GENE TRANSCRIPTION

The landscape of transcription errors in eukaryotic cells

Jean-Francois Gout,[1]* Weiyi Li,[1]* Clark Fritsch,[2,3] Annie Li,[2] Suraiya Haroon,[2] Larry Singh,[2] Ding Hua,[4] Hossein Fazelinia,[4] Zach Smith,[5] Steven Seeholzer,[4] Kelley Thomas,[6] Michael Lynch,[1†] Marc Vermulst[2†]

Accurate transcription is required for the faithful expression of genetic information. To understand the molecular mechanisms that control the fidelity of transcription, we used novel sequencing technology to provide the first comprehensive analysis of the fidelity of transcription in eukaryotic cells. Our results demonstrate that transcription errors can occur in any gene, at any location, and affect every aspect of protein structure and function. In addition, we show that multiple proteins safeguard the fidelity of transcription and provide evidence suggesting that errors that evade these layers of RNA quality control profoundly affect the physiology of living cells. Together, these observations demonstrate that there is an inherent limit to the faithful expression of the genome and suggest that the impact of mutagenesis on cellular health and fitness is substantially greater than currently appreciated.

INTRODUCTION

Biological reactions are remarkably precise. Our proteins have the unique ability to select the correct substrates out of complex mixtures of countless molecules and to do so at the right time, at the right place, and with the right partners. This precision is especially important in the context of DNA replication, transcription, and translation. Together, these three processes preserve the integrity of our genome and ensure the faithful expression of our genetic code. As a result, numerous studies have investigated the mechanisms that control the fidelity of DNA replication (*1*) and translation (*2*), but technical limitations have handicapped efforts to investigate the fidelity of transcription. Unlike genetic mutations, transcription errors are transient events that are not stably inherited from cell to cell, which makes them difficult to detect. To solve this problem, a number of novel reporter assays were recently developed that were inspired by early in vitro measurements of transcriptional fidelity (*3*–*7*). Excitingly, these reporter assays now allow transcription errors to be detected in living cells, but because they only detect transcription errors in artificial reporter constructs, it is unclear whether their findings can be extrapolated to the rest of the genome. To overcome this limitation, numerous strategies have been deployed, including the mining of RNA sequencing (RNA-seq) data for splicing errors (*8*) and the design of completely novel sequencing assays, such as the "high-resolution sequencing method" (*9*), the "replicated sequencing method" (*10*), and the "circle-sequencing method" [for an in-depth review of these methods, see the study of Gordon *et al.* (*11*)]. Conceptually, the most straightforward way to measure the fidelity of transcription is by reverse transcription of RNA, followed by complementary DNA (cDNA) sequencing. A crucial drawback of this strategy is that reverse transcriptases are notoriously error-prone and expected to make one error every ≈10,000 to 30,000 bases (*12*). In contrast, RNA polymerases are expected to make one error every 300,000 bases (*10*). Thus, a standard cDNA library will always be dominated by reverse transcription errors that mask the errors made by RNA polymerases. One solution to this problem is to reverse-transcribe the same mRNA molecule multiple times. For example, if three cDNA copies were made of a single mRNA molecule, then a true transcription error would be present at the same location in every cDNA copy of this molecule, whereas a reverse transcriptase error would appear in only one of these copies. This is the core idea behind most of these novel sequencing assays, including the "circle-sequencing" assay, which was originally designed to sequence RNA viruses (*12*, *13*). The circle-sequencing assay carries this name because a key step in its protocol is mRNA circularization. After circularization of the mRNA molecules, they are reverse-transcribed in a rolling-circle reaction so that each cDNA molecule consists of a tandem repeat of the mRNA template. These concatemers can then be sequenced to identify transcription errors and analyzed using advanced bioinformatics to distinguish true transcription errors from potential artifacts (Fig. 1). Recently, the original version of the circle-sequencing assay was applied to study the fidelity of transcription in bacteria (*14*). Here, we describe numerous modifications to the circle-sequencing assay (*12*, *13*), which allowed us to streamline the protocol, increase its sensitivity, and design a customized bioinformatic pipeline to identify transcription errors. We changed a key step in the protocol that artificially increased the detected error rate by 5- to 10-fold through direct damage to RNA targets, which could have affected the measurements made in bacteria. A more detailed discussion of these improvements and the bioinformatic pipeline we used for error discovery can be found in fig. S1 and Materials and Methods. The code for our pipeline can be downloaded at https://github.com/LynchLab.

RESULTS

Over the course of our experiments, we screened >8.5 billion bases of the yeast transcriptome and found >200,000 transcription errors in eight unique cell lines. Because previous efforts have detected only 109 transcription errors in eukaryotic cells (*10*), our experiments represent the first comprehensive analysis of the fidelity of transcription in a eukaryotic organism. The errors we detected were distributed across the entire transcriptome of *Saccharomyces cerevisiae*, indicating that our approach provides a genome-wide view of transcriptional mutagenesis in yeast (Fig. 2, A and B). Errors were found along the

[1]Department of Biology, Indiana University, Bloomington, IN 47405, USA. [2]Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, PA 19102, USA. [3]Department of Cellular and Molecular Biology, University of Pennsylvania, Philadelphia, PA 19104, USA. [4]Protein and Proteomics Core, Children's Hospital of Philadelphia, Philadelphia, PA 19102, USA. [5]Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN 47405, USA. [6]Department of Molecular, Cellular, and Biomedical Sciences, University of New Hampshire, Durham, NH 03824, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: vermulstm@email.chop.edu (M.V.); milynch@indiana.edu (M.L.)

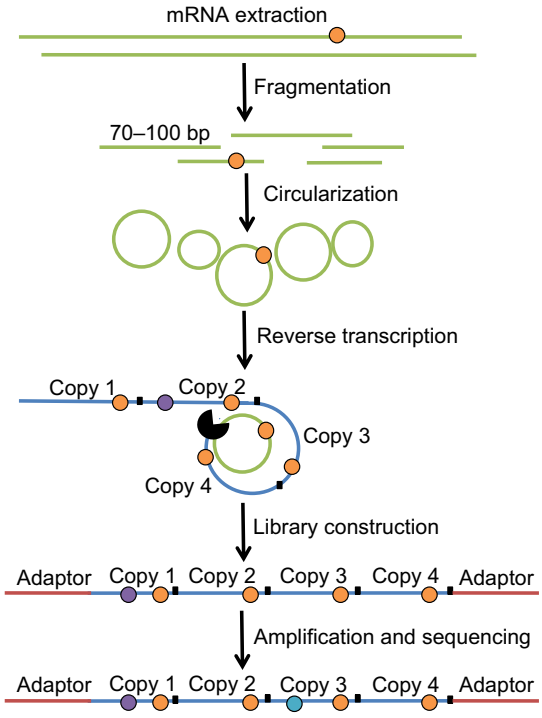


Fig. 1. A visual representation of the circle-sequencing assay. The circle-sequencing protocol identifies transcription errors (orange circles) by fragmenting RNA (green strands) into short oligonucleotides, circularizing them, and reverse-transcribing the RNA circles in a rolling-circle reaction to generate linear cDNA molecules made up of tandem repeats of the original RNA fragment (blue strands). During this step, artificial mutations may arise in the cDNA (purple circles). The cDNA is then processed to generate a library, amplified, and sequenced, during which further artifacts may arise (teal circles). However, because these artifacts are only present in one copy of the tandem repeats, they can be distinguished from true transcription errors, which are present in all tandem repeats. bp, base pair.

entire length of transcripts, indicating that they affect every aspect of RNA functionality, including the location of the start and stop codon, the stability of secondary structures, and the information that is encoded in the primary sequence. Accordingly, transcription errors also affect every aspect of protein structure and function, including residues for posttranslational modifications, catalysis, substrate binding, and structural integrity. As one illustration of these observations, we mapped a small portion of the errors we detected in the mRNA of the *ADH1* gene onto the ADH1 transcript and a larger portion on the three-dimensional structure of an ADH1 dimer (Fig. 2, C to E). Together, these experiments demonstrate that the circle-sequencing assay is a powerful new sequencing tool that can be exploited to monitor the fidelity of transcription across the entire genome with single base-pair resolution. The resultant data can then be analyzed to understand the impact of transcription errors on RNA and protein biology.

Transcription errors are not equally distributed over the transcriptome

To determine the error rate of transcription, we analyzed >2.5 billion bases from 12 biological replicates of wild-type (WT) cells and found that on average, the yeast transcriptome contains ≈4.0 errors per million base pairs. Thus, these results demonstrate that transcription errors occur >100-fold more frequently than DNA replication errors (*15*). These errors are not distributed equally over the transcriptome. mRNA molecules synthesized by RNA polymerase II (RNAPII) contain the least amount of errors (3.9×10^{-6} per base pair), followed by ribosomal RNA molecules synthesized by RNAPI (4.3×10^{-6} per base pair), mitochondrial RNA (9.3×10^{-6} per base pair), and RNA molecules synthesized by RNAPIII (1.7×10^{-5} per base pair; Fig. 3A). These results suggest that each polymerase has its own unique error rate, similar to what has been observed for DNA polymerases (*16*). Within a class of transcripts, however, the error rate was remarkably constant. For example, the error rate of transcripts synthesized by RNAPII is independent of the expression level of a gene (fig. S2), its distance from an origin of replication (fig. S3), or the position of a base along the length of the gene (fig. S4). In addition, we found that bases that are known to be subject to RNA modifications did not display an increased error rate, although we did detect a significant decrease in the coverage of these bases, indicating that they are not efficiently reverse-transcribed and thus underrepresented in our data set.

