Accurate RNA consensus sequencing for high-fidelity detection of transcriptional mutagenesis-induced epimutations

Kate S. Reid-Bayliss[a] and Lawrence A. Loeb[a,b,1]

[a]Department of Pathology, University of Washington School of Medicine, Seattle, WA 98195; and [b]Department of Biochemistry, University of Washington School of Medicine, Seattle, WA 98195

Edited by Philip C. Hanawalt, Stanford University, Stanford, CA, and approved July 10, 2017 (received for review June 2, 2017)

Transcriptional mutagenesis (TM) due to misincorporation during RNA transcription can result in mutant RNAs, or epimutations, that generate proteins with altered properties. TM has long been hypothesized to play a role in aging, cancer, and viral and bacterial evolution. However, inadequate methodologies have limited progress in elucidating a causal association. We present a high-throughput, highly accurate RNA sequencing method to measure epimutations with single-molecule sensitivity. Accurate RNA consensus sequencing (ARC-seq) uniquely combines RNA barcoding and generation of multiple cDNA copies per RNA molecule to eliminate errors introduced during cDNA synthesis, PCR, and sequencing. The stringency of ARC-seq can be scaled to accommodate the quality of input RNAs. We apply ARC-seq to directly assess transcriptome-wide epimutations resulting from RNA polymerase mutants and oxidative stress.

transcriptional mutagenesis | epimutations | RNA mutations | molecular misreading | RNAseq

Infidelity during RNA transcription, termed transcriptional mutagenesis (TM), has long been hypothesized to contribute to aging (1) and age-associated diseases, including cancer (2, 3) and neurodegeneration (4, 5). RNA mutations resulting from TM, termed epimutations, have also been implicated in bacterial and viral evolution and resistance (6–8). Studies on RNA polymerases have revealed the fidelity of in vitro transcription by multiple RNA polymerases to be on the order of 10^{-5} epimutations per nucleotide (9–13). This rate can dramatically increase during transcription of damaged templates and certain sequence contexts, such as repetitive DNA (14). Additionally, in vivo assays have revealed that TM can result in phenotypic changes in nondividing (15) and dividing cells (16–19), with the potential for TM-induced phenotypic changes to be heritable (20, 21), indicating that a single mutant transcript has the potential to have profound effects on cellular function.

The bulk of the evidence for TM has been generated using in vitro fidelity assays and highly expressed reporter genes that encompass a small number of sequence contexts, are limited in the spectrum of mutations that can be monitored, and are subject to translational errors convoluting the results (9, 22). Consequently, the results of these studies cannot be easily extrapolated to understand the extent of epimutations in cells, where transcription factors, repair enzymes, chromatin, and gene expression levels modulate transcriptional fidelity. Thus, to elucidate the roles of TM-induced epimutations in physiology, disease, and evolution, it is necessary to study individual RNA molecules transcribed in vivo in a high-throughput manner.

De novo epimutations remain a challenging target for high-throughput RNA sequencing (RNAseq). While in vitro studies estimate RNA polymerase infidelity to be on the order of one in 100,000 epimutations per nucleotide, reverse transcriptase used to generate cDNA from RNA makes approximately one error per 10,000 bases (23). Additionally, Illumina sequencers misread approximately one in 1,000 bases (24). Recent methods, such as barcoding of RNAs (25) or cDNAs (26, 27), reduce the error frequency of RNAseq. However, such methods can be of low yield (25), rely heavily on complicated bioinformatics requiring calibration for each sample (26), and do not address errors introduced during reverse transcription (26, 27). Reverse transcriptase errors can be overcome by generating multiple cDNA copies from each RNA molecule (25, 28). However, these methods can be of low yield (25), may themselves introduce errors due to harsh reaction conditions (29), and are limited by sequence read length (28). To date, these advances have proven useful for sequencing viral RNA genomes, which are inherently more error-prone, but their background errors remain too high to reliably detect TM-induced epimutations in cells.

To address the limitations of RNAseq and enable the study of epimutations in any organism, we have developed a highly accurate sequencing method, termed accurate RNA consensus sequencing (ARC-seq), to measure epimutations with single-molecule sensitivity. ARC-seq uniquely couples the use of an adaptor to barcode each RNA molecule and the generation of multiple cDNA copies per RNA molecule before sequencing. This combination enables the removal of artifacts due to cDNA synthesis, PCR errors, and sequencing errors, revealing the epimutations resulting from TM in vivo.

Results

Development of a Highly Accurate Method to Detect Epimutations. Three obstacles to accurate RNAseq include the following: (*i*) RNA must first undergo the highly error-prone process of

Significance

Epimutations arising from transcriptional mutagenesis have been hypothesized to contribute to viral and bacterial evolution, drug resistance, and age-related diseases, including cancer and neurodegeneration. However, methodology limitations have inhibited progress toward elucidating the contributions of epimutations to cellular evolution and survival in vivo. Recent efforts to overcome these limitations remain constrained by artifacts arising during RNA library preparation. We present accurate RNA consensus sequencing (ARC-seq), an accurate, high-throughput RNA sequencing method that effectively eliminates errors introduced during RNA library preparation and sequencing and represents a major advance over previous methods. ARC-seq will enable investigations of the causal roles of transcriptional fidelity and epimutations in multiple fields, including viral evolution, bacterial resistance, and age-related diseases, such as cancer and neurodegeneration.

Author contributions: K.S.R.-B. designed research; K.S.R.-B. performed experiments; K.S.R.-B. and L.A.L. analyzed data; and K.S.R.-B. and L.A.L. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The raw sequence files reported in this paper have been deposited in the National Center for Biotechnology Information's Sequence Read Archive (BioProject accession no. PRJNA396053).

