9090–9104 *Nucleic Acids Research, 2013, Vol. 41, No. 19* *Published online 7 August 2013*
doi:10.1093/nar/gkt698

Direct assessment of transcription fidelity by high-resolution RNA sequencing

Masahiko Imashimizu[1], Taku Oshima[2], Lucyna Lubkowska[1] and Mikhail Kashlev[1,*]

[1]Gene Regulation and Chromosome Biology Laboratory, Frederick National Laboratory for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, MD 21702, USA and [2]Graduate School of Biological Sciences, Nara Institute of Science and Technology, 8916-5, Takayama, Ikoma, Nara 630-0192, Japan

Received June 10, 2013; Revised July 12, 2013; Accepted July 15, 2013

ABSTRACT

Cancerous and aging cells have long been thought to be impacted by transcription errors that cause genetic and epigenetic changes. Until now, a lack of methodology for directly assessing such errors hindered evaluation of their impact to the cells. We report a high-resolution Illumina RNA-seq method that can assess noncoded base substitutions in mRNA at 10^{-4}–10^{-5} per base frequencies *in vitro* and *in vivo*. Statistically reliable detection of changes in transcription fidelity through $\sim 10^3$ nt DNA sites assures that the RNA-seq can analyze the fidelity in a large number of the sites where errors occur. A combination of the RNA-seq and biochemical analyses of the positions for the errors revealed two sequence-specific mechanisms that increase transcription fidelity by *Escherichia coli* RNA polymerase: (i) enhanced suppression of nucleotide misincorporation that improves selectivity for the cognate substrate, and (ii) increased backtracking of the RNA polymerase that decreases a chance of error propagation to the full-length transcript after misincorporation and provides an opportunity to proofread the error. This method is adoptable to a genome-wide assessment of transcription fidelity.

INTRODUCTION

Transcription infidelity by RNA polymerases (RNAPs) has been proposed to contribute to genome instability (1) and heritable phenotypic changes (2,3), which may affect aging (4) and carcinogenesis (5,6). To date, assessment of transcription fidelity *in vivo* has been performed with reporter genes targeting a small number of sequences with a limited spectrum of errors (1,7–10). To extrapolate this limited fidelity analysis to a genome-wide scale, an assumption has been made that transcription errors are randomly distributed. However, several reports have suggested that transcription errors exhibit strong sequence preferences (11–14). Fidelity analysis for the entire transcriptome has been limited by a lack of a reliable methodology. In the past decade, extensive *in vitro* analyses of transcription fidelity revealed several error-avoidance and error-correcting mechanisms based on biochemical assays for misincorporation of a unique NMP (12,13, 15–20) and single-molecule assays using optical trapping techniques (11,21). Typically, these experiments included limited or unbalanced substrate concentrations to detect misincorporation. These *in vitro* data cannot be easily extrapolated to the genetic fidelity assays involving reporter genes transcribed at high *in vivo* concentration of substrates and in the presence of transcription factors and structural proteins compacting DNA (1,7–10,22,23). Therefore, there is an urgent need for an approach that would allow simultaneous assessment of transcription fidelity *in vivo* and *in vitro* under balanced NTP concentration and on the same DNA sequences.

Deep sequencing technologies such as RNA sequencing (RNA-seq) can analyze $\geq 10^{10}$ bases in a single run, potentially allowing both a genome-wide and *in vitro* detection of transcription error rates around $10^{-5}\,b^{-1}$ rate (7,17,18). However, conventional protocols for RNA-seq generate background errors at $>10^{-5}\,b^{-1}$ frequency during the process of cDNA library/cluster formation, sequencing/detection and the mapping of the reads (24), which has made it difficult to detect transcription errors. Advanced deep sequencing techniques use tagging of individual DNA molecules by random sequences in polymerase chain reaction (PCR) primers to identify and filter out the PCR artifacts by counting only those error spots that persist throughout all DNA molecules carrying the same tag (25–27). This tag-based method substantially reduces randomly distributed PCR and sequencing errors of the deep DNA/RNA sequencing (25–27). A problem remaining in this method is that it cannot reduce the errors introduced by reverse transcriptases (RTs) that typically

To whom correspondence should be addressed. Tel: +301 846 1798; Fax: +301 846 6988; Email: kashlevm@mail.nih.gov

have lower fidelity than DNA polymerases (DNAPs) used for PCR (28,29). More recently, a deep-sequencing method was developed involving analysis of mismatches in overlapping read pairs to identify the artifact errors, but not the RT errors (30). Thus, so far there is no an approach suitable for discriminate RNA errors from the RNA-seq artifacts. Here, we present a high-resolution RNA-seq method based on a remarkable sequencing depth of 10^6 accompanied by several technical improvements reducing background errors to 10^{-5} and 10^{-4} levels. This technique enables statistically reliable detection of changes in transcription fidelity *in vitro* and in living cells, despite the presence of the artifact errors. This methodology may also be instrumental in addressing controversial noncanonical posttranscriptional RNA-editing (31–35), identification of genomic 'hotspots' for transcription errors and their contribution to the genetic diversity of viral populations (27,29,30,36).

MATERIALS AND METHODS

Reagents

NTPs, oligonucleotides and DNA purification kits were purchased from GE Healthcare, Integrated DNA Technologies and Qiagen, respectively. NTPs used in the misincorporation assay (Figure 5 and Supplementary Figure S5) were further purified as described previously (17). The high fidelity RT PrimeScript and the DNAP PrimeSTAR Max used for the cDNA preparation were purchased from Takara Bio.

Proteins

RNAP holoenzyme of *Escherichia coli* RL-916 (the strain was a kind gift from Dr Robert Landick) containing a histidine-tagged RpoC subunit was purified as described previously (37). The GreA and GreB expression plasmids pDNL278 and pMO1.4 were kind gifts from Dr Sergei Borukov. The plasmids were transformed into *E. coli* strain XL1-Blue cells (Stratagene) for overexpression. The recombinant GreA and GreB were purified according to (38) with the addition of Mono Q column (GE Healthcare) chromatography.

In vitro RNA preparation

The pPR9 plasmid containing lambda phage P_R promoter and fd phage terminator was used for the DNA template (Supplementary Figure S1A). The transcribed region is composed of refampicin-resistant *rpoB* gene that contains a 1546G→T mutation, and partial *rplL* and *rpoC* genes of *E. coli*. The plasmid DNA was purified by Qiagen mini-prep kit and phenol/chloroform/isoamylalcohol (25:24:1). The residual phenol that may affect transcription was removed by solvent extraction with diethyl ether. For transcription reaction, 400 nM holoenzyme in the absence or presence of 12 mM GreA and 4 mM GreB was incubated with 1 mM NTP and 2 nM the plasmid DNA for 15 min at 37°C in transcription buffer [TB; 20 mM Tris–HCl, pH 7.9, 5 mM $MgCl_2$ (or 1 mM $MnCl_2$), 1 mM 2-mercaptoehanol, 0.1 M KCl, 0.1 mg/ml bovine serum albumin] (Supplementary Figure S1B). The reaction was stopped by heat denaturation for 3 min at 90°C followed by DNase I (Takara Bio) treatment for 20 min at 37°C. We verified the production of a homogeneous 5.7 kb RNA by agarose-gel electrophoresis before adding DNase I (Supplementary Figure S1C). The 5.7 kb RNA was purified from the digested DNA, NTPs, abortive oligo-RNA products and proteins as shown in Supplementary Figure S1D.

In vivo RNA preparation

Total RNA was prepared from *E. coli* MG1655 strain harboring pPR9 plasmid. Cells were cultured at 28°C in LB medium containing ampicillin. The overnight cell culture was inoculated into the fresh medium at 1/70 (v/v) and was incubated for ~2 h at 28°C (OD_{600} reached 0.35) and then for 2 h at 42°C (OD_{600} reached 2.3) to induce the P_R promoter (39). The cells in 200 ml culture were harvested and resuspended with a solution containing 0.5% sodium dodecyl sulphate, 20 mM sodium acetate (pH 5.5) and 10 mM EDTA. The suspended cells were mixed with an equal volume of prewarmed saturated phenol (20 mM sodium acetate, 10 mM EDTA, pH 5.5) and incubated for 5 min at 60°C. The mixture was centrifuged, and RNA and DNA were precipitated with ethanol from the supernatant. The pellet was dissolved in DNase I buffer with 10 U of DNaseI and incubated for 30 min. RNA was separated from the digested DNA by acidic phenol extraction followed by G-50 Micro column (GE Healthcare) purification and then precipitated with ethanol. The pellet was dissolved in diethylpyrocarbonate-treated water and used for cDNA synthesis.

Library preparation

The first DNA strand was synthesized using the transcript from the P_R promoter (0.8 μg RNA synthesized *in vitro* or 5 μg total RNA purified from *E. coli* cells) and a RT PrimeScript. The RNA transcript was mixed with 1 mM dNTP and 5 μM of two specific primers (a and b, Figure 1A) that hybridizes to the RNA transcript at the most 3′ portion of the DNA segments 1 and 6 (Figure 1A) of the first PCR. A hairpin structure between the segments 1 and 2 inhibits elongation of RT on the RNA transcript to the 5′ end. The mixture was incubated for 5 min at 65°C. The PrimeScript, 1× PrimeScript buffer and RNase Inhibitor were added to the mixture according to the manufacturers' instructions, and the mixture was incubated for 45 min at 42°C, for 5 min at 37°C with RNase H and for 15 min at 70°C. The single-strand DNA product was purified with MinElute PCR purification kit and eluted with 10 μl of the elution buffer. The first PCR including the second DNA strand synthesis was performed with a DNAP PrimeSTAR Max based on the manufacturers' instructions at 5 cycles for the RNA preparation *in vitro* and 10 cycles for the RNA preparation *in vivo*. We noticed that the total RNA purified from *E. coli* cells had a concentration of the unique transcript from the P_R promoter of the pPR9 plasmid by ~30-fold less than the *in vitro* RNA preparation. Thus, the five additional cycles make almost same final concentrations

