Nascent RNA folding mitigates transcription-associated mutagenesis

Xiaoshu Chen, Jian-Rong Yang, and Jianzhi Zhang

Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109, USA

Transcription is mutagenic, in part because the R-loop formed by the binding of the nascent RNA with its DNA template exposes the nontemplate DNA strand to mutagens and primes unscheduled error-prone DNA synthesis. We hypothesize that strong folding of nascent RNA weakens R-loops and hence decreases mutagenesis. By a yeast forward mutation assay, we show that strengthening RNA folding and reducing R-loop formation by synonymous changes in a reporter gene can lower mutation rate by >80%. This effect is diminished after the overexpression of the gene encoding RNase HI that degrades the RNA in a DNA–RNA hybrid, indicating that the effect is R-loop-dependent. Analysis of genomic data of yeast mutation accumulation lines and human neutral polymorphisms confirms the generality of these findings. This mechanism for local protection of genome integrity is of special importance to highly expressed genes because of their frequent transcription and strong RNA folding, the latter also improves translational fidelity. As a result, strengthening RNA folding simultaneously curtails genotypic and phenotypic mutations.

[Supplemental material is available for this article.]

Because mutation is the ultimate source of genetic variation and evolution, measuring the mutation rate and understanding various mutational mechanisms are of vital importance (Lynch 2010a; Hodgkinson and Eyre-Walker 2011). During transcription, nascent RNA has the potential to anneal back to its template DNA after they both exit the RNA polymerase (RNAP), creating a structure known as the R-loop, consisting of a stable RNA–DNA hybrid and a single-stranded DNA (Fig. 1A). Because the single-stranded DNA is naked, it is subject to increased mutagenesis induced by mutagens. R-loops can also prime unscheduled error-prone DNA synthesis (Aguilera and García-Muse 2012). These and other mechanisms cause transcription-associated mutagenesis (TAM) (Aguilera and García-Muse 2012; Kim and Jinks-Robertson 2012). Although genomic lesions can also be repaired by transcription-coupled repair (TCR) (Hanawalt and Spivak 2008), genome-wide analyses in the bacteria *Escherichia coli* and *Salmonella typhimurium*, the budding yeast *Saccharomyces cerevisiae*, and the human germline have demonstrated that the mutation rate of a gene tends to increase with its expression level, likely because TAM exceeds TCR (Lind and Andersson 2008; Park et al. 2012; Chen and Zhang 2013, 2014).

Because for a given sequence, RNA–RNA duplexes are energetically generally more stable than RNA–DNA hybrids (Lesnik and Freier 1995), it is possible for the nascent RNA to fold on itself, allowing the template DNA strand to anneal back with the nontemplate DNA strand, effectively dissolving the R-loop (Fig. 1B). Indeed, under thermodynamic equilibrium, the greater the RNA folding strength, the more likely that the R-loop is dissolved (Supplemental Fig. S1). Because dissolving the R-loop should reduce TAM, we hypothesize that increased nascent RNA folding decreases the mutation rate of a transcribed region (Fig. 1). We first provide experimental evidence for this hypothesis using a yeast forward mutation assay. We then demonstrate the generality of this hypothesis by genomic analyses of yeast mutation accumulation strains and human intronic DNA polymorphisms. Finally, we discuss the biological implications of this finding.

Results

Yeast forward mutation assay shows that nascent RNA folding mitigates mutagenesis

To test the hypothesis that increased nascent RNA folding reduces mutagenesis, we used the yeast *CAN1* forward mutation assay (Lippert et al. 2011; Takahashi et al. 2011). In media containing the toxic arginine analog canavanine, having a functional *CAN1* gene, which codes for arginine permease, is lethal; whereas *CAN1* null mutants are viable. We synthesized two modified versions of *CAN1* by altering 5% of synonymous sites of the wild-type *CAN1* gene, one with increased and the other with reduced RNA folding, relative to the wild-type (Supplemental Table S1). To our knowledge, no existing experimental method can probe nascent RNA folding in vivo. We thus resorted to computational prediction. Specifically, we quantified the nascent RNA folding strength (F_{RNA}) by the negative of the minimum free energy of the folded structure, estimated computationally by a sliding window approach with various window sizes (see Methods); the higher the F_{RNA} value, the stronger the folding. The wild-type *CAN1* and the two modified versions span a large range of F_{RNA} of all yeast genes (Fig. 2A; Supplemental Fig. S2). The three *CAN1* versions were each coupled with one of two promoters: the endogenous *CAN1* promoter (*pCAN*) and a galactose-regulated *GAL1* promoter (*pGAL*); the latter becomes constitutive and substantially stronger than the former in strains lacking the Gal80 repressor of *pGAL* (Lippert et al. 2011). Hereinafter, the *pCAN-CAN1 gal80* strains and *pGAL-CAN1 gal80* strains are referred to as low- and high-transcription strains, respectively.

Corresponding author: jianzhi@umich.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.195164.115.

