Universally high transcript error rates in bacteria

Weiyi Li[1†], Michael Lynch[1,2*]

[1]Department of Biology, Indiana University, Bloomington, United States; [2]Center for Mechanisms of Evolution, The Biodesign Institute, Arizona State University, Tempe, United States

Abstract Errors can occur at any level during the replication and transcription of genetic information. Genetic mutations derived mainly from replication errors have been extensively studied. However, fundamental details of transcript errors, such as their rate, molecular spectrum, and functional effects, remain largely unknown. To globally identify transcript errors, we applied an adapted rolling-circle sequencing approach to *Escherichia coli*, *Bacillus subtilis*, *Agrobacterium tumefaciens*, and *Mesoplasma florum,* revealing transcript-error rates 3 to 4 orders of magnitude higher than the corresponding genetic mutation rates. The majority of detected errors would result in amino-acid changes, if translated. With errors identified from 9929 loci, the molecular spectrum and distribution of errors were uncovered in great detail. A G→A substitution bias was observed in *M. florum*, which apparently has an error-prone RNA polymerase. Surprisingly, an increased frequency of nonsense errors towards the 3′ end of mRNAs was observed, suggesting a Nonsense-Mediated Decay-like quality-control mechanism in prokaryotes.

***For correspondence:** mlynch11@asu.edu

Present address: †Joint Initiative for Metrology in Biology, SLAC National Accelerator Laboratory, Stanford University, Stanford, United States

Competing interests: The authors declare that no competing interests exist.

Funding: See page 12

Received: 06 January 2020
Accepted: 28 April 2020
Published: 29 May 2020

Reviewing editor: Christian R Landry, Université Laval, Canada

Introduction

Transcript errors refer to any inconsistencies between RNA transcripts and their corresponding genomic loci. They can occur during ribonucleotide (rNTP) incorporations by RNA polymerases and/or via post-transcriptional modifications. Errors on RNA transcripts may directly cause dysfunctions due to the regulatory roles of small RNAs and the fate determination of mRNAs by RNA structural motifs (**Strathern et al., 2012**). Such errors can also indirectly induce various effects at the protein level. Transcript errors can inactivate proteins and result in a loss-of-function (**Gordon et al., 2013**). They can also indirectly give rise to misfolded proteins and induce proteotoxic stress (**Gout et al., 2017**; **Vermulst et al., 2015**). Errors on RNA transcripts may be causal factors leading to neuron degenerative diseases (**van Leeuwen et al., 1998a**; **van Leeuwen et al., 1998b**) and tumorigenesis (**Saxowsky et al., 2008**). Therefore, transcript errors represent a significant potential mechanism influencing cellular integrity and fitness.

Reporter-construct assays have long been the major approach to evaluating the fidelity of RNA polymerases and identifying transcript errors (**Blank et al., 1986**; **Bubunenko et al., 2017**; **Nesser et al., 2006**; **Rosenberger and Foskett, 1981**; **Rosenberger and Hilton, 1983**; **Shaw et al., 2002**; **Springgate and Loeb, 1975**; **Strathern et al., 2012**), but these methods focus only on individual loci and cannot identify errors without phenotypic marker effects. Conventional high-throughput sequencing approaches have been considered to identify transcript errors at a large scale (**van Dijk et al., 2015**). However, the challenge is to distinguish the real signal of transcript errors from noise produced by technical errors resulting during reverse transcription and sequencing. To circumvent this problem, a rolling-circle amplification-based sequencing (CirSeq) method (**Acevedo and Andino, 2014**; **Acevedo et al., 2014**; **Lou et al., 2013**) was recently proposed and later applied to identify transcript errors in the whole transcriptome of prokaryotes (**Traverse and**

eLife digest Most cells contain molecules of DNA that carry instructions to make the proteins cells need to perform different tasks. When a cell requires a certain protein, the corresponding DNA sequence is first transcribed into molecules of ribonucleic acid (RNA) known as transcripts. These sequences of RNA are then read by the cell and translated into the desired protein sequence.

Errors in copying DNA before a cell divides, can lead to genetic mutations that affect the ability of the cell to carry out certain roles, influencing the overall ‘fitness’ of the cell. Similar to genetic mutations, errors that arise when forming RNA transcripts may also alter the tasks a cell performs. However, it is difficult to find out what kinds of errors cells have in their transcripts and how often these mistakes occur. This is because current methods for sequencing RNA are prone to technical inaccuracies that interfere with the ability to detect true transcript errors.

Now, Li and Lynch have adapted a method for high-throughput sequencing of RNA, which can accurately identify transcript errors in *Escherichia coli* and other species of bacteria. The experiments showed that errors in RNA molecules occurred more frequently than genetic mutations in the same sequence of DNA. Li and Lynch also found that the transcripts contained more nonsense errors – that is, mutations which prematurely stop transcripts from being translated, resulting in shorter proteins – at the end of the RNA molecule than at the beginning or middle. It is possible that transcripts with errors at the beginning or the middle are more efficiently eliminated than those at the end, suggesting that bacteria have a quality-control mechanism for removing transcripts with premature stop sequences.

These findings suggest that at any one-time cells carry thousands of transcripts with inaccuracies in their sequence, which likely impact the tasks cells perform. The next step will be to investigate how these different transcript errors affect the fitness of cells.

Ochman, 2016). We further modified this protocol to minimize RNA damage potentially introduced during the preparation of sequencing libraries (***Gout et al., 2017***).

In this study, we applied an adapted CirSeq approach, which has been demonstrated to identify transcript errors accurately and efficiently at a large scale in eukaryotes (***Gout et al., 2017***), to prokaryotes for the first time. A large number of transcript errors was detected, and transcript-error rates were revealed to be orders of magnitude higher than corresponding genetic mutation rates. Our results indicate that the bias in molecular spectra of transcript errors can be influenced by both RNA polymerases and cellular rNTP concentrations. Furthermore, the spatial distribution of transcript errors on RNAs provides novel insights into the mechanism of RNA quality-control in prokaryotes.

Results

A global view of the transcript error distribution

Applying the adapted CirSeq method (see Materials and methods) to *E. coli*, *B. subtilis*, *A. tumefaciens*, and *M. florum*, RNA sequencing libraries were made with three biological replicates for each species. Key steps of library preparations involve circularizing RNA fragments and generating cDNAs with tandem repeats by rolling-circle reverse transcription. In this way, transcript errors tend to appear on all repeats of sequencing reads, while sequencing and reverse transcription errors are nearly always revealed as singletons (***Figure 1—figure supplement 1***). The number of loci where transcript errors were identified from each species ranges from 2006 to 2942, totaling 9929 loci across all species. *M. florum* showed a per-site error rate of $1.82 \pm 0.01\ (\mathrm{SEM}) \times 10^{-5}$, the highest among the four species ($P = 0.009$, Mann-Whitney U test). The error rates in *E. coli*, *B. subtilis*, and *A. tumefaciens* were $5.84 \pm 0.10\ (\mathrm{SEM}) \times 10^{-6}$, $5.80 \pm 0.14\ (\mathrm{SEM}) \times 10^{-6}$, and $7.26 \pm 0.35\ (\mathrm{SEM}) \times 10^{-6}$, respectively. These error rates are 3 to 4 orders of magnitude higher than the corresponding genomic (DNA-level) mutation rates estimated from mutation-accumulation experiments in these species (***Lee et al., 2012***; ***Lynch et al., 2016***; ***Sung et al., 2016***; ***Sung et al., 2015***; ***Sung et al., 2012***).

With such a large number of transcript errors identified, a transcriptome-wide view of the error distribution in each species was uncovered. Based on the circular genomes of bacteria (except for *A. tumefaciens*, which has one circular chromosome, one linear chromosome, and two plasmids [**Goodner et al., 2001**]), we annotated genomic positions of transcript errors with different potential functional effects and plotted transcript-error rates in 10 kb sliding windows (1 kb for *M. florum*) (**Figure 1**). To test whether transcript errors are randomly distributed across different genes, a previously proposed test (**Long et al., 2016**) was performed to identify genes enriched with transcript errors. For each gene, the expected number of transcript errors was calculated as the product of the average transcriptome-wide error rate per base and the sequencing coverage of the gene. The Poisson probability of observing a number of errors greater than or equal to the observed number was calculated. Out of 607, 495, 586, and 186 genes with detected transcript errors in *E. coli, A. tumefaciens, B. subtilis* and *M. florum*, respectively, 1, 4, 0 and 4 genes were revealed to have significantly larger numbers of errors than random expectations (Bonferroni-corrected *P* values of 0.05, **Supplementary file 1**, Tables 2-5), suggesting that transcript errors are in general randomly distributed across genes.