[1]To whom correspondence should be addressed. Email: laloeb@uw.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1709166114/-/DCSupplemental.

reverse transcription before sequencing, (*ii*) PCR amplification of cDNA can introduce errors, and (*iii*) high-throughput sequencing itself is highly error-prone. To overcome these obstacles, we developed ARC-seq, a cDNA library preparation protocol. We start by ligating barcoded RNA adaptors to the 5′-end of fragmented RNA molecules; this adaptor contains 16 random nucleotides that uniquely identify individual RNA molecules (Fig. 1*A*). Each barcoded RNA molecule is then circularized and reverse-transcribed via rolling-circle reverse transcription. This produces a cDNA multimer containing multiple cDNA copies of the original RNA molecule. After restricting the multimeric cDNA molecule into monomers, we uniquely index each cDNA copy of the original RNA. Each indexed cDNA is then amplified by high-fidelity PCR and sequenced on an Illumina HiSeq instrument. After sequencing, using bioinformatics, the cDNA indexes are used to generate a PCR consensus sequence, eliminating artifacts due to sequencing and PCR errors (Fig. 1*B*). Finally, the RNA barcode is used to generate a cDNA consensus sequence, eliminating reverse transcription and damage-induced artifacts; thus, we are able to regenerate the original RNA sequence.

The upper estimate of next-generation sequencing error is one in 100 nucleotides sequenced (30); thus, the theoretical background of ARC-seq approaches 0.01^n, where n is the number of cDNA copies produced from each RNA molecule. By increasing the length of the rolling-circle reverse transcription reaction, we can generate more cDNA copies per RNA molecule, thus increasing the stringency of the error correction or ARC-seq. This enables accurate sequencing of even highly damaged RNA molecules.

ARC-Seq Effectively Corrects Reverse Transcription, PCR, and Sequencing Artifacts. To validate the power of ARC-seq to eliminate artifacts due to reverse transcription, PCR, and sequencing errors, we synthesized three types of RNAs by in vitro transcription, using T7 RNA polymerase (Fig. S1): (*i*) high-fidelity RNA, generated using a pristine DNA template [expected epimutation frequency of 3×10^{-5} (12)]; (*ii*) damaged RNA, generated by treating the high-fidelity RNA with hydrogen peroxide (H_2O_2) [expected epimutation frequency is the same as the high-fidelity RNA (3×10^{-5}) because no new mutations are introduced]; and (*iii*) mutated RNA, which was generated from a DNA template oxidatively damaged with H_2O_2 to induce mistakes during transcription (expected to have an elevated epimutation frequency). These RNAs were then sequenced via ARC-seq. At a cDNA family size of one, which corresponds to RNAseq with tag-based error correction (e.g., ref. 27), the error frequency of the high-fidelity RNA is $\sim 2 \times 10^{-4}$, ~10-fold higher than the expected epimutation frequency (Fig. 1*C*); the error frequency of the damaged RNA template is elevated approximately threefold greater than the high-fidelity RNA, consistent with the high error rate of conventional RNAseq, especially on damaged RNA templates (23).

In contrast, by requiring five unique cDNA copies per RNA molecule, ARC-seq reveals the true epimutation frequency of the high-fidelity RNA to be $\sim 2 \times 10^{-5}$. Furthermore, by requiring six cDNA copies per RNA molecule to form a consensus sequence, and therefore increasing the stringency of its error correction, ARC-seq fully corrects for damage-induced artifacts and reveals the true epimutation frequency of the damaged RNA to be equivalent to the undamaged high-fidelity RNA. In contrast, even with a high stringency of eight cDNA copies per RNA molecule, the mutation frequency of the mutated RNA remains more than 10-fold greater than the high-fidelity RNA, consistent with ARC-seq eliminating errors without mistakenly removing true epimutations. Thus, by repeatedly sequencing the same RNA molecule, ARC-seq eliminates damage-induced and sequencing artifacts, revealing the TM-induced epimutations present in the original RNA molecule.

ARC-Seq Reveals the Frequency and Spectrum of Epimutations in Vivo. Several mutants of *Saccharomyces cerevisiae* (yeast) have been shown to have reduced in vitro RNA synthesis fidelity. *Rpb1 E1103G* is a point mutant of the catalytic domain of RNA polymerase II and confers dependence on transcription factor S-II (13). *ΔRpb9* is a deletion mutant of a transcription factor that enhances the fidelity of mRNA transcription in yeast (31). To establish ARC-seq's utility for measuring in vivo epimutations, we applied the