26:50–59 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/16; www.genome.org

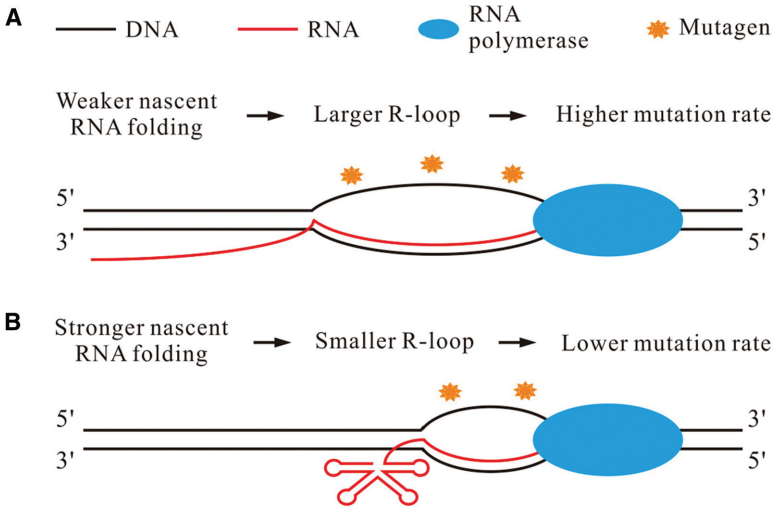

Figure 1. A schematic diagram of the hypothesis that nascent RNA folding mitigates transcription-associated mutagenesis. (*A*) With weaker RNA folding, an R-loop accumulates, which increases the exposure time of the naked nontemplate DNA and error-prone DNA synthesis, leading to a higher mutation rate. (*B*) With stronger RNA folding, the R-loop is dissolved, which reduces the exposure time of the naked nontemplate DNA and error-prone DNA synthesis, resulting in a lowered mutation rate.

To examine whether the weak and strong F_{RNA} versions of *CAN1* have different probabilities of R-loop formation, we used DNA:RNA immunoprecipitation (DRIP) followed by quantitative polymerase chain reaction (PCR). Taking advantage of the high specificity and affinity of the S9.6 monoclonal antibody toward DNA–RNA hybrids of various lengths, DRIP efficiently purifies R-loops. Quantitative PCR is then used to measure the DNA component in the R-loops formed in specific regions of the genome. We probed two nonoverlapping segments of *CAN1* with 91 and 112 nt, respectively (Supplemental Table S1). For each of these segments, the strong F_{RNA} version of *CAN1* has stronger predicted nascent RNA folding than the weak F_{RNA} version. We found that, in both segments, the relative R-loop concentration is significantly lower for the strong F_{RNA} version than the weak F_{RNA} version ($P = 0.002$) (Fig. 2B), as hypothesized (Fig. 1) and computationally predicted (Supplemental Fig. S1).

To compare the mutation rates among the three versions of *CAN1* requires that they have similar mutational target sizes. We confirmed this by estimating the relative probability that a random point mutation is a missense (Supplemental Fig. S3A) or nonsense (Supplemental Fig. S3B) in each version. We also confirmed that these versions have similar numbers of AT/TA and TC/CT dinucleotide repeats (Supplemental Fig. S3C), which are major deletion hotspots (Lippert et al. 2011). Sequencing of canavanine-resistant (CAN^R) mutants detected no significant difference among the three versions of *CAN1* in the fraction of mutations that are insertions/deletions (Supplemental Fig. S4).

We estimated the null mutation frequency of *CAN1* by quantifying the fraction of CAN^R mutants in a cell population after three generations of growth in a nonselective medium, followed by a correction for potential false positives (see Methods). Among the low-transcription strains, the mutation frequency of the weak F_{RNA} strain and that of the strong F_{RNA} strain are 22% higher ($P = 9 \times 10^{-3}$, Mann-Whitney *U* test) and 17% lower ($P = 0.02$), respectively, compared with that of the wild-type strain, which has the intermediate F_{RNA} (Fig. 3A). These mutation frequency differences are in the direction predicted by our hypothesis. Among the high-transcription strains, the mutation frequency of the weak F_{RNA} strain and that of the strong F_{RNA} strain are 36% higher ($P = 6 \times 10^{-5}$) and 76% lower ($P = 6 \times 10^{-14}$), respectively, compared with that of the wild-type strain (Fig. 3B). Thus, the strong F_{RNA} version has a mutation frequency that is only 18% of that of the weak F_{RNA} version ($P = 6 \times 10^{-14}$) (Fig. 3B). We estimated that the mutation frequency of each *CAN1* version is 19–72 times higher in the high-transcription strain than in the low-transcription strain (see Methods), comparable to previous reports (Lippert et al. 2011; Takahashi et al. 2011). Using quantitative reverse transcription PCR (RT-PCR) (see Methods), we confirmed that, for each *CAN1* version, the expression level in the high-transcription strain is 36–50 times that in the low-expression strain (Supplemental Fig. S5), as expected (Takahashi et al. 2011). Interestingly, for the

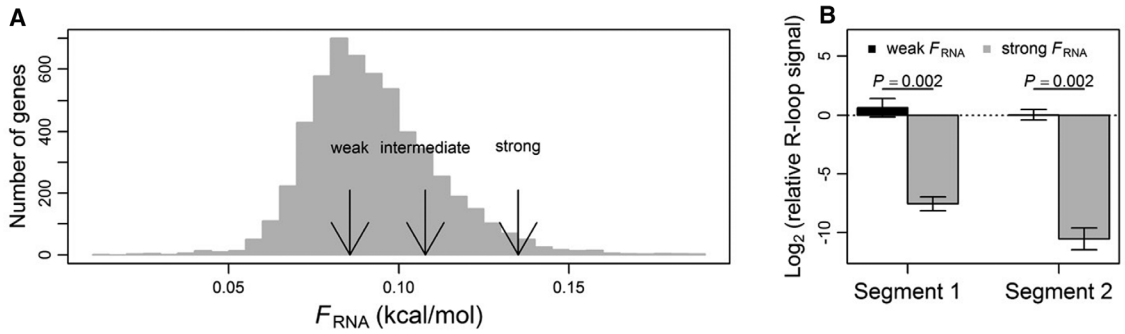

Figure 2. Predicted nascent RNA folding and measured R-loop signals for the three versions of *CAN1*. (*A*) Frequency distribution of the RNA folding strength (F_{RNA}) of all yeast genes. The three versions of *CAN1*, with weak, intermediate (wild-type), and strong F_{RNA} values, respectively, are indicated by arrows. F_{RNA} is computationally predicted using sliding windows of 26 nt and then standardized to a per site value. Computational predictions based on other window sizes are shown in Supplemental Figure S2. (*B*) Experimentally determined R-loop signals, relative to that of *ACT1*, for the weak and strong F_{RNA} versions of *CAN1* in two probed segments. Error bars indicate standard error. *P*-values are based on two-tailed *t*-test.

