpine, D. M. & Wang, J. D. Co-orientation of replication and transcription preserves genome integrity. *PLoS Genet.* **6,** e1000810 (2010).
16. Kim, N., Abdulovic, A. L., Gealy, R., Lippert, M. J. & Jinks-Robertson, S. Transcription-associated mutagenesis in yeast is directly proportional to the level of gene expression and influenced by the direction of DNA replication. *DNA Repair* **6,** 1285–1296 (2007).
17. Veaute, X. & Fuchs, R. P. Greater susceptibility to mutations in lagging strand of DNA replication in *Escherichia coli* than in leading strand. *Science* **261,** 598–600 (1993).
18. Fijalkowska, I. J., Jonczyk, P., Tkaczyk, M. M., Bialoskorska, M. & Schaaper, R. M. Unequal fidelity of leading strand and lagging strand DNA replication on the *Escherichia coli* chromosome. *Proc. Natl Acad. Sci. USA* **95,** 10020–10025 (1998).
19. Maliszewska-Tkaczyk, M., Jonczyk, P., Bialoskorska, M., Schaaper, R. M. & Fijalkowska, I. J. SOS mutator activity: unequal mutagenesis on leading and lagging strands. *Proc. Natl Acad. Sci. USA* **97,** 12678–12683 (2000).
20. Szczepanik, D. *et al.* Evolution rates of genes on leading and lagging DNA strands. *J. Mol. Evol.* **52,** 426–433 (2001).
21. Lin, C. H., Lian, C. Y., Hsiung, C. A. & Chen, F. C. Changes in transcriptional orientation are associated with increases in evolutionary rates of enterobacterial genes. *BMC Bioinformatics* **12** (suppl. 9), S19 (2011).
22. Juurik, T. *et al.* Mutation frequency and spectrum of mutations vary at different chromosomal positions of *Pseudomonas putida*. *PLoS ONE* **7,** e48511 10.1371/journal.pone.0048511 (2012).
23. Dennis, G. Jr *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4,** 3 (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge the work of others that has contributed to our understanding of the topic, which we were unable to cite here. We thank B. Brewer, J. Mougous, F. Fang, M. Parsek and S. Miller for reading the manuscript. We thank C. Merrikh for comments on the manuscript and technical help with sequencing of revertants. We thank E. Robleto for the auxotrophic YB955 strain. H.M. was supported by start-up funds from the Department of Microbiology at the University of Washington. S.M.-W. was supported by the University of Washington Biophysics Training Grant. E.S., S.P, and S.C. were supported by National Institutes of Health (NIH) grants RC4 AI092828 and R01 GM084318.

Author Contributions H.M., E.S., S.P. and S.M.-W. designed experiments, S.P., S.C. and S.M.-W. performed the bioinformatics analysis and the experiments, H.M., E.S., S.P., S.C. and S.M.-W. analysed the data. H.M. and E.S. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.S. (evs@u.washington.edu) or H.M. (merrikh@uw.edu).

METHODS

Strains. For the bioinformatics analysis, fully assembled genomes of five *B. subtilis* strains were downloaded from NCBI GenBank. Those strains were *B. subtilis* subsp. *subtilis* str. 168, *B. subtilis* BSn5, *B. subtilis* subsp. *subtilis* str. RO-NN-1, *B. subtilis* subsp. *spizizenii* TU-B-10 and *B. subtilis* subsp. *spizizenii* str. W23 (Supplementary Table 1).

All experiments were performed using isogenic derivatives of *Bacillus subtilis* strain YB955 (*ile*, *leu*, *his*, *met*) (Supplementary Table 7). To construct a strain with an exogenous copy of histidine synthetase oriented head-on to replication, *hisC952* was first amplified from genomic DNA from YB955 using the primers HM386 (AAA AGT CGA CAA GGA GGT ATA CAT TTG CGT ATC AAA GAA CAT TTA AAA C) and HM396 (AAA GCA TGC TCC TTA TGA ATG GGA CTT ATA AAA TTT CAG CTA AAA TGG). The PCR fragment was digested with SalI and SphI and ligated into the integration vector pDR111 to generate the plasmid pHM392. This plasmid was introduced by double crossover into YB955 at *amyE* to generate strain HM419 (*leuC427 metB5 hisC952 amyE*::{P$_{spank(hy)}$-*hisC952* Head-On *spc*}). To construct a strain with the *hisC952* allele oriented co-directionally with replication, a fragment containing *hisC952* and P$_{spank(hy)}$ was amplified off of pHM392 using primers HM392 (AAA GAA TTC TAA CTC ACA TTA ATT GCG TTG C) and HM393 (AAA GGA TCC TGG CAA GAA CGT TGC TCG AGG). This fragment was digested with EcoRI and BamHI, inverted and ligated into pDR111 to generate plasmid pHM394. This plasmid was introduced by transformation into strain YB955, generating strain HM420 (*leuC427 metB5 hisC952 amyE*::{P$_{spank(hy)}$-*hisC952* Co-Directional *spc*}). See Supplementary Table 8 for a list of the plasmids. Candidate spectinomycin-resistant transformants were screened for disruption of the *amyE* locus by starch agar plate assay. Genomic DNA was isolated from amylase-deficient transformants and the chromosomal region containing *amyE* was amplified using primers HM205 (TCC AAA CTG GAC ACA TGG AA) and HM207 (AAA GAG GCG TAC TGC CTG AA). The resulting PCR fragments were sequenced to confirm integration of *hisC952* in the proper orientation.