Multiple RNA polymerase subunits control the error rate of transcription

RNAPII is further known to be associated with RNAPII subunit RPB9 and transcription factor II S (TFIIS), two proteins that were previously shown to improve the fidelity of RNAPII on genetically engineered DNA templates (*17*, *18*). We found that *rpb9*Δ and *dst1*Δ cells (*Dst1* encodes the TFIIS protein) displayed a 5- to 10-fold increase in the error rate of mRNA synthesis, indicating that these proteins are responsible for the high fidelity of transcription by RNAPII (Fig. 3B). RPB9 is further known to interact with the trigger loop of RNAPII (*17*), a highly dynamic structure that is thought to function as a kinetic selector for correct nucleoside triphosphate substrates (*19*). Accordingly, a single-point mutation in the major catalytic subunit of RNAPII that directly affects the trigger loop ($Rbp1^{E1103G}$) (*20*) increased the error rate of mRNA synthesis fivefold (Fig. 3B), further cementing the role of the trigger loop in the fidelity of RNAPII (*17*, *20*, *21*). Other RNA species were not affected by these interventions, confirming that these alleles only regulate the fidelity of RNAPII (Fig. 3B). RNAPI is associated with RPA12, a protein that is partially homologous to both RPB9 and TFIIS (*22*, *23*). Accordingly, our results suggest that RPA12 may regulate the fidelity of RNAPI. To test this hypothesis, we measured the error rate of *rpa12*Δ cells and found that *rpa12*Δ cells display an 11-fold increase in the error rate of transcription by RNAPI, whereas the error rate of RNAPII remained constant (Fig. 3B), revealing parallels in the mechanisms responsible for the fidelity of different RNAP. Because subunit C11 of RNAPIII is homologous to RPB9, TFIIS, and RPA12 (*22*, *23*), it would be interesting to determine whether this protein is responsible for the fidelity of RNAPIII.

Finally, we found that none of the error-prone mutants we tested displayed a higher genomic mutation rate than WT cells, excluding this possibility as a potential explanation for our findings (fig. S5). Note that additional safety mechanisms are built into our bioinformatic pipeline that also prevent genetic mutations from affecting our measurements on WT cells. Instead, these observations strongly support the idea that the fidelity of RNA polymerases is maintained by the inherent design of the catalytic subunits and the accessory subunits that directly interact with the holoenzyme (*19*).

The spectrum of transcription errors in yeast cells

To gain more insight into the molecular mechanisms that drive the error rate of transcription by RNAPII, we examined its error spectrum in

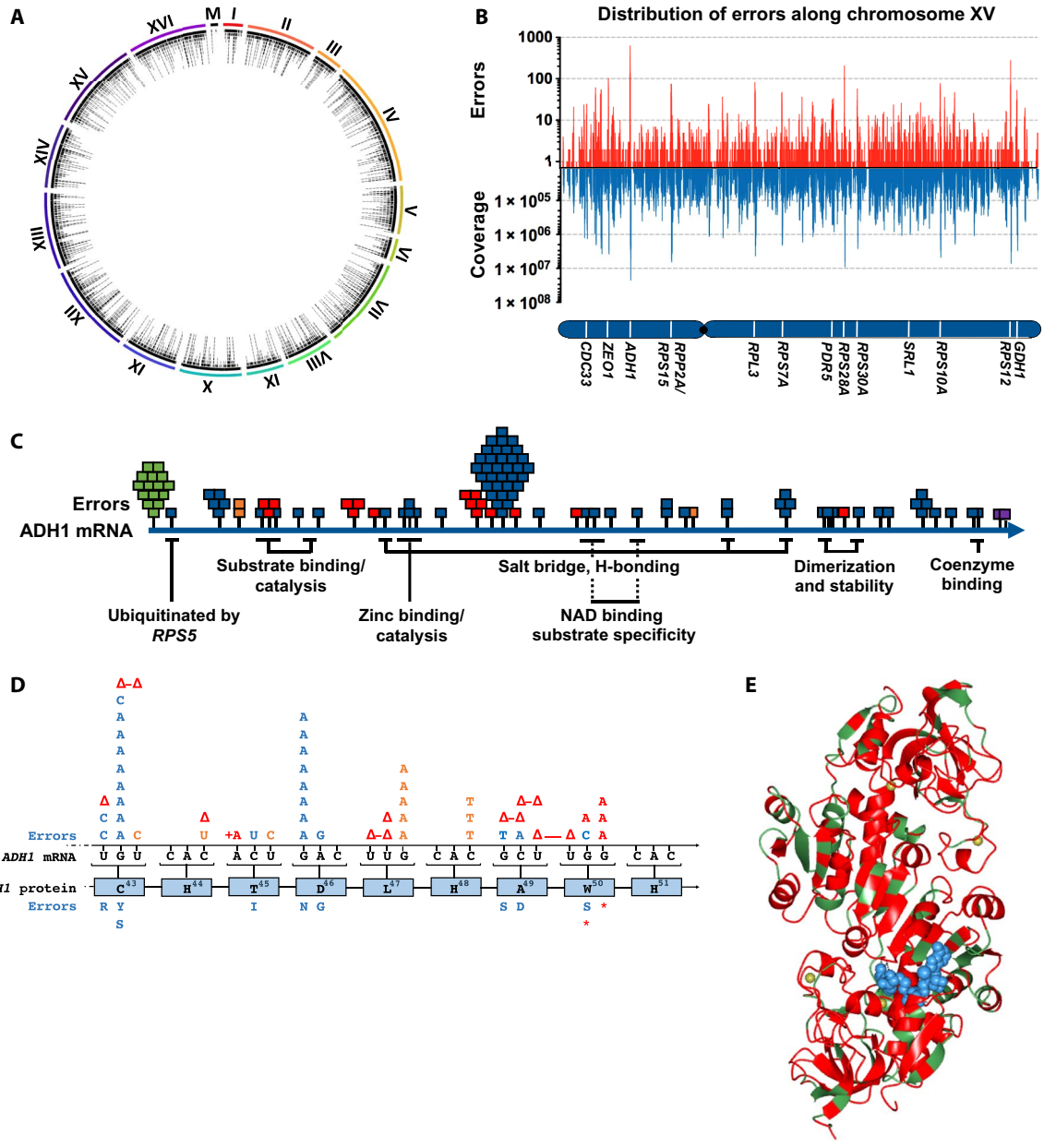


Fig. 2. Overview of transcriptional mutagenesis in yeast. Over the course of our experiments, we detected >200,000 transcription errors. Here, we provide a broad overview of our results at increasing levels of detail. (**A**) The transcription errors detected were distributed across the entire genome of yeast. (**B**) Although transcription errors occurred randomly across the length of a chromosome, most errors were detected in highly transcribed genes. These genes do not display an increased error rate per nucleotide but were simply sequenced at a greater frequency and thus provided the greatest amount of information to our data set. "Errors" indicate the total number of errors detected within a 100-bp interval. "Coverage" indicates the number of times a base pair in that interval was sequenced. (**C**) Depiction of a subset of the errors that were detected in the *ADH1* gene. More than 2000 errors were detected in the *ADH1* gene, affecting approximately 50% of all possible nucleotides. Each block represents a single error. Green blocks represent errors that changed the start codon of the *ADH1* gene, purple errors changed its stop codon, and red errors generated premature termination codons. We also detected synonymous (orange) and nonsynonymous errors (blue), which altered almost every aspect of protein function and structure. (**D**) Individual errors detected in a small region of the ADH1 mRNA. (**E**) All errors detected in the ADH1 mRNA that are mapped onto the protein structure. All amino acids in which errors were detected are shown in red. For clarity, NAD is depicted in blue, and zinc is depicted in yellow.

greater detail. In WT cells, RNAPII primarily makes C→U and G→A transitions and G→U transversions (Fig. 3C). This error spectrum overlaps with RNAPI, suggesting that similar mechanisms control the fidelity of these polymerases. We further found that these errors occur in a wide variety of genetic contexts, which display several interesting patterns (Fig. 3D). For example, cytosine is most mutable when flanked at the 3′ end by a purine base, whereas guanine is most mutable when flanked at the 3′ end by a pyrimidine base, suggesting that the transition between purines and pyrimidines can be problematic. A clear pattern emerged for uracil as well, which is most likely to be mutated when flanked on the 5′ end by a guanine. Most likely, multiple mechanisms contribute to these error rates, including the rate at which nucleotides