The whole bacterial transcriptome is synthesized by a single type of RNA polymerase. However, RNA products from protein-coding and noncoding RNA (ncRNA) regions undergo distinct co- and post-transcriptional processes. mRNAs are mature upon transcription and ready for translation, while ncRNAs, such as ribosomal RNAs (rRNA) and transfer RNAs (tRNA), need to be further processed to be functional (**Cooper, 2000**). To evaluate whether transcript-error rates of these two genomic regions are different, we calculated the error rates of protein-coding and ncRNA transcripts by dividing the number of errors by the number of nucleotides assayed in corresponding regions. Transcript-error rates of these two regions are similar in *E. coli* and *A. tumefaciens*, but the error rate of ncRNA transcripts is higher than that of protein-coding transcripts in *B. subtilis* and lower in *M. florum* ($p<0.05$, paired t-test) (**Figure 2**).

The molecular spectra of transcript errors are biased to C→U and G→A substitutions

A transition/transversion bias of genetic mutations has been widely observed in different species, with the molecular spectrum mostly dominated by G:C→A:T substitutions (**Hershberg and Petrov, 2010**; **Hildebrand et al., 2010**; **Lynch, 2010**). However, knowledge on the molecular spectrum of transcript errors in prokaryotes remains limited (**Imashimizu et al., 2015**; **Traverse and Ochman, 2016**; **Traverse and Ochman, 2018**). In this study, we calculated the error rate of all twelve categories of substitutions for each species (**Figure 3**), revealing a general bias of transitions over transversions. This bias has been thought to be driven solely by C→U substitutions (**Traverse and Ochman, 2016**), which may mainly result from post-transcriptional cytosine deaminations. However, the transition/transversion bias here even holds after C→U substitutions are excluded ($P < 0.005$, χ^2 test, **Supplementary file 1**, Table 6). This observation indicates that the transcriptional machinery in bacteria, similar to the replication machinery, tends to have a low ability to distinguish rNTPs within the same structural class of nitrogenous bases (**Keightley et al., 2009**; **Kucukyildirim et al., 2016**; **Lee et al., 2012**; **Long et al., 2015a**; **Long et al., 2015b**; **Lynch, 2007**; **Lynch, 2010**; **Lynch et al., 2008**; **Ossowski et al., 2010**; **Sung et al., 2015**). Of all transitions, the C→U substitution rate is consistently high in all four species. In addition, an unexpectedly high G→A substitution rate is revealed in *M. florum*, which displayed the highest transcript-error rates among four species in the present study. Intriguingly, this substitution bias was also recently observed in yeast and *E. coli* transcription-machinery mutants with decreased fidelity (**Gout et al., 2017**; **Imashimizu et al., 2015**; **Traverse and Ochman, 2018**). Thus, the G→A substitution bias may be a signature of error-prone RNA polymerase in both eukaryotes and prokaryotes.

Characterization of transcript errors

To evaluate potential functional effects of transcript errors, we categorized transcript errors within protein-coding regions into synonymous, missense, and nonsense substitutions using SnpEff (**Cingolani et al., 2012**; **Table 1**). Based on the bias of rNTP substitution rates (**Figure 3**) and codon usages of each bacterium, we also calculated the expected percentages of each error type under the assumption that transcript errors are randomly generated across the genome without error-

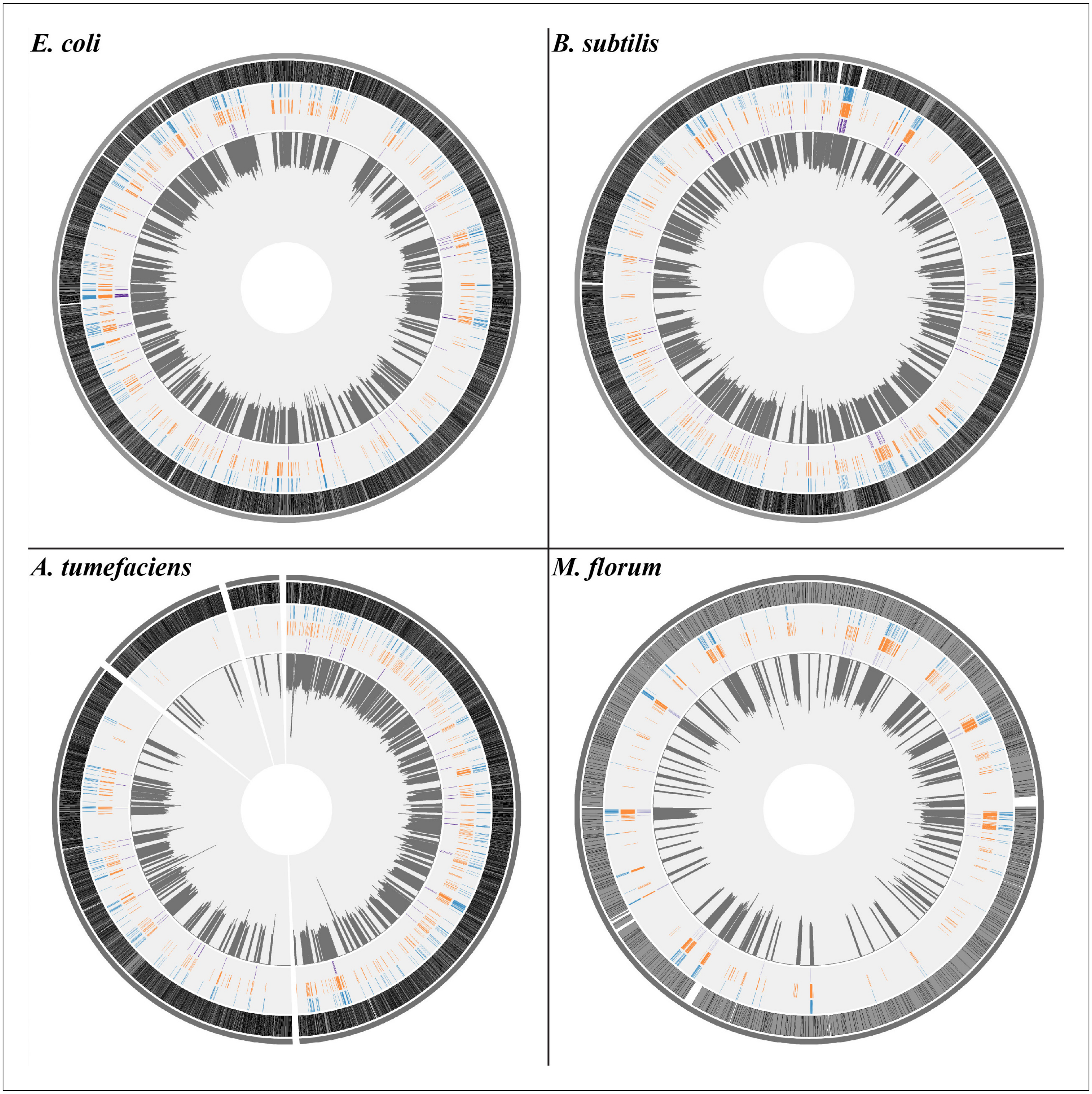

Figure 1. The distribution of transcript errors across the whole transcriptomes of *E. coli*, *B. subtilis*, *A. tumefaciens*, and *M. florum*. The first nucleotide of the circular chromosome starts at the 12 o'clock position. For *A. tumefaciens*, chromosomes and plasmids are arranged from the largest to smallest size in a clockwise orientation. From the outer ring to the inner ring: bacterial chromosomes (dark gray), protein-coding region (grey, black strokes indicate gene densities), synonymous substitutions (blue), missense substitutions (orange), nonsense substitutions (purple) and average transcript-error rates (plots in dark gray) in a 10 kb sliding window with a step size of 1 bp (1 kb windows for *M. florum*). Windows without sufficient sequencing coverages to detect transcript errors are left blank.

The online version of this article includes the following source data and figure supplement(s) for figure 1:

Source data 1. Numerical data that are represented as a graph in ***Figure 1***.

Figure supplement 1. The flowchart of CirSeq method.

Figure supplement 2. An overview of the bioinformatic pipeline to process CirSeq reads to identify transcript errors.