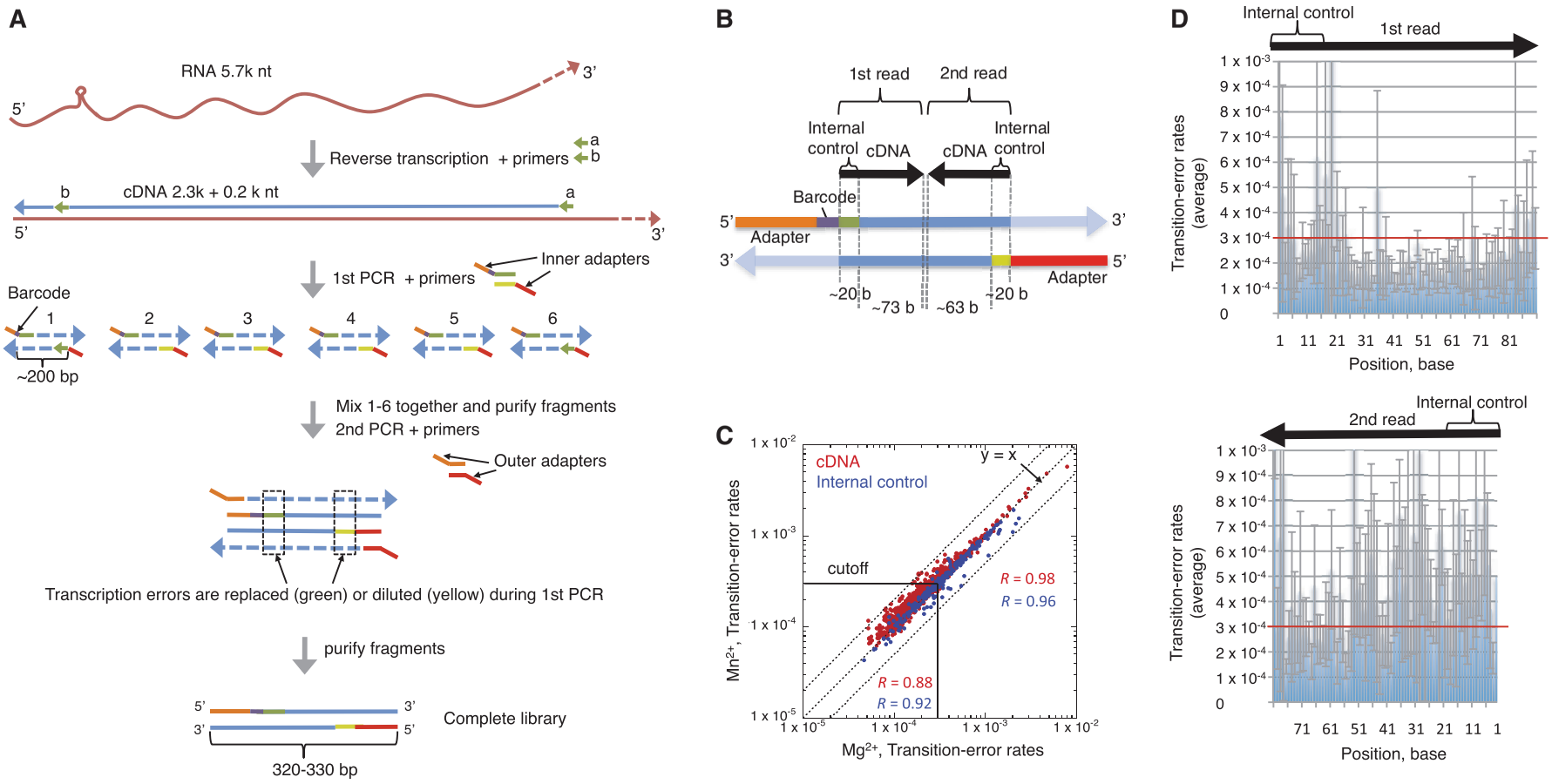

Figure 1. Experimental setup for transcription-error analysis. (**A**) Schematic representation of reverse transcription and two PCR steps used to produce barcoded cDNA libraries. The five libraries were made from each of the RNA samples corresponding to the five transcription conditions (Mg^{2+}, Mn^{2+}, GreAB/Mg^{2+} and GreAB/Mn^{2+} for *in vitro* and *E.coli* cell). The RT primers 'a' and 'b' (the green arrowheads) replace transcription errors with the chemical oligonucleotide synthetic errors during reverse transcription step. Similarly, in a course of PCR, the first PCR primers (green and yellow lines) replace (green lines) or dilute by >10-fold (yellow lines) transcription errors in the corresponding regions to which these primers hybridize (shown by empty boxes). A six-bases barcode (purple line) and Illumina-specific sequencing adapters (orange and red lines) are introduced to the libraries during first and second PCR steps. (**B**) The cDNA and internal control regions in the PCR fragment used for Illumina paired-end sequencing. The lengths and directions of the first and the second sequencing reads are indicated. Both sequencing reads contain ~20 bases of the primer-hybridizing regions where transcription errors are significantly depleted during cDNA preparation (internal controls). All colors are the same as in panel A, except that the DNA regions lacking the original mRNA are white-shaded. (**C**) Scatter plot of transition-error rates for Mn^{2+} and Mg^{2+} RNA products *in vitro*. Positions in cDNA and internal control are indicated by red and blue colors. The diagonal dotted lines represent $y = 2x$ (upper), $y = x$ (middle) and $y = 1/2x$ (lower). Correlation coefficient (R) of the two samples with or without cutoff value $>3 \times 10^{-4}$ b^{-1} is shown. (**D**) Transition-error rates in the second read (lower) of the paired-end sequencing are higher than those in the first read (upper). Transition-error rates averaged by the five different RNA preparations and the six sequence segments (see panel A) are plotted against DNA positions with the standard deviations. Red line indicates the cutoff value.

of cDNA libraries derived from the *in vitro* and *in vivo* RNA preparations. One-tenth total reaction volume of the single strand DNA purified by the Qiagen kit and each primer pair including a barcode and the inner Illumina sequence adapters in the 5′ tails (Figure 1A and Supplementary Table S5) were used for the PCR. This PCR amplified the six different DNA segments comprising the cDNA (transcript) and the internal control (primer) (Figure 1A and B) for the five libraries with respective barcodes. We confirmed that no first PCR product was obtained in each primer pair when RNA solution without RT was used as a template. The six DNA segments obtained from each reaction tube of the first PCR were mixed and purified by the Qiagen kit, and eluted with 10 µl of the elution buffer. The second PCR was performed with one-fifth total reaction volume of the obtained PCR products, a primer pair containing the outer sequencing adapters in the 5′ tails (Supplementary Table S5), and the same PCR enzyme as the first PCR at 6 cycles. The presence of the full-length Illumina sequence adapter and barcode sequence (Illumina TruSeq Index 1–5) in each of the five cDNA libraries was confirmed by Sanger sequencing.

Illumina sequencing

Quantifications of the numbers of amplifiable molecules in the libraries were performed by qPCR using a Library Quantification Kit (KK4824, Kapa Biosystems) and Agilent 2100 Bioanalyzer. The cluster generation on a paired-end flow cell and sequencing were performed with cBot and HiSeq 2000, respectively, according to the user guides of Illumina. The summary of sequencing data is shown in Supplementary Table S1. Raw sequencing data and processed data are available for download at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE46479.

Data analysis for the sequencing

The initial data processing, including reads separation by the barcodes and the generation of fastq files, was performed with the CASAVA software (Illumina). A single large fastq format file of high quality reads (Phred score $Q \geq 30$, see Supplementary Table S1) was split into about 10 smaller files by using a shell script splitReads.sh (http://code.google.com/p/perm/downloads/detail?name=splitReads.sh) to use SAMtools mpileup –A commands for the following error analysis in the sequence reads (see below). The obtained reads were aligned and mapped to the pPR9 plasmid DNA sequences (1056 bp) using a program Bowtie 0.12.7 that does not allow insertion and deletion in the alignment (40). We chose Bowtie parameter that allows three mismatches. To calculate error rates, we counted the numbers of 4 bases A, T, G, C, and N (not determined) in each position of the mapped reads by using the program SAMtools 0.1.18 (41) with supplemental use of a Perl script, parse_samtools_mpileup.pl (a kind gift from Dr Wei Shao). Each type of error rates per position was determined as the number of sequence reads with a particular type of base-substitution divided by the number of the reads with the reference base in each DNA position.

Ternary elongation complex formation and biochemical transcription assays

The ternary elongation complexes (TECs) carrying 5′-labelled RNAs (see Supplementary Table S6) were assembled and immobilized on Ni^{2+}-NTA agarose (Qiagen) in TB as described previously (42,43). Five to ten picomoles of RNAP was incubated with 7.5–15 pmols of the preannealed RNA–DNA hybrid in 25–50-µl volume for 10 min at room temperature. Next, 15–50 pmols of the nontemplate DNA strand (NDS) were added for 10 min. The immobilized TEC9s were washed with TB containing 1 M KCl. The TECs were eluted from Ni^{2+}-NTA agarose by TB(-$MgCl_2$) with 100 mM imidazole as described previously (17) and diluted with TB(-$MgCl_2$). TEC18s were obtained from TEC9s by walking on the template of 170 G and 474 G sequences (37). To allow the 1 or 2 steps of walking, the 14 G and 17 G on the 170 G and 474 G sequences, respectively, were substituted with C (see Figures 4A, 5B and Supplementary Figure S3A). The typical concentration of TEC is ~1 nM (44). All reactions were performed in TB at 37°C. The reactions were stopped by gel-loading buffer (5 M urea; 25 mM EDTA final concentrations). The RNA products were analyzed as described previously (17). Details about the experimental setups for misincorporation, mismatch-extension, RNA cleavage and NTP competition assays are described in the corresponding figures or supplementary figures.

Exonuclease III footprinting

The rear-end Exonuclease III (ExoIII) footprinting was performed as described previously (17,44). TECs were assembled on the 5′ end-labeled template DNA strand and the unlabeled NDS (Supplementary Table S6). The reaction was started by mixing 15 µl TB containing 10 U of ExoIII (New England Biolabs) with 15 µl of the elongation complex at 30°C. To prevent digestion of the NDS by ExoIII for the rear-end footprinting, the NDS carries phosphorothioate bond at the 3′ end. The active state (pretranslocated state) and backtracked states of TECs were determined by shifting the boundaries of RNAP due to stepwise extension of RNA in TECs (17,44).

RESULTS

Strategy for the assessment of transcription fidelity by RNA-seq

The strategy is based on two key assumptions: (i) combined error rates of RT and DNAPs used for the RNA-seq can be reduced to $\sim 10^{-5}$ b^{-1} range by using high-fidelity RT and DNAPs. The estimated error rates are based on information provided by the manufacturers, which are consistent with our data (see Supplementary Table S3), (ii) multi-subunit RNAPs generate errors with sequence preferences different from those of the structurally unrelated RTs/DNAPs as suggested previously (11–14,36,45). We did not use the tag-based error-correction method to reduce artifact errors because this approach cannot identity/correct RT errors and typically

increases the number of PCR cycles owing to the loss of the original templates for PCR in a course of the DNA tagging (25–27). Instead, we significantly reduced the number of PCR cycles to minimize the DNAP errors during the library production (see 'Materials and Methods' section). To identify the sequence sites dominated by transcription errors, we used error-prone and error-proof transcription conditions *in vitro* to increase and decrease transcription errors, respectively. In this system, transcription error rates are changed in a controlled manner for the sequencing reads, whereas the artifact errors remain constant. Detection of transcription errors should be possible at sequence sites favoring transcription and disfavoring the artifact errors even when the averaged enzymatic artifact rates exceed those of transcription. Widespread existences of such sites through the analyzed sequences should be also statistically evaluated. We also reduced the nonenzymatic sequencing errors caused by incorrect base-calling (46) and misalignment (47) of the Illumina reads by setting an appropriate filter eliminating these artificial error hotspots (see below). Finally, we significantly increased the read depth to average 3×10^6 to cover the predicted $\sim 10^{-5}$ b^{-1} rate for transcription errors.