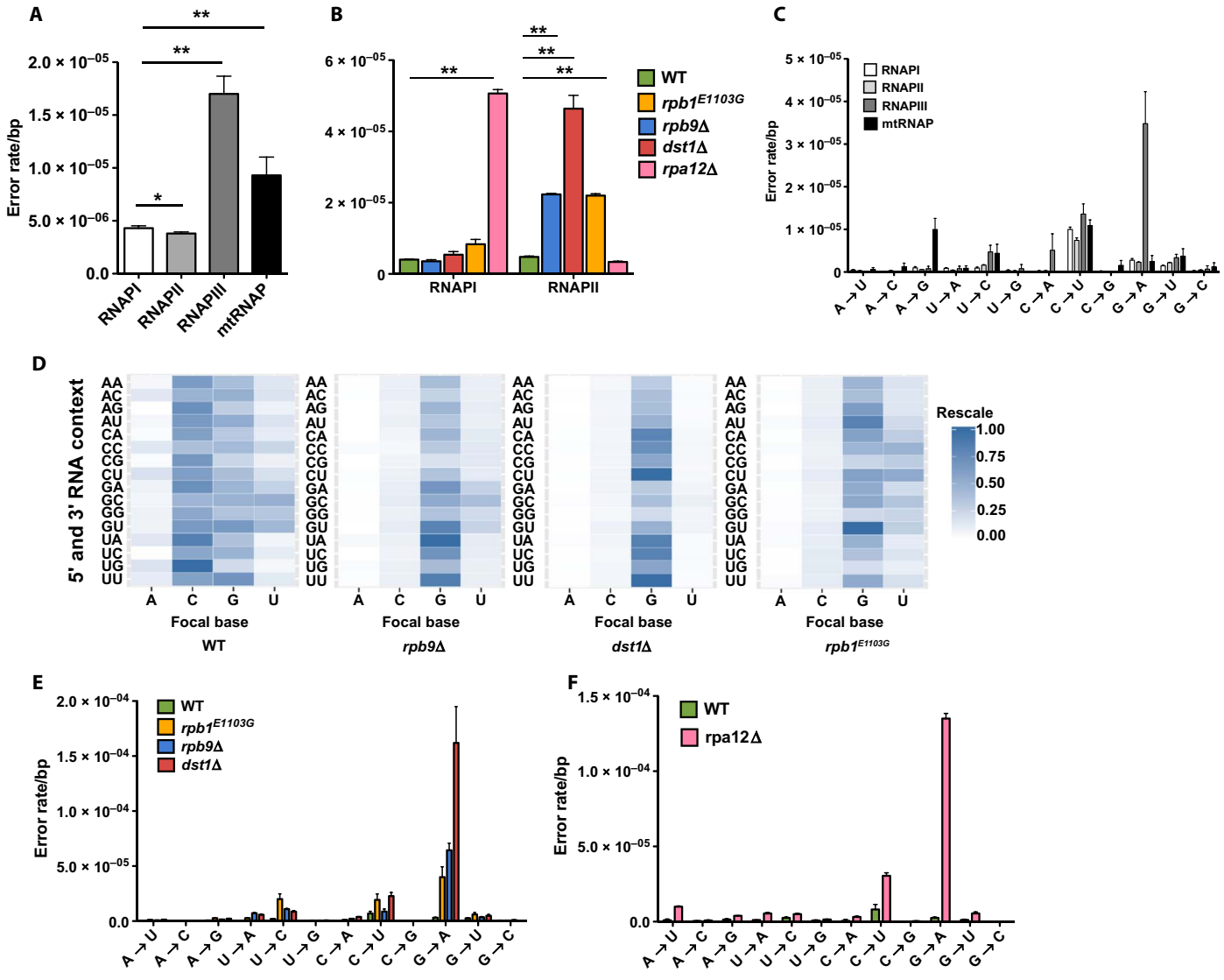


Fig. 3. The error rate and error spectrum of transcription in yeast. (**A**) Error rate of transcripts generated by all major RNA polymerases in yeast cells. Because the error rate of transcription is >10-fold higher than the genetic mutation frequency, <1% of these errors are likely due to genetic mutation. Additional safety mechanisms have been built into our bioinformatic pipeline to identify these genetic mutations and remove them from further analysis. (**B**) Loss of *Rpb9* and *Dst1* or introduction of the *rpb1*^{E1103G} allele results in error-prone transcription by RNAPII. Loss or *Rpa12* results in error-prone transcription by RNAPI. (**C**) Error spectrum of transcripts generated by RNAPI, RNAPII, RNAPIII, and mtRNAP (mitochondrial RNAP) (**D**) Matrices depicting the genetic context that transcription errors occur in WT cells and three error-prone cell lines. The focal base is the base where the error occurred. The first base on the *y* axis is directly upstream of the focal base, whereas the second base is directly downstream. (**E**) All error-prone alleles that we tested resulted in a marked increase in G→A transitions by RNAPII. (**F**) Loss or *Rpa12* results in a similar increase in G→A transitions by RNAPI.

are misincorporated, extended, and proofread (*24*), and the impact of DNA damage on transcriptional fidelity (*25*). Similar errors commonly occur in bacteria (*14*) and *Caenorhabditis elegans* (*10*), suggesting the existence of conserved mechanisms of transcriptional mutagenesis across the tree of life. We further found that the error spectrum of RNAPII strongly depends on the trigger loop and the function of TFIIS because the error spectra of *rpb9*Δ, *dst1*Δ, and *rbp1*^{E1103G} cells are primarily dominated by G→A transitions (Fig. 3E). Then, each of these alleles seems to have evolved in ways that primarily prevent just a single base-pair substitution, although they do so in slightly different genetic contexts (Fig. 3D). The error spectrum of *rpa12*Δ cells was also strongly biased toward G→A transitions, further underlining the functional similarities between RPA12, RPB9, and TFIIS (Fig. 3F). The error spectrum of RNAPIII was also dominated by G→A transitions (fig. S6), which, in combination with its increased error rate, suggests that it functions like an error-prone version of RNAPI and RNAPII. The error spectrum of the mitochondrial RNAP was completely unique, most likely due to its evolutionary origin as a phage polymerase (Fig. 3C).

In addition to single base-pair substitutions, RNAPII also commits insertions (7.4×10^{-7} per base pair) and deletions (2.1×10^{-7} per base pair), almost all of which were either one or two bases in length. As expected, RNAPI commits these errors as well but does so at a slightly higher rate than RNAPII (8.8×10^{-7} per base pair for insertions and 3.4×10^{-7} per base pair for deletions; Fig. 4A). The frameshifts committed

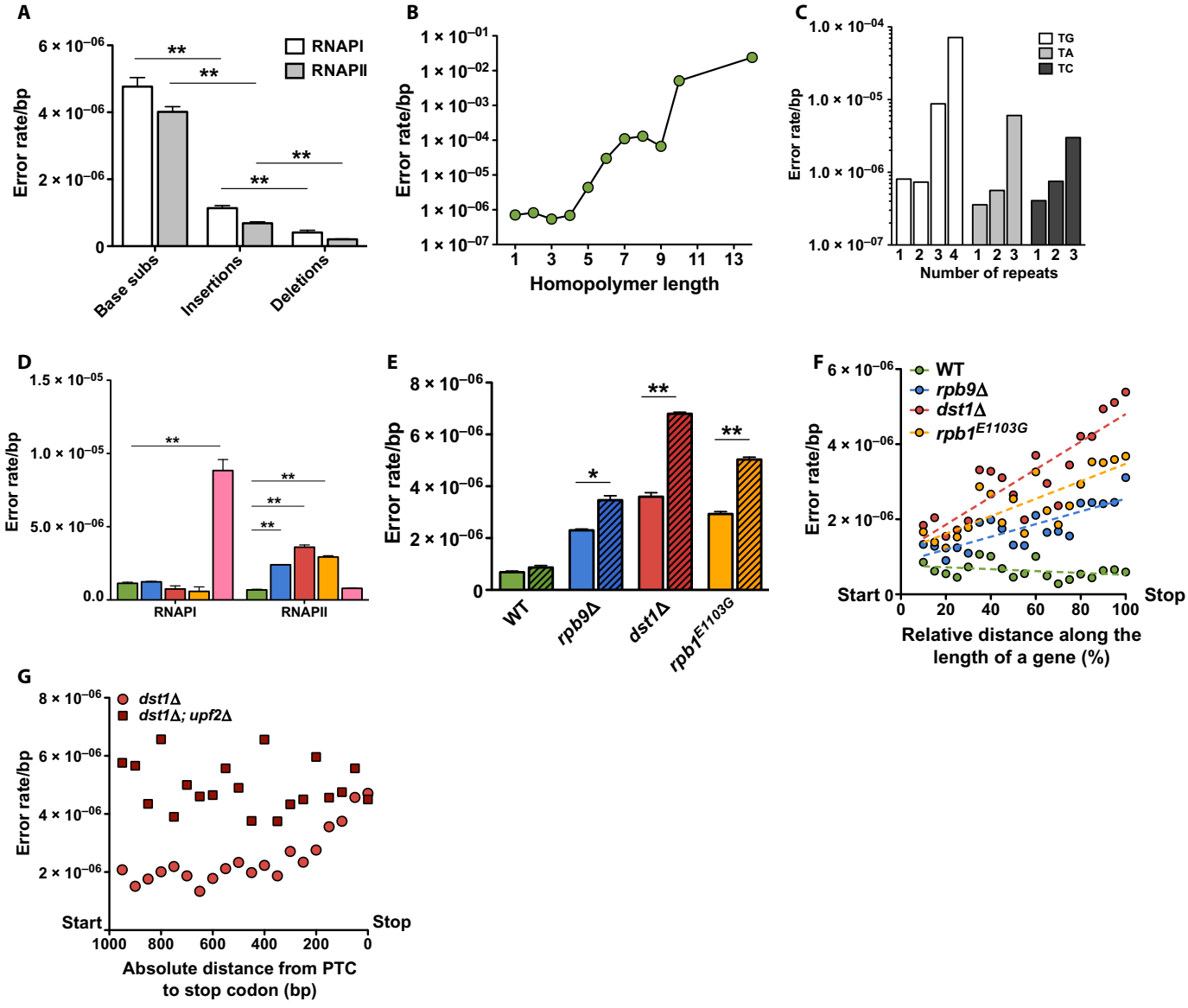


Fig. 4. Frameshifts arise during transcription in yeast. (**A**) Insertions and deletions occur less frequently than base pair substitutions in yeast. (**B**) Homopolymeric tracts are hotspots for frameshift errors in yeast. Here, all possible homopolymer tracts (A, C, G, and T) were combined. (**C**) Tracts of dinucleotides are hotspots for frameshift errors in yeast as well. (**D**) Loss of *Rpb9* and *Dst1* or introduction of the *rpb1*^{E1103G} allele results in an increase in frameshift errors in molecules transcribed by RNAPII, but not by RNAPI. (**E**) Loss of *Upf2* increased the frequency of insertions in the error-prone cell lines. (**F**) Insertions were detected primarily at the 3′ end of genes. "Start" indicates the first codon of the transcript, whereas "Stop" indicates the stop codon. (**G**) Loss of *Upf2* abolished the relationship between insertions and distance along a gene.

by RNAPII preferentially occurred on homonucleotide and dinucleotide tracts of DNA (Fig. 4, B and C), and their frequency increased exponentially with the length of the tract, closely matching observations on genetically engineered templates (*21*, *26*, *27*). Similar transcriptional frameshifts occur on dinucleotide tracts inside the β-amyloid precursor gene in patients with nonfamilial Alzheimer's disease (*28*, *29*), which result in short, aggregation-prone peptides that actively contribute to disease progression, indicating that these tracks are of direct medical relevance. We further found that all of the error-prone alleles increased the insertion rate by 5- to 10-fold (Fig. 4D), whereas only *dst1*Δ cells displayed an increased deletion rate (fig. S6).