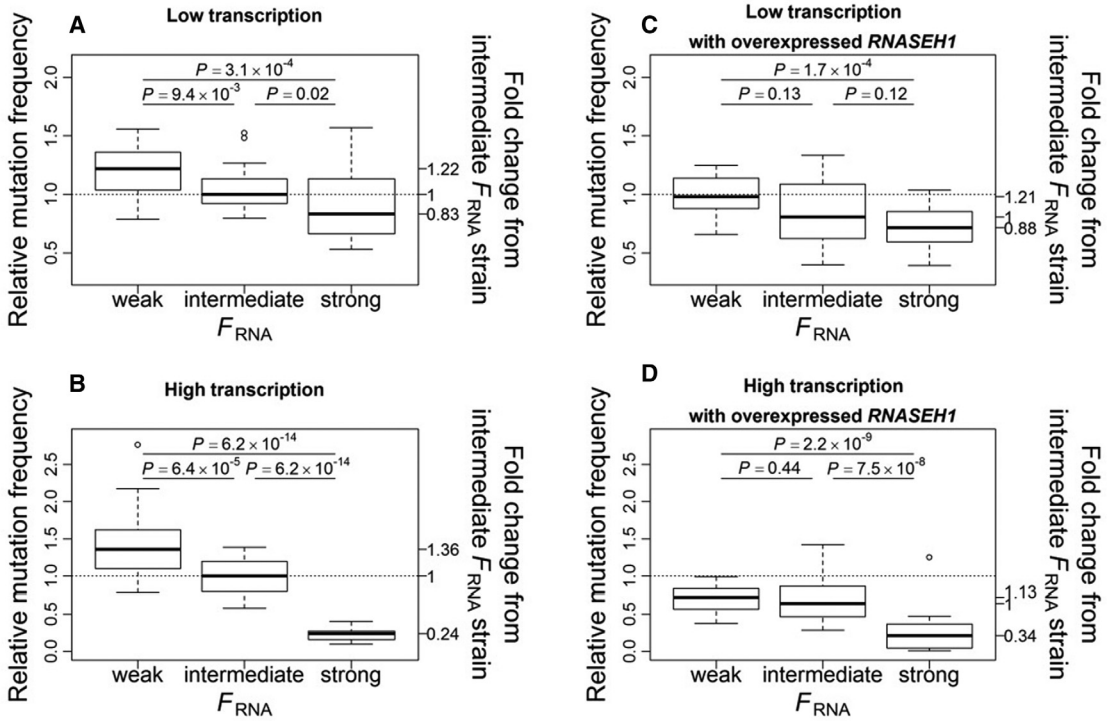

Figure 3. Null mutation frequency at *CAN1* decreases with its nascent RNA folding strength (F_{RNA}). The mutation frequency of a strain is presented relative to that of the strain with the same promoter and wild-type (i.e., intermediate F_{RNA}) *CAN1* without overexpressed *RNASEH1* (dotted line). (*A*) Relative mutation frequencies in low-transcription strains (carrying the promoter *pCAN*). (*B*) Relative mutation frequencies in high-transcription strains (carrying the promoter *pGAL*). (*C*) Relative mutation frequencies in low-transcription strains with overexpressed *RNASEH1*. (*D*) Relative mutation frequencies in high-transcription strains with overexpressed *RNASEH1*. In each panel, the left *y*-axis shows the mutation frequency relative to the dotted line, whereas the right *y*-axis shows the mutation frequency relative to the wild-type *CAN1* in the same panel. The *bottom* and *top* of each box are the first and third quartiles, and the band *inside* the box shows the median. The whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range from the box edges. Circles show outliers, which lie outside the range shown by the whiskers. *P*-values are based on Mann-Whitney *U* tests.

three strains with the same promoter of either *pCAN* or *pGAL*, we observed a monotonic increase in expression level with F_{RNA} (Supplemental Fig. S5). Because transcription is mutagenic, these expression differences render our estimate of the impact of F_{RNA} on mutation frequency conservative. They also make the estimate of the difference in R-loop signal between weak and strong F_{RNA} versions (Fig. 2B) conservative.

TAM has the distinctive feature of higher frequencies at G/C sites than A/T sites (Lippert et al. 2011; Takahashi et al. 2011). If the above mutation rate difference between the strong and weak F_{RNA} versions of *CAN1* is due to TAM rather than other mechanisms, such as different levels of engagement of translesion DNA polymerases (Goodman and Woodgate 2013), the mutation rate per site at G/C sites relative to that at A/T sites (γ) should be higher for the weak F_{RNA} version than the strong F_{RNA} version. Because the exact size of the mutational target for generating CAN^R is unknown, we chose to focus on G/C and A/T sites where point mutations can be nonsense. We sequenced CAN^R mutants from the two versions of *CAN1* under low transcription and found that γ for the weak F_{RNA} version is more than twice that of the strong F_{RNA} version ($P = 0.023$, simulation test) (Table 1), supporting the hypothesis that the higher mutation rate of the weak F_{RNA} version relative to the strong F_{RNA} version is owing to different levels of TAM.