The RNA samples for the RNA-seq were generated by transcription of pPR9 plasmid (39) by *E. coli* RNAP *in vitro* and *in vivo*. The plasmid contains an ~5.7-kb fragment of *E. coli rpoBC* operon that is transcribed from a strong lambda phage P_R promoter and terminated at an fd phage transcription terminator (Supplementary Figure S1A). A multi-round transcription by the purified RNAP holoenzyme generated $\sim 10^{15}$ RNA molecules with a uniform length of ~5.7 kb (Supplementary Figure S1 B-D). The reference transcription reaction was performed in a TB (42) with 5 mM $MgCl_2$ to determine the standard error rate. To reduce fidelity (the error-prone condition), we replaced Mg^{2+} with Mn^{2+} (48–50). To increase fidelity (the error-proof condition), we added GreA/GreB proteins (51) for proofreading activity. We kept a balanced high concentration of NTPs (1 mM) in all conditions to avoid forced nonphysiological misincorporation, although the actual concentrations of NTPs *in vivo* may not be uniform and vary under different growth conditions (52). For the *in vivo* fidelity measurement, we purified total RNA from the wild-type *E. coli* strain harboring the same pPR9 plasmid after 2 h induction of the P_R promoter at 42°C (39).

We established a method for preparing five different cDNA libraries each with its own barcode for Illumina sequencing (Figure 1A). Each 6-nt barcode allows multiplexing all five *in vitro* and *in vivo* preparations in a single sequencing analysis. The 5′ fragment of the 5.7 kb RNA transcripts was reverse transcribed, and the product was subjected to PCR reactions that generated six ~200 bp segments (Figure 1A). The primers contained a specific barcode for each of the five starting preparations and the inner Illumina-sequencing adapters (Figure 1A). The second step of PCR generated the final cDNA libraries for the Illumina sequencing by using the first-step PCR product as a template and primers containing the outer sequencing adapters in the 5′ tails (Figure 1A). In the first cycle and the remaining 4 cycles of the first PCR, chemical synthetic errors in the DNA primers replace and steadily dilute transcription errors by 2-fold in mRNA segments to which these primers hybridize. Thus, transcription errors in the corresponding cDNA segments were replaced or 16-fold reduced by 5 cycles of the first PCR (Figure 1A, the empty boxes). Consequently, contribution of transcription errors in these segments becomes negligible in the final cDNA libraries compared with the rest of cDNA. Importantly, we used these outer segments (shown by green and yellow lines, Figure 1A and B) as internal controls to compare error rates in these sequences with those in the embedded cDNA segment carrying intact transcription errors (Figure 1B, blue lines). Base substitution errors made during synthesis of primer DNA are reported at 10^{-4}–10^{-5} b^{-1} rates (based on the manufacturer information), which is consistent with our data (see Supplementary Figure S2 and Table S4).

We obtained 191 099 124 reads with high base-calling quality [Phred score $Q \geq 30$ (46)] by the paired-end sequencing (Supplementary Table S1). Each sequenced read included the cDNA and the internal control sequence (Figure 1B). The uniquely mapped sequence reads covered 1056 bp with an average 3×10^6 read depth (Supplementary Table S2). To assess types and rates of RNA/DNA changes per position, we excluded insertion and deletion errors to avoid reads misalignment (47) during bioinformatic analysis. In the mapped sequence reads, we found a few positions with abnormally high background of transversions A→C (first read) and G→T (second read) with 10^{-2} or 10^{-3} b^{-1} frequency, which are unlikely due to transcription errors. These errors are probably caused by the relatively close emission spectra of the corresponding fluorophores and their incomplete separation by optical filters in the Illumina platform at these particular positions (24). These rare positions have been ignored.

We plotted transition error rates for the cDNA sequences in the standard Mg^{2+} transcription condition against the error-prone Mn^{2+} condition (Figure 1C, red dots). We compared this plot with the corresponding plot derived from the internal controls where transcription errors were replaced or diluted with the oligo DNA-synthetic errors (Figure 1C, blue dots). If there are no differences between the Mg^{2+}/Mn^{2+}sets (indicating a failure in detection of transcription errors), the data should fall along the y = x line as is observed for the internal control positions ($R > 0.9$). In contrast, for the cDNA positions, the plots of the lower error rates were localized in the y > x area ($R < 0.9$ for the $\leq 3 \times 10^{-4}$ b^{-1} rates). Two-tailed F-test for the Mg^{2+}/Mn^{2+} RNA samples confirmed that the error rates $\leq 3 \times 10^{-4}$ b^{-1} are not equally distributed in cDNA [$P = 2 \times 10^{-4}$ ($n = 540$)] as opposed to their equal distribution in the internal controls [$P = 0.5$ ($n = 125$)]. Notably, the transition errors occurring at $> 3 \times 10^{-4}$ b^{-1} rates were primarily observed in the second read that required an additional strand synthesis step (Figure 1D). Therefore, these errors mostly derived from the artifact of paired-end sequencing. We used this information to set a cutoff value of 3×10^{-4} b^{-1} error rate in our statistical analysis of transcription errors.

The high-resolution RNA-seq detects changes in transition error rates *in vitro*

Next, we separately compared each type of transition error between the two *in vitro* RNA samples representing the standard and the error-prone transcription conditions (Mg^{2+}/Mn^{2+} plot, Figure 2, left column). We observed an up to 2-fold Mn^{2+}-dependent increase in errors for G→A and T(U)→C transitions in a majority of cDNA positions in the error range from 3×10^{-4} to 6×10^{-5} b^{-1}. A nonparametric *t*-test between the two samples provided a significant difference in means of the two samples ($P < 0.05$). We observed slight Mn^{2+}-dependent increase in C→T(U) transition rate ($P = 0.09$) and no difference in A→G transition ($P = 0.7$). Because the detected mean rate of A→G transition was the lowest among the four types of transitions (Supplementary Figure S2), the A→G transcription errors appeared to be masked by the artifacts even in the error-prone conditions for RNAP. Note that the internal control showed no significant effect of Mn^{2+} on any type of transition error (Figure 2, left column). Thus, the RNA-seq detected Mn^{2+}-

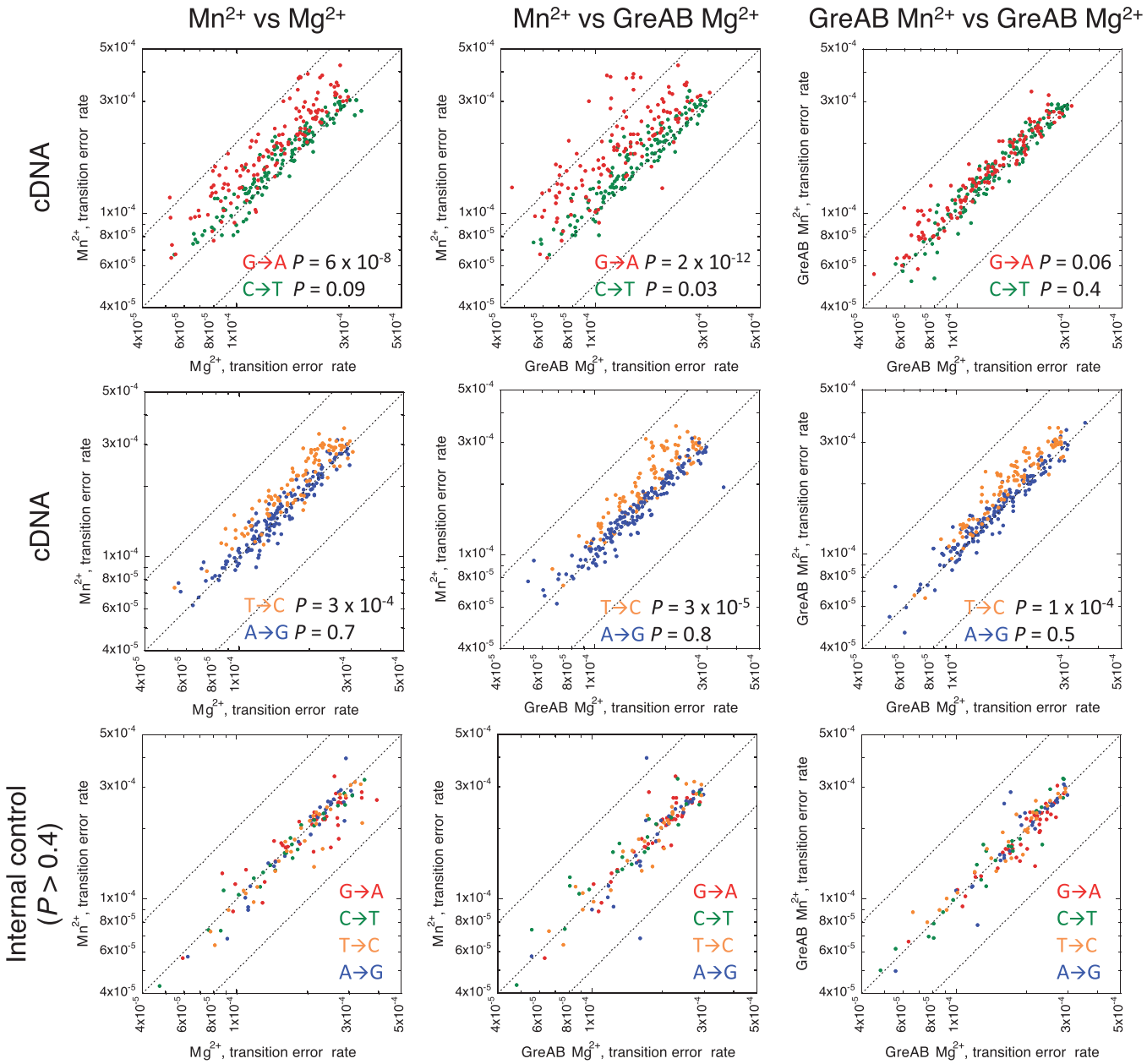

Figure 2. Scatter plots of transition-error rates. The error rates per position in the cDNA and internal control are plotted for error-prone/standard (left column), error-prone/error-proof (middle column) and moderate-error-proof/error-proof (right column) sets of conditions as shown on the top. The error rates $\leq 3 \times 10^{-4}$ b^{-1} were used for the statistical analysis. *P* value of two-tailed nonparametric *t*-test for the two samples is shown. For the cDNA, $n = 132$ (G→A), $n = 142$ (C→T), $n = 104$ (T→C) and $n = 162$ (A→G). For the internal control, $n = 39$ (G→A), $n = 26$ (C→T), $n = 30$ (T→C) and $n = 30$ (A→G).

dependent increase in three types of transcription errors made by *E. coli* RNAP *in vitro* at the 10^{-5}–10^{-4} b^{-1} rates.

To further validate the difference of the error rates in the Mn^{2+}/Mg^{2+} samples, we added GreA/B to our standard reaction (Mg^{2+}). GreA/B are expected to reduce errors by its proofreading activity (15,53). As expected, GreA/B amplified the differences in the rates of G→A, T(U)→C and C→T(U) transitions between the Mn^{2+} and Mg^{2+} samples (Figure 2, middle column), and significantly reduced the means of the three transition-error rates in Mg^{2+} and Mn^{2+} samples (Figure 2, right column), indicating that we detected proofreading activity of GreA/B in both Mg^{2+} and Mn^{2+} conditions. GreA/B did not affect the error rates in the internal control. A→G transitions were also not affected by GreA/B in the cDNA and in the internal control regions, again suggesting the artifact origin of the majority of A→G errors. In a fraction of cDNA positions for the other three transition types, we also did not observe significant changes in the error rates between the error-prone and error-proof transcription conditions (Figure 2).