Nonsense-mediated RNA decay becomes less efficient in the 3′ end of transcripts

Because frameshifts are more disruptive than single-base substitutions, it is likely that additional safeguards have evolved to prevent them. For example, frameshifted mRNA molecules typically contain premature termination codons (PTCs), triggering their elimination by the nonsense-mediated RNA decay (NMD) pathway (*30*). Accordingly, we found that the insertion rate increased almost twofold when an essential component of the NMD pathway (*Upf2*) was knocked out (Fig. 4E) (*31*). We observed a similar trend in single-base substitutions that generate premature stop codons, whereas errors that cause synonymous

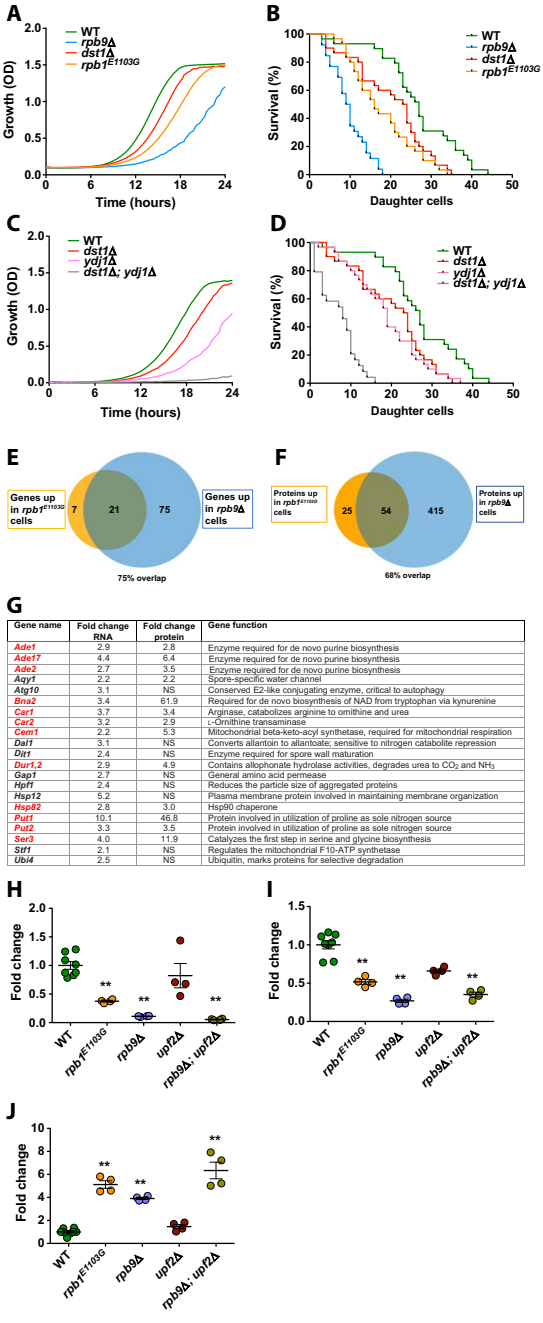
or missense mutations were unaffected by *Upf2* deletion (table S1). It is further thought that the ability of NMD to identify a PTC diminishes the closer it is to the polyadenylate [poly(A)] tail (*30*), although a detailed map of the efficacy of NMD along the length of a gene has not been established. In agreement with this idea, we found that PTCs are relatively rare in the 5′ end of transcripts but that their frequency increases markedly in the final 400 bases preceding the 3′ end of transcripts (Fig. 4G). Deletion of *upf2* abolished this pattern in all of the error-prone cell lines (Fig. 4G), confirming that NMD was responsible for this asymmetric distribution and thereby outlining the practical limitations of this pathway in yeast. Similar observations were made for PTCs generated by single-base substitutions.

Transcription errors reduce proteostasis

Transcription errors play an important role in protein stability. In humans, transcription errors generate toxic versions of the Aβ protein in patients with nonfamilial Alzheimer's disease (*28*, *29*) and faulty ubiquitin-B (UBB) proteins in patients with Down syndrome (*28*, *29*). In addition, transcription errors induce proteotoxic stress and accelerate cellular aging in yeast (*32*). To better understand the link between transcription errors and protein instability, we examined the impact of transcription errors in proteins in greater detail using the ADH1 protein as a model for our observations (Fig. 2, C and E). We found that transcription errors can affect the function of ADH1 in every conceivable way. Because most amino acids support the structural integrity of proteins, transcription errors affected the structural integrity of the ADH1 protein the most. For example, some errors prevented the formation of hydrogen bonds and salt bridges that normally mediate the internal stability to ADH1 monomers, whereas other errors changed amino acids that allow ADH1 to form stable dimers and tetramers (Fig. 2C) (*33*). Most likely, these observations directly underlie the link between transcription errors and misfolded proteins inside cells. We previously showed, and confirm here, that these misfolded proteins can affect both the growth rate and life span of yeast cells (Fig. 5, A and B) (*32*), forcing them to depend on molecular chaperones, such as YDJ1, to prevent greater toxicity (Fig. 5, C and D) (*32*).

Transcription errors may affect multiple biological processes

To further explore the consequences of transcription errors on cellular health, we used an independent RNA-seq analysis to compare the transcriptome of WT cells to two cell lines suffering from increased levels of transcription errors (*rbp1*^{E1103G} and *rpb9*Δ cells) and identified 21 genes that were significantly up-regulated >2-fold in both of the error-prone cell lines (Fig. 5E and tables S2 and S3). Five of these genes play a role in protein quality control, consistent with the idea that transcription errors result in proteotoxic stress. Surprisingly, we found that the remaining genes were involved in various metabolic pathways. A whole-proteome analysis that detected >4000 proteins demonstrated that 12 of these 21 genes were also up-regulated >2-fold at the protein level in both error-prone cell lines (Fig. 5, F and G, and table S4). Of these 12 genes, *Ade1*, *Ade2*, and *Ade17* regulate sequential steps in purine biosynthesis; *Car1*, *Car2*, *Dur1,2*, *Put 1*, and *Put2* play a role in the urea cycle; and *Bna2* regulates the degradation of tryptophan to kynurenine, a building block for the synthesis of nicotinamide adenine dinucleotide (NAD). These results suggest that transcription errors may deplete several cellular resources, including nitrogen, purines, and NAD. To test this hypothesis, we performed a metabolomic analysis and found that both error-prone cell lines display a significant reduction in guanosine, guanine, and 2′-deoxyguanosine (Fig. 5H and fig. S7), as well as citrulline and



Gene name	Fold change RNA	Fold change protein	Gene function
Ade1	2.9	2.8	Enzyme required for de novo purine biosynthesis
Ade17	4.4	6.4	Enzyme required for de novo purine biosynthesis
Ade2	2.7	3.5	Enzyme required for de novo purine biosynthesis
Agy1	2.2	2.2	Spore-specific water channel
Atg10	3.1	NS	Conserved E2-like conjugating enzyme, critical to autophagy
Bna2	3.4	61.9	Required for de novo biosynthesis of NAD from tryptophan via kynurenine
Car1	3.7	3.4	Arginase, catabolizes arginine to ornithine and urea
Car2	3.2	2.9	L-Ornithine transaminase
Cem1	2.2	5.3	Mitochondrial beta-keto-acyl synthetase, required for mitochondrial respiration
Dal1	3.1	NS	Converts allantoin to allantoate; sensitive to nitrogen catabolite repression
Dit1	2.4	NS	Enzyme required for spore wall maturation
Dur1,2	2.9	4.9	Contains allophonate hydrolase activities, degrades urea to CO_2 and NH_3
Gap1	2.7	NS	General amino acid permease
Hpf1	2.4	NS	Reduces the particle size of aggregated proteins
Hsp12	5.2	NS	Plasma membrane protein involved in maintaining membrane organization
Hsp82	2.8	3.0	Hsp90 chaperone
Put1	10.1	46.8	Protein involved in utilization of proline as sole nitrogen source
Put2	3.3	3.5	Protein involved in utilization of proline as sole nitrogen source
Ser3	4.0	11.9	Catalyzes the first step in serine and glycine biosynthesis
Stf1	2.1	NS	Regulates the mitochondrial F10-ATP synthetase
Ubi4	2.5	NS	Ubiquitin, marks proteins for selective degradation

