Rates and mechanisms of bacterial mutagenesis from maximum–depth sequencing

Justin Jee1,2, Aviram Rasouly1,3, Ilya Shamovsky1, Yonatan Akivis1, Susan R. Steinman1, Bud Mishra2§ & Evgeny Nudler1,3§

In 1943, Luria and Delbrück used a phage-resistance assay to establish spontaneous mutation as a driving force of microbial diversity1. Mutation rates are still studied using such assays, but these can only be used to examine the small minority of mutations conferring survival in a particular condition. Newer approaches, such as long-term evolution followed by whole-genome sequencing2,3, may be skewed by mutational ‘hot’ or ‘cold’ spots4,4. Both approaches are affected by numerous caveats5–7. Here we devise a method, maximum–depth sequencing (MDS), to detect extremely rare variants in a population of cells through error-corrected, high-throughput sequencing. We directly measure locus-specific mutation rates in *Escherichia coli* and show that they vary across the genome by at least an order of magnitude. Our data suggest that certain types of nucleotide misincorporation occur 104-fold more frequently than the basal rate of mutations, but are repaired in *vivo*. Our data also suggest specific mechanisms of antibiotic-induced mutagenesis, including downregulation of mismatch repair via oxidative stress, transcription–replication conflicts, and, in the case of fluoroquinolones, direct damage to DNA.

*De novo* mutations in bacteria remain a notoriously difficult target for high-throughput sequencing. Whereas *E. coli* mutate fewer than 1 in 109 bases per generation, high-fidelity polymerases used for library preparation polymerase chain reaction (PCR) cause errors in ~4 out of 108 bases8. Illumina machines misread ~1 in 108 bases9. Recent methods, such as barcoding of reads from the same original DNA molecule10, have lowered the error rate of sequencing. However, such methods can have low yields10,10 and do not address errors introduced by PCR. PCR errors can be overcome using duplex barcoding, which forms a consensus from both strands of a DNA template molecule11. However, even when a small region is targeted12, duplexing lowers yield even further. The mutational landscape of an RNA virus with mutation rate 104-fold greater than *E. coli* was recently mapped using ‘circle sequencing’. However, this technique is not designed for targeted coverage of a single locus, and its accuracy is limited by sequence read length10,13.

We introduce maximum–depth sequencing (MDS) for detecting extremely rare variants in any region of interest (ROI) in a population of cells (see Methods, Fig. 1a). By synthesizing unique barcodes directly onto the ROI of an original genomic DNA molecule and then copying that molecule using linear amplification, we increase yield (Fig. 1b) and substantially reduce both polymerase and sequencing errors (Fig. 1c). On mock cultures with single-nucleotide mutants spiked in at known concentrations, MDS reliably recovers the expected proportion of mutants at the lowest frequency tested, 10−6 (Extended Data Fig. 1). On *in vitro* synthesized DNA templates, MDS reduces the error rate to less than 5 × 10−10 per nucleotide sequenced (Fig. 1c, Extended Data Fig. 2). By increasing the number of reads used to call a consensus sequence (R), MDS can lower error rate indefinitely, given sufficient coverage (see Methods, error rate of MDS). Application of a second barcode after linear PCR increases accuracy at an even sharper rate and was used here to demonstrate library preparation efficiency (Extended Data Fig. 2; Supplementary Information: testing sample preparation and PCR efficiency.)

We used MDS to investigate mutation rates in *MG1655 E. coli* grown for ≤120 generations. We investigated six ~100-nucleotide ROIs: (1) part of the coding sequence (CDS) of the β subunit of RNA polymerase (rpoB), which confers rifampicin resistance when mutated; (2) the 3′ untranslated region (UTR) of rpoB; (3) the RNA polymerase ω subunit, rpoZ; (4) the CDS of cold-shock response gene *cspE*; (5) the centre of the CDS of penicillin-binding protein gene *mrcA*; and (6) the 3′ end of the CDS of *mrcA*. The last three genes, when knocked out, do not affect cell growth14,15. Whereas rpoB, rpoZ, and *cspE* are highly transcribed, *mrcA* is one of the least-transcribed genes in *E. coli* under normal conditions16. All ROIs have balanced AT and CG content, are transcribed on the leading strand, and lack homopolymers >8 nucleotides (nt).

Mutation rates in *E. coli* have been reported from 0.2 × 10−10 to 5 × 10−10 nt per generation16,17. Our calculated rate of mutation in *rpoB* CDS using synonymous substitutions is 4.1 × 10−10 nt per generation, comparable to the rate obtained in ref. 17 and at least one long-term evolution experiment using MG1655 (ref. 2). Yet it is also higher than rates calculated by fluctuation assay and long-term evolution on other strains (Fig. 2a, Extended Data Fig. 3). We performed fluctuation assays and recovered a similar spectrum and low rate of mutation to others using such approaches16. It is likely that the higher rate of mutation in *rpoB* obtained with MDS indicates a rate uninfluenced by negative selection, phenotypic lag, or imperfect plating efficiency15.