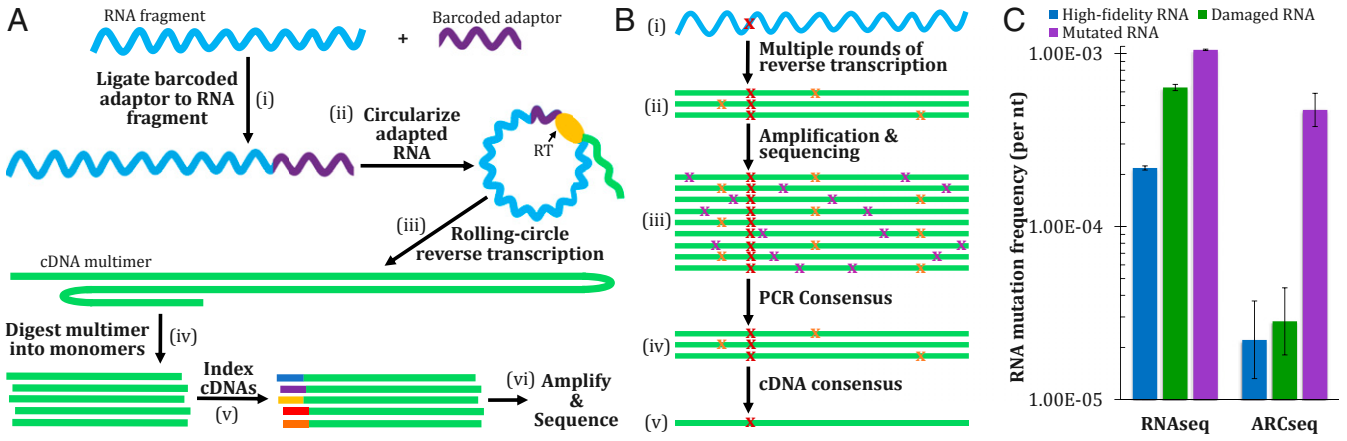

Fig. 1. (*A*) Overview of the ARC-seq method. (*i*) Each RNA is ligated to an adaptor containing a unique barcode. Ligated RNAs are then circularized (*ii*) and subjected to rolling-circle reverse transcription (*iii*), generating a multimeric cDNA from each RNA molecule. (*iv*) cDNA multimers are then restricted into monomers, which are cDNA copies of the original RNA molecule. Each cDNA is then tagged with a unique index (*v*), amplified (*vi*), and sequenced. (*B*) Error correction by ARC-seq. (*i*) Single RNA molecule containing a true epimutation (red); this molecule is barcoded. (*ii*) Rolling-circle reverse transcription generates multiple cDNA copies from each ligated RNA molecule, introducing random errors (orange). (*iii*) Amplification and sequencing amplify the existing errors and introduce new errors (purple), further obscuring the true epimutation. Artifacts present in standard RNAseq data are illustrated at this level. (*iv*) After sequencing, cDNA tags are bioinformatically matched and a consensus sequence is generated for each cDNA copy, eliminating many amplification and sequencing artifacts. (*v*) Finally, the RNA barcodes are matched, and a consensus sequence is generated from the cDNA copies, which regenerates the original RNA molecule's sequence, revealing the true epimutation. (*C*) ARC-seq eliminates damage-induced, reverse transcription, PCR, and sequencing artifacts, revealing true epimutations. High-fidelity (blue), damaged (green), and mutated (purple) RNAs were generated by in vitro transcription by T7 RNA polymerase and sequenced via ARC-seq. While conventional RNAseq has a high level of artifacts, with increased artifacts observed in the damaged RNA template, ARC-seq is able to fully correct damage-induced artifacts, revealing the true epimutation frequency to be $\sim 2 \times 10^{-5}$, without removing true epimutations. Error bars represent Wilson scores of 95% confidence.

method to study TM in these yeast mutants. When we analyze the epimutation frequencies obtained at increasing cDNA copy number per RNA molecule, we find that the mRNA and rRNA mutation frequencies of all three yeast strains plateau with just three cDNA copies per RNA molecule (Fig. 2*A*); the mRNA mutation frequency of stationary phase wild-type yeast is 4.21×10^{-5}, more than