Fig. 5. Biological effects of transcription errors in eukaryotic cells. (**A**) Error-prone cell lines display a reduced growth rate. (**B**) Error-prone cells display a reduced life span. (**C** and **D**) Deletion of the molecular chaperone *Ydj1* in *Dst1*Δ cells markedly decreases growth rate and life span, indicating that the error-prone cells exhibit proteotoxic stress. Previously, we made similar observations for *rpb9*Δ and *rpb1*^{E1103G} cells (*32*). (**E**) A transcriptome analysis of two error-prone cell lines indicates that 75% of the genes that are overexpressed >2-fold in *rpb1*^{E1103G} cells are also overexpressed in *rpb9*Δ cells. (**F**) A proteomic analysis of two error-prone cell lines indicates that 68% of the proteins that are up-regulated >2-fold in *rpb1*^{E1103G} cells are also up-regulated in *rpb9*Δ cells. (**G**) List of all the genes that are up at the transcriptome level in both error-prone cell lines. Genes that were up-regulated at the protein level as well in both of the error-prone cell lines are listed in red. NS, not significant. (**H** and **I**) Metabolomic analysis of pathways that are up-regulated at the protein and transcriptome level using guanine, citrulline, and kynurenine as examples. Each point represents one biological replicate.

arginosuccinate, two components of the urea cycle regulated by the *Car* and *Put* genes (Fig. 5I and fig. S8). Finally, we detected greatly increased stocks of kynurenine in the error-prone cells, which is directly regulated by BNA2 (biosynthesis of nicotinic acid protein 2), as well as decreased stocks of NAD, nicotinamide, and nicotinamide riboside (Fig. 5J and fig. S9). These metabolites were rarely altered in *upf2*Δ cells (which only display a very small increase in transcription errors) but were altered in *upf2*Δ cells that also lacked *Rpb9*. Together, these experiments provide evidence for the idea that in addition to proteotoxic stress, transcription errors can also lead to widespread changes in the metabolism of eukaryotic cells due to the depletion of vital resources. However, further experimentation is required to fully test this hypothesis and to rule out any alternative explanations, including the possibility that other features of the error-prone alleles drove these phenotypes, such as their use of alternative transcriptional start sites.

DISCUSSION

The genome provides a precise biological blueprint of life. To implement this blueprint correctly, the genome must be transcribed with great precision. Here, we demonstrate that this process is inherently error-prone and that transcription errors can occur in any gene, at any location, and affect every aspect of protein structure and function. In addition, we describe how numerous proteins maintain the fidelity of transcription, including proteins associated with RNAPI, RNAPII, and the NMD. These observations provide the first comprehensive analysis of the fidelity of transcription in eukaryotic cells. Furthermore, with the modified protocol of the circle-sequencing assay we describe here, it will be possible to examine transcriptional fidelity in an even greater detail. For example, by mimicking our analysis of *Rpa12*Δ, $Rpb1^{E1103G}$, *Rpb9*Δ, and *Dst1*Δ cells, it will be possible to identify every gene that controls the fidelity of transcription—for all four major RNA polymerases in eukaryotic cells—in any organism of choice. Similar experiments could determine how age, nutrition, genotype, or exposure to chemicals affects the error rate of transcription or whether transcriptional fidelity is perturbed in the context of human disease. Our experiments also allow new cell types to be studied in the context of mutation research. For example, postmitotic cells tend to resist genetic mutation because they do not undergo DNA replication. As a result, it is thought that most mutations in peptide sequences arise during transcription and translation. With the technology we describe here, it will be possible to define the transcriptional component of these nongenetic mutations for the first time and to understand how this molecular noise affects cellular function. Together, these considerations indicate that our experiments open up a new field of mutagenesis to widespread experimentation. One of the most challenging aspects of this field will be to define the impact of transcription errors on cellular health. Our experiments (*32*), as well as those of others (*28*, *29*), now suggest that transcription errors are particularly detrimental to cellular proteostasis. For example, in patients that suffer from nonfamilial cases of Alzheimer's disease, transcription errors can generate toxic versions of the amyloid precursor protein, whereas similar errors generate mutated versions of the UBB protein (*28*, *29*). In both cases, these errors occur on tracts of GA repeats that are present in the coding regions of the affected genes. These observations suggest that transcription errors can directly contribute to human pathology if they occur repeatedly at the same location. However, in addition to these highly specific transcription errors, it has long been suspected that a much larger population of errors may exist that has thus far evaded detection because it consists of errors that occur randomly throughout the genome. Our experiments now confirm this suspicion and describe the landscape of these errors in great detail. Moreover, we found that these random transcription errors seem to affect proteostasis as well and do so in a way that is complementary to specific transcription errors. More specifically, because most amino acids support the structural integrity of proteins, random transcription errors tend to cause protein misfolding. Accordingly, error-prone cells up-regulate various aspects of the protein quality control machinery to alleviate this stress, which is essential to the health of the error-prone cells. These observations build on the results of a previous study (*32*) in which we used genetic analyses, biochemistry, fluorescence microscopy, proteomics, and electron microscopy to demonstrate a similar effect. We went on to show that by overwhelming the protein quality control machinery, random transcription errors can allow other alleles, which are normally targets of this machinery, to evade degradation (*32*). For example, we found that Aβ(1–42) is degraded less efficiently in cells that display error-prone transcription because the attention of the protein quality control machinery was diverted by the misfolded proteins that were generated by random transcription errors. As a result, Aβ(1–42) aggregated at lower concentrations into more numerous foci in error-prone cells compared to WT cells. Similar observations were made when TDP-43 (transactive response DNA binding protein 43 kDa; which is implicated in amyotrophic lateral sclerosis), Htt103Q (Huntington's disease), and a yeast prion were expressed (*32*). Thus, these observations suggest that transcription errors do not only generate highly specific disease-related peptides but also provide the very conditions that allow these proteins to survive inside cells and seed aggregates. As a result, transcription errors may provide a new mechanism by which the severity, progression, and age of onset of multiple protein misfolding diseases can be affected. Our RNA-seq analysis further suggested that transcription errors could also perturb other biological processes, including nucleotide synthesis, nitrogen metabolism, and tryptophan degradation. An unbiased proteomic screen supported these findings, and a metabolomic analysis subsequently suggested that these processes were modulated to compensate for the loss of vital resources, including purine, nitrogen, and NAD metabolites. Similar to the relationship between transcription errors and proteotoxic stress, we suspect that these observations were the result of countless transcription errors acting together to enable a specific physiological change. For example, because transcription errors cause widespread protein misfolding, they up-regulate several molecular chaperones. In addition to maintaining proteostasis, these chaperones are also involved in the regulation of purinosomes (*34*), protein complexes that seem to control purine biosynthesis, suggesting that the purine-related changes seen in the error-prone cells are indirectly related to reduced proteostasis. Similarly, overexpression of an out-of-frame UBB protein in yeast (which is generated by transcription errors in Alzheimer's patients) modulates *Put1*, arginine, and ornithine availability (*35*), three key components of nitrogen metabolism that were both directly and indirectly implicated in the error-prone cells. Finally, NAD is deeply intertwined with cellular life span (*36*, *37*), suggesting that the reduced life span of the error-prone cell lines may have precipitated altered NAD metabolism. It is important to note that further experiments are required to test these preliminary hypotheses and to rule out alternative explanations, including the possibility that other features of the error-prone alleles drove these phenotypes, such as their use of alternative transcriptional start sites.

In addition to numerous transcription errors acting in concert to influence global biological processes, we hypothesize that it is also possible

for individual errors to affect cellular function, particularly if they occur repeatedly at the same location. For example, repeated transcription errors can activate green fluorescent protein, luciferase, and the oncogenic mitogen-activated protein kinase pathway in cells in culture (*25*, *38*–*40*). In each of these experiments, a precisely placed DNA lesion provoked the same transcription error during multiple transcription events. Accordingly, these observations suggest that until it is repaired, DNA damage can have the same effect on cellular function as a genetic mutation. A recent study further suggests that single-transcription errors can affect cellular function as well. Here, it was shown that a single-transcription error in bacteria can switch the state of a bistable network of genes and cause a heritable change in the fate of the cell. Because transcription errors are ubiquitous throughout the genome and can affect any gene at any location, we suspect that the molecular noise created by these errors could be substantial. An important challenge in the future will be to connect these errors directly to the changes in cellular function and monitor their effect on cellular health. We anticipate that these experiments will ultimately lead to the discovery of a wide range of unexpected phenomena, including new mutagens, new mutational mechanisms, and new disease processes that could help us understand how the environment and our lifestyle choices affect our overall health, as well as our predisposition to diseases that are caused by protein aggregation.

MATERIALS AND METHODS

Cell growth and RNA extraction

Single colonies of each genotype were inoculated in yeast extract/adenine/peptone/dextrose (YAPD) and incubated for approximately 20 hours at 30°C. The optical density (OD_{600nm}) of each culture was measured using a NanoDrop 2000C (Thermo Fisher Scientific), and the cells were reinoculated to an OD_{600} of 0.1 in 50 ml of YAPD. The cells were then reincubated at 30°C until they reached an OD_{600} of 0.8 and harvested by centrifugation. The cells were lysed with the RiboPure Yeast kit from Ambion (PN1926M) according to the manufacturer's protocol, with the exception that the RNA was not exposed to temperatures higher than 70°C. After isolation of total RNA, we purified mRNA with either one or two rounds of poly(A) purification using the Sigma-Aldrich GenElute mRNA Miniprep kit (MRN70-1KT) according to the manufacturer's protocol, again with the exception that the RNA was not exposed to temperatures higher than 70°C and no longer than 2.5 min.

Library preparation

Five hundred nanograms of RNA was fragmented with the NEBNext RNase III RNA Fragmentation Module (E6146S) for 90 min at 37°C. RNA fragments were then purified with an Oligo Clean and Concentrator kit (D4061) by Zymo Research and circularized with RNA ligase 1 (M0204S, NEB) according to the manufacturer's guidelines. These RNA molecules were then reverse-transcribed in a rolling-circle reaction according to the protocol described by Acevedo *et al.* (*12*), with the exception that the incubation time at 42°C was extended from 2 to 20 min. Second-strand synthesis and the remaining steps for library preparation were then performed with the NEBNext Ultra RNA Library Prep kit for Illumina (E7530L, NEB) and the NEBNext Multiplex Oligos for Illumina (E7335S and E7500S, NEB) according to the manufacturer's protocols.