To verify the role of R-loop in the influence of nascent RNA folding on mutagenesis, we inserted into the yeast genome an RNase H1 gene controlled by a strong promoter (see Methods). RNase H1 hampers R-loop formation by degrading the RNA in an RNA–DNA hybrid (Wahba et al. 2011). Under our hypothesized mechanism of the influence of nascent RNA folding on mutagenesis, adding a highly expressed *RNASEH1* should reduce not only the mutation rate but also the impact of nascent RNA folding on mutation rate. Indeed, introducing *RNASEH1* decreased the *CAN1* mutation frequency in all examined strains. Specifically, in the low *CAN1* transcription strains, *RNASEH1* reduced the mutation frequency by 20% ($P = 2 \times 10^{-4}$, Mann-Whitney *U* test), 19% ($P = 9 \times 10^{-3}$), and 14% ($P = 0.02$) in strains with the weak, intermediate, and strong F_{RNA}, respectively (Fig. 3C). In the high *CAN1* transcription strains, *RNASEH1* reduced the mutation frequency by 46% ($P = 1 \times 10^{-11}$), 36% ($P = 2 \times 10^{-4}$), and 11% ($P = 0.46$) for the three versions, respectively (Fig. 3D). Further, after the introduction of *RNASEH1*, the mutation frequency is no longer significantly different between the weak and intermediate F_{RNA} versions in both low-transcription ($P = 0.13$) (Fig. 3C) and high-transcription ($P = 0.44$) (Fig. 3D) strains. Similarly, the mutation frequency is no longer significantly different between the strong and intermediate F_{RNA} versions in low-transcription strains ($P = 0.12$) (Fig. 3C). In the high-transcription strains, although the mutation frequency difference between the strong and intermediate F_{RNA} versions remains significant ($P = 8 \times 10^{-8}$), the difference has shrunk from 4.2-fold (Fig. 3B) to 2.9-fold (Fig. 3D). Together, these experiments strongly suggest that

Table 1. Relative point mutation rates at nonsense target sites of low-transcription *CAN1* without *RNASEH1* overexpression

	Strong F_{RNA} version				Weak F_{RNA} version				
Types of sites	Number of mutations	Number of sites	Mutability (r)	Ratio γ ($r_{G/C}$: $r_{A/T}$)	Number of mutations	Number of sites	Mutability (r)	Ratio γ ($r_{G/C}$: $r_{A/T}$)	*P*-value[a]
G/C sites	67	127	0.528	3.08	86	138	0.623	6.54	
A/T sites	18	105	0.171		10	105	0.095		0.023

[a]Probability that γ from the weak F_{RNA} version is equal to or smaller than that from the strong F_{RNA} version, determined by computer simulation.

nascent RNA folding reduces the mutability at the *CAN1* locus by dissolving R-loops.

Yeast mutation accumulation genomic data support that nascent RNA folding mitigates mutagenesis

To confirm that the impact of F_{RNA} on mutagenesis is not limited to *CAN1*, we analyzed a set of yeast mutation accumulation lines derived from a strain deficient in mismatch repair (Fares et al. 2013). However, it is unknown what window size is most relevant for folding nascent RNAs. Transcription by RNAP II is known to be intermittent (Churchman and Weissman 2011), with rapid elongations interrupted by long pauses that play roles in nascent RNA folding (Pan and Sosnick 2006). It is thus appropriate to fold nascent RNAs using windows corresponding to RNA segments between consecutive pauses. However, reliable information on individual pauses is lacking for many genes. As an approximation, we used yeast native elongating transcript sequencing (NET-seq) data (Churchman and Weissman 2011) to estimate the median distance between pauses in each gene and then estimate the median value across all genes, which should be insensitive to the imprecision of the pause data from individual genes. We found the median to be 26 bases (Supplemental Fig. S6A) and thus computationally estimated F_{RNA} using sliding windows of 26 bases (see Methods). In support of our hypothesis, the mutation rate per site for a gene is significantly negatively correlated with the average F_{RNA} of the gene ($\rho = -0.047$, $P = 5 \times 10^{-4}$) (Fig. 4A), and similar correlations were observed when window sizes of 10–40 bases were used in RNA folding (Supplemental Fig. S7). This correlation remains significant after the control of potential confounding factors such as gene expression level (Park et al. 2012) (partial correlation $\rho = -0.043$, $P = 1 \times 10^{-3}$) (see Fig. 4B for the comparison among a subset of genes with similar expression levels), nucleosome occupancy (Chen et al. 2012) (partial correlation $\rho = -0.051$, $P = 1 \times 10^{-4}$) (see also Fig. 4C), and replication timing (Stamatoyannopoulos et al. 2009; Koren et al. 2010; Lang and Murray 2011) (partial correlation $\rho = -0.046$, $P = 6 \times 10^{-4}$) (see also Fig. 4D).

To verify that the genome-wide signal of the impact of F_{RNA} on mutation rate is mediated by R-loops, we analyzed a microarray-based data set of RNA–DNA hybrid propensity (Chan et al. 2014). Specifically, for each gene, we calculated partial Pearson's correlation between the RNA–DNA hybrid signal (i.e., R-loop score) of a probe and the RNA folding strength of the probe (F_{RNA}) among all probes, after controlling the GC content of the probe, a known confounding factor for microarray intensity signal (Xia 2010). We also calculated the same partial correlation after randomly shuffling the F_{RNA} values of all probes within a gene. We found that the mean correlation for all genes was more negative from the actual data than from the shuffled data in each of 1000 sets of random shuffling ($P < 10^{-3}$) (Fig. 4E), supporting our hypothesis that strong nascent RNA folding weakens R-loop formation.