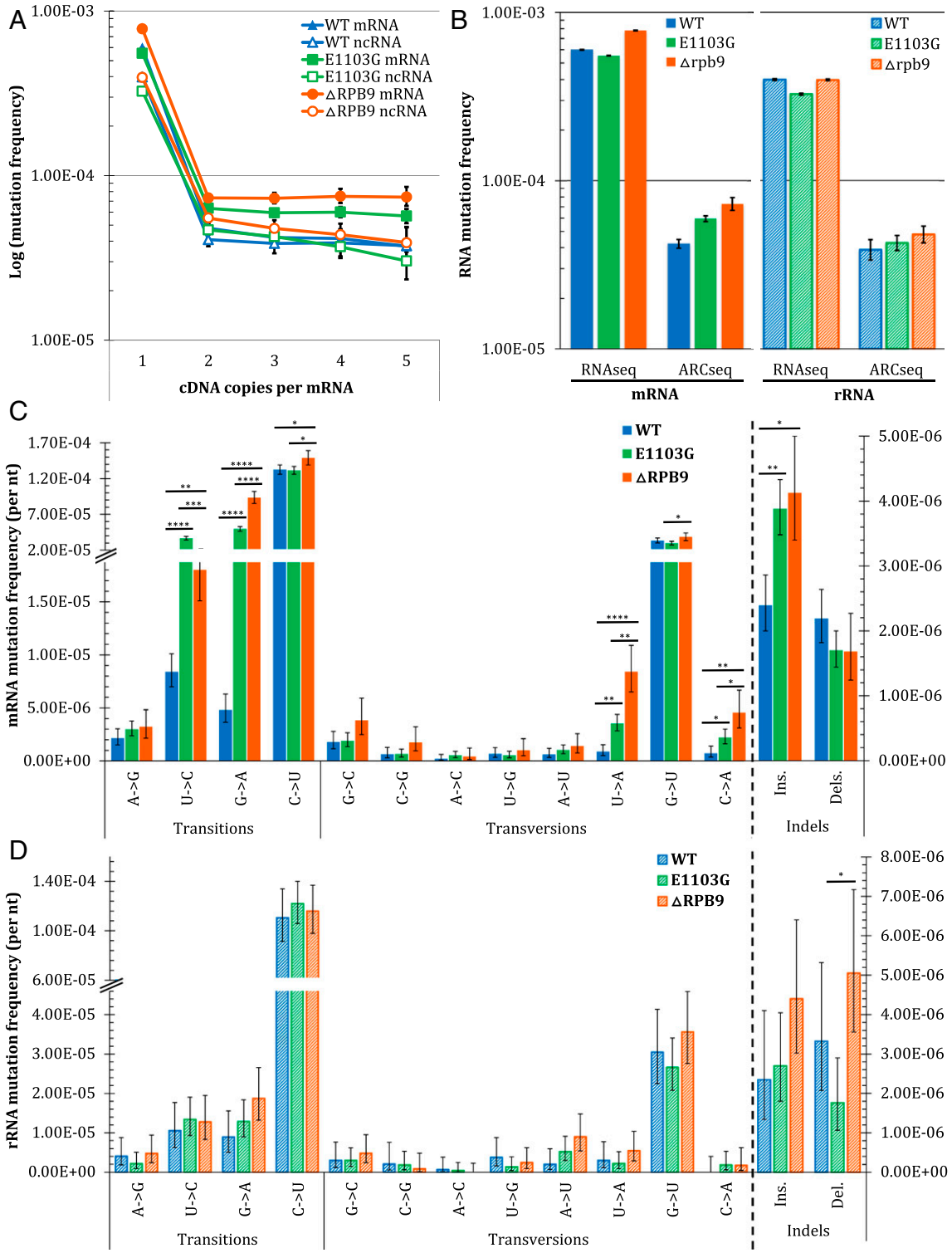

Fig. 2. ARC-seq reveals differences in epimutation frequencies and in the spectrum between yeast RNA polymerase mutants. RNAs from wild-type (WT), E1103G, and ΔRpb9 yeasts were sequenced via ARC-seq, with the number of cDNA copies per RNA molecule required to generate a consensus sequence varied from one through five. (*A*) Epimutation frequency stabilizes at three cDNAs per RNA molecule, revealing epimutation frequency differences between WT and the two mutants. (*B*) Comparison of epimutation frequencies observed with one cDNA copy per RNA molecule, corresponding to conventional RNAseq with tag-based error correction, and three cDNA copies per RNA molecule for mRNA (*Left*) and rRNA (*Right*). Differences between the mRNA (*C*) and rRNA (*D*) mutation spectrums of WT and mutant yeasts are shown. Error bars represent Wilson scores of 95% confidence. $*P < 0.01$, $**P < 10^{-5}$, $***P < 10^{-10}$, $****P < 10^{-15}$.

an order of magnitude lower than the error frequency obtained with conventional RNAseq (Fig. 2*B*). Additionally, both RNA polymerase mutants have mRNA mutation frequencies elevated over wild type: 5.94×10^{-5} ($P < 2.2 \times 10^{-16}$) and 7.28×10^{-5} ($P < 2.2 \times 10^{-16}$) for E1103G and ΔRpb9, respectively. In contrast, consistent with the yeast mutants having error-prone RNA polymerase II transcription, the frequency of mutations in both mutants' rRNAs, which are transcribed by RNA polymerases I and III, are not significantly different from the frequency of mutations in wild-type yeast. Furthermore, the mutation spectrums reveal differences between the types of mRNA mutations induced in the three yeast strains (Fig. 2*C* and Table S1). While C→U mutations are the most frequently observed epimutation in all three yeasts, both mutants show elevated frequencies of U→C, G→A, U→A, and C→A mutations, as well as single-base insertions, in their mRNAs relative to wild type. Additionally, E1103G has a greater elevation in U→C mutations than ΔRpb9 in its mRNA, whereas ΔRpb9 has greater elevations in G→A, U→A, and C→A mutations, relative to E1103G, in its mRNA. In contrast, in the rRNAs, no mutation subtype of either mutant differed significantly from wild type (Fig. 2*D* and Table S2), consistent with the defects of E1103G and ΔRpb9 being restricted to RNA polymerase II transcription.

Oxidative Stress Induces TM in Vivo. DNA damage due to oxidative stress is well known to induce DNA mutations, and in vitro studies of RNA polymerase activity at DNA lesions indicate that it behaves similar to DNA polymerases (14). Thus, to determine if oxidative stress induces elevated TM in vivo, we treated log-phase wild-type yeast with 50 μM H_2O_2 for 30 min, extracted the RNAs, and sequenced them via ARC-seq. Following oxidative stress, the mRNA mutation frequency increases from 5.6×10^{-5} to 1.3×10^{-4} (Fig. 3*A*). While oxidative stress induces elevations in multiple mutation subtypes, the most frequent changes observed are G→A and U→G substitutions, induced 80-fold, and C→A substitutions, induced 164-fold (Fig. 3*B* and Table S3). In rRNA, nearly all mutation subtypes increase following oxidative stress, with the most frequent change again being C→A substitutions (Fig. 3*C* and Table S4), induced 217-fold. The dramatic increases in C→A mutations are consistent with TM of the 8-oxodG lesion in template DNA, which is the most common form of oxidative DNA damage in cells (32–35). Additionally, the large increase in G→A mutations in mRNA is consistent with TM across from deaminated cytosines in the DNA template, a common consequence of oxidative stress in cells (34, 36).