Bioinformatics analysis of circle-sequencing data

We developed a pipeline to analyze RNA-seq reads generated by circle sequencing. Briefly, this pipeline started by identifying repeats within each read based on sequence similarity (minimum repeat size, 30 nucleotide (nt); minimum identity between repeats, 90%). Then, a consensus sequence of the repeat unit was built by summing the quality score of all four possible base calls (A, T, C, or G) from the repeats at each position and retaining the one with the highest total quality score. The next step consisted of identifying the position in the consensus sequence that corresponded to the 5′ end of the RNA fragment (because reverse transcription is randomly primed, the cDNA—and therefore, the read sequence—can start anywhere on the circularized RNA). This was carried out by searching for the longest continuous mapping region in a BLAT mapping of a tandem copy of the consensus sequence against the reference transcriptome. The consensus sequence was then reorganized to start from the identified ligation point (that is, the 5′ end of the original RNA fragment). This reorganized consensus sequence was then mapped against the genome with TopHat (version 2.1.0 with bowtie 2.1.0), and all nonperfect hits went through an algorithm of refining the search for the location of the ligation point before being mapped again. Finally, every mapped nucleotide was inspected and must pass a number of thresholds to be retained: (i) The mapped nucleotide must be supported by at least three repeats from the original sequence reads, (ii) all repeats must support the same base call, (iii) the sum of base call qualities at this position is above 100, (iv) the nucleotide must be more than 5 nt away from the end of the consensus sequence (to minimize false-positives induced by mapping errors), and (iv) the nucleotide must also be at a genomic position covered by at least 20 reads and with less than 5% of these reads supporting a base call different than that of the reference genome (this allows for filtering out polymorphic sites). For each read containing at least one mismatch passing these thresholds, sequences corresponding to all possible versions of the position of the ligation point were generated and mapped against the genome with TopHat. If at least one of these sequences finds a perfect match, then the original read is discarded. This last test only removes a small fraction of the error-containing reads (typically less than 5%), but it ensures that errors in calling the position of the ligation point cannot produce false-positives. Every mapped nucleotide that passes all these thresholds was considered as an event of transcription for which the transcribed nucleotide was known with certainty, and the total transcription error rate was calculated as the number of mismatches divided by the total number of mapped nucleotides that passed all quality thresholds. Because the RNA-seq library preparation used here did not preserve strand information, we relied on the genome annotation (Ensembl, R64-1-1, version 84) to polarize the mismatches (mismatches outside annotates transcripts or in regions where multiple transcripts from opposite strands overlap cannot be polarized).

Yeast growth assay

Single colonies of each genotype were inoculated in YAPD and incubated for approximately 20 hours at 30°C. The OD_{600nm} of each culture was measured using a Nanodrop 2000C (Thermo Fisher Scientific). Cells of each genotype were then diluted to an OD_{600} of 0.01 in a total volume of 200 µl in YAPD in Falcons 96-well plate (reference no. 353075). Quadruple biological replicates were used for each genotype with wells of YAPD as blanks. Growth rate was measured using Molecular Devices' SpectraMax Paradigm Multi-Mode detection platform with the SoftMax Pro 6.3 software. Measurements at 600 nm were taken at every 15-min interval for 24 hours at 30°C and set to orbital shaking at medium intensity for 20 s before the first and between each read. Raw numbers were extracted for data analysis.

Strain list

BY4741: MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0
BY4742: MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0
*Rpb9*Δ: MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0 *rpb9*::KanMX
*Dst1*Δ: MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0 *dst1*::KanMX
$Rpb1^{E1103G}$: MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0 $rpb1^{E1103G}$
*Upf2*Δ: MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0 *upf2*::KanMX
*Rpb9*Δ; *Upf2*Δ: MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0 *rpb9*::KanMX *upf2*::KanMX
*Dst1*Δ; *Upf2*Δ: MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0 *dst1*::KanMX *upf2*::KanMX

All strains, except for the $rpb1^{E1103G}$ strain, were constructed by standard mating and sporulation protocols, with strains commercially available from the MATa and MATα deletion libraries. The $rpb1^{E1103G}$ strain was a gift from J. Strathern and M. Kashlev and was backcrossed 15 times into the BY4741 background by our laboratory and Strathern's laboratory.

Can1 mutation assay

Single colonies were inoculated in 5 ml of YAPD medium for approximately 20 hours at 30°C. These cultures were then washed twice in 1× phosphate-buffered solution (PBS), resuspended in 5 ml of 1× PBS, and incubated at room temperature for 2 hours to deplete intracellular arginine levels. Approximately 2×10^7 cells were then plated onto SC-Arg (synthetic complete medium minus arginine) plates containing canavanine (100 μg/ml) and incubated at 30°C for up to 7 days to allow for canavanine-resistant cells to grow. In addition, cells were serially diluted onto SC plates lacking arginine and canavanine to determine the total number of cells that were plated. All measurements were obtained in at least quadruplicate and analyzed in GraphPad Prism 7 using an unpaired *t* test. In two cases, outliers were identified using the ROUT method with a desired maximum false discovery rate (FDR; *Q*) of 1% to remove samples that contained a "jackpot" mutation.

Protein extraction

Single colonies of each genotype were inoculated in YAPD and incubated for approximately 20 hours at 30°C. The OD_{600nm} of each culture was measured using a Nanodrop 2000C (Thermo Fisher Scientific), and the cells were reinoculated to an OD_{600} of 0.1 in 50 ml of YAPD. The cells were then reincubated at 30°C until they reached an OD_{600} of 0.5 and harvested by centrifugation. Proteins were then extracted with the YPX Yeast Protein Extraction kit from Expedeon (44102) according to the manufacturer's protocol.

Protein hydrolysis

Five hundred micrograms of yeast protein isolates was diluted with 20 mM tris-HCl (pH 8.0) to a final volume of 200 μl, reduced with 5 mM dithiothreitol (40 min at 37°C), and alkylated with 20 mM iodoacetamide (40 min at 37°C). Protein was then precipitated by the addition of four volumes of cold acetone overnight at −20°C. The precipitated samples were spun at 15,000 relative centrifugal force (rcf), and the protein pellet was washed twice with 80% cold acetone. The acetone was removed, and the pellet was dissolved in 200 μl of sodium deoxycholate (SDC) [0.1% SDC and 75 mM tris-HCl (pH 8)]. Trypsin was prepared by dissolving trypsin (catalog no. V5111, Promega) in 50 mM acetic acid at a concentration of 1 μg/μl and adding 10 μg to each sample. After incubation overnight at 37°C, SDC was precipitated by the addition of trifluoroacetic acid (TFA). After centrifugation, the peptides in the supernatant were desalted using an Oasis HLB 96-well plate (particle size, 30 μm; catalog no. 186000128, Waters). Briefly, the Oasis HLB plate was conditioned by the addition of 200 μl of acetonitrile under vacuum at 5 in Hg, equilibrated twice with 200 μl of 0.1% TFA. The peptides from each sample were loaded into each individual well, washed twice with 200 μl of 0.1% TFA, and eluted three times with 100 μl of 80% acetonitrile/0.1% TFA to a 96-well Protein LoBind plate (catalog no. 951032107, Eppendorf). The eluted peptides were transferred to a microtube, lyophilized, and stored at −80°C until further use.

Multidimensional high-performance liquid chromatography–mass spectrometry analysis

First, dimension separation was carried out with an H-Class UPLC instrument (Waters) using a Zorbax 300Å Extend-C18 column (2.1 × 100 mm; 3.5 μm) (catalog no. 76177s02, Agilent). Mobile phases were 2% acetonitrile/5 mM ammonium formate (pH 10) (solvent A) and 90% acetonitrile/5 mM ammonium formate (pH 10) (solvent B). Tryptic peptides were dissolved in solvent A and spun at 20,000 rcf for 5 min. The peptide concentrations were measured by ultraviolet spectrophotometry at 280 nm with an assumed extinction coefficient of 1.1 ml/cm per mg. Fifty micrograms of peptides from each sample was separated (0.3 ml/min at 30°C) using the following gradient (time, % B): 3 min, 0% B; 5 min, 6% B; 12 min, 14% B; 23 min, 26% B; 27 min, 34% B; 28 min, 65% B; 28.1 min, 100% B. The column was equilibrated for 10 min before the next gradient run started. A total of 33 1-min fractions were collected with a 96-well Protein LoBind plate. These were reduced to six fractions by concatenated recombination of every sixth fraction, lyophilized, and dissolved in 0.1 % TFA. The HRM (Hyper Reaction Monitoring) standard (catalog no. Pp-2001, Biognosys) was added to each sample before liquid chromatography–mass spectrometry (LC-MS) analysis.