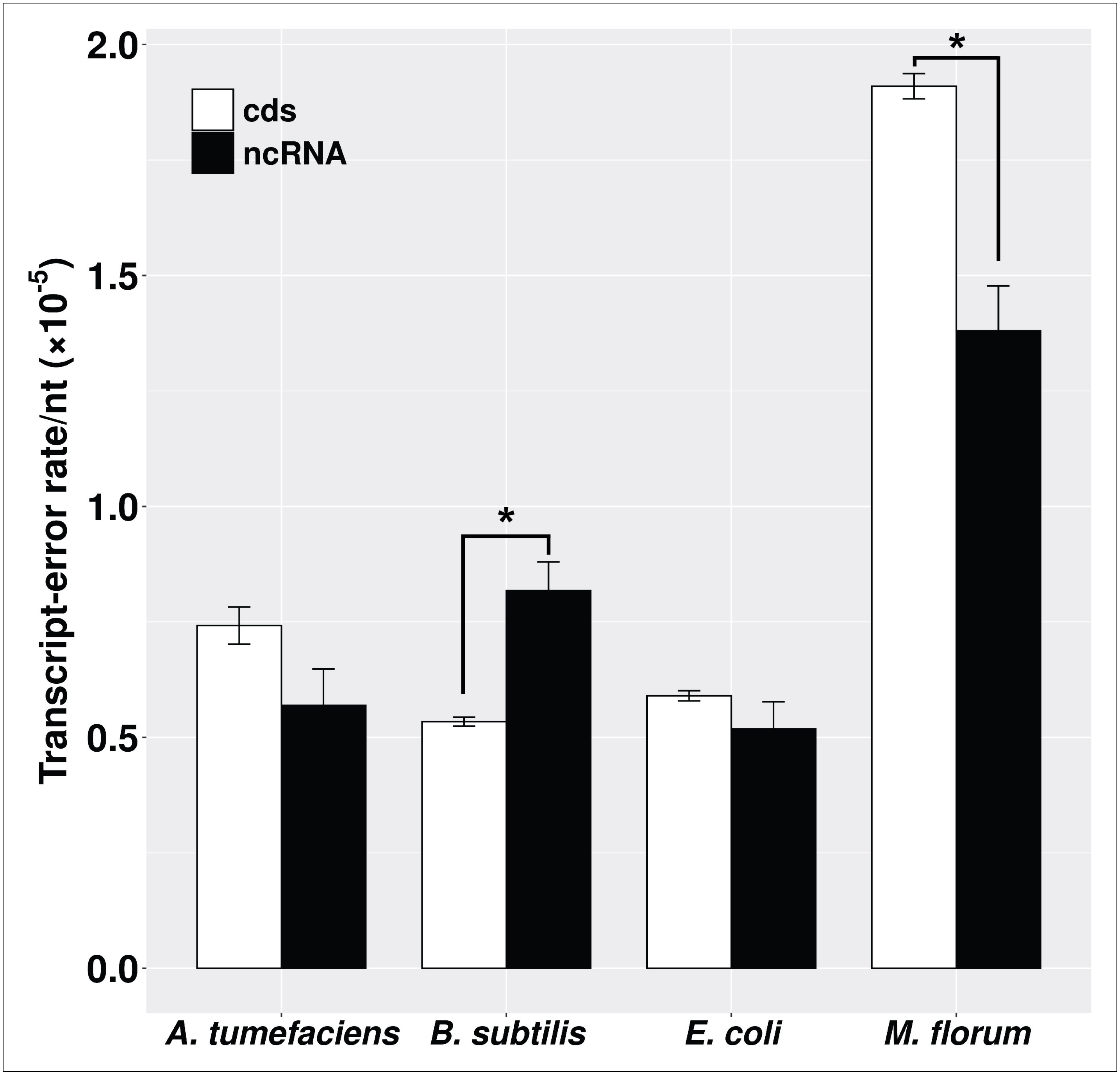

Figure 2. Transcript-error rates of protein-coding and ncRNA regions. cds includes all protein-coding genes that were sequenced in this study. ncRNA refers to RNAs that are functional but not translated into proteins, for example tRNA and rRNA. Transcript-error rates were calculated by dividing the number of errors by the number of nucleotides assayed in corresponding regions. Error bars indicate standard errors. The level of significance difference is indicated by asterisks (*p<0.05, paired t-test).

The online version of this article includes the following source data for figure 2:

Source data 1. Numerical data that are represented as a graph in ***Figure 2***.

correction processes (see Materials and methods, and ***Supplementary file 1***, Table 7). Consistent with observations, the majority of transcript errors are expected to result in amino-acid changes, if translated (***Table 1***). For nonsense errors, the observed percentages are close to or significantly lower than the random expectation ($P < 0.005$, χ^2 test, ***Table 1***).

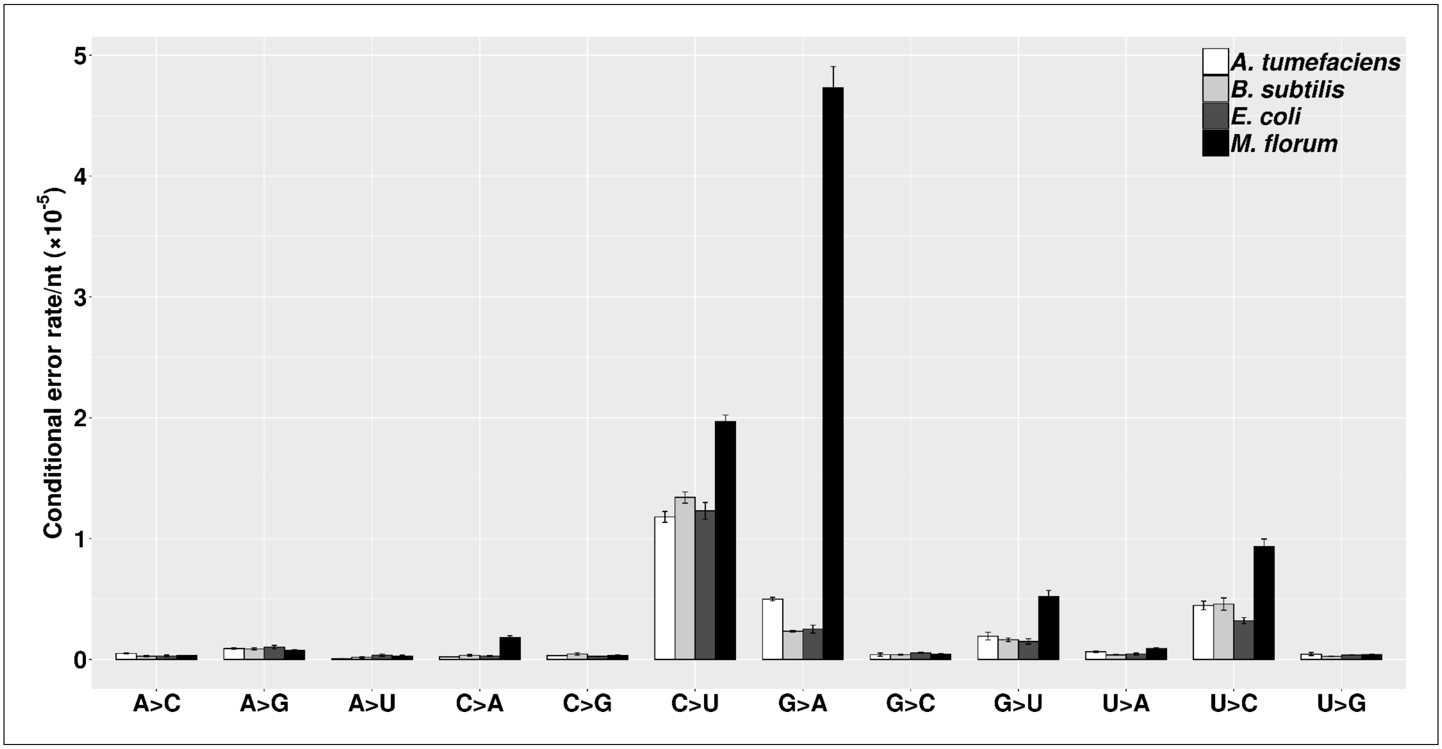

Figure 3. The molecular spectra of transcript errors for four bacterial species. The conditional error rates of each type of substitutions were calculated from the number of particular transcript errors, divided by the number of corresponding ribonucleotides assayed. Error bars indicate standard errors.
The online version of this article includes the following source data for figure 3:

Source data 1. Numerical data that are represented as a graph in ***Figure 3***.