Mutation rate in nonessential *rpoZ*, and *cspE*, as well as *rpoB* UTR, is only slightly higher than that in essential *rpoB* CDS, but our calculated rate of mutation in the middle of *mrcA* is 3.5 × 10−9 nt per generation, an order of magnitude higher than the observed rate in *rpoB* CDS and significantly higher than the rates of mutation in all other ROIs (*P < 0.001* by ANOVA). The 3′ end of *mrcA* also has a higher rate of mutation than all other ROIs considered except for the middle of *mrcA*, suggesting spatial clustering of mutation rates. Comparison of genomes from several *E. coli* strains has suggested that clustered, highly transcribed genes are protected from mutation by an unknown mechanism13, a finding that has since been challenged16,18. Our results demonstrate that at least one gene with low transcription rate has significantly higher mutation rate than three others with high transcription rate.

The mutational spectrum from MDS matches that found in long-term sequencing experiments, with transition mutations favoured over transversions (Fig. 3a, Extended Data Figs 4, 5a). We also note an unexpected high frequency of C→A substitutions. These do not appear to be lasting mutations, as complementary G→T substitutions emerged with less than 0.1-fold frequency. A similar effect was found to a lesser extent for G→A and C→T substitutions. Increasing R did not significantly reduce these high substitution frequencies (Fig. 3b, Supplementary Information: model of damaged base pairs).

1Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, New York 10016, USA. 2Courant Institute of Mathematical Sciences, New York University, New York, New York 10012, USA. 3Howard Hughes Medical Institute, New York University School of Medicine, New York, New York 10016, USA. §These authors jointly supervised this work.
sugest that the majority of in vivo C→A substitutions are not due to damaged nucleotides. We found that in vitro templates synthesized with 8-oxoguanine (8-oxoG) resulted in low C→A substitution rates (Extended Data Fig. 3c), and treatment of in vivo DNA with formamidopyrimidine DNA glycosylase (FPG) did not change the observed substitution frequency (Extended Data Fig. 3c), further confirming that these C→A substitutions are probably not due to 8-oxoG. It is possible that As, or ribonucleotide As, are misincorporated into the genome at C sites in vivo. We found that neighbouring Cs are predictive of a higher frequency of C→A substitutions, suggesting that these transient substitutions cluster spatially along the genome, unlike polymerase or sequencing errors (Fig. 3c, Extended Data Figs 3b, 4, 5b).

In vivo, these misincorporations must be reversed after genome replication. However, our observations represent a snapshot of this dynamic process before repair can occur. Although these events would be invisible to conventional methods, the frequency of these substitutions, at ~10^{-5} per nucleotide, is over 10^4 times more frequent than the true rate of mutation.

To clarify which substitutions are transient rather than involved in ‘true’ mutation, we analysed DNA from bacteria collected after ≤20 generations, a short enough time period to expect few true mutations, given our sample size (Fig. 3a). We observed enrichment for most types of substitutions in our ≤120 generation trial over our ≤20 generation control, as would be expected from true mutations. However, C→A, A→G, and C→T substitutions occur in comparable frequency in the 20 and 120 generation trials, suggesting these substitutions reflect a continual process of base misincorporation and repair. We did not include these abundant A and T substitutions in our calculation of mutation rates. However, these findings suggest that the mechanism underlying the increase of AT content in E. coli grown for long periods is a dynamic process of misincorporation and repair.

We calculated short (≤12 base pairs (bp)) indel rates in mrcA, rpoB UTR, rpoZ, and cspE ROIs (Fig. 2b). Indel rate varies widely by position and size. As might be expected, 100% of the observed 1-bp indels occur at a site adjacent to a homopolymer. The frequency of 1-bp indels also increases with homopolymer length, potentially explaining why cspE, with an 8-bp T homopolymer, has the highest 1-bp indel rate. Longer indels are not localized to homopolymers and are positively correlated with substitution rates across all ROIs (Extended Data Fig. 6), supporting previous work suggesting that indels and substitutions spatially cluster in comparisons of genomes from divergent bacterial species. In all ROIs, deletions were detected at ≥10-fold frequency of insertions.