Tryptic digests were analyzed by LC–tandem MS (LC-MS/MS) on a Q Exactive HF mass spectrometer (Thermo Fisher Scientific) coupled to an UltiMate 3000 RSLCnano UPLC system (Dionex). Peptides were separated by reversed-phase high-performance LC (RP-HPLC) on a nanocapillary column, Acclaim PepMap column (75-μm inside diameter × 25 cm; 2 μm). Mobile phase A consisted of 0.1% formic acid (Thermo Fisher Scientific), and mobile phase B consisted of 0.1% formic acid/acetonitrile. Peptides were eluted into the mass spectrometer at 300 nl/min with each RP-HPLC run comprising a 90-min gradient from 10 to 25% B in 65 min and from 25 to 40% B in 25 min. The mass spectrometer was set to repetitively scan mass/charge ratio (*m*/*z*) from 300 to 1400 (*R* = 240,000), followed by data-dependent MS/MS scans on the 20 most abundant ions, a minimum AGC value of 1×10^4, a dynamic exclusion with a repeat count of 1, a repeat duration of 30 s (*R* = 15,000). The Fourier transform-based MS full-scan AGC target value was 3×10^6, whereas the MSn AGC value was 1×10^5. MSn injection time was 160 ms; microscans were set to 1. Rejection of unassigned and 1+, 6-8 charge states was set.

Raw MS files were processed using MaxQuant (version 1.5.5.1) for the identification of peptides and proteins. The peptide MS/MS spectra were searched against the UniProtKB/Swiss-Prot Yeast Reference Proteome database. Fragment ion tolerance was set to 0.5 Da, with full tryptic specificity required and a maximum of two missed tryptic cleavage sites. Precursor ion tolerance was 7 parts per million. Oxidation of methionine, acetylation of the protein N terminus, and conversion of glutamine to pyroglutamic acid were used as variable modifications, whereas carbamidomethylation of cysteine was set as a fixed modification. The minimal length required for a peptide was seven amino acids. Target-decoy approach was used to control FDR. A maximum FDR of 1% at

both the peptide and the protein level was allowed. Protein groups containing matches to decoy database or contaminant proteins were discarded. The MaxQuant match-between-runs feature was enabled, and iBAQ (intensity based absolute quantification) values were used for quantification.

Metabolomics sample preparation

Samples were prepared as described by Beattie *et al.* (*41*). Briefly, samples were prepared using a MicroLab STAR system from Hamilton Company. For each experiment, numerous recovery standards were added for quality control (QC) purposes. To remove proteins and small molecules and to recover a diverse array of metabolites, proteins were precipitated with methanol by shaking and centrifugation in a GenoGrinder 2000 (Glen Mills). Extracts were then divided into five fractions. Two of these fractions were analyzed by two separate RP/ultrahigh-performance LC-MS/MS (RP/UPLC-MS/MS) methods with positive ion mode electrospray ionization (ESI), one fraction with RP/UPLC-MS/MS with negative ion mode ESI, and one fraction with hydrophilic interaction liquid chromatography (HILIC)/UPLC-MS/MS with negative ion mode ESI. Finally, one fraction was reserved for backup. Organic solvents were removed by a TurboVap (Zymark), and sample extracts were stored overnight in liquid nitrogen before analysis.

Quality assurance/quality control

Several types of controls were analyzed in concert with the experimental samples: A pooled matrix sample generated by taking a small volume of each experimental sample (or alternatively, use of a pool of well-characterized human plasma) served as a technical replicate throughout the data set; extracted water samples served as process blanks; and a cocktail of QC standards that were carefully chosen not to interfere with the measurement of endogenous compounds were spiked into every analyzed sample, allowed instrument performance monitoring, and aided chromatographic alignment. Instrument variability was determined by calculating the median relative SD (RSD) for the standards that were added to each sample before injection into the mass spectrometers. Overall process variability was determined by calculating the median RSD for all endogenous metabolites (that is, noninstrument standards) present in 100% of the pooled matrix samples. Experimental samples were randomized across the platform run, with QC samples spaced evenly among the injections.

UPLC-MS/MS spectroscopy

Similar to the study of Beattie *et al.* (*41*), all four methods used an ACQUITY UPLC instrument (Waters) and a Q-Exactive high-resolution mass spectrometer that was interfaced with a heated ESI source, and an Orbitrap mass analyzer operated at 35,000 mass resolution. Sample extracts were dried and reconstituted in solvents compatible with all four methods. Each solvent contained predetermined standards to ensure consistency from experiment to experiment. One aliquot was analyzed under acidic positive ion conditions and was chromatographically optimized for hydrophilic compounds, whereas another aliquot was optimized for hydrophobic compounds. A third aliquot was analyzed under basic negative ion optimized conditions with separate C18 columns, and extracts were eluted with methanol and water. A fourth aliquot was also analyzed with negative ionization, which was followed by elution from a HILIC column (2.1 × 150 mm; 1.7 μm) with the help of a gradient that consisted of water and acetonitrile with 10 mM ammonium formate (pH 10.8). MS analysis alternated between MS and data-dependent MSn scans with dynamic exclusion, with a scan range that covered 70 to 1000 *m/z*.

Metabolite quantification and data normalization

All peaks were analyzed by quantifying the area under the curve. In addition, we included a data normalization step to correct for possible day-to-day variation resulting from subtle tuning differences in the tuning of instruments. To do so, compounds were corrected by registering the medians to 1.00 and normalizing each data point proportionately between experiments performed on different days. In certain instances, biochemical data may have been normalized to an additional factor [for example, cell counts, total protein (as determined by Bradford assay), osmolality, etc.] to account for differences in metabolite levels due to differences in the amount of material present in each sample.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/3/10/e1701484/DC1

fig. S1. Optimizing the circle-sequencing assay.
fig. S2. The error rate of transcription is not affected by the expression level of a gene.
fig. S3. The error rate of transcription is not affected by the vicinity of a gene to an origin of replication.
fig. S4. The error rate of transcription is equal along the length of a gene.
fig. S5. Cell lines that display error-prone transcription do not exhibit elevated mutation frequencies.
fig. S6. Transcriptional deletion rate in WT and error-prone cell lines.
fig. S7. Multiple components of the purine synthesis and salvage pathways are affected in error-prone cells.
fig. S8. Multiple components of nitrogen metabolism are affected in error-prone cells.
fig. S9. Multiple components of NAD metabolism are affected in error-prone cells.
table S1. Distribution of synonymous, missense, and nonsense errors in WT and error-prone cell lines.
table S2. Genes significantly up-regulated >2-fold at the RNA level in $rpb1^{E1103G}$ cells.
table S3. Genes significantly up-regulated >2-fold at the RNA level in *rpb9*Δ cells.
table S4. Proteins significantly up-regulated >2-fold at the protein level in both $rpb1^{E1103G}$ and *rpb9*Δ cells.
References (*42*, *43*)

REFERENCES AND NOTES

1. L. A. Loeb, R. J. Monnat Jr., DNA polymerases and human disease. *Nat. Rev. Genet.* **9**, 594–604 (2008).
2. H. S. Zaher, R. Green, Fidelity at the molecular level: Lessons from protein synthesis. *Cell* **136**, 746–762 (2009).
3. A. Blank, J. A. Gallant, R. R. Burgess, L. A. Loeb, An RNA polymerase mutant with reduced accuracy of chain elongation. *Biochemistry* **25**, 5920–5928 (1986).
4. L. de Mercoyrol, Y. Corda, C. Job, D. Job, Accuracy of wheat-germ RNA polymerase II. General enzymatic properties and effect of template conformational transition from right-handed *B*-DNA to left-handed *Z*-DNA. *Eur. J. Biochem.* **206**, 49–58 (1992).
5. R. F. Rosenberger, J. Hilton, The frequency of transcriptional and translational errors at nonsense codons in the lac*Z* gene of *Escherichia coli*. *Mol. Gen. Genet.* **191**, 207–212 (1983).
6. R. J. Shaw, N. D. Bonawitz, D. Reines, Use of an in vivo reporter assay to test for transcriptional and translational fidelity in yeast. *J. Biol. Chem.* **277**, 24420–24426 (2002).
7. C. F. Springgate, L. A. Loeb, On the fidelity of transcription by *Escherichia coli* ribonucleic acid polymerase. *J. Mol. Biol.* **97**, 577–591 (1975).
8. L. B. Carey, RNA polymerase errors cause splicing defects and can be regulated by differential expression of RNA polymerase subunits. *eLife* **4**, e09945 (2015).
9. M. Imashimizu, T. Oshima, L. Lubkowska, M. Kashlev, Direct assessment of transcription fidelity by high-resolution RNA sequencing. *Nucleic Acids Res.* **41**, 9090–9104 (2013).
10. J.-F. Gout, W. K. Thomas, Z. Smith, K. Okamoto, M. Lynch, Large-scale detection of in vivo transcription errors. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 18584–18589 (2013).
11. A. J. E. Gordon, D. Satory, J. A. Halliday, C. Herman, Lost in transcription: Transient errors in information transfer. *Curr. Opin. Microbiol.* **24**, 80–87 (2015).
12. A. Acevedo, L. Brodsky, R. Andino, Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505**, 686–690 (2014).
13. A. Acevedo, R. Andino, Library preparation for highly accurate population sequencing of RNA viruses. *Nat. Protoc.* **9**, 1760–1769 (2014).
14. C. C. Traverse, H. Ochman, Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 3311–3316 (2016).