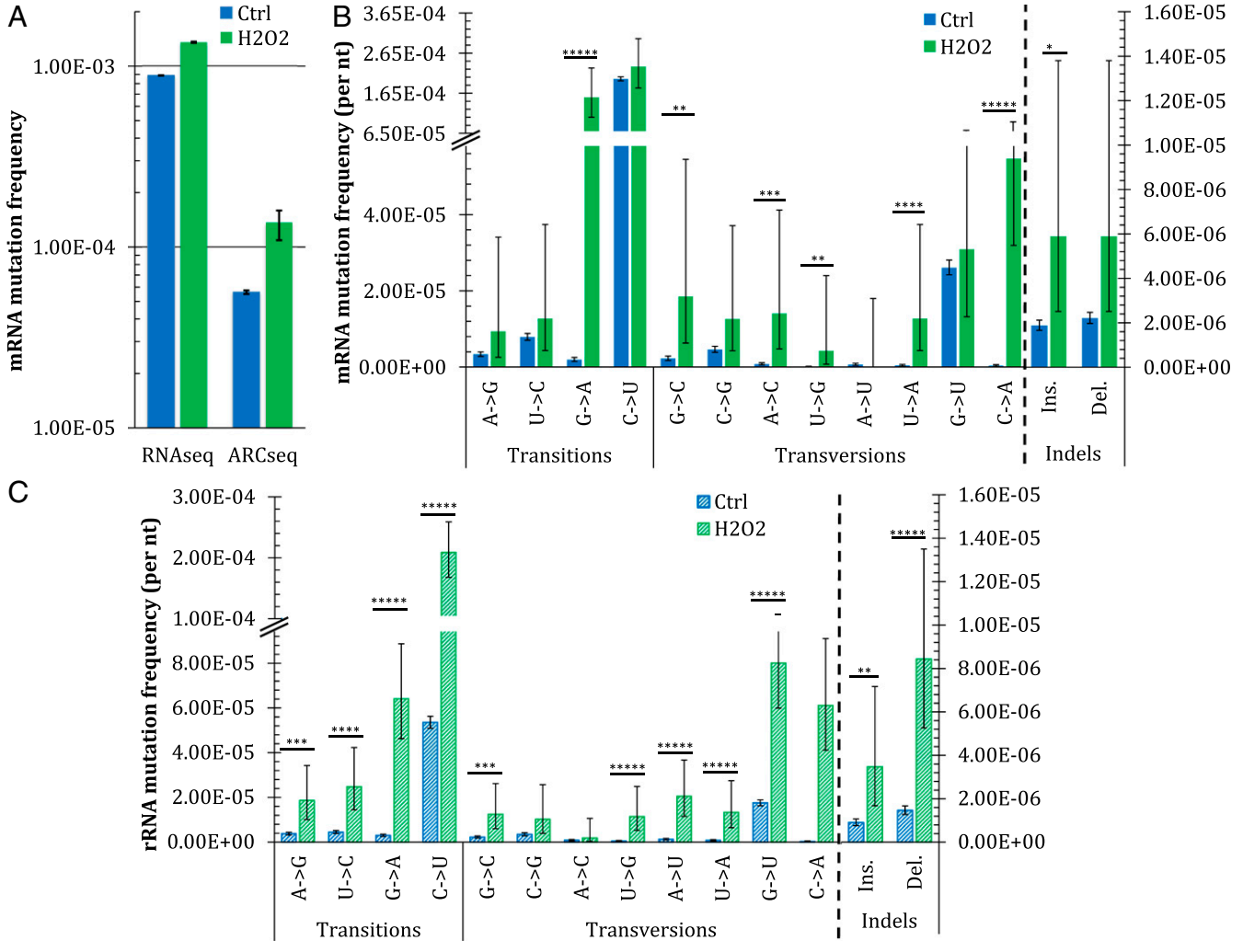

Fig. 3. ARC-seq reveals differences in TM after oxidative stress between yeast RNA fidelity mutants and between RNA types. Wild-type yeast was exposed to H_2O_2, and its RNAs were then sequenced via ARC-seq. (*A*) Comparison of mRNA mutation frequencies observed with one cDNA copy per RNA molecule, corresponding to conventional RNAseq with tag-based error correction, and with three cDNA copies per RNA molecule. (*B*) mRNA frequency and spectrum in untreated (ctrl) and 50 μM-treated (H_2O_2) yeasts. (*C*) Frequency and spectrum of rRNA in ctrl and 50 μM-treated (H_2O_2) yeasts. Error bars represent Wilson scores of 95% confidence. *$P < 0.05$, **$P < 10^{-2}$, ***$P < 10^{-5}$, ****$P < 10^{-10}$, *****$P < 10^{-15}$.

Discussion

TM is hypothesized to play roles in aging, cancer, neurodegeneration, viral evolution, and drug resistance (14, 37–39). However, little progress has been made in elucidating the contribution of TM to human health and disease, because methods for detecting epimutations in vivo have been limiting. The requirement to reverse-transcribe RNA before sequencing, as well as the high error rate of next-generation sequencing itself, constrains the accuracy of conventional RNAseq. Recent efforts to overcome these limitations have made progress toward more accurate RNA sequencing (25–29, 40, 41) but do not adequately remove the artifacts arising from these sources of error.

In developing ARC-seq, we reasoned that by generating multiple cDNA copies per RNA molecule, we could markedly reduce reverse transcription, PCR, and sequencing errors. We used a molecular barcode strategy to uniquely identify each RNA molecule before generating and sequencing multiple cDNA copies of each original RNA molecule. Furthermore, to distinguish between cDNA duplicates of a single RNA molecule and PCR duplicates of a single cDNA copy, and thereby eliminate PCR errors, we introduced an additional index sequence to each cDNA molecule. This unique combination enables the elimination of artifacts due to cDNA synthesis, PCR errors, and sequencing errors, revealing the sequence of the original RNA molecule.