Single nucleotide indels and longer frameshifting mutations were also observed in rpoB CDS, albeit at low frequency, even though such mutations should be deleterious. As expected, the rate of in-frame indels was higher than the rate of frameshift indels of >1-bp length (Fig. 2b). Because of the low rate of indel errors from in vitro polymerases used here, it is plausible that the observed frameshift mutations are from inviable bacteria, as DNA from such cells may still enter our protocol. The recovery of frameshift indels, as well as the nonsignificant difference between rates of synonymous and nonsynonymous
The high frequency of C→A substitutions is consistent even as R increases. If these substitutions were polymerase errors due to damaged nucleotides, they should decline with increasing R faster than the line representing a model in which the polymerase makes C→A errors with 50% frequency for a subpopulation of DNA molecules (see Supplementary Information: model of damaged base pairs). c, C→A substitutions in vivo cluster in nucleotides with at least two neighbouring Cs within a 2-bp radius (P < 0.01 by t-test), unlike polymerase errors. Error bars are 95% CI upper bound.

Exposing E. coli to sub-inhibitory doses of multiple classes of antibiotics increases the rate at which bacteria acquire resistance to rifampicin. Whether this increase is caused by nucleotide oxidation, downregulation of mismatch repair, or an unrelated pathway, has become a topic of interest. We investigated the effect of sub-inhibitory doses of ampicillin and norfloxacin—a β-lactam and fluoroquinolone respectively—on mutation rate using MDS of rpoB CDS and mrcA CDS, as well as detailed fluctuation assays (Fig. 4a). Addition of ampicillin increased the rate of transition mutations in rpoB, a signature indicative of downregulated mismatch repair. In cells overexpressing catalase, basal mutation rate decreased by a factor of 8 (Fig. 4b), indicating that background oxidation contributes significantly to the basal mutation rate under non-stressed conditions. Addition of ampicillin during catalase overexpression did not increase this low rate (Fig. 4b). Overexpression of a catalase with inactivating point mutation H106Y during catalase overexpression did not increase this low rate (Fig. 4b).

Induction itself may thus be an important mechanism of stress-induced mutagenesis. The low translation rate of mrcA, coupled with our finding that rpoB UTR has a higher rate of mutation than the CDS, suggests that translation may be protective for highly transcribed genes. We constructed an additional strain in which IPTG-regulated mrcA has a canonical Shine–Dalgarno sequence and start codon, rather than its low-translation endogenous sequence. Increasing translation decreased substitution rate in the IPTG-induced state by a factor of 50% and a factor of 75% when high-frequency (C→A, for example) substitutions are excluded (Fig. 4c). Although translation does not lower the mrcA mutation rate to rpoB levels, it probably contributes to protection of highly transcribed genes (Supplementary Information: relationship between transcription, translation, and mutation rate).

There is debate as to whether highly transcribed genes in bacteria have a higher18,27 or lower4 mutation rate than other genes. Our analysis in E. coli shows that mrcA has a higher basal rate of mutation than more highly transcribed genes. Yet interestingly, addition of ampicillin increased transversions and indel formation in mrcA, but not in rpoB CDS (Fig. 4a). It is known that mrcA undergoes mild induction upon addition of ampicillin24. To study the effect of transcription on mutagenesis further, we created a strain in which a chromosomal copy of mrcA is regulated by an isopropyl-β-D-thiogalactopyranoside (IPTG) promoter. Induction of mrcA transcription increased the frequency of all classes of mrcA substitution and indel~8-fold more than when wild-type cells were exposed to ampicillin (Fig. 4c). These results suggest that although, in basal conditions, cells may have a means of protecting the most highly transcribed genes, co-directional collisions between transcription and replication machinery, which can cause double-strand breaks, are themselves mutagenic. Induction itself may thus be an important mechanism of stress-induced mutagenesis30.

Outside extensions to MDS would allow for analysis of many ROIs simultaneously and assembly of longer ROIs (Supplementary Information: MDS protocol). MDS may also be useful in detection of genetic abnormalities in cell-free DNA due to foetal mutations or cancer.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.
3. Lee, H., Popodi, E., Tang, H. & Foster, P. L. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc. Natl Acad. Sci. USA* **109**, E2774–E2783 (2012).

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

5. Lencki, R. E., Statak, M. & Ayala, F. J. Mutation and selection in bacterial populations: alternatives to the hypothesis of directed mutation. *Proc. Natl Acad. Sci. USA* **86**, 2775–2778 (1989).

6. Wielgosz, S. et al. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc. Natl Acad. Sci. USA* **110**, 222–227 (2013).

7. Drake, J. W. Contrasting mutation rates from specific-locus and long-term mutation-accumulation procedures. *G3 (Bethesda)* **2**, 483–485 (2012).

8. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl Acad. Sci. USA* **108**, 9530–9535 (2011).

9. Minoche, A. E., Dohm, J. C. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.* **12**, R112 (2011).

10. Lou, D. I. et al. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc. Natl Acad. Sci. USA* **110**, 19872–19877 (2013).