To verify the mutagenic effect of R-loops at the genomic scale, we obtained the R-loop score for a gene by the intensity of each probe set (see Methods). Because gene expression level affects the microarray-based R-loop score, we correlated the R-loop score of a gene with its mutation rate for a set of genes whose expression levels are within 0.95 and 1.05 times the mean expression level of all genes. We found this correlation ($\rho = 0.119$) to be significantly positive ($P = 0.02$) (Fig. 4F), confirming the mutagenic effect of R-loops. This result was further validated ($\rho = 0.070$, $P = 1 \times 10^{-7}$) using a recently published data set of yeast R-loops based on RNase H targets (El Hage et al. 2014). Together, these analyses in yeast provide genome-wide evidence for the role of nascent RNA folding in reducing mutagenesis by weakening R-loops.

Human population genomic data support that nascent RNA folding mitigates mutagenesis

To examine if the effect of F_{RNA} on mutation rate exists in multicellular organisms, especially humans, we studied human intronic single nucleotide polymorphisms (SNPs) in HapMap (The International HapMap 3 Consortium et al. 2010). We analyzed 542,575 SNPs in 7852 genes that passed various filters for data quality and selective neutrality (see Methods). For each SNP, two non-SNP flanking sites in the same intron were used for comparison, each randomly picked from each side of the SNP with a distance from the SNP between 100 and 200 nt. Because mutation rate is expected to be higher at SNP sites than at non-SNP control sites under the assumption that intron mutations are neutral, our hypothesis predicts that, within a gene, SNP sites have lower F_{RNA} than non-SNP control sites. F_{RNA} at each nucleotide was calculated with 48-base sliding windows, because human global run-on sequencing (GRO-seq) data (Core et al. 2008) showed a median distance of 48 bases between pauses of RNAP II in introns (Supplemental Fig. S6B). We found 1159 genes to support our prediction at the nominal *P*-value of 0.05 (Mann-Whitney *U* test), compared to 252 genes that are against our prediction; the former is significantly greater than the latter ($P < 10^{-99}$, binomial test). To further evaluate the relationship between human nascent RNA folding and mutation rate, we constructed a 2×2 table for each gene by respectively classifying its SNPs and non-SNP control sites into two categories based on whether their F_{RNA} values are higher or lower than the mean F_{RNA} of all SNPs and non-SNP control sites of the gene that was analyzed. We then calculated an odds ratio (OR_1) from the table (see Methods); a gene is supportive of our hypothesis if its OR_1 is lower than 1. We combined the OR_1s from all genes using the Mantel-Haenszel (MH) procedure and found the overall OR_1 to be significantly smaller than 1 ($P < 10^{-99}$) (Fig.