Applying ARC-seq to sequencing in vitro-transcribed (IVT) RNAs demonstrates its unique ability to modulate the stringency of the method's error correction. This increased stringency selectively eliminates artifacts and enables even highly damaged RNAs to be sequenced accurately, which will likely prove crucial to many in vivo applications, such as analyses of biopsies or postmortem tissues, where RNAs may be partially degraded and highly damaged.

Applying ARC-seq to the study of RNAs generated in vivo by yeast RNA polymerase mutants demonstrates its ability to sensitively detect mutation frequency and spectrum differences. While several mutation types are elevated in the mRNA of both mutants, relative to wild type, there are no elevations in rRNA mutations in either mutant. These results not only confirm the specificity of the yeast mutants' defects in the fidelity of RNA polymerase II transcription but also serve as confirmation that ARC-seq accurately reveals epimutations. Importantly, we determined the frequency of each mutation type by the number of mutations observed over the total number of observations of the wild-type nucleotide; therefore, the differences observed are not due to differences in nucleotide distribution between the three strains. Of note, while not differing dramatically between the three yeast strains, C→U is the most frequent mutation observed; an unknown fraction of these mutations could be the result of deamination, either spontaneously or due to the action of cytosine deaminases on RNA rather than transcriptional infidelity. Further cell-based studies altering the expression of various cytosine deaminases could elucidate the extent to which TM versus RNA deamination contributes to C→U mutations in RNA.

Finally, applying ARC-seq to the study of the transcriptional mutagenic consequences of oxidative stress demonstrates its utility for addressing important biological questions. We see that oxidative stress induces high levels of epimutations not only in mRNA but also in rRNA. These results suggest that oxidative DNA damage, whether due to exogenous agents or endogenous perturbations, could have profound yet unappreciated consequences for cells. Of interest, the untreated rRNA mutation frequency of wild-type yeast was approximately twofold lower than its mRNA mutation frequency, largely due to decreased C→U mutations. Two possible explanations may account for this difference: (*i*) rDNA may be more readily repaired than protein-coding gene regions in the genome or (*ii*) the fidelity of rRNA synthesis is higher than the fidelity of mRNA synthesis. Given that rRNA is longer lived and involved in protein translation, either of these possibilities has merit. While a mutated mRNA may be translated multiple times, yielding a pool of mutant proteins, codon redundancy limits the impact of an individual mutation, and even if a codon change results, there is still only that one protein species affected by the TM event. In contrast, a mutated rRNA could disrupt the function or fidelity of the ribosome, potentially creating many more mutant proteins, which would be a worse consequence for the cell. Therefore, rDNA genome regions may be more closely protected against the persistence of DNA damage, or RNA polymerases I and III may have higher fidelity than RNA polymerase II. Further studies combining measurement of DNA damage distribution coupled with TM studies of rRNA and mRNA may help distinguish between these possibilities.

An important consideration in studying TM using ARC-seq is the scale of study desired. While we herein presented whole-transcriptome data, such a broad view may not always be desired or feasible. Two potential modifications to ARC-seq are possible that enable focusing on specific loci. First, one could enrich transcript regions of interest after ARC-seq library preparation, before sequencing, via either via single (42) or double capture (43). Such methods have been instrumental in enabling studies of small genomic regions in mammalian systems and would be an easy addition to ARC-seq; however, they require gene capture sets and additional steps after the initial library preparation. The second option is to use transcript-specific primers (44) during rolling-circle reverse transcription instead of the primer against the RNA adaptor. Transcript-specific reverse transcription represents a minor modification to ARC-seq as presented and would enable targeting of specific transcripts with only a primer, greatly minimizing the expense of targeting, relative to the capture approach.

Conclusions

We have developed a highly accurate RNA sequencing method that effectively eliminates artifacts due to reverse transcription, PCR, and sequencing errors. ARC-seq represents a major advance over previous methods. First, the method itself uses low temperatures, neutral solutions, and short incubations whenever possible, thereby minimizing the damage to the RNA template that limits other methods (28, 29). Next, it reliably generates multiple cDNA copies from each RNA molecule with high yield, a significant advance over prior attempts that were significantly limited by low yields (25). This high-yield cDNA copy generation also accounts for its scalable stringency, which enables highly accurate detection of TM-induced epimutations even from highly damaged sources. Finally, because it is not limited by sequence length, ARC-seq can be applied to any sample, without limitations on accuracy; with minor modifications, it could be used to look at transcriptome-wide TM, as we have demonstrated, or gene-specific TM to drive studies of the role of epimutations at specific loci in disease processes.

The accuracy, sensitivity, and scalable stringency of ARC-seq make it advantageous for application to numerous biological questions that have remained intractable to date. Future studies of TM in model systems, such as the yeast mutants studied here, could explore how perturbing or enhancing various aspects of transcription, including transcription factors or the nucleotide pool, affects transcriptional fidelity (22). Such studies could perhaps not only provide greater insight into the basic biology of transcription but also potentially lead to studies on how perturbing the transcriptional apparatus may potentially be useful as a therapeutic target in cancer and microbial diseases (2, 6, 45). Applying ARC-seq to studies of RNA viral populations could provide greater insight into the nature of quasispecies and how viral populations evolve and under which conditions, and potentially provide insight into how to prevent therapeutic resistance or even directly manipulate viral transcriptional apparati to induce lethal mutagenesis (7, 8). Additionally, applying ARC-seq to studies of TM in aging and neurodegeneration could elucidate whether or not epimutations underlie the pathologies of age-related disease, such as sporadic Alzheimer's disease (4, 15) and cancer (6, 20), and, finally, address the long-standing hypothesis of protein synthesis errors driving aging and disease (37–39).