Biased distribution of nonsense errors in RNA transcripts

As shown in ***Table 1***, nonsense errors represent only a small percentage of all errors. However, they are of particular interest because they will result in the formation of a premature termination codon (PTC) and thus truncated proteins if not degraded. To ameliorate the potential severe fitness effects resulting from such errors, eukaryotes have evolved the Nonsense Mediated Decay (NMD) mechanism (***Losson and Lacroute, 1979***; ***Maquat, 1995***; ***Peltz et al., 1993***) to facilitate the degradation of RNA transcripts carrying PTCs. A key to the success of NMD is distinguishing a PTC from the original stop codon (***Amrani et al., 2004***; ***Le Hir et al., 2001***), and the ability of the NMD machinery to identify a PTC is thought to diminish as the PTC approaches the 3′ end of a mRNA (***Isken and***

Table 1. Percentages of transcript errors in mRNAs that are synonymous, missense, or nonsense (other potential types of transcript errors with small percentages, such as start/stop codon loss-errors, are not shown).
Observed and expected (in parentheses) percentages are presented. Based on the bias of observed rNTP substitution rates and codon usages of each bacterium, expected percentages are calculated assuming a random generation of errors and an absence of error-correction processes. The level of significant difference is indicated by asterisks (**P* < 0.05, ** *P* < 0.005, χ^2 test).

Species	Synonymous	Missense	Nonsense
E. coli	40.18 (34.35) **	56.25 (59.79) *	3.57 (5.62) **
B. subtilis	32.76 (31.86)	61.69 (61.63)	5.15 (6.15)
A. tumefaciens	40.68 (36.76) *	56.36 (59.17)	2.96 (3.86)
M. florum	17.58 (24.12) **	79.27 (70.58) **	2.37 (4.85) **

Maquat, 2007). This hypothesis is supported by yeast transcript-error data that show a marked increase in the frequency of PTCs towards the 3′ end of mRNAs (**Gout et al., 2017**).

Although no analog of the eukaryotic NMD system is known in prokaryotes, a destabilizing effect of PTCs on mRNA stability has been observed in bacteria (**Arnold et al., 1998**; **Braun, 1998**; **Morse and Yanofsky, 1969**; **Nilsson et al., 1987**). Evaluating the distribution of nonsense errors across the whole length of mRNA transcripts, we observed an increased frequency of nonsense errors at the 3′ end of transcripts, although the trend is not statistically significant in *A. tumefaciens* (***Figure 4A***). Compared to other three species, a smaller number of nonsense errors were detected in *A. tumefaciens* (***Supplementary file 1***, Table 7), which may result in a low statistical power to reveal a potential pattern for the distribution of nonsense errors. We further modified the analysis by dividing the frequency of nonsense errors by that of all errors. This ratio tends to be higher at the 3′ end of mRNAs (***Figure 4—figure supplement 1***), excluding the possibility that the enrichment of nonsense errors results mainly from a higher overall transcript-error rate at the 3′ end of mRNAs.

Of all types of genetic codons, those with one nucleotide difference from a stop codon (one-off codons) have a higher probability of mutating into PTCs. We further normalized the frequency of nonsense errors by the abundance of one-off codons at corresponding loci. This still revealed an increased frequency of nonsense errors towards 3′ ends of transcripts (***Figure 4—figure supplement 2***), suggesting the higher frequency of nonsense errors is not caused by more abundant one-off codons at the 3′ end of transcripts.

The increased frequency of PTCs at the 3′ end of mRNA transcripts suggests the presence of an NMD-like process, albeit by a likely different mechanism than in eukaryotes, which largely rely on the poly-A tail or exon-exon junction complex (**Amrani et al., 2004**). One speculative model for the degradation of PTCs in eukaryotes, the ribosome-release model (**Brogna and Wen, 2009**), in which the degradation of RNAs with PTCs depends on the degree of ribosome coverage on RNA molecules, has the potential to hold true in prokaryotes. Ribosomes can load on to nascent transcripts immediately after RNA synthesis. Therefore, a whole transcript with a normal stop codon can be covered by multiple ribosomes towards its 3′ end, with these ribosomes protecting the transcript from degradation by blocking ribonuclease cleavage sites. In contrast, a PTC upstream of the original stop codon will stall the ribosomes, leaving the ribonucleotides between the PTC and the site of the original stop codon unprotected by ribosomes, potentially promoting degradation by cellular ribonucleases (***Figure 4B***).

Discussion

A key to accurately identifying *bona fide* transcript errors is to distinguish them from technical errors and low-frequency genetic mutations. With previous efforts on method development to eliminate sequencing errors (**Acevedo and Andino, 2014**; **Acevedo et al., 2014**; **Lou et al., 2013**) and to evaluate the error rate of the reverse transcriptase (**Gout et al., 2013**), it is now possible to ensure that contributions from such technical errors are orders of magnitudes lower than true transcript-error rates by the CirSeq approach (See Materials and methods). Except for *M. florum*, transcript-error rates in bacteria estimated by the current study are about one order of magnitude lower than those from a previous study (**Traverse and Ochman, 2016**). Specifically in *E. coli*, our error-rate estimates for each type of substitutions tend to be lower than those from **Traverse and Ochman (2016)**, the most striking difference involving the C→U substitution rate, which could be partly due to the use of a metal ion-based RNA fragmentation approach in the previous work vs. enzymatic RNA fragmentation in the present study. The latter minimizes RNA damage (**Gout et al., 2017**), in particular cytosine deaminations, introduced during the preparation of the sequencing library.

Besides base-substitution errors, a small portion of transcript errors can occur in other forms such as insertions and deletions. Estimates of transcript insertion/deletion (indel) error rates from species in this study are 0.1 to 0.2 of the corresponding base-substitution error rates (***Supplementary file 1***, Table 1).

Bacterial transcriptomes predominantly consist of ncRNA transcripts, such as rRNAs and tRNAs (**Westermann et al., 2012**). However, only a small portion of the whole ncRNA transcripts was evaluated in the present study (***Supplementary file 1***, Table 8) because of technical limitations. The rRNA depletion procedure in the sequencing library preparation protocol removes the majority of rRNAs. Secondary structures and nucleotide modifications of tRNAs interfere the cDNA synthesis and