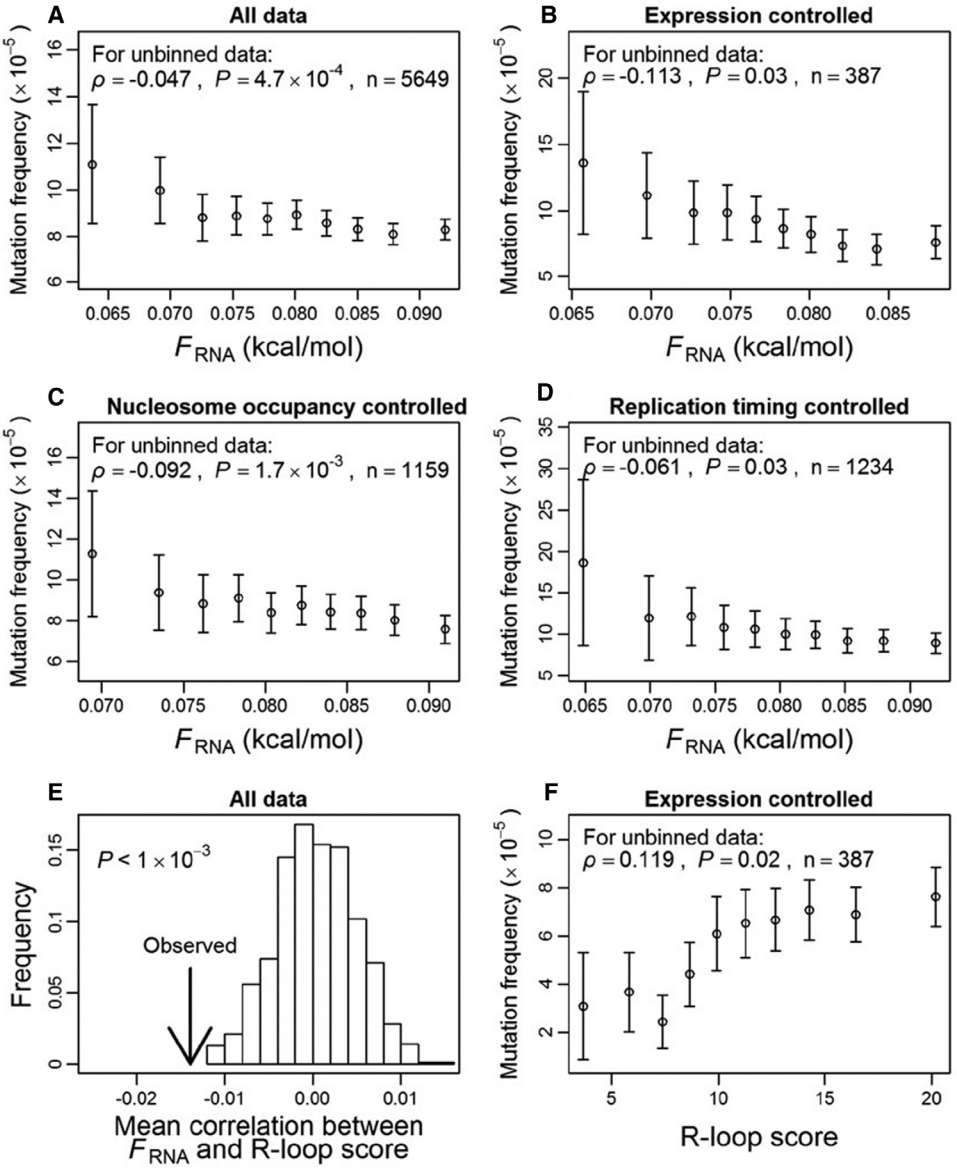

Figure 4. Genome-wide correlations among the mutation rate of a gene, its nascent RNA folding strength (F_{RNA}), and R-loop score in yeast. For each gene, mutation rate per site is estimated from a set of mutation accumulation lines. (*A*) Mutation rate decreases with F_{RNA}. (*B*) Mutation rate decreases with F_{RNA} for the subset of genes whose expression levels are between 0.95 and 1.05 times the mean expression level of all genes. (*C*) Mutation rate decreases with F_{RNA} for the subset of genes whose nucleosome occupancy levels are between 0.95 and 1.05 times the mean nucleosome occupancy level of all genes. (*D*) Mutation rate decreases with F_{RNA} for the subset of genes whose replication timings are between −0.2 and 0.2 (in a standard normal distribution). (*E*) The among-gene average of the within-gene partial Pearson's correlation between F_{RNA} and R-loop score after controlling for GC content is significantly more negative than the random expectation. The arrow indicates the actual observation, whereas the bars show the frequency distribution of the corresponding value derived from 1000 sets of genes with F_{RNA} values randomly shuffled within genes. (*F*) Mutation rate increases with R-loop score for the subset of genes whose expression levels are between 0.95 and 1.05 times the mean expression level of all genes. In *A*–*D*, dots from left to right, respectively, contain the 10%, 20%, 30%, …, and 100% of genes with the lowest F_{RNA} values. In *F*, dots from left to right, respectively, contain the 10%, 20%, 30%, …, and 100% of genes with the lowest R-loop scores. In all panels, error bars show one standard error.

5A), supporting that strong nascent RNA folding reduces human germline mutation rate. A similar trend was observed when G/C and A/T sites were analyzed separately (Fig. 5A). Because mutation rates were compared between intronic SNPs and their flanking control sites, other potentially confounding factors such as the replication timing and expression level were automatically controlled.

To confirm that the impact of nascent RNA folding on human mutation rate is R-loop-dependent, we estimated human R-loop scores using DNA–RNA immunoprecipitation sequencing data (Ginno et al. 2013). We correlated F_{RNA} with R-loop score across intronic SNPs and non-SNP control sites in each gene. The same correlation was also calculated after we randomly shuffled F_{RNA} among all these sites within the gene. The mean correlation of all genes from the actual data is significantly more negative than the expectation derived from 1000 sets of randomly shuffled data ($P = 0.037$) (Fig. 5B), supporting that strong nascent RNA folding weakens R-loops.

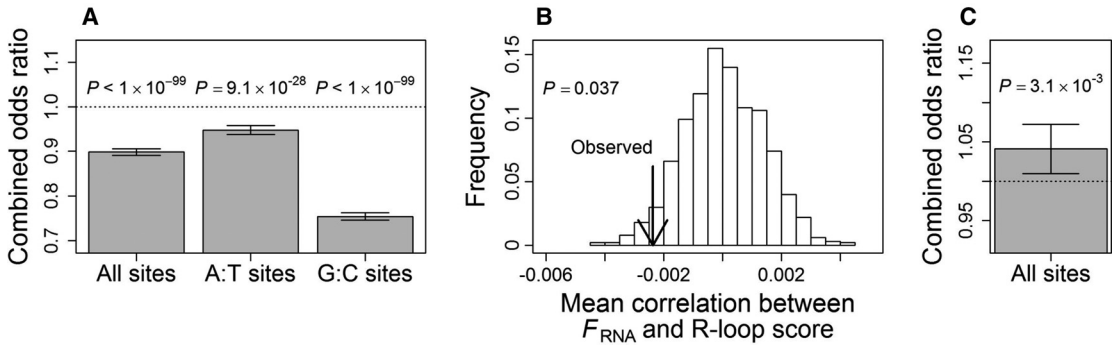

Figure 5. Human population genomic data of introns reveal the negative impact of nascent RNA folding strength (F_{RNA}) on R-loop formation and mutation rate. SNPs are considered to have higher mutation rates than non-SNP control sites. (*A*) Evidence for a negative impact of F_{RNA} on mutation rates at all sites, A/T sites, and G/C sites. Odds ratio <1 indicates that SNP sites have lower F_{RNA} than non-SNP control sites within a gene. Shown are the results combined from all genes by the MH procedure. Error bars represent 95% confidence intervals estimated by bootstrapping the genes 1000 times. The dotted line indicates an odds ratio of 1. (*B*) The among-gene average of the within-gene Pearson's correlation between F_{RNA} and R-loop score is significantly more negative than the random expectation. The arrow indicates the actual observation, whereas the bars show the frequency distribution of the corresponding value derived from 1000 sets of genes with F_{RNA} values randomly shuffled within genes. (*C*) Evidence for a positive impact of R-loop on mutation rate. Odds ratio >1 indicates that SNP sites have higher R-loop scores than non-SNP control sites within a gene. Shown are the results combined from all genes by the MH procedure. Error bars represent 95% confidence intervals, estimated by bootstrapping the genes 1000 times. The dotted line indicates an odds ratio of 1.