Comparison of transition-error rates *in vitro* and *in vivo*

It is broadly assumed that RNAP has similar intrinsic fidelity *in vivo* and *in vitro* (7,17). However, *in vitro* fidelity assessed by single NMP misincorporation assay does not account for error propagation to full-length RNA. Moreover, the *in vitro* fidelity, defined as a ratio of k_{pol}/K_d for cognate and noncognate NTP (17,18), does not take into account for proofreading activity of RNAP that requires backtracking of the enzyme. The *in vivo* fidelity could also be affected by local DNA structures, DNA damage and promoter strength of the gene (1,54). Therefore, the *in vitro* fidelity is not exactly related to the *in vivo* fidelity.

To evaluate these differences, we used the RNA-seq to compare the error rates for the same RNA produced either *in vitro* or *in vivo*. Scatter plots visualized the differences in transition-error rates between *in vivo* sample and *in vitro* Mg^{2+} samples ± GreA/B (Figure 3). In the cDNA positions, we observed significant differences between the *in vivo* and the standard ($+Mg^{2+}$) *in vitro* samples: C→T(U) transitions were overrepresented in the *in vivo* samples ($P < 0.05$), whereas G→A and T(U)→C transitions were underrepresented ($P \leq 0.05$) (Figure 3). For G→A and T(U)→C transitions, addition of GreA/B to the *in vitro* reaction reduced the differences between the *in vivo* and *in vitro* samples (Figure 3). This result indicates an extensive RNA proofreading by GreA/B in the living cells as was suggested previously (55). Thus, our data suggest that transcription in the wild-type *E. coli* cells containing functional GreA/B proteins has similar fidelity as transcription *in vitro* in the presence of Gre factors.

The increased rate of C→T(U) transition in the *in vivo* sample was insensitive to GreA/B, suggesting that these errors may be introduced by DNAP during the five additional cycles of the first PCR used only for cDNA synthesis with the *in vivo* RNA sample. The same increased background may dilute G→A and T(U)→C errors for the *in vivo* sample. The detected difference in C→T(U) error rates might be caused by a modest cellular stress during shift to 42°C for induction of P_R promoter of the *rpoBC* gene (see 'Materials and Methods' section), decrease of intrinsic fidelity of RNAP for certain types of errors at elevated temperature or due to a spontaneous cytosine deamination before or during RNA purification from *E. coli* cells. Although the *in vivo* frequency of a spontaneous deamination of cytosines in DNA is known to occur at 10^{-9} order (56), the corresponding rate for the RNA is unknown. We also observed minor increases in the error rates for G→A and C→T(U) transitions in the internal controls for the *in vivo* sample compared with the standard *in vitro* sample (Figure 3). This was likely due to the errors introduced during the DNA oligonucleotides synthesis rather than the DNAP errors during PCR [see Supplementary Figure S2, the error rates for the oligo-DNA synthesis are slightly higher (by $\sim 1 \times 10^{-4}$) than those of transcription for all four types of transition].

Backtracking controls mismatch extension

We performed a hierarchical clustering analysis (57) of the error rates in all positions used for the statistical analysis of the errors at the lower than the threshold value, 3×10^{-4} b^{-1}. This analysis connects by a series of branches the DNA positions and the fluctuation patterns/levels of error rates depending on transcription conditions. Thus, this analysis identifies DNA positions exhibiting the similar error-rate profiles under the standard *in vitro*, error-prone Mn^{2+}, error-proof GreA/B and the *in vivo* conditions. We chose G→A error because of its highest sensitivity to Mn^{2+} and GreA/B (Figure 2). The G→A errors were clustered into major C–G and minor A, B groups where the error rates were increased and not affected by Mn^{2+}, respectively (Figure 4A). The majority of the Mn^{2+}-sensitive errors in groups C–G was also reduced by GreA/B (Figure 4A). This significant overlap strongly indicates transcription origin of these errors, which were susceptible to the chemical and protein factors specifically targeting transcription fidelity. We further argue that the Mn^{2+}-insensitive errors belonged to the RNA-seq artifacts that become more prominent at the sequences exhibiting relatively low transcription error rates. Alternatively, these sequences might generate 'true' transcription errors with an intrinsic resistance to Mn^{2+}. Note, that the averaged error rate in group A (1.1×10^{-4}) was lower than in group B (1.8×10^{-4}), suggesting that transcription errors from the former group are more diluted with the artifact errors.

In each cluster, we aligned the 9-nt sequences located immediately upstream from the G→A error site (Figure 4A). This was based on the assumption that the catalytic properties of RNAP are mainly determined in 9-bp RNA–DNA hybrid of a TEC (58). Interestingly, the RNA–DNA hybrid sequences for Mn^{2+}-insensitive errors were strongly enriched with short A/T(U) tracts (group A in Figure 4A) as opposed to the more balanced sequence content of the sites affected by Mn^{2+} and GreA/B (the representative Mn^{2+}-sensitive group F in Figure 4A). A/U-rich sequences in the RNA–DNA hybrid have been shown to promote RNAP backtracking on DNA (59) as

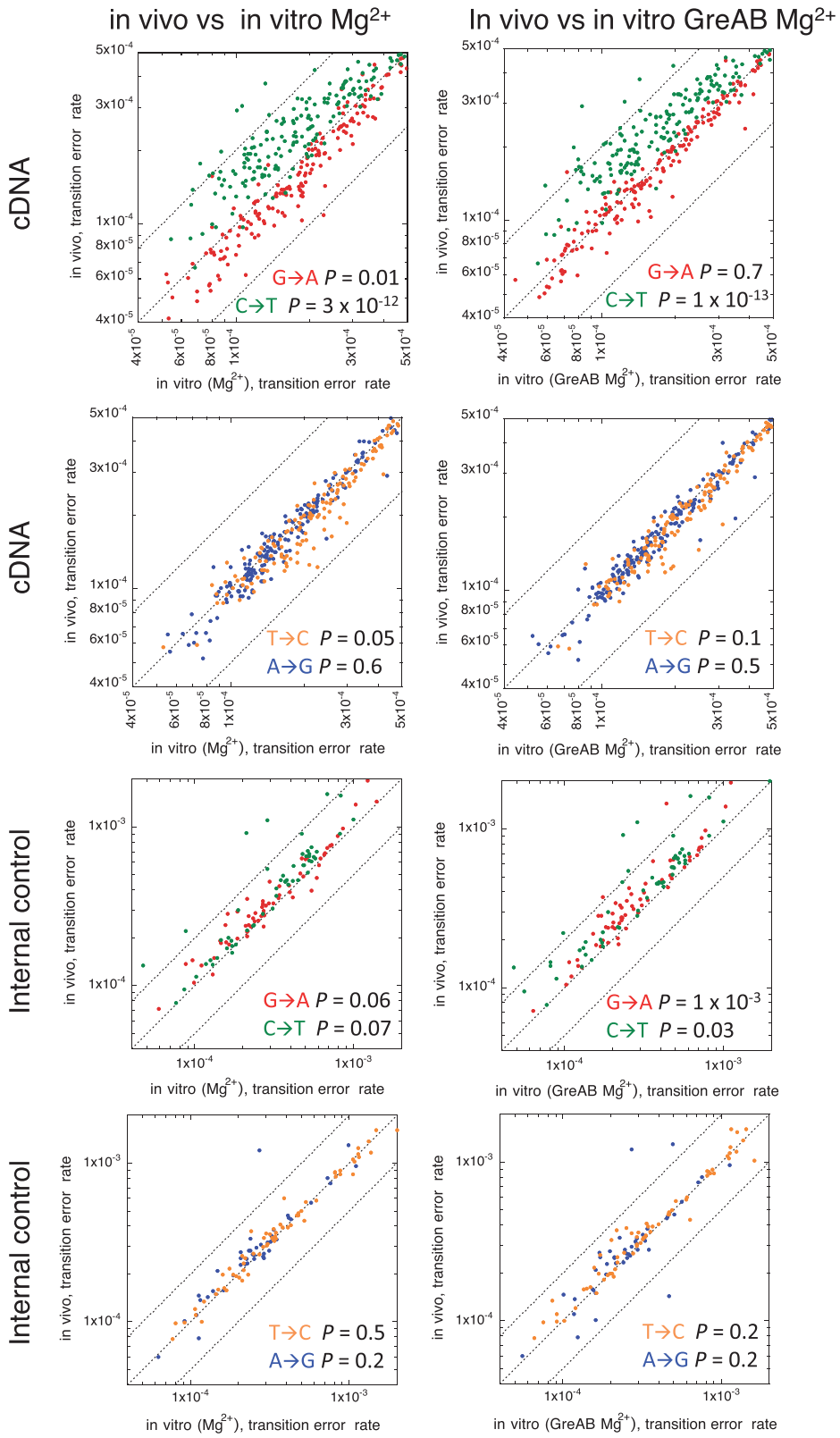

Figure 3. Scatter plot of transition-error rates for *in vivo* and *in vitro* Mg^{2+} samples with (left) or without (right) GreA/B in the cDNA (top) and internal control (bottom). All symbols are the same as in Figure 2. The cutoff for the error rates is applied for two-tailed nonparametric *t*-test, but not for the scatter plots. The *n* for the *t*-test is same as in Figure 2.