Methods

IVT RNAs were generated from a single-stranded m13mp18 DNA template via an established protocol (46), using T7 RNA polymerase. To generate damaged IVT RNA, following transcription, the high-fidelity RNA was treated with 100 μM H_2O_2 and $FeCl_3$ to induce oxidative DNA damage, according to an established protocol (47). To generate mutated IVT RNA, the m13mp18 DNA template was treated with 1 mM H_2O_2 before transcription.

Wild-type yeast and E1103G yeast were a gift from Mikhail Kashlev at the NIH/National Cancer Institute (NCI), Bethesda, and ΔRpb9 yeast was a gift from Jeffrey Strathern at the NIH/NCI. To measure TM in yeast, log-phase yeast or stationary-phase yeast was pelleted, washed with cold 1× PBS, and repelleted. The cell walls were then digested by incubating cells in a buffer containing sorbitol and 100 units of Zymolyase, according to an established protocol (48). RNAs were then extracted, enriching for mRNA, using the Dynabead mRNA Direct Kit from Ambion. Extracted RNAs were stored in 10 mM Tris and 0.1 mM EDTA buffer (pH 8.0) made with diethyl pyrocarbonate (DEPC)-treated nuclease-free water, with 100 units of murine RNase inhibitor from New England Biolabs (NEB) added, at −80 °C.

RNA Library Preparation. RNA libraries were prepared via the ARC-seq protocol, as detailed in *SI Materials and Methods*. Briefly, fragmented RNAs were end-repaired, preadenylated, and ligated to ARC-seq adaptors. Adapted RNAs were circularized and then subjected to rolling-circle reverse transcription to generate multimeric cDNAs. The cDNA multimers were restricted into cDNA monomers, each of which was subsequently indexed via 5′-overhang extension PCR. Indexed cDNA monomers were amplified and sequenced on an Illumina HiSeq 2200 instrument, using the dual-indexing protocol.

Data Processing. Reads were filtered for those containing properly located tag sequences, and the 16-nt RNA barcode and 8-nt cDNA index were combined to create a 24-nt tag for each read. Reads containing identical tag sequences were grouped together to form PCR consensus reads. PCR consensus reads sharing identical 16-nt RNA barcodes were then grouped together to form cDNA consensus families. The cDNA consensus for any position is considered undefined if the position is represented by fewer than *n* instances in the family or if less than 70% of the sequences at that position in the read are in agreement; *n* represents the number of cDNA copies generated from each RNA molecule and can be adjusted to increase assay stringency if the RNA template is damaged. Further details are provided in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Scott Kennedy for assistance with bioinformatics; Edward Fox, Scott Kennedy, Gwen Garden, Peter Rabinovitch, and Alan Weiner for helpful discussions; and Tom Walsh and Ming Lee for assistance with sequencing. This work was supported by NIH NCI Grants P01-CA77852 and R01-CA160674 (to L.A.L.) and National Institute on Aging Grant T32 AG000057 (to K.S.R.-B).