To verify that weakening R-loops decreases mutation rate, we constructed a 2 × 2 table for each gene by respectively classifying the intronic SNPs and non-SNP control sites based on whether their R-loop scores exceed the mean value of all of these sites. An odds ratio (OR_2) is then calculated (see Methods); the higher the OR_2 relative to 1, the stronger the evidence for our hypothesis. Indeed, the combined OR_2 from all genes significantly exceeds 1 ($P = 3 \times 10^{-3}$) (Fig. 5C), supporting that weakening R-loops reduces mutagenesis. Note that, due to the lack of R-loop data from the human germline, we used the R-loop data from a pluripotent human embryonal carcinoma cell line (Ntera2). However, the fact that a significant correlation still exists between this R-loop signal and the mutation rate estimated from polymorphisms suggests that the true correlation between mutation rate and germline R-loop signal is likely stronger.

Discussion

Taken together, our forward mutation assay in *CAN1* and genome-wide analyses of yeast mutation accumulation lines and human intronic SNPs provide consistent evidence that strong nascent RNA folding during transcription reduces the mutation rate of the transcribed region. Thus, the mutagenesis of DNA is modulated locally, not only by the biochemical activities of DNA such as replication and transcription, but also those of its transcriptional product. Whether RNA-binding proteins also play a role in this process remains to be studied. Note that although strong nascent RNA folding would predict strong folding of the nontemplate ssDNA, our findings are not attributable to ssDNA folding because a previous study did not find ssDNA folding to reduce TAM in yeast (Park et al. 2012). Because of the difficulty in determining nascent RNA folding experimentally, we resorted to computational predictions with fixed window sizes, which are at best moderately accurate (Park et al. 2013). This inaccuracy explains at least in part why many of the genomic correlations detected are small in magnitude and suggests that the true impact of F_{RNA} on mutation rate is greater, as demonstrated in the *CAN1* case.

When comparing the nascent RNA folding strength among genes, we discovered that highly expressed genes tend to have strong folding in both yeast ($\rho = 0.167$, $P = 1 \times 10^{-36}$) and human ($\rho = 0.261$, $P < 10^{-99}$). This correlation could have resulted from differential natural selection for minimized mutational load, because highly expressed genes are functionally more important (Zhang and He 2005) and subject to severer TAM (Park et al. 2012) than lowly expressed genes. Nonetheless, selection for reduced mutational load is extremely weak (Chen and Zhang 2013), and our calculation indicates that the selective advantage of a single-nucleotide change that reduces the mutation rate at a few sites by enhancing local nascent RNA folding is one to several orders of magnitude smaller than what natural selection can detect in yeast and human (see Methods). Strong nascent RNA folding is known to alleviate the backtracking of RNAP II during transcription and thus may enhance the speed of RNA synthesis (Zamft et al. 2012), but it may also decrease splicing efficiency (Braberg et al. 2013). Hence, these potential effects do not unambiguously explain why highly expressed genes should have stronger nascent RNA folding. Interestingly, it was recently discovered that strong mRNA folding is selected for in highly expressed genes (Zur and Tuller 2012; Park et al. 2013), likely because it enhances translational accuracy (Yang et al. 2014). Having a high translational accuracy could reduce the harm arising from translational error-associated protein misfolding, misinteraction, and energy waste (Drummond and Wilke 2008; Yang et al. 2012, 2014; Zhang and Yang 2015) and is of special importance to abundantly produced proteins simply due to their large amounts. Calculations suggested that differential selection for translational accuracy can lead to stronger mRNA folding for more highly expressed genes (Yang et al. 2014). Because mRNA folding strength and nascent RNA folding strength are positively correlated (yeast: $\rho = 0.549$, $P < 10^{-99}$; human: $\rho = 0.220$, $P < 10^{-99}$), it is likely that selection for translational accuracy results in enhanced mRNA folding, of which strengthened nascent RNA folding is a byproduct. Although the relatively strong nascent RNA folding of highly expressed genes may have been fortuitous, this property does enlarge the benefit of nascent RNA folding in lowering the mutational load. To the best of our knowledge, this is the first example in which a demand for accurate processing of genetic information (i.e., translational fidelity) also results in accurate transmission of the information (i.e.,

low mutation rate); and RNA folding is the first known mechanism that simultaneously and concordantly regulates the genotypic mutation rate and phenotypic mutation rate (Bürger et al. 2006). The full biological ramifications of the intriguing coupling between the processing and transmission accuracies of genetic information via RNA folding await further explorations.

Methods

Computational prediction of nascent RNA folding strength

The secondary structures of yeast RNAs were predicted using RNAfold (http://www.tbi.univie.ac.at/~ivo/RNA/) (Hofacker 2009) under 30°C, following a previous study (Park et al. 2013) with minor modifications. Our prediction used overlapping windows with a window size of L nucleotides and a step size of $S = 1$ nucleotide (Lange et al. 2012). The folding strength of a window is defined by $-\Delta G$, the negative of the minimum free energy of the folded RNA. The folding strength for a nucleotide is calculated by the mean folding strength of all windows covering the nucleotide divided by L. The folding strength of an RNA is the mean folding strength of all nucleotides of the RNA. In the yeast RNA–DNA hybrid microarray, the folding strength of a probe is simply the negative of the minimum free energy of the folded 25-nt RNA sequence for the probe divided by 25. The secondary structures of human RNAs were predicted in the same way as yeast RNAs, except that a different L was used under 37°C. To predict nascent RNA folding, we used $L = 26$ for yeast, 48 for human introns, and 21 for human exons, unless otherwise noted. They represent the genome-wide median distances between RNAP II pauses in yeast, human introns, and human exons, respectively.