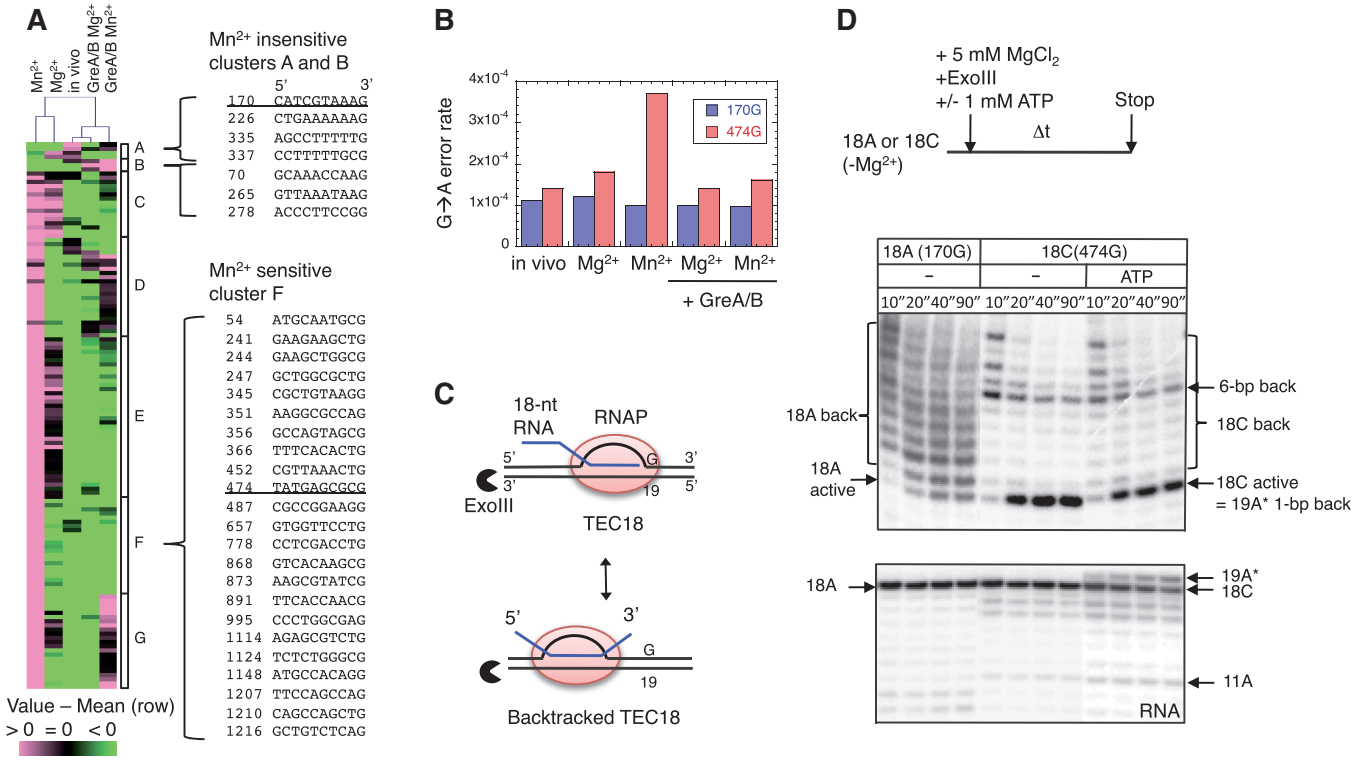

Figure 4. Mn^{2+}-sensitivity and frequency of G→A errors depend on propensity of RNAP to backtrack at the error site. (**A**) Hierarchical clustering was performed with MeV v4.7.0. G→A error rates exceeding 3×10^{-4} b^{-1} at 132 DNA positions are used to generate the clustering diagram. Each error rate is subtracted by the mean of five different RNA preparations to distinguish the error rate difference among the transcription conditions per position. Clusters A–G are indicated by boxes. The 10-nt DNA sequences (nontranscribed strand, 5′-to-3′ direction) where the G→A error occurred at the 3′ RNA end are shown. The number on the left side of each sequence indicates position of G residue analyzed by the RNA-seq. Two sequences from clusters A and F (170 G and 474 G, respectively) that were used for biochemical analyses are underlined. (**B**) G→A transition rates at the positions 170 G and 474 G in the five RNA preparations analyzed by the RNA-seq. (**C**) Schematic representations of reversible backtracking of TEC18 bearing the 10-nt sequence of 170G or 474G from +10 to +19 position, where +1 is 5′ end of the RNA. The access of Exo III form the rear end of RNAP is also shown. (**D**) ExoIII footprinting of the TEC18A and TEC18C. The reaction scheme is shown on the top. The rear-end boundaries of RNAP in the active and backtracked states are shown. The bottom panel shows the 18-nt RNA transcripts in the TECs. The capital letter following the number indicates the base of the 3′ RNA and the RNA length in TEC. AMP-misincorporation at the position 19 is marked by asterisk.

one of the mechanisms increasing RNAP fidelity (21). Thus, we assumed that a relatively low frequency of Mn^{2+}-insensitive transcription errors was related to increased backtracking of RNAP.

To address backtracking as a potential error-correcting mechanism during processive elongation, we arbitrarily selected one sequence from each group: 170G (Mn^{2+}-insensitive group A, relatively lower error rate) and 474 G (Mn^{2+}-sensitive group F, relatively high error rate) (Figure 4B) and analyzed RNAP backtracking at these sequences by ExoIII footprinting (17,44,60). A dynamic pattern of DNA digestion by ExoIII provides information of distance and stability for individual backtracked states of RNAP. The TEC was assembled with a 9-nt RNA hybridized to the DNA template containing 170 G or 474 G sequence with a modification that is required for the TEC walking (see 'Materials and Methods' section) (37). The 9-nt RNA was elongated to 18-nt length with NTPs, making TEC18A (corresponding to the 170 G sequence) or 18C (corresponding to the 474 G sequence), which has the new 3′RNA end located immediately 5′ of the site where G→A error was detected (Figure 4C). When RNAP reversibly backtracks, ExoIII that digests DNA from the rear-end of RNAP (Figure 4C) produces the expanded rear-end boundary(ies) of the backtracked state(s), which converts to a boundary of the active state on prolonged incubation with the nuclease (44). We observed two differences in backtracking at the 170 G and 474 G sequences (Figure 4D). TEC18A/170G was equilibrated between the active state and 1–10 bp stably backtracked states within 90 s of incubation with ExoIII, whereas TEC18C/474 G was equilibrated primarily in the active state with a minor 6-bp backtracked state. Thus, backtracking from the active state was more strongly induced or stabilized in TEC18A/170G compared with TEC18C/474 G as was predicted from the difference in their A/T(U) sequence contents (Figure 4A). Interestingly, AMP misincorporation in TEC18C/474 G (mimicking the G→A error during processive elongation) did not cause RNAP to advance 1 bp forward on the DNA (Figure 4D). This result indicates that the TEC19A remains in a 1 bp backtracked state after the misincorporation, which is consistent with the previous findings on the effect of misincorporation on backtracking (15,21). We concluded that the higher backtracking potential and thus better proofreading on the

170 G as opposed to the 474 G sequence is responsible for the relatively lower error rate detected by the RNA-seq (see Supplementary text and Supplementary Figure S3).

Next, we tested if backtracking on 474 G sequence affects the G→A misincorporation in the presence of Mn^{2+} as was indicated by the clustering analysis. To mimic the G→A misincorporation at 474 G site during processive elongation, we measured the rate of AMP-misincorporation in TEC18C in the presence of Mg^{2+} or Mn^{2+} (Figure 5A and B). As expected, Mn^{2+}-sensitive TEC18C misincorporated AMP more rapidly in the presence of Mn^{2+} than in the presence of Mg^{2+} (Figure 5C). Remarkably, we detected high level of an endonucleolytic RNA cleavage at 7-nt upstream of the 3′ RNA end in the presence of Mg^{2+}, and to a substantially less extent in the presence of Mn^{2+} (Figure 5C). This cleavage was consistent with the ExoIII footprinting data (Figure 4D), showing backtracking of this complex at 6 bp distance. We propose that backtracking after the misincorporation generated a substrate for the cleavage in this complex with or without GreA/B. At the longest 6–24 min incubation time with noncognate ATP, Mn^{2+} also appeared to decrease extension of the 3′ error (19 A* product) with the next cognate substrate (21 A* and 22 A** products) (Figure 5C and E). One would expect these opposite effects of Mn^{2+} on error correction and extension to counteract one another leading to a net zero impact of Mn^{2+} to fidelity. However, at the shorter 10–90 s, where an impact of the slow intrinsic RNA cleavage was negligible, Mn^{2+} stimulated rather than inhibited the error extension (Figure 5C and E). This effect was confirmed in the experiment with the TEC

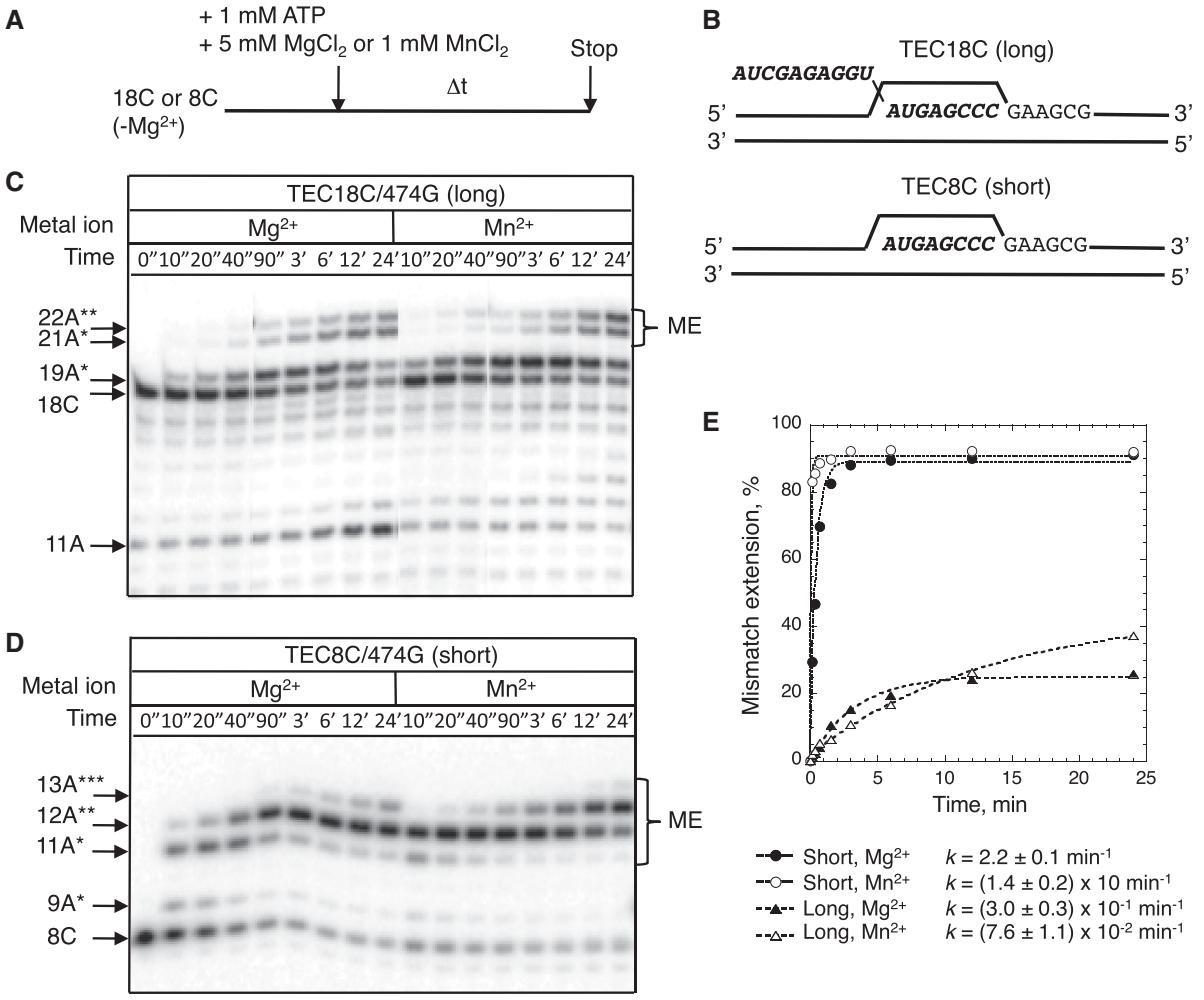

Figure 5. Effects of backtracking on the efficiencies of mismatch extension (ME) and intrinsic transcript cleavage, and their dependences on Mn^{2+}. (**A**) Reaction scheme for AMP misincorporation followed by ME. (**B**) RNA and downstream nontemplate DNA sequences in the TECs with long (18 nt) and short (8 nt) transcripts used in the assay. (**C**) Incubations of TEC18C/474 G with the noncognate ATP in the presence of Mg^{2+} or Mn^{2+}. Arrows indicate the original 18-nt RNA, misincorporation (marked by asterisks) and ME. (**D**) 5′ RNA shortening to 8-nt length in TEC18C (making TEC8C) increases ME. (**E**) Quantification of the ME (% of the total fraction in each detection time) from the panels C and D. The curves represent the single-exponential fit of the data; apparent rate constants (k) are shown. Note for lager k in 'Long, Mg^{2+}' compared with 'Long, Mn^{2+}' condition: This difference is due to the intrinsic transcript cleavage of 19 A* product of misincorporation, which occurs substantially faster in Mg^{2+} compared with Mn^{2+}. The faster cleavage in Mg^{2+} leads to apparent earlier than expected saturation of the ME reaction under these conditions. Although the plotting of ME appeared to follow single exponential kinetics, they result from a superposition of 3 different processes of 19 A* misincorporation, 19 A* cleavage and 19 A* extension with the next cognate NMP.

disregarding the cleavage activity described below. Thus, Mn^{2+} appeared to decrease transcription fidelity on 474 G sequence by suppressing intrinsic RNA proofreading activity in the backtracked complex and by promoting extension of the error with the next cognate NMP, which allowed the error propagation into a full-length RNA.