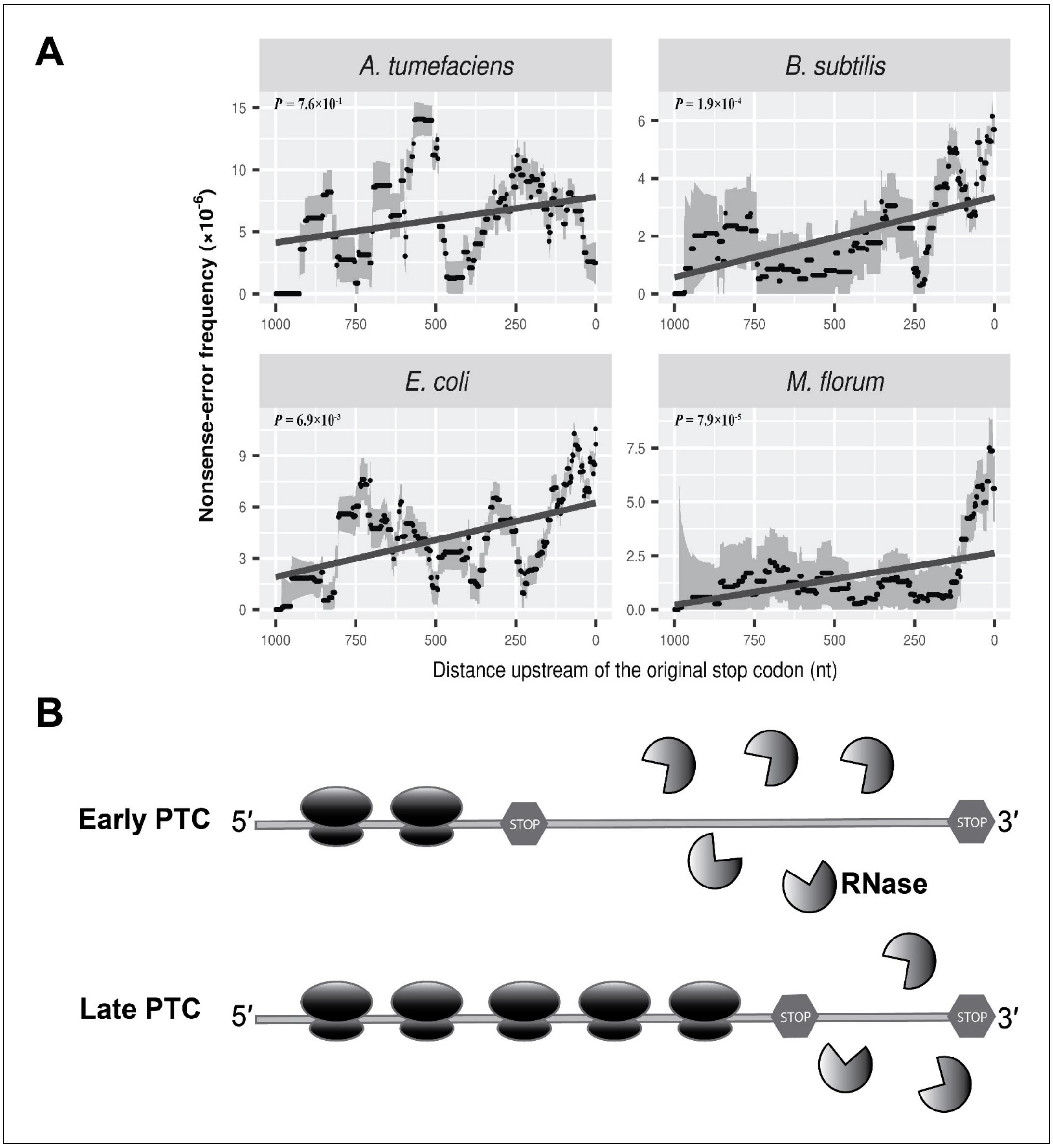

Figure 4. Nonsense errors in prokaryotic transcripts. (A) Distributions of nonsense errors across mRNA transcripts. The frequency of nonsense errors is calculated in a 100-nt sliding window with a step size of 1 nt for data visualization. Grey intervals represent standard deviations assuming the number of errors at each locus follows a binomial distribution. Linear regression between the distance to the original stop codon and the frequency of nonsense errors of each window is indicated in dark grey lines. *P* values were calculated from weighted linear regressions of individual data points before binning into a window. (B) The ribosome-release model for PTCs degradation in prokaryotes. Compared to a late PTC, an early PTC results in a larger portion of ribonucleotides unprotected by ribosomes, and therefore a higher probability of being digested by cellular ribonuclease.
The online version of this article includes the following source data and figure supplement(s) for figure 4:

Source data 1. Numerical data that are represented as a graph in ***Figure 4A***.
Figure supplement 1. Distributions of the ratio of nonsense error frequency to total error frequency across mRNA transcripts.
Figure supplement 1—source data 1. Numerical data that are represented as a graph in ***Figure 4—figure supplement 1***.
Figure supplement 2. Distribution of PTCs across the length of mRNA transcripts.
Figure supplement 2—source data 1. Numerical data that are represented as a graph in ***Figure 4—figure supplement 2***.

sequencing adapter ligations. In the future, to achieve a better measurement of transcript-error rates of ncRNA transcripts, total RNAs can be mixed with rRNA-depleted RNAs at a certain ratio to increase the abundance of rRNAs in the sequencing library. Demethylase enzymes and thermophilic reversetranscriptase can be used to remove nucleotide modifications of tRNAs and to improve processivity in generating cDNAs from highly structured RNA templates (**Schwartz et al., 2018**; **Zheng et al., 2015**).

The molecular spectrum of transcript errors revealed in our work indicates a general C→U substitution bias, which has been proposed to be due to spontaneous deamination (**Imashimizu et al., 2013**; **Traverse and Ochman, 2016**) owing to the chemical instability of cytosine (**Alberts et al., 2015**). Besides this widely accepted mechanism, non-Watson-Crick base pairing during rNTP incorporations may also contribute to this bias. Because dG and rU can form a base pair (**Sugimoto et al., 2000**; **Sugimoto et al., 1997**), mispairing between a template DNA (dG) and an RNA (rU) during rNTP incorporations likely also contributes to the C→U substitution bias.

Another intriguing observation from the molecular spectra in the present study is the G→A substitution bias in *M. florum*. One source for this substitution may be unrepaired uracils on the DNA antisense strand, which pair with rATPs during transcription, resulting in a G→A substitution on the RNA transcript. Although *M. florum* has a diminutive genome (0.79 Mb) and lacks many genes (RefSeq NC_006055.1), a uracil-DNA glycosylase (UDG) ortholog whose product presumably removes uracils (**McCullough et al., 1999**) does exist in the genome. Therefore, the extent to which mismatches between the unrepaired uracil and rATP can explain the G→A bias remains unclear.

Taking data from previous studies (**Gout et al., 2017**; **Imashimizu et al., 2015**; **Traverse and Ochman, 2018**) and this work together, G→A substitution bias seems to be a general pattern in cells with error-prone transcription machineries. What might be the underlying mechanism? The error spectrum is shaped by two factors. One is the ability of an RNA polymerase to distinguish correct rNTPs from incorrect ones. The other factor, which is sometimes neglected, is the rNTP pool within a cell. The error rate of competitive binding of rNTPs to the template can be expressed as, $(k_{incorrect} \cdot C_{incorrect-rNTPs})/(k_{correct} \cdot C_{correct-rNTPs})$, where k refers to the rNTP incorporation rate and C indicates the concentration of rNTPs. As suggested by this equation, a biased cellular rNTP concentration might present an additional challenge to transcriptional fidelity for certain categories of rNTPs. Based on observations that RNA polymerases have a low ability to distinguish rNTPs with the same structural class of nitrogenous bases and that the cellular concentration of rATPs is the highest among all types of nucleotides in both eukaryotes and prokaryotes (**Bennett et al., 2009**; **Buckstein et al., 2008**; **Traut, 1994**), it is reasonable to speculate that the high cellular concentration of rATPs contribute to the observed bias towards G→A substitutions.

An additional cellular process influencing transcript errors is RNA quality-control. Because genes involved in NMD, such as up-frameshift (UPF) genes, have not been identified in prokaryotes, evidence for the existence of NMD in prokaryotes is still lacking. However, previous studies based on single gene-reporters (**Baker and Mackie, 2003**; **Braun, 1998**; **Nilsson et al., 1987**) and our transcriptome-wide survey suggest a Nonsense-Mediated Decay-like quality-control mechanism in prokaryotes. A key implication of the increased frequency of nonsense errors at the 3′ end of mRNAs (**Figure 4A**) is that the degradation of RNAs carrying nonsense errors may simply result from a higher degree of exposure to cellular ribonucleases rather than from a reliance on specific protein-based systems.

Current models of mRNA surveillance mechanisms mostly focus on stop codon-related errors (**Deutscher, 2006**; **Richards et al., 2008**), which are expected to represent only a small portion of the total transcript errors in a cell. It is largely unknown whether, and if so by which mechanisms the major transcript errors (missense errors) get degraded. To resolve this, future research will be required to evaluate the rate at which transcript errors are degraded after initially being generated during transcription. This might be possible by comparing transcript errors on nascent transcripts bound to RNA polymerases with those on mature transcripts associated with ribosomes.

Materials and methods

Bacteria strains and growth conditions

All bacteria strains were inoculated into liquid culture from single colonies and grew to mid-exponential growth phase upon harvest. *E. coli* MG1655 and *B. subtilis* NCIB 3610 were grown at 37˚C in LB liquid medium. *M. florum* L1 (ATCC #33453) was grown at 30˚C in SNE liquid medium. *A. tumefaciens* C58 was grown at 28˚C in LB liquid medium.

RNA extraction

Bacteria were harvested from liquid culture media by centrifugation and total RNA was extracted and purified using the FastRNA Blue Kit (MPBiomedicals), RNase-free DNase set (Qiagen), and the RNeasy Mini Kit (Qiagen). rRNA was depleted by the Ribo-Zero rRNA Removal Kit (Bacteria) (Illumina) for the following library preparations.

Library preparation and sequencing

We followed a refined protocol of CirSeq (**Gout et al., 2017**) to prepare libraries for transcript error identifications. Five hundred nanograms of rRNA-depleted RNAs were firstly fragmented with the NEBNext RNase III RNA Fragmentation Module (New England Biolabs) for 90 min at 37˚C. After a clean-up using the Oligo Clean and Concentrator kit (Zymo Research), RNA fragments were circularized with RNA ligase 1 (New England Biolabs) according to the manufactuer's guidelines. cDNA with tandem repeats was generated by the rolling-circle reverse transcription as described in the refined CirSeq protocol. Synthesis of the second strand of cDNA and sequencing library preparation were performed using the NEBNext Ultra RNA Library Prep Kit and NEBNext Multiplex Oligos for Illumina (New England Biolabs). The size selection and clean-up during sequencing library preparations were performed by Agencourt AMPure XP Beads (Beckman Coulter) according the NEB guideline that is optimized for approximately 200nt RNA inserts. A final gel-based size selection was performed to enrich PCR amplified products that are longer than 300nt. Single-end reads (300nt) were then generated using Illumina HiSeq 2500 System. The sequencing data were deposited in NCBI with the BioProject Number PRJNA592142.