11. Schmitt, M. W. et al. Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl Acad. Sci. USA* **109**, 14508–14513 (2012).

12. Schmitt, M. W. et al. Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat. Methods* **12**, 423–425 (2015).

13. Acedo, A., Brodsky, L. & Andino, R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505**, 686–690 (2014).

14. Baba, T. et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).

15. Taniguchi, Y. et al. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010).

16. Ganbayan, L. et al. Use of the pO6 gene to determine the specificity of base substitution mutations on the Escherichia coli chromosome. *DNA Repair (Amst.)* **2**, 593–608 (2003).

17. Drake, J. W. A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl Acad. Sci. USA* **88**, 7160–7164 (1991).

18. Chen, X. & Zhang, J. No gene-specific optimization of mutation rate in Escherichia coli. *Mol. Biol. Evol.* **30**, 1559–1562 (2013).

19. McDonald, M. J., Wang, W.-C., Huang, H.-D. & Leu, J.-Y. Clusters of nucleotide insertions/deletions and the nucleotide diversity in bacteria. *Mol. Biol. Evol.* **26**, 2353–2361 (2009).

20. Zhu, L., Wang, Q., Tang, P., Araki, H. & Tian, D. Genomewide association between insertions/deletions and the nucleotide diversity in bacteria. *Mol. Biol. Evol.* **26**, 311–320 (2010).

21. Dwyer, D. J. et al. Antibiotics induce redox-related physiological alterations as part of their lethality. *Proc. Natl Acad. Sci. USA* **111**, E2100–E2109 (2014).

22. Gutierrez, A. et al. β-Lactam antibiotics promote bacterial mutagenesis via an RpoS-mediated reduction in replication fidelity. *Nat. Commun.* **4**, 1610 (2013).

23. Liu, Y. & Imlay, J. A. Cell death from antibiotics without the involvement of reactive oxygen species. *Science* **339**, 1210–1213 (2013).

24. Lang, G. I. & Murray, A. W. Estimating the per-base-pair mutation rate in the yeast Saccharomyces cerevisiae. *Genetics* **178**, 67–82 (2008).

25. Pohlhaus, J. R. & Kreuzer, K. N. Norfloxacin-induced DNA gyrase cleavage complexes block Escherichia coli replication forks, causing double-stranded breaks in vivo. *Mol. Microbiol.* **56**, 1416–1429 (2005).

26. Merrick, H., Zhang, Y., Grossman, A. D. & Wang, J. D. Replication-transcription conflicts in bacteria. *Nat. Rev. Microbiol.* **10**, 449–458 (2012).

27. Sangurdekar, D. P., Srienc, F. & Khodursky, A. B. A classification based framework for quantitative description of large-scale microarray data. *Genome Biol.* **7**, R32 (2006).

28. Dutta, D., Shatalin, K., Epstein, V., Gottesman, M. E. & Nudler, E. Linking RNA polymerase backtracking to genome instability in E. coli. *Cell* **146**, 533–543 (2011).

29. Rosenberg, S. M. Evolving responsively: adaptive mutation. *Nat. Rev. Genet.* **2**, 504–515 (2001).

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

Acknowledgements We thank A. Heguy and the NYU Genome Technology Center, which is partially supported by the Cancer Center Support Grant, P30CA016087, at the Laura and Isaac Perlmutter Cancer Center. This work used computing resources at the High Performance Computing Facility of the Center for Health Informatics and Bioinformatics at the NYU Langone Medical Center. We thank D. Dwyer and K. Shankarling for materials, and T. Artemyev for his contribution. This work was supported by NIH grant R01GM107329 and HHMI (E.N.) and NCI PSOC grant U54 CA193313 (B.M.). J.J. was supported by the NYU Medical Scientist Training Program and a National Defense Science and Engineering Graduate Fellowship.

Author Contributions J.J. and I.S. designed the MDS protocols. J.J., A.R., and Y.A. performed the experiments. J.J., B.M., S.S., and I.S. performed the data analysis. J.J. and E.N. designed the biological experiments. J.J., A.R., and Y.A. performed the experiments. J.J., B.M., S.S., and I.S. performed the data analysis. J.J. and E.N. wrote the manuscript with input from all co-authors. B.M. and E.N. supervised the research.

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.N. (evgeny.nudler@nyumc.org) or B.M. (mishra@nyu.edu).

Reviewer Information Nature thanks N. Luscombe, I. Martincorena, J. Wang and the other anonymous reviewer(s) for their contribution to the peer review of this work.*
METHODS

Data reporting. No statistical methods were used to predetermine sample size.