To prove that backtracking was the major error correction mechanism at the 474 G site, we assembled a version of TEC18C/474 G but reducing the RNA length from 18 to 8 nt by removing 10 nt from the 5′ end (TEC8C, Figure 5B). The shortening of the nascent RNA has been shown to prevent backtracking (44). Interestingly, elimination of backtracking dramatically enhanced G→A error and extension of the error with the next cognate AMP at this sequence (Figure 5C and D). The efficiency and the rate of mismatch extension increased 4- and >10-fold, respectively, in TEC8C compared with the original TEC18C in the presence of Mg^{2+} or Mn^{2+} (Figure 5E). Thus, backtracking followed by error correction by an intrinsic RNA cleavage represents a major mechanism for control of the G→A error at the 474 G site during processive transcription. The slow rate of the intrinsic cleavage at the 474 G site indicated that a substantial fraction of the 3′RNA misincorporation are not able to propagate to the full-length transcript due to backtrack pausing of RNAP after misincorporation. Applying the RNA-seq to nascent transcripts, isolated from backtracked elongation complexes of RNAP containing a 3′ error, by the previously established NET-seq method warrants addressing the effect of mismatch extension on the detected error rates (61).

Error rate depends on the nucleotide at the 3′ end of the transcript

We noted that not all short A/T tracts followed by a 3′ guanine residue identified by the RNA-seq exhibit low frequency of G→A errors, suggesting that high propensity for backtracking may not be the only parameter to control transcription fidelity (12). In search of another fidelity parameter embedded into sequence context, we aligned the sequences surrounding the G→A sites composed of the top 10% of the either lowest or highest error rate group, each of which was displayed by sequence logo (62,63). This analysis revealed a strong preference for adenine in n − 1 position for the low error rate sites and cytosine in the same position for high error rate sites when n is a position for the error (Figure 6A). This analysis also revealed sequence preferences for n−1 and n−2 DNA positions for the other types of transition errors (Supplementary Figure S4).

Next, we tested if a residue in the DNA or the RNA in n − 1 position affects RNAP misincorporation rate.

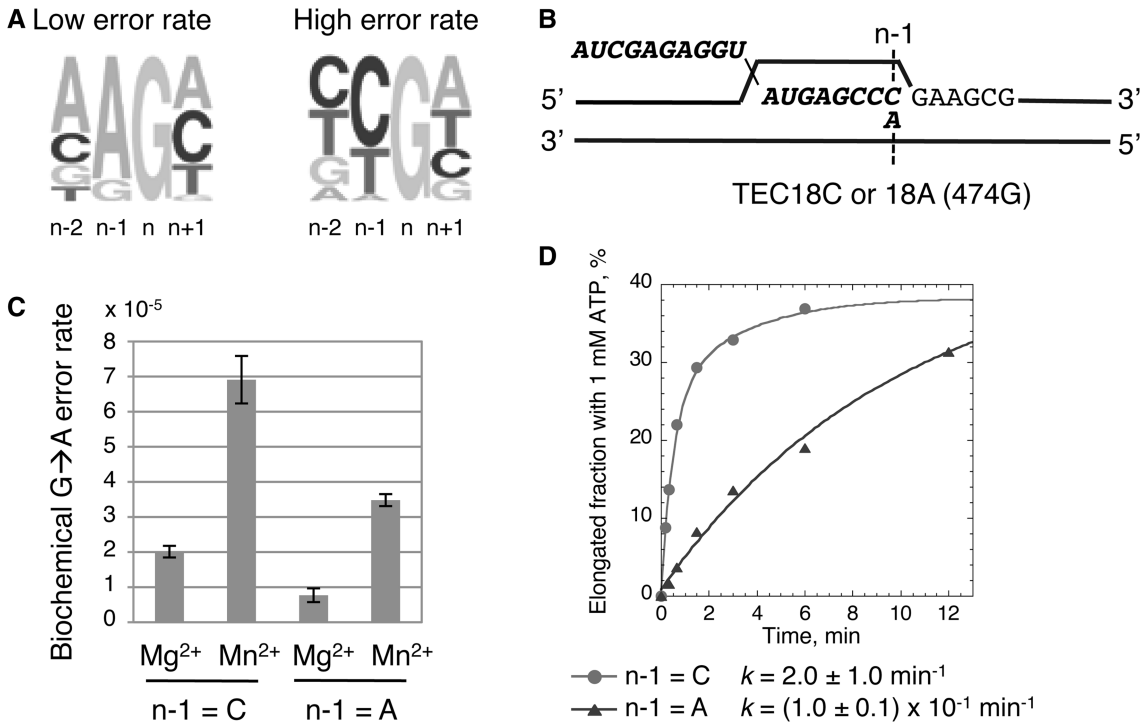

Figure 6. A 3′ residue in the nascent transcript determines the G→A error rate. (**A**) DNA logo derived from a sequence alignment around the dG residues coding for the low or high G→A error rate. Top lowest 10% (left) and top highest 10% (right) of all G→A error rates ($<1 \times 10^{-3}$) averaged by five different RNA preparations are used for the analysis. The residue frequencies from $n-2$ to $n+1$ (G→A error occurs at n site) were plotted with WebLogo (63). Y-axis is not shown as typical log base 2, but it represents the actual number to depict the residue types. (**B**) DNA/RNA scaffold for testing the effect of dC→dA substitution in the $n-1$ site of DNA. TEC18C ($n-1$ = C) and TEC18A ($n-1$ = A) on the 474 G sequence are shown. (**C**) Biochemical G→A error rates in TEC18C or 18 A as determined by NTP competition assay (see text for more details) (64,65). (**D**) Time course of AMP misincorporation for GMP in TEC18C or TEC18A. The curves represent the double exponential (TEC18C) or single exponential (TEC18A) fit of the data; apparent rate constants (k) are shown. The slower misincorporation rate obtained from the double-exponential fitting curve for TEC18C data was related to the intrinsic cleavage of 3′ RNA in this complex.

We used a previously developed NTP competition assay monitoring a single NMP misincorporation in the presence of a mixture of a cognate and noncognate NTP (64,65). We compared the G→A error rates in TEC18C/474 G and TEC18A/474 G carrying C18 (n − 1)→A substitution in DNA (Figure 6B, C and Supplementary Figure S5 A–C). As predicted by the sequence logo (Figure 6A), the C18 ($n-1$)→A substitution decreased the biochemical G19→A error rate in both Mg^{2+} and Mn^{2+} on 474 G sequence (Figure 6C) without affecting RNAP backtracking and intrinsic transcript cleavage (Supplementary Figure S5 D and E). The $n-1$ mutation also caused a 10-fold reduction of the rate of G19→A transcription error in a single AMP misincorporation assay lacking the cognate GTP (Figure 6D) Interestingly, the $n-1$ mutation stimulated AMP misincorporation without a strong effect on the cognate GMP incorporation (data not shown). This difference suggests that a chemical nature of the 3′ RNA–DNA base pair plays a major role in binding or addition of a noncognate NTP with only minor contribution to the same processes with a cognate substrate.

DISCUSSION

Our work provides the first evidence that RNA-seq can assess physiological transcription error rates even in the presence of artifact errors. We directly detected changes in the transition-error rates in the range of 4×10^{-5} to 3×10^{-4} b^{-1} (Figures 2 and 3). These limits identify a lower baseline of the standard transition-error rates at 10^{-5} order or less. Our findings that GreA/B increases the fidelity of processive transcription *in vitro* to a level of the fidelity *in vivo* (Figures 2 and 3) provide an ample opportunity for application of the RNA-seq for evaluation of transition-type transcription errors in *E. coli* cells harboring viable *greA*/*greB* deletions, mutations in RNAP subunit that reduce fidelity *in vitro* (19,66) and in the wild-type cells under different growth conditions including biological stresses/DNA damages (25,67). Although we did not detect any obvious hotspots (29) for the RNAP errors within the tested sequence, our results do not exclude that these hotspots exist genome-wide. The transversion errors by RNAP seem to occur with lower rates than transitions hindering their detection by the RNA-seq (Supplementary Figure S6). The transversions appear to favor conversion to thymine or adenine rather than to cytosine or guanine (Supplementary Figure S6), suggesting that RNAP has preferences for transversion errors that are similar to RT and/or DNAP.

Although transcription fidelity has been extensively studied by a single NMP incorporation in elongation complexes deprived of NTP substrates, the mechanism of fidelity control during processive transcription awaits development of an appropriate methodology. Our new approach that combines RNA-seq and biochemical analyses of transcription errors propagating to the full-length RNA revealed two sequence-specific mechanisms used during processive transcription under physiological NTPs concentration: (i) NTP selection related to the chemical nature of DNA–RNA base pair immediately upstream from the error site, and (ii) postincorporation error correction by the intrinsic transcript cleavage in the backtracked RNAP (Figure 7). A recent biochemical study suggested that a noncognate NTP is rejected from RNAP by formation of a stressed sugar–phosphate backbone in the template DNA strand, which involves

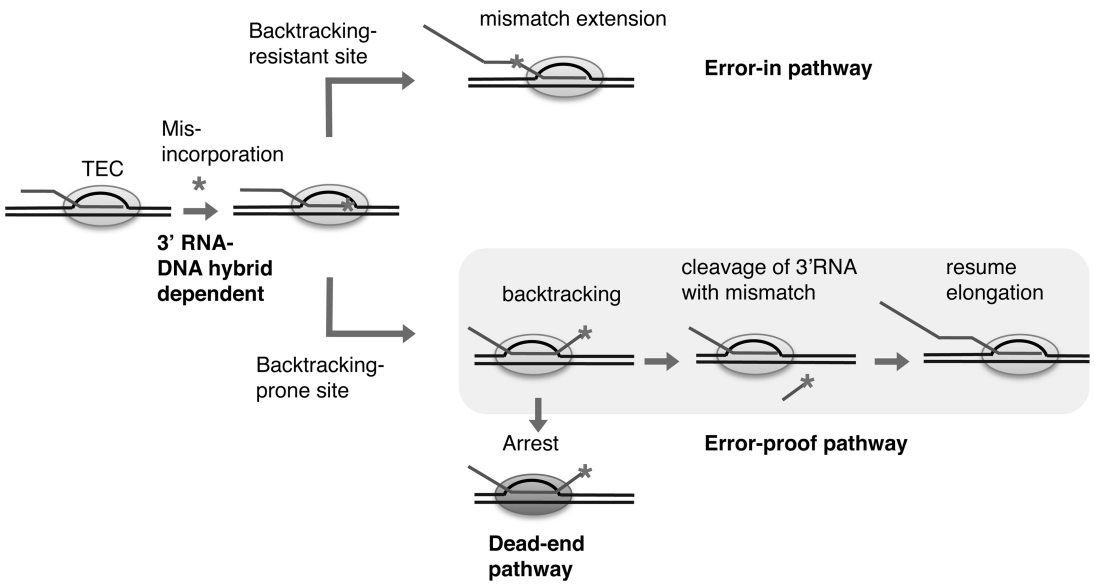

Figure 7. Multiple pathways for control of RNAP fidelity. Transcription error rate is determined by the 3′ RNA–DNA base pair in TEC (preincorporation substrate selection) and by backtracking propensity of RNAP (postincorporation proofreading). The 3′ RNA–DNA base pair controls misincorporation rate of a noncognate substrate (indicated by an asterisk). The DNA sequences such as A/T-rich tracts and protein factors that promote backtracking increase fidelity by decreasing extension of the 3′ RNA error with the next cognate NMP (shaded). The error is corrected by the intrinsic or Gre-assisted transcript cleavage in backtacked TEC. The irreversible backtrack arrest of TEC carrying the 3′ RNA error may derive from the inefficient transcript cleavage in the backtracked complex (the dead-end pathway).