Genome references and annotation files

The accession numbers of genome references for *E. coli*, *B. subtilis*, and *M. florum* are NC_000913.3, NZ_CM000488.1, and NC_006055.1. For *A. tumefacien*, accession numbers are NC_003062.2, NC_003063.2, NC_003064.2 and NC_003065.3. The corresponding genome annotation files are from RefSeq.

Data analysis

Several analysis pipelines already existed to process reads with multiple tandem repeats and call transcript errors, but with their own limitations. The CirSeq_v2 pipeline (**Acevedo and Andino, 2014**; **Acevedo et al., 2014**) can only analyze reads with exactly three repeats and reads generated by CirSeq approach can contain more than four repeats if the original RNA template is smaller than 75 nt. Another pipeline described in a recent work in yeast (**Gout et al., 2017**) cannot generate consensus calls and recalculate the quality score from a site where not all base calls are identical. Therefore, we developed Python scripts following the methods outlined by **Lou et al. (2013)** (**Figure 1—figure supplement 2**). The structure of repeats within one read was identified by an autocorrelation-based method, in which the length of one potential repeat P is detected by the maximum fraction of identical base calls that are separated by a distance P within one read. The consensus sequence was constructed and the corresponding new quality score was calculated by a Bayesian approach where an inferred consensus call is taken with the maximum posterior probability given all observed base calls. This approach also allows the processing of varied numbers and types of base calls at one site. To identify the ligation junction of circular templates and to reorganize the consensus sequence, a tandem duplicate of the consensus sequence was constructed and then mapped back to the reference genome by BWA (**Li and Durbin, 2009**). The longest continuous mapped regions of the duplicated consensus sequences therefore correspond to original RNA fragments. We also excluded the 4 nucleotides at both ends of the reorganized consensus sequence to minimize potential confusions, because mapping can be ambiguous at the two ends of RNA fragments. After mapping of reconstructed consensus sequences, reads uniquely mapped to protein-coding regions and all reads mapped to ncRNA regions were kept. Transcript errors were called if a mismatch between a consensus call and the reference was supported by less than 1% of reads at corresponding loci. To exclude false positives of transcript errors from genetic mutations in multiple copies of ncRNA genes (such as rRNA and tRNA genes), an additional filter was included to exclude an error call that is supported by genetic variations from different copies of ncRNA genes. The transcript error rate of a given region was calculated as the number of transcript errors divided by the total number of rNTPs assayed from the corresponding region. The code for the bioinformatic pipeline can be found at https://github.com/LynchLab/CirSeq4TranscriptErrors (***Li, 2020***; copy archived at https://github.com/elifesciences-publications/CirSeq4TranscriptErrors).

Strategies to distinguish transcript errors from other types of errors

First, reverse transcription and sequencing errors need to be filtered out in the analysis. Because the rate of transcript error is generally 10^{-6} ~ 10^{-5} /nt , the recalculated probability of an erroneous base call at 10^{-7} or lower was required to minimize contaminations from sequencing errors. Because the error rate of the reverse transcriptase used here is ~10^{-4} /nt (***Gout et al., 2013***), at least two tandem repeats were required in the analysis to minimize false positives from reverse transcription errors.

Second, genetic mutations (DNA level) can arise during cell culture and low frequency mutations can behave like transcript errors in the sequencing data. The probability of capturing a genetic mutation can be calculated by dividing the expected number of genetic mutations generated during cell propagations by the total transcriptome size at the time point of sample collection, $\frac{\mu \cdot g \cdot T \cdot n}{T \cdot n}$, in which μ is the per site per generation mutation rate, g is the number of generations during cell culture, T is the size of genome regions get transcribed, and n is the average expression level per site. This equation can be further simplified as μg. Because we know the mutation rate from mutation accumulation experiments (***Lee et al., 2012***; ***Lynch et al., 2016***; ***Sung et al., 2016***; ***Sung et al., 2015***; ***Sung et al., 2012***) and the number of generations from culture-growth dynamics (~30 generations), Low frequency genetic mutation can only inflate the transcript-error rate we calculated here by ~1‰ -1%.

To calculate the expected percentages of transcript errors with different effects

Take the calculation for synonymous substitution as one example. The percentage can be calculated by summing the probabilities of each codon to have a synonymous change, $P(syn) = \sum_{i=1}^{64} P_i \cdot P_{i(syn)}$. P_i refers to the probability of having codon i based on the codon usage of a specific genome and there are 64 codons in total. $P_{i(syn)}$ is the probability that codon i has a synonymous substitution and it can be calculated from, $P_{i(syn)} = \sum_{j=1}^{9} \mu_j \cdot 1_{\{j \text{ results in syn}\}}$. μ_j denotes the substitution probability of 1 of the 9 single-base substitutions that can happen in one codon. And it can be calculated by, $\mu_j = \frac{e_j}{\sum_{j=1}^{9} e_j}$, in which e_j refers to the error rate of 1 of the 9 substitutions in one codon. Estimates of e_j are displayed in ***Figure 3***.

The sliding window analysis and weighted linear regression to evaluate the distribution of nonsense errors on mRNA transcripts

The sliding window analysis (window size = 100nt and step size = 1nt) of the distribution of nonsense errors across mRNAs was used for data visualization. To evaluate whether or not the negative correlation between the frequency of nonsense errors and the corresponding distance from a nonsense error to the original stop codon is statistically significant, a weighted linear regression method was used. The weight was calculated as the reciprocal of a variance of a nonsense error frequency. Because the observed number of transcript errors at each locus is expected to follow a binomial distribution, the variance of the nonsense error frequency can be estimated as $\frac{p(1-p)}{n}$, where p is the estimated frequency of errors and n refers to the read coverage at the corresponding locus.

Acknowledgements

We thank Stephen Simpson, W Kelley Thomas, Samuel F Miller, Jiaqi Zheng, Jie Huang, and James Ford for technical support, and Daniel Kearns and Clay Fuqua for providing *B. subtilis* NCIB 3610 and *A. tumefaciens* C58 strains. We also thank Hongan Long, Michelle Marasco, Chi-Chun Chen, Parul Johri, and Jean-Francois Gout for helpful discussions. This work was supported by National Institutes of Health Awards R01-GM036827 and R35-GM122566, and Multidisciplinary University

Research Initiative Award W911NF-09-1-0444 and W911NF-14-1-0411 from the US Army Research Office (to ML).

Additional information

Funding

Funder	Grant reference number	Author
National Institutes of Health	R01-GM036827	Michael Lynch
National Institutes of Health	R35-GM122566	Michael Lynch
Army Research Office	W911NF-09-1-0444	Michael Lynch
Army Research Office	W911NF-14-1-0411	Michael Lynch

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Weiyi Li, Conceptualization, Resources, Data curation, Software, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing - original draft, Writing - review and editing; Michael Lynch, Conceptualization, Supervision, Funding acquisition, Validation, Investigation, Writing - original draft, Project administration, Writing - review and editing

Author ORCIDs

Weiyi Li https://orcid.org/0000-0002-1168-7093
Michael Lynch https://orcid.org/0000-0002-1653-0642

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.54898.sa1
Author response https://doi.org/10.7554/eLife.54898.sa2

Additional files

Supplementary files

- Supplementary file 1. Supplementary Table 1. Estimates of transcript indel error rates of four bacterial species. Standard error of the mean from three biological replicates are displayed. Supplementary Table 2—5. The Observed and expected numbers of transcript errors in genes of four bacterial species. Supplementary Table 6. The transition/transversion error bias revealed in four bacterial species. Supplementary Table 7. The observed and expected numbers/percentages of transcript errors that are synonymous, nonsynonymous, or nonsense. Supplementary Table 8. The numbers and rates of detected transcript errors that are from ncRNA transcripts in four bacterial species.
- Transparent reporting form

Data availability

Sequencing data of this study are available at NCBI with the BioProject Number PRJNA592142.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Weiyi Li, Michael L	2020	Transcript error studies on Escherichia coli, Bacillus subtilis, Agrobacterium tumefaciens, and Mesoplasma florum	http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA592142	NCBI BioProject, PRJNA592142

References

Acevedo A, Brodsky L, Andino R. 2014. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505**:686–690. DOI: https://doi.org/10.1038/nature12861

Acevedo A, Andino R. 2014. Library preparation for highly accurate population sequencing of RNA viruses. *Nature Protocols* **9**:1760–1769. DOI: https://doi.org/10.1038/nprot.2014.118, PMID: 24967624

Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, Walter P. 2015. *Molecular Biology of the Cell*. Garland Science.