Maximum-depth sequencing. First, genomic DNA is treated with a restriction enzyme, which cleaves at the 3' end of the ROI. A single PCR cycle is performed with barcoded primers annealing to the 3' end of the ROI. Because of the exposed 3' site on the genomic DNA molecule left by the restriction enzyme, the genomic DNA molecule acts as a 'primer', causing the barcode and an adaptor to be synthesized onto the end of the ROI. This synthesis effectively attaches the barcode to the original genomic DNA molecule. Unused barcodes from primers are removed, and N cycles of linear amplification are performed using only primers to the forward adaptor sequence. This step is important for screening polymerase errors. The polymerase may make an error in any single round of synthesis, increasing the probability of generating a faulty read by N, but by copying the same original DNA molecule multiple times, the probability of recovering a defective copy after analysis is reduced by a factor of N^k, where k is the number of independent reads used to build a consensus sequence. Thus the total error reduction is 1/N^N^k ·1/k (see below). In this study, typically N = 12 and R = 3, although the empiric value of N after accounting for inefficiencies in PCR is somewhat lower (see Extended Data Fig. 2 and Supplementary Information: testing sample preparation and PCR efficiency). Note that one could also attach a second barcode to each read after linear amplification but before exponential amplification—doing so could allow one to reduce the error rate even further by ensuring multiple reads from the linear amplification step are used in the analysis. By targeting a ROI, we can also use paired-end sequencing to increase yield. Detailed error rate spectra for both Phusion and Q5 polymerase are measured and reported in Supplementary Table 1. It should be noted that when Phusion and Q5 polymerase are measured and reported in Supplementary Table 1, the error rate for those shown in Fig. 1 are derived almost entirely from transition substitutions typical of PCR polymerases, and that for other kinds of substitutions, error rate is virtually nonexistent. In MDS, each read represents additional 1/3 reads produced by the linear amplification step is 10^9-fold coverage using an Illumina HiSeq machine. For details on the specific enzymes, primers, and PCR conditions used in this study see Supplementary Information: MDS protocol. For details on consensus base calling, see Supplementary Information: analysis.

Error rate of MDS. Sources of error include damaged DNA during extraction, polymerase errors during PCR, and sequencing errors. Because our goal is to identify rare mutants, we consider error as the rate of false positives, which affect mutant frequency to a much larger extent than false negatives. If the probability of a single nucleotide X being misread as Y owing to polymerase error is Ppol,XY, and the rate of the corresponding sequencing error is Pseq,XY, then the probability that X will be read as Y owing to either source of error in a standard sequencing protocol is

$$P_{XY} = P_{pol,XY} + P_{seq,XY}$$

As discussed briefly in the main text, in our assay, the total polymerase error rate Epol,XY can be derived as follows (for visual aid, see Extended Data Fig. 10). For convenience, Ppol,XY will hereafter be referred to as p. After exponential PCR, there are N pools of reads, each derived from one of the original linear amplification steps. The probability of having k pools derive from an original polymerase error is binomially distributed. Furthermore, because N^k < 1, the distribution is Poisson.

$$\sum_{k=0}^{N} \left( \frac{N^k}{k!} \right) \cdot (1-p)^{N-k} \approx \left( \frac{N^k}{k!} \right) e^{-Np}$$

The probability of a false positive is the probability that all R reads used to form a consensus came from one of the k' error' pools

$$\sum_{k=0}^{N} \left( \frac{N^k}{k!} \right) \cdot (1-p)^{N-k} \approx \left( \frac{N^k}{k!} \right) e^{-Np} = \frac{B_0(Np)}{N^R}$$

Where M^k is the Rth moment of the Poisson distribution in equation (2) and B_0 is the Rth Bell polynomial. Because N^k < 1, an upper bound on this error form can be written as follows:

$$E_{pol,XY} = \frac{B_0(Np)}{N^R} \leq p \cdot B_0(1) \leq p \cdot \frac{0.792R}{ln(R+1)}$$

Where the upper bound of the Rth Bell number B_0(1) is from ref. 31. These bounds will decrease rapidly as R increases, given that R ≤ N.

We note that in practice, the probability that the same error would emerge in k > 1 reads produced by the linear amplification step is ~10^-12, so low that the expected number of such multi-errors for all the nucleotides sequenced in this study is <1. With this in mind, it is possible to simplify equation (4) so that the Bell number term is a non-contributor to the total error. Under this assumption, the probability of false positive is

$$E_{pol,XY} \approx \frac{p}{N^R}$$

The above formula for Epol,XY only takes into account errors introduced during linear amplification. However, the maximum error that could be contributed during a subsequent round of doubling, or exponential, PCR (D) can be found by substituting N2^D for N in the equation above. The sum of all possible errors from all rounds of PCR would thus be

$$E_{total,XY} \approx \sum_{D} \frac{p}{(N2^D)^R}$$

For R = 2, this will be a geometric series with sum no greater than 2Epol,XY. For R > 2, the sum will be closer to Epol,XY.