angling of a 3′ RNA–DNA base pair to align the NTP for catalysis (16). The authors argued that the angling may weaken stacking of the 3′ base pair and ribose contacts with a noncognate NTP to induce its preferential rejection from the active center without a significant effect on the properly paired cognate NTP. We speculate that 3′ rA-dT and 3′ rC-dG base pairs could have an unequal stacking potential to differently affect the noncognate NTP rejection. Further analysis is warranted for generalizing this sequence-specific mechanism for the NTP selection. When the preincorporation NTP selection fails, the enhanced backtracking that interferes with an extension of the 3′ RNA mismatch provides an additional time to proofread the error by RNA cleavage. In this scenario, *trans*-acting factors that promote backtracking like DNA-bound proteins or nucleosomes may increase fidelity, whereas factors that interfere with backtracking like trailing RNAP, ribosomes (in prokaryotes) or secondary structure in the nascent RNA may decrease fidelity.

In eukaryotic transcription, Nesser et al. previously analyzed transcription errors throughout ~450-bp cDNA of *CAN1* transcript in yeast by Sanger sequencing (9). Their work reported a much higher 1.3×10^{-3}/bp rate of substitutions compared to the rate observed for *E. coli* RNAP in our study and the rates determined for yeast RNAP II *in vitro* (17). The authors claimed transcription rather than an artificial origin of these errors by showing that the rate was increased to 1.7×10^{-3}/bp in the mutant cell lacking Rpb9 subunit of RNAP II. Rpb9 is linked to transcription fidelity based on the results of *in vitro* misincorporation assays (49). Surprisingly, deletion of yeast *DST1* gene coding for RNA proofreading factor TFIIS (GreA/B analog) had almost no effect on the error rate *in vivo* (9). Once again, this result was different from our observation of a major impact of GreA/B proteins on transcription fidelity in *E. coli*. A source of these differences requires additional investigation.

The future application of the RNA-seq will allow monitoring genome-wide transcription fidelity under different growth conditions and in different cell types. Would 10^5 read depth (instead of 10^6 read depth used here) that covers 10^{-5} b^{-1} frequency be sufficient for detection of changes in transition-type transcription errors? This depth reduction could allow determination of fidelity against $\sim 10^5$ bases of transcriptome by a single Illumina sequencing analysis and lead to a significant cost reduction. We positively answer this question by showing that, for a voluntarily chosen position 578 of the *rpoBC* transcript, G→A transition-error rate was not significantly affected by a 10-fold decrease of the depth to 2×10^5 (Supplementary Figure S7). At this reduced depth, we successfully detected the responses of the error rates to Mn^{2+} and GreA/B *in vitro*. Thus, the 10^5 read depth appears to be sufficient to assess increases in transition errors from the basal level across an entire transcriptome with caution that the sensitivity to transcription error is varied by mRNA levels among different genes. Another potential issue for the genome-wide RNA-seq is the additional PCR cycles and barcoding bias (68), accompanied by the adapter ligation to cDNA during the library preparation. Our cDNA preparation included 11 cycles of PCR for the *in vitro* transcription samples and 16 cycles for the *in vivo* sample. However, we found that the *in vivo* sample had significantly lower G→A and T(U)→C errors than the standard (Mg^{2+}) *in vitro* sample, indicating no significant contribution of the PCR artifacts to the types of transcription errors detected in this work (Figure 3 and Supplementary Figure S2). This suggests that such an increase in the PCR cycles appeared not to dilute transcription errors beyond the detection limit. Our data also clearly indicate that the most significant technical improvement allowing reduction of the artifact transition errors in the Illumina platform is a suppression of the errors during the second read of the paired-end sequencing, which typically occur with $>3 \times 10^{-4}$ b^{-1} frequency (Figure 1D). The tag-based method (25–27) and overlapping read pairs method (30) have a strong potential in identification of the sequencing errors. Additional improvement of the RNA-seq bioinformatics has the potential to discriminate between transcription frame-shift errors (not analyzed in this work) and insertion/deletions artifacts associated with the Illumina platform and reads mapping. This approach may enable detection of physiologically relevant transcription slippage at short homopolymeric tracts and dinucleotide repeats broadly present in transcribed genes and contributing to slippage-associated diseases in humans (23,69).

It is worth mentioning that transcription errors at 10^{-5} error rate may have a deleterious effect on genome stability by inducing a prolonged stalling of RNAP at multiple sites across $>10^5$ bp transcribed region in a genome. The irreversibly arrested TEC should block DNA replication and subsequent rounds of transcription leading to double-strand DNA breaks and cessations of gene expression (64,70). The prolonged RNAP stalling is exemplified by almost irreversible loss of RNAP catalytic activity after an NMP misincorporation to TEC18, which was not accompanied by dissociation of RNAP from DNA (Figure 5). This mechanism is different from the previously proposed production of toxic proteins due to transcription errors, which requires an assumption that error rate of transcription is comparable with that of translation. Transcription misreading may have an impact on cell physiology comparable with translation misreading owing to multi-round translation of an erroneous mRNA molecule.

ACCESSION NUMBERS

GEO number: GSE46479.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We thank D.L. Court of NCI for discussions and critical reading of the manuscript; W. Tang, T.D. Schneider and Y. Zhao of NCI and T. Sakamoto of NAIST for technical comments and support in sequence analysis; S.H. Hughes,

M.L. Kireeva, S. Kakar, M. Bubunenko and J.N. Strathern of NCI for discussions and comments on the manuscript; W. Shao of NCI for the Perl script. We also thank NCI sequencing facility for Illumina sequencing and early bioinfomaic analysis.

FUNDING

Fellowship from JSPS (to M.I. in part); Intramural Research Program of the National Institutes of Health, National Cancer Institute, Center for Cancer Research (to M.K). Funding for open access charge: The Intramural Research Program of the National Institutes of Health, National Cancer Institute, Center for Cancer Research.

Conflict of interest statement. None declared.