Amrani N, Ganesan R, Kervestin S, Mangus DA, Ghosh S, Jacobson A. 2004. A faux 3′-UTR promotes aberrant termination and triggers nonsense- mediated mRNA decay. *Nature* **432**:112–118. DOI: https://doi.org/10.1038/nature03060

Arnold TE, Yu J, Belasco JG. 1998. mRNA stabilization by the ompA 5′ untranslated region: two protective elements hinder distinct pathways for mRNA degradation. *RNA* **4**:319–330. PMID: 9510333

Baker KE, Mackie GA. 2003. Ectopic RNase E sites promote bypass of 5′-end-dependent mRNA decay in *Escherichia coli*. *Molecular Microbiology* **47**:75–88. DOI: https://doi.org/10.1046/j.1365-2958.2003.03292.x, PMID: 12492855

Bennett BD, Kimball EH, Gao M, Osterhout R, Van Dien SJ, Rabinowitz JD. 2009. Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nature Chemical Biology* **5**:593–599. DOI: https://doi.org/10.1038/nchembio.186

Blank A, Gallant JA, Burgess RR, Loeb LA. 1986. An RNA polymerase mutant with reduced accuracy of chain elongation. *Biochemistry* **25**:5920–5928. DOI: https://doi.org/10.1021/bi00368a013, PMID: 3098280

Braun F. 1998. Ribosomes inhibit an RNase E cleavage which induces the decay of the rpsO mRNA of *Escherichia coli*. *The EMBO Journal* **17**:4790–4797. DOI: https://doi.org/10.1093/emboj/17.16.4790

Brogna S, Wen J. 2009. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nature Structural & Molecular Biology* **16**:107–113. DOI: https://doi.org/10.1038/nsmb.1550

Bubunenko MG, Court CB, Rattray AJ, Gotte DR, Kireeva ML, Irizarry-Caro JA, Li X, Jin DJ, Court DL, Strathern JN, Kashlev M. 2017. A *cre* Transcription Fidelity Reporter Identifies GreA as a Major RNA Proofreading Factor in *Escherichia coli*. *Genetics* **206**:179–187. DOI: https://doi.org/10.1534/genetics.116.198960, PMID: 28341651

Buckstein MH, He J, Rubin H. 2008. Characterization of nucleotide pools as a function of physiological state in *Escherichia coli*. *Journal of Bacteriology* **190**:718–726. DOI: https://doi.org/10.1128/JB.01020-07, PMID: 17965154

Cingolani P, Platts A, Wang leL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single Nucleotide Polymorphisms, SnpEff: snps in the genome *of Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**:80–92. DOI: https://doi.org/10.4161/fly.19695, PMID: 22728672

Cooper GM. 2000. *The Cell: A Molecular Approach*. ASM Press.

Deutscher MP. 2006. Degradation of RNA in Bacteria: comparison of mRNA and stable RNA. *Nucleic Acids Research* **34**:659–666. DOI: https://doi.org/10.1093/nar/gkj472, PMID: 16452296

Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, Qurollo B, Goldman BS, Cao Y, Askenazi M, Halling C, Mullin L, Houmiel K, Gordon J, Vaudin M, Iartchouk O, Epp A, Liu F, Wollam C, Allinger M, Doughty D, et al. 2001. Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science* **294**:2323–2328. DOI: https://doi.org/10.1126/science.1066803, PMID: 11743194

Gordon AJ, Satory D, Halliday JA, Herman C. 2013. Heritable change caused by transient transcription errors. *PLOS Genetics* **9**:e1003595. DOI: https://doi.org/10.1371/journal.pgen.1003595, PMID: 23825966

Gout JF, Thomas WK, Smith Z, Okamoto K, Lynch M. 2013. Large-scale detection of in vivo transcription errors. *PNAS* **110**:18584–18589. DOI: https://doi.org/10.1073/pnas.1309843110, PMID: 24167253

Gout JF, Li W, Fritsch C, Li A, Haroon S, Singh L, Hua D, Fazelinia H, Smith Z, Seeholzer S, Thomas K, Lynch M, Vermulst M. 2017. The landscape of transcription errors in eukaryotic cells. *Science Advances* **3**:e1701484. DOI: https://doi.org/10.1126/sciadv.1701484, PMID: 29062891

Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in Bacteria. *PLOS Genetics* **6**:e1001115. DOI: https://doi.org/10.1371/journal.pgen.1001115, PMID: 20838599

Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in Bacteria. *PLOS Genetics* **6**:e1001107. DOI: https://doi.org/10.1371/journal.pgen.1001107, PMID: 20838593

Imashimizu M, Oshima T, Lubkowska L, Kashlev M. 2013. Direct assessment of transcription fidelity by high-resolution RNA sequencing. *Nucleic Acids Research* **41**:9090–9104. DOI: https://doi.org/10.1093/nar/gkt698, PMID: 23925128

Imashimizu M, Takahashi H, Oshima T, McIntosh C, Bubunenko M, Court DL, Kashlev M. 2015. Visualizing translocation dynamics and nascent transcript errors in paused RNA polymerases in vivo. *Genome Biology* **16**:98. DOI: https://doi.org/10.1186/s13059-015-0666-5, PMID: 25976475

Isken O, Maquat LE. 2007. Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes Development* **21**:1833–3856. DOI: https://doi.org/10.1101/gad.1566807, PMID: 17671086

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Research* **19**:1195–1201. DOI: https://doi.org/10.1101/gr.091231.109, PMID: 19439516

Kucukyildirim S, Long H, Sung W, Miller SF, Doak TG, Lynch M. 2016. The Rate and Spectrum of Spontaneous Mutations in *Mycobacterium smegmatis* , a Bacterium Naturally Devoid of the Postreplicative Mismatch Repair Pathway . *G3: Genes|Genomes|Genetics* **6**:2157–2163. DOI: https://doi.org/10.1534/g3.116.030130

Le Hir H, Gatfield D, Izaurralde E, Moore MJ. 2001. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *The EMBO Journal* **20**:4987–4997. DOI: https://doi.org/10.1093/emboj/20.17.4987, PMID: 11532962

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. *PNAS* **109**:E2774–E2783. DOI: https://doi.org/10.1073/pnas.1210309109, PMID: 22991466

Li W. 2020. Analysis pipelines to call and analyze transcript errors from CirSeq data. *GitHub*. 606267b. https://github.com/LynchLab/CirSeq4TranscriptErrors

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760. DOI: https://doi.org/10.1093/bioinformatics/btp324, PMID: 19451168

Long H, Kucukyildirim S, Sung W, Williams E, Lee H, Ackerman M, Doak TG, Tang H, Lynch M. 2015a. Background mutational features of the Radiation-Resistant bacterium *Deinococcus radiodurans*. *Molecular Biology and Evolution* **32**:2383–2392. DOI: https://doi.org/10.1093/molbev/msv119, PMID: 25976352

Long H, Sung W, Miller SF, Ackerman MS, Doak TG, Lynch M. 2015b. Mutation rate, spectrum, topology, and Context-Dependency in the DNA mismatch Repair-Deficient *Pseudomonas fluorescens* ATCC948. *Genome Biology and Evolution* **7**:262–271. DOI: https://doi.org/10.1093/gbe/evu284

Long H, Miller SF, Strauss C, Zhao C, Cheng L, Ye Z, Griffin K, Te R, Lee H, Chen CC, Lynch M. 2016. Antibiotic treatment enhances the genome-wide mutation rate of target cells. *PNAS* **113**:E2498–E2505. DOI: https://doi.org/10.1073/pnas.1601208113, PMID: 27091991

Losson R, Lacroute F. 1979. Interference of nonsense mutations with eukaryotic messenger RNA stability. *PNAS* **76**:5134–5137. DOI: https://doi.org/10.1073/pnas.76.10.5134, PMID: 388431

Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, Sawyer SL. 2013. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *PNAS* **110**:19872–19877. DOI: https://doi.org/10.1073/pnas.1319590110, PMID: 24243955

Lynch M. 2007. *The Origins of Genome Architecture*. Sinauer Associates.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, Thomas WK. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *PNAS* **105**:9272–9277. DOI: https://doi.org/10.1073/pnas.0803466105, PMID: 18583475