The error rate of sequencing after forming a barcode, as discussed thoroughly in other texts^{10} is the probability that the same error happens R times

$$E_{seq,XY} = (P_{seq,XY})^R$$

Where R is the number of 'not necessarily independent' reads used to form a consensus (that is, overlapping paired-end sequences of the same read are included). If single-end sequencing is used, R = 1. If paired-end sequencing is used, a maximum of R = 2R not necessarily independent reads are used.

Alternatively, one could estimate Eseq,XY based on the sum of the quality scores of the R reads contributing to the consensus, but in practice we find this to be unnecessary because sequencing errors are not the major contributor to overall error when R > 2.

The total error rate for any given nucleotide position is the sum of all E_{XY}, X = Y, for a given X. The values reported in the main text and Fig. 1c are total error. Raw polymerase and sequencing error rates^{8} are shown in Supplementary Table 1. Note that this model is also the basis for the damaged base-pair analysis presented in Fig. 3b and the Supplementary Information.

Growth and mutation rate analysis. E. coli were streaked onto Luria-Bertani (LB) agar from freezer stocks and grown at 30 °C for 24 h. According to plating and colony-forming unit (c.f.u.) counting, the average number of cells in such colonies is 3 x 10^8 (thus the number of generations is ln(3 x 10^8) = 19.5. Bacteria from a single colony were used to inoculate a small liquid culture (1 ml LB broth in a round-bottom tube). For the purposes of generation counting, it is assumed that after the transition to growing in liquid, growth occurs for only ~3 generations. The culture was grown in a 37 °C shaker to allow for the transition to growth in broth for 12 h, after which a measurable optical density could be reliably detected. 4 ml (~10^8 bacteria) were transferred to a fresh 100 ml LB liquid culture (in a 250 ml Erlenmeyer flask). Liquid cultures were grown for 24 h on a 37 °C shaker, to a density of 2.5 x 10^9 bacteria according to cell counts (for a total of 2.5 x 10^11 bacteria). This process was repeated 9 times. The average number of generations a bacterium would have grown in each liquid culture is

$$\ln(2.5 x 10^{11}) \approx 10.1 \text{ generations}$$

Thus the average total number of generations is 19.5 + 3 + 9 x 10.1 = 113.

In addition to the large passage size, we stop passaging hundreds of generations before selective sweeps are expected to occur^{23} and, importantly, long before selection for a hyper-mutating strain might be expected^{6}. We also performed simulations to test the effects the probability that any two bacteria have the same founder given expected conditions of passage size (see Supplementary Information: calculation of mutation rate).

Mutation rates $\mu$ in our assay are chosen to maximize the likelihood of recovering the mean mutant frequency for substitutions of a given type X → Y, which we find are well approximated by a Poisson process over a certain number of generations (in this case 113).

$$\mu_{XY} \approx f(Y) - E_{XY}$$

More precisely, the frequency is defined as the number of barcode groups with a given mutation divided by the total number of barcode groups under consideration. For example, if R = 3, then f(Y) is the number of read families of size ≥ 3 with mutation Y divided by the total number of read families of size ≥ 3. Mutation rates given in Fig. 2 are computed from the average across all X of $\sum_{Y \neq X} \mu_{XY}$. With C → A, G → T, C → T, and G → A substitutions, excluded for aforementioned reasons. In their place, a correction term the average transversion or transition...
rate based on all other substitutions) is used so that the mutation rate is not systematically underestimated.

Four biological replicates of each condition were grown. All liquid cultures, including the small founding culture, had the possible addition of 1 μl ml⁻¹ ampicillin or 15 ng ml⁻¹ norfloxacin. Cultures for the short-term growth assay and mock culture were grown similarly except without passaging (Supplementary Information: mock culture and short growth assay).

**Strains.** MG1655 *E. coli* were used as wild-type cells for all experiments. The IPTG-regulated *mrcA* strain MG1655 and IPTG-regulated strain with modified Shine–Dalgarno were recombined according to ref. 33. For details, see Supplementary Information: strains. For details on the catalase overexpression mutant and inactive H106Y catalase overexpression mutant see ref. 22. In the *mutS* knockout strain, MG1655 *mutS* was replaced with a kanamycin resistance cassette.

**Preparation of DNA samples**

*Genomic DNA.* Up to 5 ml of bacterial liquid culture were spun down (see later section for specific growth conditions). Cells were resuspended in 500 μl Tris-EDTA buffer (pH 7.5), and 1,000 units of Ready-Lyse (Epigenetec) added, before incubation at room temperature for 1 h and freezing at ~80 °C overnight. Genomic DNA extraction was performed using Qiagen genomic tip (100G), but without lysozyme and quantified using Nanodrop.