1. Paoloni-Giacobino A, Rossier C, Papasavvas MP, Antonarakis SE (2001) Frequency of replication/transcription errors in (A)/(T) runs of human genes. *Hum Genet* 109:40–47.
2. Rodin SN, Rodin AS, Juhasz A, Holmquist GP (2002) Cancerous hyper-mutagenesis in p53 genes is possibly associated with transcriptional bypass of DNA lesions. *Mutat Res* 510:153–168.
3. Hubbard K, Catalano J, Puri RK, Gnatt A (2008) Knockdown of TFIIS by RNA silencing inhibits cancer cell proliferation and induces apoptosis. *BMC Cancer* 8:133.
4. van Leeuwen FW, et al. (1998) Frameshift mutants of beta amyloid precursor protein and ubiquitin-B in Alzheimer's and Down patients. *Science* 279:242–247.
5. van Leeuwen FW, Burbach JP, Hol EM (1998) Mutations in RNA: A first example of molecular misreading in Alzheimer's disease. *Trends Neurosci* 21:331–335.
6. Morreall JF, Petrova L, Doetsch PW (2013) Transcriptional mutagenesis and its potential roles in the etiology of cancer and bacterial antibiotic resistance. *J Cell Physiol* 228:2257–2261.
7. Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R (2006) Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* 439:344–348.
8. Coffey LL, et al. (2008) Arbovirus evolution in vivo is constrained by host alternation. *Proc Natl Acad Sci USA* 105:6970–6975.
9. Blank A, Gallant JA, Burgess RR, Loeb LA (1986) An RNA polymerase mutant with reduced accuracy of chain elongation. *Biochemistry* 25:5920–5928.
10. Rosenberger RF, 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.
11. Ninio J (1991) Connections between translation, transcription and replication error-rates. *Biochimie* 73:1517–1523.
12. Remington KM, Bennett SE, Harris CM, Harris TM, Bebenek K (1998) Highly mutagenic bypass synthesis by T7 RNA polymerase of site-specific benzo[a]pyrene diol epoxide-adducted template DNA. *J Biol Chem* 273:13170–13176.
13. Kireeva ML, et al. (2008) Transient reversal of RNA polymerase II active site closing controls fidelity of transcription elongation. *Mol Cell* 30:557–566.
14. Doetsch PW (2002) Translesion synthesis by RNA polymerases: Occurrence and biological implications for transcriptional mutagenesis. *Mutat Res* 510:131–140.
15. Viswanathan A, You HJ, Doetsch PW (1999) Phenotypic change caused by transcriptional bypass of uracil in nondividing cells. *Science* 284:159–162.
16. Brégeon D, Doddridge ZA, You HJ, Weiss B, Doetsch PW (2003) Transcriptional mutagenesis induced by uracil and 8-oxoguanine in Escherichia coli. *Mol Cell* 12:959–970.
17. Pastoriza-Gallego M, Armier J, Sarasin A (2007) Transcription through 8-oxoguanine in DNA repair-proficient and Csb(-)/Ogg1(-) DNA repair-deficient mouse embryonic fibroblasts is dependent upon promoter strength and sequence context. *Mutagenesis* 22:343–351.
18. Saxowsky TT, Meadows KL, Klungland A, Doetsch PW (2008) 8-Oxoguanine-mediated transcriptional mutagenesis causes Ras activation in mammalian cells. *Proc Natl Acad Sci USA* 105:18877–18882.
19. Burns JA, Dreij K, Cartularo L, Scicchitano DA (2010) O6-methylguanine induces altered proteins at the level of transcription in human cells. *Nucleic Acids Res* 38:8178–8187.
20. Gordon AJ, Satory D, Halliday JA, Herman C (2013) Heritable change caused by transient transcription errors. *PLoS Genet* 9:e1003595.
21. Gordon AJ, et al. (2009) Transcriptional infidelity promotes heritable phenotypic change in a bistable gene network. *PLoS Biol* 7:e44.
22. Strathern JN, Jin DJ, Court DL, Kashlev M (2012) Isolation and characterization of transcription fidelity mutants. *Biochim Biophys Acta* 1819:694–699.
23. Ji JP, Loeb LA (1992) Fidelity of HIV-1 reverse transcriptase copying RNA in vitro. *Biochemistry* 31:954–958.
24. Minoche AE, Dohm JC, Himmelbauer H (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol* 12:R112.
25. Gout JF, Thomas WK, Smith Z, Okamoto K, Lynch M (2013) Large-scale detection of in vivo transcription errors. *Proc Natl Acad Sci USA* 110:18584–18589.
26. Imashimizu M, Oshima T, Lubkowska L, Kashlev M (2013) Direct assessment of transcription fidelity by high-resolution RNA sequencing. *Nucleic Acids Res* 41:9090–9104.
27. Jabara CB, Jones CD, Roach J, Anderson JA, 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. Acevedo A, Brodsky L, Andino R (2014) Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* 505:686–690.
29. Acevedo A, Andino R (2014) Library preparation for highly accurate population sequencing of RNA viruses. *Nat Protoc* 9:1760–1769.
30. Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA (2014) Accuracy of next generation sequencing platforms. *Next Gener Seq Appl* 1:106.
31. Walmacq C, et al. (2009) Rpb9 subunit controls transcription fidelity by delaying NTP sequestration in RNA polymerase II. *J Biol Chem* 284:19601–19612.
32. Kasai H, Tanooka H, Nishimura S (1984) Formation of 8-hydroxyguanine residues in DNA by X-irradiation. *Gan* 75:1037–1039.
33. Lee DH, Pfeifer GP (2008) Translesion synthesis of 7,8-dihydro-8-oxo-2′-deoxyguanosine by DNA polymerase eta in vivo. *Mutat Res* 641:19–26.
34. De Bont R, van Larebeke N (2004) Endogenous DNA damage in humans: A review of quantitative data. *Mutagenesis* 19:169–185.
35. Cooke MS, et al. (2003) Oxidative DNA damage: Mechanisms, mutation, and disease. *FASEB J* 17:1195–1214.
36. Lindahl T (1979) DNA glycosylases, endonucleases for apurinic/apyrimidinic sites, and base excision-repair. *Prog Nucleic Acid Res Mol Biol* 22:135–192.
37. Orgel LE (1963) The maintenance of the accuracy of protein synthesis and its relevance to ageing. *Proc Natl Acad Sci USA* 49:517–521.
38. Orgel LE (1970) The maintenance of the accuracy of protein synthesis and its relevance to ageing: A correction. *Proc Natl Acad Sci USA* 67:1476.
39. Martin GM, Bressler SL (2000) Transcriptional infidelity in aging cells and its relevance for the Orgel hypothesis. *Neurobiol Aging* 21:897–900; discussion 903–904.
40. Zhou S, et al. (2015) Primer ID validates template sampling depth and greatly reduces the error rate of next-generation sequencing of HIV-1 genomic RNA populations. *J Virol* 89:8540–8555.
41. Traverse CC, Ochman H (2016) Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proc Natl Acad Sci USA* 113:3311–3316.
42. Mamanova L, et al. (2010) Target-enrichment strategies for next-generation sequencing. *Nat Methods* 7:111–118.
43. Schmitt MW, et al. (2015) Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat Methods* 12:423–425.
44. Frohman MA, Dush MK, Martin GR (1988) Rapid production of full-length cDNAs from rare transcripts: Amplification using a single gene-specific oligonucleotide primer. *Proc Natl Acad Sci USA* 85:8998–9002.
45. Brégeon D, Doetsch PW (2011) Transcriptional mutagenesis: Causes and involvement in tumor development. *Nat Rev Cancer* 11:218–227.
46. Korencic D, Soll D, Ambrogelly A (2002) A one-step method for in vitro production of tRNA transcripts. *Nucleic Acids Res* 30:e105.
47. McBride TJ, Preston BD, Loeb LA (1991) Mutagenic spectrum resulting from DNA damage by oxygen radicals. *Biochemistry* 30:207–213.
48. Klassen R, et al. (2008) A modified DNA isolation protocol for obtaining pure RT-PCR grade RNA. *Biotechnol Lett* 30:1041–1044.
49. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25:1754–1760.