Detection of common orthologous genes. To identify the core genes shared between all five strains of *B. subtilis*, we used *B. subtilis* subsp. *subtilis* str. 168 as the reference genome and extracted homologues from the other four strains with nucleotide sequence identity and length coverage values of at least 95%. We excluded the annotated pseudogenes and genes with either internal stop codons or non-ACGT characters.

Identification of leading- and lagging-strand genes. We considered the origin of replication to be positioned between ribosomal protein L34 and chromosomal replication initiation protein (DnaA), whereas the terminus region before replication terminator protein (Rtp) for all five strains. Depending on these locations, we defined the genes present in leading- and lagging strands for each strain. For analysis, we considered those genes that were present in the same strand (either leading or lagging) for all five strains. In total we found 759 core polymorphic genes: 627 in the leading strand and 132 in the lagging strand (Supplementary Table 2).

Molecular evolutionary analysis. Alignment of every gene set was performed by clustalW[24]. The rates of nonsynonymous (dN) and synonymous (dS) mutations for each gene were computed by the ratio of number of nonsynonymous and synonymous changes to the total number of nonsynonymous and synonymous sites respectively, using the mutation-fraction method[25]. To assess the statistical significance between any two sample sets, a *Z*-test was applied to dS and dN values[26].

Detection of convergent mutations was performed by Zonal Phylogeny software[27]. Zonal Phylogeny reconstructs an unrooted protein phylogram on the basis of nucleotide sequences where several alleles differentiated by synonymous changes are collapsed into single protein variant, followed by the mapping of phylogenetically unlinked repeated mutations at specific amino-acid positions (that is, convergent amino-acid mutations). DnaSPv5 was used to calculate the $2 \times 2\ \chi^2$ statistics[28].

Media and growth conditions. For all experiments, cells were grown at 37 °C, shaking at 260 r.p.m. Liquid cultures of *Bacillus subtilis* were grown in rich medium (Luria-Bertain (LB)) supplemented with spectinomycin (50 μg ml^{-1}) where appropriate. *E.* coli was grown in LB supplemented with ampicillin at 100 μg ml^{-1}. To determine viabilities, *B. subtilis* cells were plated on solid agar plates containing Spizizen's Minimal Medium (0.2 mg ml^{-1} ammonium sulphate, 1.4 mg ml^{-1} monobasic potassium phosphate, 0.6 mg ml^{-1} dibasic potassium phosphate, 0.1 mg ml^{-1} sodium citrate dihydrate, 0.02 mg ml^{-1} magnesium sulphate heptahydrate, 100 μg ml^{-1} glutamic acid, 5 μg ml^{-1} glucose), supplemented with isoleucine, methionine, leucine and histidine at 50 μg ml^{-1}. To select for reversion to prototrophy, cells were plated on solid agar plates containing Spizizen's Minimal Medium supplemented with isoleucine, methionine and leucine at 50 μg ml^{-1}.

Reversion assays. Fluctuation assays were performed to determine the mutation rate. Strains HM419 and HM420 were grown overnight in LB supplemented with spectinomycin at 50 μg ml^{-1}. Saturated cultures were diluted 1:10^4 into 2 ml LB medium and 2 ml LB medium supplemented with 1 mM IPTG. For these experiments, 12 parallel cultures for each strain and condition were grown for 5 h at 37 °C with shaking at 260 r.p.m. When cells reached saturation, 100 μl of culture was withdrawn, diluted 1:10^6 and spread onto solid agar plates containing Spizizen's Minimal Medium supplemented with all required amino acids to determine the number of viable cells present in the culture at plating. The remainder of the culture was pelleted by centrifugation, re-suspended in Spizizen's Minimal Salts and spread onto solid agar plates lacking histidine. Revertant colonies were counted after 48 h of growth at 37 °C. The experiment was repeated three times on different days.

Confirmation of reversions in the *hisC* gene at *amyE*. To determine whether reversions were occurring in the copy of *hisC* at *amyE*, 28 colonies, 24 head-on and four co-directional, from plates with IPTG were sequenced. The revertant colonies were grown up in selective (histidine− spectinomycin+) media and genomic DNA was prepared. The *amyE* region was amplified by PCR with primers HM205 (TCC AAA CTG GAC ACA TGG AA) and HM207 (AAA GAG GCG TAC TGC CTG AA). The PCR products were sequenced using the primer HM396 (AAA GCA TGC TCC TTA TGA ATG GGA CTT ATA AAA TTT CAG CTA AAA TGG). Changes at position 952 were analysed (Supplementary Table 9).

Calculation of mutation rates. Mutation rates were calculated using the Ma–Sandri–Sarkar maximum-likelihood method[29,30], with the aid of the Fluctuation Analysis Calculator. Statistical significance was determined using a standard two-tailed *t*-test.

24. Chenna, R. *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31,** 3497–3500 (2003).
25. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3,** 418–426 (1986).
26. Suzuki, Y. & Gojobori, T. in *The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny* 1st edn (eds Salemi, M & Vandamme, A.-M.) 283–311 (Cambridge Univ. Press, 2003).
27. Chattopadhyay, S., Dykhuizen, D. E. & Sokurenko, E. V. ZPS: visualization of recent adaptive evolution of proteins. *BMC Bioinformatics* **8,** 187 (2007).
28. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25,** 1451–1452 (2009).
29. Sarkar, S., Ma, W. T. & Sandri, G. H. On fluctuation analysis: a new, simple and efficient method for computing the expected number of mutants. *Genetica* **85,** 173–179 (1992).
30. Hall, B. M., Ma, C. X., Liang, P. & Singh, K. K. Fluctuation analysis CalculatOR: a web tool for the determination of mutation rate using Luria-Delbruck fluctuation analysis. *Bioinformatics* **25,** 1564–1565 (2009).