In vitro DNA. Single-stranded oligonucleotides with sequences corresponding to MG1655 rpoB at position 1511–1632 and *mrcA* at 1258–1379 were ordered from IDT and resuspended in deionized water. These oligonucleotides were used directly as input to the Extreme-depth sequencing protocol above for calculation of error rate in Fig. 1 and the ‘negative control’ rows in Supplementary Table 1. Note, as expected from quality control reports from IDT, we found a large number of indels in the *in vitro* synthesized templates (~1% of molecules had some type of indel). However, the fact that we recovered a low substitution rate could be used to confirm the chemical purity of the mononucleotide pools used for synthesis by IDT.

Separately, 10 ng of the same DNA oligonucleotides were used as templates for a standard 20-cycle exponential PCR reaction with only the ROI-annealing component of the forward and reverse primers above using either Q5 or Phusion polymerase. The amplified DNA was used as input into the MDS protocol and used to calculate the intrinsic substitution error rate of those two polymerases as reported in Supplementary Table 1.

**Sequencing depth.** On average, we divide single HiSeq Rapid Runs of ~240 M reads into four different ‘conditions’, each corresponding to a particular ROI from bacteria grown under a certain condition. The ~60 M reads of each condition are further subdivided in order to process triplicate or quadruplicate trials.

We recover ~2.5 M total barcode ‘families’ for each condition using our threshold of R ≥ 3 (for the purposes of calculating total yield, we divide by 2 since each read is pair-end sequenced). We examine ~100 bp per ROI, thus providing a significant pool from which to observe mutations. There is an interesting level of variability across quadruplicates, likely due to stochastic variation when combining and purifying DNA samples and in binding to the HiSeq flowcell itself (Extended Data Fig. 1b, c). Note that when mutation frequencies are averaged over multiple trials, each trial is weighted according to its relative representation in terms of number of families.

**Fluctuation assays.** Fluctuation assays were carried out as in ref. 16. We picked single colonies of *E. coli* as above and grew them in 1 ml Luria-Bertani (LB) broth overnight. 0.1 μl from this starter culture was used to inoculate 25 separate trial cultures. Each trial culture was grown (in a 37 °C shaker) to either an optical density (OD₆₀₀) of 0.3 (for exponential growth trials) in 2 ml LB broth or for 24 h (for saturation) in 0.2 ml LB broth and plated cultures on Petri dishes containing LB agar with 100 mg ml⁻¹ rifampicin. Colonies were grown for 48 h in 30 °C and c.f.u. were counted. The rpoB region conferring rifampicin resistance was sequenced and used to compute the mutational profiles in Fig. 4. Number of bacteria per culture was calculated by serial dilution, plating on LB agar, and counting c.f.u. and used to compute the mutational profiles in Fig. 4. Number of bacteria per culture was calculated by serial dilution, plating on LB agar, and counting c.f.u.

**Availability.** Raw sequence data are available from Sequence Read Archive (SRA301985). Code is available from https://github.com/justinjee/MDS and https://github.com/susinmotion/barcode_tries.

31. Berend, D. & Tassa, T. Improved bounds on Bell numbers and on moments of sums of random variables. *Probab. Math. Stat.* 30, 185–205 (2010).
32. Woods, R. J. et al. Second-order selection for evolvability in a large *Escherichia coli* population. *Science* 331, 1433–1436 (2011).
33. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* 97, 6640–6645 (2000).
34. 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).
Extended Data Figure 1 | MDS accuracy and yield. a, Mock culture composed of rpoB point mutants of known concentration was sequenced using MDS. Output concentrations of each point mutant recovered from $R = 2$ analysis are plotted against its input concentration (see Supplementary Information Table 2 for details). b, c, Distribution of the sizes of barcode families in four trials, shown as $\log_{10}$ (number of barcode families) per trial versus size of barcode family in reads ($R$). b, Trials used for the calibration run shown in a ($\sim 100$ M reads total, divided into four trials). c, Representative quadruplicate trials (from rpoB of wild-type bacteria grown in LB broth with no antibiotics) taking up a total of one quarter of the output of a HiSeq rapid run, a total of $\sim 60$ M reads.
Extended Data Figure 2 | Dual-barcode MDS. a, Barcodes are attached to original DNA molecules as per MDS protocol. After linear amplification, a second barcode is attached to the opposite end of each read (see Supplementary Information: testing sample preparation and PCR efficiency). Exponential PCR is then performed. In the analysis phase, reads can be grouped both by primary barcode (that is, a classic MDS barcode family) and a second barcode corresponding to a ‘subfamily’ of reads with the same parent from a particular linear amplification step before exponential amplification. b, The probability that for a given family only reads of one subfamily are recovered (a ‘homogenous’ barcode) decreases exponentially with $R$. For example, for $R = 3$, the probability all 3 reads are of the same subfamily is 0.02. c, We show the number of reads in each subfamily, sorted within each column by subfamily size, for the 1,500 largest primary barcode families in the experiment. For families of such size, it is unlikely that a single subfamily will account for more than 25% of the total number of reads recovered from that family.
Extended Data Figure 3 | Substitution frequency controls.