REFERENCES

1. Strathern,J.N., Jin,D.J., Court,D.L. and Kashlev,M. (2012) Isolation and characterization of transcription fidelity mutants. *Biochim. Biophys. Acta*, **1819**, 694–699.
2. Gordon,A.J., Halliday,J.A., Blankschien,M.D., Burns,P.A., Yatagai,F. and Herman,C. (2009) Transcriptional infidelity promotes heritable phenotypic change in a bistable gene network. *PLoS Biol.*, **7**, e44.
3. Goldsmith,M. and Tawfik,D.S. (2009) Potential role of phenotypic mutations in the evolution of protein expression and stability. *Proc. Natl Acad. Sci. USA*, **106**, 6197–6202.
4. Paoloni-Giacobino,A., Rossier,C., Papasavvas,M.P. and Antonarakis,S.E. (2001) Frequency of replication/transcription errors in (A)/(T) runs of human genes. *Hum. Genet.*, **109**, 40–47.
5. Rodin,S.N., Rodin,A.S., Juhasz,A. and Holmquist,G.P. (2002) Cancerous hyper-mutagenesis in p53 genes is possibly associated with transcriptional bypass of DNA lesions. *Mutat. Res.*, **510**, 153–168.
6. Hubbard,K., Catalano,J., Puri,R.K. and Gnatt,A. (2008) Knockdown of TFIIS by RNA silencing inhibits cancer cell proliferation and induces apoptosis. *BMC Cancer*, **8**, 133.
7. Blank,A., Gallant,J.A., Burgess,R.R. and Loeb,L.A. (1986) An RNA polymerase mutant with reduced accuracy of chain elongation. *Biochemistry*, **25**, 5920–5928.
8. Taddei,F., Hayakawa,H., Bouton,M., Cirinesi,A., Matic,I., Sekiguchi,M. and Radman,M. (1997) Counteraction by MutT protein of transcriptional errors caused by oxidative damage. *Science*, **278**, 128–130.
9. Nesser,N.K., Peterson,D.O. and Hawley,D.K. (2006) RNA polymerase II subunit Rpb9 is important for transcriptional fidelity *in vivo*. *Proc. Natl Acad. Sci. USA*, **103**, 3268–3273.
10. Shaw,R.J., Bonawitz,N.D. and Reines,D. (2002) Use of an *in vivo* reporter assay to test for transcriptional and translational fidelity in yeast. *J. Biol. Chem.*, **277**, 24420–24426.
11. Larson,M.H., Zhou,J., Kaplan,C.D., Palangat,M., Kornberg,R.D., Landick,R. and Block,S.M. (2012) Trigger loop dynamics mediate the balance between the transcriptional fidelity and speed of RNA polymerase II. *Proc. Natl Acad. Sci. USA*, **109**, 6555–6560.
12. Sydow,J.F., Brueckner,F., Cheung,A.C., Damsma,G.E., Dengl,S., Lehmann,E., Vassylyev,D. and Cramer,P. (2009) Structural basis of transcription: mismatch-specific fidelity mechanisms and paused RNA polymerase II with frayed RNA. *Mol. Cell*, **34**, 710–721.
13. Kashkina,E., Anikin,M., Brueckner,F., Pomerantz,R.T., McAllister,W.T., Cramer,P. and Temiakov,D. (2006) Template misalignment in multisubunit RNA polymerases and transcription fidelity. *Mol. Cell*, **24**, 257–266.
14. Rosenberger,R.F. and Hilton,J. (1983) The frequency of transcriptional and translational errors at nonsense codons in the lacZ gene of *Escherichia coli*. *Mol. Gen. Genet.*, **191**, 207–212.
15. Erie,D.A., Hajiseyedjavadi,O., Young,M.C. and von Hippel,P.H. (1993) Multiple RNA polymerase conformations and GreA: control of the fidelity of transcription. *Science*, **262**, 867–873.
16. Sosunova,E., Sosunov,V., Epshtein,V., Nikiforov,V. and Mustaev,A. (2013) Control of transcriptional fidelity by active center tuning as derived from RNA polymerase endonuclease reaction. *J. Biol. Chem.*, **288**, 6688–6703.
17. Kireeva,M.L., Nedialkov,Y.A., Cremona,G.H., Purtov,Y.A., Lubkowska,L., Malagon,F., Burton,Z.F., Strathern,J.N. and Kashlev,M. (2008) Transient reversal of RNA polymerase II active site closing controls fidelity of transcription elongation. *Mol. Cell*, **30**, 557–566.
18. Roghanian,M., Yuzenkova,Y. and Zenkin,N. (2011) Controlled interplay between trigger loop and Gre factor in the RNA polymerase active centre. *Nucleic Acids Res.*, **39**, 4352–4359.
19. Holmes,S.F., Santangelo,T.J., Cunningham,C.K., Roberts,J.W. and Erie,D.A. (2006) Kinetic investigation of *Escherichia coli* RNA polymerase mutants that influence nucleotide discrimination and transcription fidelity. *J. Biol. Chem.*, **281**, 18677–18683.
20. Bar-Nahum,G., Epshtein,V., Ruckenstein,A.E., Rafikov,R., Mustaev,A. and Nudler,E. (2005) A ratchet mechanism of transcription elongation and its control. *Cell.*, **120**, 183–193.
21. Shaevitz,J.W., Abbondanzieri,E.A., Landick,R. and Block,S.M. (2003) Backtracking by single RNA polymerase molecules observed at near-base-pair resolution. *Nature*, **426**, 684–687.
22. Wagner,L.A., Weiss,R.B., Driscoll,R., Dunn,D.S. and Gesteland,R.F. (1990) Transcriptional slippage occurs during elongation at runs of adenine or thymine in *Escherichia coli*. *Nucleic Acids Res.*, **18**, 3529–3535.
23. Baranov,P.V., Hammer,A.W., Zhou,J., Gesteland,R.F. and Atkins,J.F. (2005) Transcriptional slippage in bacteria: distribution in sequenced genomes and utilization in IS element gene expression. *Genome Biol.*, **6**, R25.
24. Kircher,M., Heyn,P. and Kelso,J. (2011) Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics*, **12**, 382.
25. Schmitt,M.W., Kennedy,S.R., Salk,J.J., Fox,E.J., Hiatt,J.B. and Loeb,L.A. (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl Acad. Sci. USA*, **109**, 14508–14513.
26. Kinde,I., Wu,J., Papadopoulos,N., Kinzler,K.W. and Vogelstein,B. (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl Acad. Sci. USA*, **108**, 9530–9535.
27. Jabara,C.B., Jones,C.D., Roach,J., Anderson,J.A. and Swanstrom,R. (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc. Natl Acad. Sci. USA*, **108**, 20166–20171.
28. Ji,J.P. and Loeb,L.A. (1992) Fidelity of HIV-1 reverse transcriptase copying RNA *in vitro*. *Biochemistry*, **31**, 954–958.
29. Hu,W.S. and Hughes,S.H. (2012) HIV-1 reverse transcription. *Cold Spring Harb. Perspect. Med.*, **2**, a006882.
30. Chen-Harris,H., Borucki,M.K., Torres,C., Slezak,T.R. and Allen,J.E. (2013) Ultra–deep mutant spectrum profiling: improving sequencing accuracy using overlapping read pairs. *BMC Genomics*, **14**, 96.
31. Ramaswami,G., Zhang,R., Piskol,R., Keegan,L.P., Deng,P., O'Connell,M.A. and Li,J.B. (2013) Identifying RNA editing sites using RNA sequencing data alone. *Nat. Methods*, **10**, 128–132.
32. Li,M., Wang,I.X., Li,Y., Bruzel,A., Richards,A.L., Toung,J.M. and Cheung,V.G. (2011) Widespread RNA and DNA sequence differences in the human transcriptome. *Science*, **333**, 53–58.
33. Lin,W., Piskol,R., Tan,M.H. and Li,J.B. (2012) Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*, **335**, 1302; author reply 1302.
34. Kleinman,C.L. and Majewski,J. (2012) Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*, **335**, 1302; author reply 1302.
35. Pickrell,J.K., Gilad,Y. and Pritchard,J.K. (2012) Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*, **335**, 1302; author reply 1302.
36. Abram,M.E., Ferris,A.L., Shao,W., Alvord,W.G. and Hughes,S.H. (2010) Nature, position, and frequency of mutations

made in a single cycle of HIV-1 replication. *J. Virol.*, **84**, 9864–9878.
37. Kashlev,M., Nudler,E., Severinov,K., Borukhov,S., Komissarova,N. and Goldfarb,A. (1996) Histidine-tagged RNA polymerase of *Escherichia coli* and transcription in solid phase. *Methods Enzymol.*, **274**, 326–334.
38. Borukhov,S. and Goldfarb,A. (1996) Purification and assay of *Escherichia coli* transcript cleavage factors GreA and GreB. *Methods Enzymol.*, **274**, 315–326.
39. Kashlev,M.V., Bass,I.A., Lebedev,A.N., Kaliaeva,E.S. and Nikiforov,V.G. (1989) [Deletion-insertion mapping of the region non-essential for functioning of the beta-subunit of *Escherichia coli* RNA polymerase]. *Genetika*, **25**, 396–405.
40. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
41. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
42. Kireeva,M.L., Lubkowska,L., Komissarova,N. and Kashlev,M. (2003) Assays and affinity purification of biotinylated and nonbiotinylated forms of double-tagged core RNA polymerase II from Saccharomyces cerevisiae. *Methods Enzymol.*, **370**, 138–155.
43. Komissarova,N., Kireeva,M.L., Becker,J., Sidorenkov,I. and Kashlev,M. (2003) Engineering of elongation complexes of bacterial and yeast RNA polymerases. *Methods Enzymol.*, **371**, 233–251.
44. Imashimizu,M., Kireeva,M.L., Lubkowska,L., Gotte,D., Parks,A.R., Strathern,J.N. and Kashlev,M. (2013) Intrinsic Translocation Barrier as an Initial Step in Pausing by RNA Polymerase II. *J. Mol. Biol.*, **425**, 697–712.
45. Sousa,R. (1996) Structural and mechanistic relationships between nucleic acid polymerases. *Trends Biochem. Sci.*, **21**, 186–190.
46. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
47. Grant,G.R., Farkas,M.H., Pizarro,A.D., Lahens,N.F., Schug,J., Brunk,B.P., Stoeckert,C.J., Hogenesch,J.B. and Pierce,E.A. (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, **27**, 2518–2528.
48. Niyogi,S.K. and Feldman,R.P. (1981) Effect of several metal ions on misincorporation during transcription. *Nucleic Acids Res.*, **9**, 2615–2627.
49. Walmacq,C., Kireeva,M.L., Irvin,J., Nedialkov,Y., Lubkowska,L., Malagon,F., Strathern,J.N. and Kashlev,M. (2009) Rpb9 subunit controls transcription fidelity by delaying NTP sequestration in RNA polymerase II. *J. Biol. Chem.*, **284**, 19601–19612.
50. Imashimizu,M., Tanaka,K. and Shimamoto,N. (2011) Comparative study of cyanobacterial and *E. coli* RNA polymerases: misincorporation, abortive transcription, and dependence on divalent cations. *Genet. Res. Int.*, 2011, 572689.
51. Borukhov,S., Sagitov,V. and Goldfarb,A. (1993) Transcript cleavage factors from *E. coli*. *Cell.*, **72**, 459–466.
52. Petersen,C. and Moller,L.B. (2000) Invariance of the nucleoside triphosphate pools of *Escherichia coli* with growth rate. *J. Biol. Chem.*, **275**, 3931–3935.
53. Orlova,M., Newlands,J., Das,A., Goldfarb,A. and Borukhov,S. (1995) Intrinsic transcript cleavage activity of RNA polymerase. *Proc. Natl Acad. Sci. USA*, **92**, 4596–4600.
54. Nakano,T., Ouchi,R., Kawazoe,J., Pack,S.P., Makino,K. and Ide,H. (2012) T7 RNA polymerases backed up by covalently trapped proteins catalyze highly error prone transcription. *J. Biol. Chem.*, **287**, 6562–6572.
55. Stepanova,E., Lee,J., Ozerova,M., Semenova,E., Datsenko,K., Wanner,B.L., Severinov,K. and Borukhov,S. (2007) Analysis of promoter targets for *Escherichia coli* transcription elongation factor GreA *in vivo* and *in vitro*. *J. Bacteriol.*, **189**, 8772–8785.
56. Drake,J.W. (1991) A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl Acad. Sci. USA*, **88**, 7160–7164.
57. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
58. Komissarova,N. and Kashlev,M. (1998) Functional topography of nascent RNA in elongation intermediates of RNA polymerase. *Proc. Natl Acad. Sci. USA*, **95**, 14699–14704.
59. Komissarova,N., Becker,J., Solter,S., Kireeva,M. and Kashlev,M. (2002) Shortening of RNA:DNA hybrid in the elongation complex of RNA polymerase is a prerequisite for transcription termination. *Mol. Cell*, **10**, 1151–1162.
60. Artsimovitch,I., Chu,C., Lynch,A.S. and Landick,R. (2003) A new class of bacterial RNA polymerase inhibitor affects nucleotide addition. *Science*, **302**, 650–654.
61. Churchman,L.S. and Weissman,J.S. (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, **469**, 368–373.
62. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
63. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
64. Walmacq,C., Cheung,A.C., Kireeva,M.L., Lubkowska,L., Ye,C., Gotte,D., Strathern,J.N., Carell,T., Cramer,P. and Kashlev,M. (2012) Mechanism of translesion transcription by RNA polymerase II and its role in cellular resistance to DNA damage. *Mol Cell*, **46**, 1–12.
65. Kireeva,M.L., Opron,K., Seibold,S.A., Domecq,C., Cukier,R.I., Coulombe,B., Kashlev,M. and Burton,Z.F. (2012) Molecular dynamics and mutational analysis of the catalytic and translocation cycle of RNA polymerase. *BMC Biophys.*, **5**, 11.
66. Nedialkov,Y.A., Opron,K., Assaf,F., Artsimovitch,I., Kireeva,M.L., Kashlev,M., Cukier,R.I., Nudler,E. and Burton,Z.F. (2013) The RNA polymerase bridge helix YFI motif in catalysis, fidelity and translocation. *Biochim. Biophys. Acta*, **1829**, 187–198.
67. Saxowsky,T.T., Meadows,K.L., Klungland,A. and Doetsch,P.W. (2008) 8-Oxoguanine-mediated transcriptional mutagenesis causes Ras activation in mammalian cells. *Proc. Natl Acad. Sci. USA*, **105**, 18877–18882.
68. Alon,S., Vigneault,F., Eminaga,S., Christodoulou,D.C., Seidman,J.G., Church,G.M. and Eisenberg,E. (2011) Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Res.*, **21**, 1506–1511.
69. Atkins,J.F. and Bjork,G.R. (2009) A gripping tale of ribosomal frameshifting: extragenic suppressors of frameshift mutations spotlight P-site realignment. *Microbiol. Mol. Biol. Rev.*, **73**, 178–210.
70. Dutta,D., Shatalin,K., Epshtein,V., Gottesman,M.E. and Nudler,E. (2011) Linking RNA polymerase backtracking to genome instability in *E. coli*. *Cell*, **146**, 533–543.