Lynch M. 2010. Rate, molecular spectrum, and consequences of human mutation. *PNAS* **107**:961–968. DOI: https://doi.org/10.1073/pnas.0912629107, PMID: 20080596

Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics* **17**:704–714. DOI: https://doi.org/10.1038/nrg.2016.104, PMID: 27739533

Maquat LE. 1995. When cells stop making sense: effects of nonsense codons on RNA metabolism in vertebrate cells. *RNA* **1**:453–465. PMID: 7489507

McCullough AK, Dodson ML, Lloyd RS. 1999. Initiation of base excision repair: glycosylase mechanisms and structures. *Annual Review of Biochemistry* **68**:255–285. DOI: https://doi.org/10.1146/annurev.biochem.68.1.255, PMID: 10872450

Morse DE, Yanofsky C. 1969. Polarity and the degradation of mRNA. *Nature* **224**:329–331. DOI: https://doi.org/10.1038/224329a0, PMID: 4898925

Nesser NK, Peterson DO, Hawley DK. 2006. RNA polymerase II subunit Rpb9 is important for transcriptional fidelity in vivo. *PNAS* **103**:3268–3273. DOI: https://doi.org/10.1073/pnas.0511330103, PMID: 16492753

Nilsson G, Belasco JG, Cohen SN, von Gabain A. 1987. Effect of premature termination of translation on mRNA stability depends on the site of ribosome release. *PNAS* **84**:4890–4894. DOI: https://doi.org/10.1073/pnas.84.14.4890, PMID: 2440033

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**:92–94. DOI: https://doi.org/10.1126/science.1180677, PMID: 20044577

Peltz SW, Brown AH, Jacobson A. 1993. mRNA destabilization triggered by premature translational termination depends on at least three cis-acting sequence elements and one trans-acting factor. *Genes & Development* **7**:1737–1754. DOI: https://doi.org/10.1101/gad.7.9.1737, PMID: 8370523

Richards J, Sundermeier T, Svetlanov A, Karzai A. 2008. Quality control of bacterial mRNA decoding and decay. *Biochimica Et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1779**:574–582. DOI: https://doi.org/10.1016/j.bbagrm.2008.02.008

Rosenberger RF, Foskett G. 1981. An estimate of the frequency of in vivo transcriptional errors at a nonsense Codon in *Escherichia coli*. *Molecular and General Genetics MGG* **183**:561–563. DOI: https://doi.org/10.1007/BF00268784, PMID: 7038382

Rosenberger RF, Hilton J. 1983. The frequency of transcriptional and translational errors at nonsense codons in the lacZ gene of *Escherichia coli*. *Molecular and General Genetics MGG* **191**:207–212. DOI: https://doi.org/10.1007/BF00334815

Saxowsky TT, Meadows KL, Klungland A, Doetsch PW. 2008. 8-Oxoguanine-mediated transcriptional mutagenesis causes ras activation in mammalian cells. *PNAS* **105**:18877–18882. DOI: https://doi.org/10.1073/pnas.0806464105, PMID: 19020090

Schwartz MH, Wang H, Pan JN, Clark WC, Cui S, Eckwahl MJ, Pan DW, Parisien M, Owens SM, Cheng BL, Martinez K, Xu J, Chang EB, Pan T, Eren AM. 2018. Microbiome characterization by high-throughput transfer RNA sequencing and modification analysis. *Nature Communications* **9**:5353. DOI: https://doi.org/10.1038/s41467-018-07675-z, PMID: 30559359

Shaw RJ, Bonawitz ND, Reines D. 2002. Use of an *in* vivo reporter assay to test for transcriptional and translational fidelity *in* yeast. *Journal of Biological Chemistry* **277**:24420–24426. DOI: https://doi.org/10.1074/jbc.M202059200, PMID: 12006589

Springgate CF, Loeb LA. 1975. On the fidelity of transcription by *Escherichia coli* ribonucleic acid polymerase. *Journal of Molecular Biology* **97**:577–591. DOI: https://doi.org/10.1016/S0022-2836(75)80060-X, PMID: 1102716

Strathern JN, Jin DJ, Court DL, Kashlev M. 2012. Isolation and characterization of transcription fidelity mutants. *Biochimica Et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1819**:694–699. DOI: https://doi.org/10.1016/j.bbagrm.2012.02.005

Sugimoto N, Yasumatsu I, Fujimoto M. 1997. Stabilities of internal rU-dG and rG-dT pairs in RNA/DNA hybrids. *Nucleic Acids Symposium Series* **37**:199–200.

Sugimoto N, Nakano M, Nakano S. 2000. Thermodynamics-structure relationship of single mismatches in RNA/DNA duplexes. *Biochemistry* **39**:11270–11281. DOI: https://doi.org/10.1021/bi000819p, PMID: 10985772

Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. 2012. Drift-barrier hypothesis and mutation-rate evolution. *PNAS* **109**:18488–18492. DOI: https://doi.org/10.1073/pnas.1216223109, PMID: 23077252

Sung W, Ackerman MS, Gout JF, Miller SF, Williams E, Foster PL, Lynch M. 2015. Asymmetric Context-Dependent mutation patterns revealed through Mutation-Accumulation experiments. *Molecular Biology and Evolution* **32**:1672–1683. DOI: https://doi.org/10.1093/molbev/msv055, PMID: 25750180

Sung W, Ackerman MS, Dillon MM, Platt TG, Fuqua C, Cooper VS, Lynch M. 2016. Evolution of the Insertion-Deletion mutation rate across the tree of life. *G3: Genes|Genomes|Genetics* **6**:2583–2591. DOI: https://doi.org/10.1534/g3.116.030890

Traut TW. 1994. Physiological concentrations of purines and pyrimidines. *Molecular and Cellular Biochemistry* **140**:1–22. DOI: https://doi.org/10.1007/BF00928361, PMID: 7877593

Traverse CC, Ochman H. 2016. Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *PNAS* **113**:3311–3316. DOI: https://doi.org/10.1073/pnas.1525329113, PMID: 26884158

Traverse CC, Ochman H. 2018. A Genome-Wide Assay Specifies Only GreA as a Transcription Fidelity Factor in *Escherichia coli* . *G3: Genes|Genomes|Genetics* **8**:2257–2264. DOI: https://doi.org/10.1534/g3.118.200209

van Dijk D, Dhar R, Missarova AM, Espinar L, Blevins WR, Lehner B, Carey LB. 2015. Slow-growing cells within isogenic populations have increased RNA polymerase error rates and DNA damage. *Nature Communications* **6**:7972. DOI: https://doi.org/10.1038/ncomms8972, PMID: 26268986

van Leeuwen FW, Burbach JP, Hol EM. 1998a. Mutations in RNA: a first example of molecular misreading in Alzheimer's disease. *Trends in Neurosciences* **21**:331–335. DOI: https://doi.org/10.1016/S0166-2236(98)01280-6, PMID: 9720597

van Leeuwen FW, de Kleijn DP, van den Hurk HH, Neubauer A, Sonnemans MA, Sluijs JA, Köycü S, Ramdjielal RD, Salehi A, Martens GJ, Grosveld FG, Peter J, Burbach H, Hol EM. 1998b. Frameshift mutants of beta amyloid precursor protein and ubiquitin-B in Alzheimer's and Down patients. *Science* **279**:242–247. DOI: https://doi.org/10.1126/science.279.5348.242, PMID: 9422699

Vermulst M, Denney AS, Lang MJ, Hung CW, Moore S, Moseley MA, Mosely AM, Thompson JW, Thompson WJ, Madden V, Gauer J, Wolfe KJ, Summers DW, Schleit J, Sutphin GL, Haroon S, Holczbauer A, Caine J, Jorgenson J, Cyr D, et al. 2015. Transcription errors induce proteotoxic stress and shorten cellular lifespan. *Nature Communications* **6**:8065. DOI: https://doi.org/10.1038/ncomms9065, PMID: 26304740

Westermann AJ, Gorski SA, Vogel J. 2012. Dual RNA-seq of pathogen and host. *Nature Reviews Microbiology* **10**:618–630. DOI: https://doi.org/10.1038/nrmicro2852, PMID: 22890146

Zheng G, Qin Y, Clark WC, Dai Q, Yi C, He C, Lambowitz AM, Pan T. 2015. Efficient and quantitative high-throughput tRNA sequencing. *Nature Methods* **12**:835–837. DOI: https://doi.org/10.1038/nmeth.3478, PMID: 26214130