a. Empirically, average substitution frequency (with high frequency substitutions such as C→A excluded) stabilizes as R increases. Note, substitution frequencies are not normalized by number of generations.

b. Empirical sequencing C→A error rate at C→A mutational hotspots with neighbouring Cs (same as those in Fig. 3c) versus all other positions.

c. C→A substitution frequencies when 10% 8-oxoG is synthetically added to in vitro DNA and in FPG-treated samples. Frequencies are reported from ROI positions with potential 8-oxoG incorporations as described in template 'rpoB_reverse_complement_8-oxo-Dg'. Frequencies are reported at $R = 2$ level. For $R > 2$, no C→A substitutions were found in 72,646 in vitro template sites. Data represent biological triplicates. Error bars are standard deviation.
Extended Data Figure 4 | Substitution rates per locus. Positive frequencies denote synonymous substitutions. Negative frequencies denote nonsynonymous substitutions. a, c, Values are averaged across quadruplicate trials. b, d, In vitro synthesized DNA has undergone 20-cycle PCR amplification using Q5 polymerase.
**Extended Data Figure 5 | Mutational spectra and contexts.**

**a.** Substitution frequencies of all ROIs after ~120 generations of growth. Note that values are not normalized for the number of generations and are thus true frequencies, rather than mutation rates. **b.** Mutation frequencies are shown in context of their 5′ (A, C, G, or T on the x axis) and 3′ (A, C, G, or T on the y axis) neighbours. **c.** The relative relationship between in vivo substitution frequencies and expected errors due to sequencing and PCR (from in vitro DNA assays) is poorly described by a linear approximation ($R^2 = 0.27$). Furthermore, the recovered frequency from in vivo substitutions ($R = 3$) is higher than the rate of error (equivalent frequencies would be represented by the dotted line), even with the relatively relaxed read-cutoff threshold of $R = 2$ (the sequencing + PCR error with an $R = 3$ cutoff is approximately an order of magnitude lower). Templates are rpoB CDS and mrcA ROIs.
Extended Data Figure 6 | Comparing substitution rate and indel rate across 5 ROIs reveals a positive correlation. Pearson correlation coefficient $= 0.76$. 

$R^2 = 0.57938$
Extended Data Figure 7 | Rate of rifampicin resistance per generation.

a–d. As calculated in fluctuation assays in wild-type cells grown in exponential phase only (a), wild-type cells grown to saturation (b), katG overexpression mutant grown to saturation (c) and inactive katG (H106Y point mutation) overexpression mutant grown to saturation (d). Growth in LB broth was supplemented with possible subinhibitory doses of ampicillin (amp), norfloxacin (nor), or gentamycin (gen). Rates are mean. Error bars are 95% CI. N = 25 (see Methods: fluctuation assays).
Extended Data Figure 8 | Transversion and transition rates (per nucleotide-generation). As calculated in fluctuation assays in anaerobic conditions (a) and in a mutS knockout (b). Note that because the transition (Ts) rate was high in MutS strains, transversion mutations could not be detected. Rates are mean. Error bars are 95% CI. N = 25 (see Methods: fluctuation assays).
Extended Data Figure 9 | Rates of rpoB and mrcA substitutions in the presence of antibiotics as calculated by MDS. Asterisks indicate cultures grown separately and prepared with Phusion rather than Q5. Although not shown, we note that only in-frame (3×) indels were observed in rpoB in fluctuation assays, as expected since frameshift indels would be deleterious. These increased in frequency by a factor of 10 on addition of norfloxacin.
Extended Data Figure 10 | Schematic depicting the mathematical derivation of the false positive rate of MDS due to polymerase error. 
a. The origin of various terms used in equations (2)–(7). b. Illustration of an example calculation of false positive rate given more ‘intuitive’ values of N, R and P. The false positive rate is calculated in a way that accounts for the possibility that an error in one or more ‘linear’ cycles propagates to a whole family of reads. The number of reads with an error (k) is Poisson distributed according to equation (2). The probability of a false positive is the sum of the probabilities that all R reads come from one of k families, for all possible k, according to equation (3). Note that in practice, $P < 10^{-6}$, and in our study $N = 12$, $R > 2$, making the false positive rate much lower (see Fig. 1).