15. M. Lynch, M. S. Ackerman, J.-F. Gout, H. Long, W. Sung, W. K. Thomas, P. L. Foster, Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* **17**, 704–714 (2016).
16. T. A. Kunkel, DNA replication fidelity. *J. Biol. Chem.* **279**, 16895–16898 (2004).
17. C. Walmacq, M. L. Kireeva, J. Irvin, Y. Nedialkov, L. Lubkowska, F. Malagon, J. N. Strathern, M. Kashlev, Rpb9 subunit controls transcription fidelity by delaying NTP sequestration in RNA polymerase II. *J. Biol. Chem.* **284**, 19601–19612 (2009).
18. C. Jeon, K. Agarwal, Fidelity of RNA polymerase II transcription controlled by elongation factor TFIIS. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13677–13682 (1996).
19. P. Cramer, K.-J. Armache, S. Baumli, S. Benkert, F. Brueckner, C. Buchen, G.E. Damsma, S. Dengl, S.R. Geiger, A.J. Jasiak, A. Jawhari, S. Jennebach, T. Kamenski, H. Kettenberger, C.-D. Kuhn, E. Lehmann, K. Leike, J.F. Sydow, A. Vannini, Structure of eukaryotic RNA polymerases. *Annu. Rev. Biophys.* **37**, 337–352 (2008).
20. M. L. Kireeva, Y. A. Nedialkov, G. H. Cremona, Y. A. Purtov, L. Lubkowska, F. Malagon, Z. F. Burton, J. N. Strathern, M. Kashlev, Transient reversal of RNA polymerase II active site closing controls fidelity of transcription elongation. *Mol. Cell* **30**, 557–566 (2008).
21. Y. N. Zhou, L. Lubkowska, M. Hui, C. Court, S. Chen, D. L. Court, J. Strathern, D. J. Jin, M. Kashlev, Isolation and characterization of RNA polymerase *rpoB* mutations that alter transcription slippage during elongation in *Escherichia coli*. *J. Biol. Chem.* **288**, 2700–2710 (2013).
22. A. J. Jasiak, K.-J. Armache, B. Martens, R.-P. Jansen, P. Cramer, Structural biology of RNA polymerase III: Subcomplex C17/25 X-ray structure and 11 subunit enzyme model. *Mol. Cell* **23**, 71–81 (2006).
23. C.-D. Kuhn, S. R. Geiger, S. Baumli, M. Gartmann, J. Gerber, S. Jennebach, T. Mielke, H. Tschochner, R. Beckmann, P. Cramer, Functional architecture of RNA polymerase I. *Cell* **131**, 1260–1272 (2007).
24. J. F. Sydow, F. Brueckner, A. C. M. Cheung, G. E. Damsma, S. Dengl, E. Lehmann, D. Vassylyev, P. Cramer, Structural basis of transcription: Mismatch-specific fidelity mechanisms and paused RNA polymerase II with frayed RNA. *Mol. Cell* **34**, 710–721 (2009).
25. T. T. Saxowsky, P. W. Doetsch, RNA polymerase encounters with DNA damage: Transcription-coupled repair or transcriptional mutagenesis? *Chem. Rev.* **106**, 474–488 (2006).
26. J. N. Strathern, D. J. Jin, D. L. Court, M. Kashlev, Isolation and characterization of transcription fidelity mutants. *Biochim. Biophys. Acta* **1819**, 694–699 (2012).
27. J. Strathern, F. Malagon, J. Irvin, D. Gotte, B. Shafer, M. Kireeva, L. Lubkowska, D. J. Jin, M. Kashlev, The fidelity of transcription: RPB1 (RPO21) mutations that increase transcriptional slippage in S. cerevisiae. *J. Biol. Chem.* **288**, 2689–2699 (2013).
28. F. W. van Leeuwen, E. M. Hol, P. H. Burbach, Mutations in RNA: A first example of molecular misreading in Alzheimer's disease. *Trends Neurosci.* **21**, 331–335 (1998).
29. F. W. van Leeuwen, D. P. V. de Kleijn, H. H. van den Hurk, A. Neubauer, M. A. F. Sonnemans, J. A. Sluijs, S. Köycü, R. D. J. Ramdjielal, A. Salehi, G. J. M. Martens, F. G. Grosveld, J. P. H. Burbach, E. M. Hol, Frameshift mutants of β amyloid precursor protein and ubiquitin-B in Alzheimer's and Down patients. *Science* **279**, 242–247 (1998).
30. O. Isken, L. E. Maquat, Quality control of eukaryotic mRNA: Safeguarding cells from abnormal mRNA function. *Genes Dev.* **21**, 1833–1856 (2007).
31. F. He, A. H. Brown, A. Jacobson, Upf1p, Nmd2p, and Upf3p are interacting components of the yeast nonsense-mediated mRNA decay pathway. *Mol. Cell. Biol.* **17**, 1580–1594 (1997).
32. M. Vermulst, A. S. Denney, M. J. Lang, C.-W. Hung, S. Moore, M. A. Moseley, J. W. Thompson, V. Madden, J. Gauer, K. J. Wolfe, D. W. Summers, J. Schleit, G. L. Sutphin, S. Haroon, A. Holczbauer, J. Caine, J. Jorgenson, D. Cyr, M. Kaeberlein, J. N. Strathern, M. C. Duncan, D. A. Erie, Transcription errors induce proteotoxic stress and shorten cellular lifespan. *Nat. Commun.* **6**, 8065 (2015).
33. S. B. Raj, S. Ramaswamy, B. V. Plapp, Yeast alcohol dehydrogenase structure and catalysis. *Biochemistry* **53**, 5791–5803 (2014).
34. A. M. Pedley, S. J. Benkovic, A new view into the regulation of purine metabolism: The purinosome. *Trends Biochem. Sci.* **42**, 141–154 (2017).
35. R. J. Braun, C. Sommer, C. Leibiger, R. J.G. Gentier, V. I. Dumit, K. Paduch, T. Eisenberg, L. Habernig, G. Trausinger, C. Magnes, T. Pieber, F. Sinner, J. Dengjel, F. W. van Leeuwen, G. Kroemer, F. Madeo, Accumulation of basic amino acids at mitochondria dictates the cytotoxicity of aberrant ubiquitin. *Cell Rep.* **10**, 1557–1571 (2015).
36. M. S. Bonkowski, D. A. Sinclair, Slowing ageing by design: The rise of NAD^+ and sirtuin-activating compounds. *Nat. Rev. Mol. Cell Biol.* **17**, 679–690 (2016).
37. M. B. Schultz, D. A. Sinclair, Why NAD^+ declines during aging: It's destroyed. *Cell Metab.* **23**, 965–966 (2016).
38. T. T. Saxowsky, K. L. Meadows, A. Klungland, P. W. Doetsch, 8-Oxoguanine-mediated transcriptional mutagenesis causes Ras activation in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 18877–18882 (2008).
39. A. Viswanathan, H. J. You, P. W. Doetsch, Phenotypic change caused by transcriptional bypass of uracil in nondividing cells. *Science* **284**, 159–162 (1999).
40. D. Brégeon, Z. A. Doddridge, H. J. You, B. Weiss, P. W. Doetsch, Transcriptional mutagenesis induced by uracil and 8-oxoguanine in *Escherichia coli*. *Mol. Cell* **12**, 959–970 (2003).
41. S. R. Beattie, K. M. K. Mark, A. Thammahong, L. N. A. Ries, S. Dhingra, A. K. Caffrey-Carr, C. Cheng, C. C. Black, P. Bowyer, M. J. Bromley, J. J. Obar, G. H. Goldman, R. A. Cramer, Filamentous fungal carbon catabolite repression supports metabolic plasticity and stress responses essential for disease progression. *PLOS Pathog.* **13**, e1006340 (2017).
42. N. K. Nesser, D. O. Peterson, D. K. Hawley, RNA polymerase II subunit Rpb9 is important for transcriptional fidelity in vivo. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 3268–3273 (2006).
43. H. Koyama, T. Ito, T. Nakanishi, K. Sekimizu, Stimulation of RNA polymerase II transcript cleavage activity contributes to maintain transcriptional fidelity in yeast. *Genes Cells* **12**, 547–559 (2007).

Acknowledgment: We thank D. Taylor for the help in the graphical representation of Fig. 2A and F. van Leeuwen, D. Erie, and M. Kashlev for the critical comments on this manuscript and data representations. We also thank C.-C. Chu for his suggestion of using the rolling-circle method. **Funding:** This work was supported by National Institute on Aging grants R00AG041809 (to M.V.), R01AG054641 (to M.V.), and R35GM122566 (to M.L.); NIH grant ROI-GM036827 to (to M.L.); and NSF Major Research Instrumentation grant DBI-1229361 (to K.T.).
Author contributions: J.-F.G. and W.L wrote and executed the bioinformatic pipeline. J.-F.G., W.L., and M.V. analyzed the sequencing data. W.L. and M.V. optimized the circle-sequencing assay. J.-F.G., W.L, C.F., and M.V. generated the sequencing libraries. C.F. performed the mutation analyses. A.L. and M.V. generated the cell lines and growth curves and performed the proteomics analysis. L.S. and M.V. performed the standard RNA-seq analysis. D.H., H.F., and S.S. generated and analyzed the proteomics data. S.H. and M.V. generated the RNA for RNA-seq. Z.S. and K.T. generated sequencing data. M.V. and M.L. conceived the project. M.V. generated the cell lines, performed the aging assays, harvested the proteins, RNA, and cells for downstream processing, and performed the metabolomic analyses. M.V. and M.L oversaw the project. J.-F.G. and M.V. wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 5 May 2017
Accepted 21 September 2017
Published 20 October 2017
10.1126/sciadv.1701484

Citation: J.-F. Gout, W. Li, C. Fritsch, A. Li, S. Haroon, L. Singh, D. Hua, H. Fazelinia, Z. Smith, S. Seeholzer, K. Thomas, M. Lynch, M. Vermulst, The landscape of transcription errors in eukaryotic cells. *Sci. Adv.* **3**, e1701484 (2017).

Science Advances

The landscape of transcription errors in eukaryotic cells

Jean-Francois Gout, Weiyi Li, Clark Fritsch, Annie Li, Suraiya Haroon, Larry Singh, Ding Hua, Hossein Fazelinia, Zach Smith, Steven Seeholzer, Kelley Thomas, Michael Lynch and Marc Vermulst

Sci Adv **3** (10), e1701484.
DOI: 10.1126/sciadv.1701484

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.