Preference of RNA folding over R-loop formation

To quantify the thermodynamic preference of nascent RNA folding over R-loop formation during transcription, we used a sliding window of $L = 26$ nt and $S = 1$ nt to scan each of the yeast intronless genes. Within each window, we used RNAfold (Hofacker 2009) to predict the minimum free energies of RNA folding (ΔG_1) (Mathews et al. 2004), DNA duplex (ΔG_2) (SantaLucia and Hicks 2004; Turner and Mathews 2010), RNA/DNA hybrid duplex (ΔG_3) (Lorenz et al. 2012), and folding of the transcribed (nontemplate) ssDNA (ΔG_4) (SantaLucia and Hicks 2004; Turner and Mathews 2010) under 30°C. The preference of RNA folding over R-loop is measured by $\Delta G_3 + \Delta G_4 - \Delta G_1 - \Delta G_2$, which is presented as per-site minimum free energy along with F_{RNA} ($= -\Delta G_1$) in Supplemental Figure S1.

Modified versions of *CAN1* and strain construction

We changed 5% of synonymous sites in the coding region of the wild-type *CAN1* to create a strong F_{RNA} version and a weak F_{RNA} version of *CAN1*, respectively. The details of the design of these two sequences are provided in Supplemental Methods. The haploid *S. cerevisiae* strain BY4741 was used for *CAN1* forward mutation assays. See Supplemental Methods for experimental details of strain construction.

DRIP followed by quantitative PCR

The DRIP experiment was performed following a published protocol (El Hage et al. 2014) with minor modifications. Quantitative PCR was performed on a 7500 Fast Real-Time PCR System (Applied Biosystems). The experimental protocol is detailed in Supplemental Methods.

Forward mutation assay

The assay follows the standard protocol (Lippert et al. 2011; Takahashi et al. 2011) with modifications. Experimental details are provided in Supplemental Methods.

Sequencing CAN^R colonies

Seven CAN^R colonies from each of three replicated cultures of each strain were isolated and directly amplified by PCR (Ex Taq, Takara). The amplified fragments were Sanger sequenced. Primer sequences are shown in Supplemental Table S2. Identical mutations observed more than once in a single culture were counted only once to avoid multiple counting of the same mutation when estimating the fraction of mutations that are insertions/deletions (indels). False positives are colonies in which the *CAN1* sequence bears no mutation. We found no false positives among low-transcription colonies, but some among high-transcription colonies. The higher false positive rate in high-transcription strains was probably because these strains had a high concentration of CAN1 (see the next section) and thus accumulated a high concentration of cellular arginine before being transferred to the selective medium, where some cells with the functional *CAN1* may still be able to grow into colonies owing to the availability of a large amount of cellular arginine (Gong 2008). The false positives do not affect our conclusion, because our conclusion is based on comparisons among strains carrying the same promoter. Given our observation that the expression level of *CAN1* under the control of the same promoter increases slightly with F_{RNA} (see the next section), our assumption of the same true positive rate across the three strains carrying *pGAL* probably causes an overestimation of the mutation frequency in the strain with strong F_{RNA}, rendering our conclusion of the negative impact of nascent RNA folding on mutagenesis conservative. The absence of false positive colonies of low-transcription *CAN1* strains verified the presumption that even low transcription of wild-type *CAN1* is lethal.

To better estimate the fraction of mutations that are indels in the three F_{RNA} versions of *CAN1* in high-transcription strains, strains were grown in YPGE for seven generations to increase the number of mutations. Seven CAN^R colonies from each of three replicated cultures of each strain were isolated, directly amplified by PCR as above, and Sanger sequenced.

Quantifying *CAN1* expression level by quantitative RT-PCR

The expression level of *CAN1* was measured using quantitative RT-PCR, with experimental details provided in Supplemental Methods.

Relative mutation rates at G/C and A/T sites

To compare the mutation rate at G/C sites relative to that at A/T sites between the strong and weak F_{RNA} versions of *CAN1*, we used low-transcription strains (due to their zero false positive rates and relatively low fractions of mutations that are indels) without overexpressing *RNASEH1*. Three CAN^R colonies from each of 112 replicated cultures of each strain were isolated, directly amplified by PCR as described above, and Sanger sequenced. Only nonsense point mutations were considered. Mutational target size for G/C sites (or A/T sites) was the number of G/C sites (or A/T sites) where a point mutation could be nonsense. We respectively calculated the mutability at G/C sites ($r_{G/C}$) and A/T sites ($r_{A/T}$) for each *CAN1* version, and then computed their ratio $\gamma = r_{G/C}:r_{A/T}$.

To test the null hypothesis that the ratio in γ between the weak and strong versions of *CAN1* is equal to or smaller than 1, we conducted a computer simulation. Let the mutational target

size at G/C sites and A/T sites be a and b for the strong F_{RNA} version, and c and d for the weak F_{RNA} version, respectively. We first generated a binomial random number x following B(a, $r_{G/C}$) and another random number y following B(b, $r_{A/T}$), where $r_{G/C}$ and $r_{A/T}$ are both from the strong F_{RNA} version. We then generated two binomial random numbers, z and w, following B(c, $r_{G/C}$) and B(d, $r_{A/T}$), respectively, where $r_{G/C}$ and $r_{A/T}$ are from the weak F_{RNA} version. We then calculated $g = [(z/c)/(w/d)]/[(x/a)/(y/b)] = (yzad)/(xwbc)$. We repeated this process 10,000 times and calculated the fraction of times (P) in which $g \leq 1$. P is the probability that the null hypothesis is true.

Mutation rates in yeast mutation accumulation (MA) lines

Fares and colleagues accumulated mutations in several lines of a mismatch repair deficient *S. cerevisiae* strain (BY4741; *Mata; his3D1; leu2D0; met15D0; ura3D0; msh2::kanMX4*) for approximately 2200 generations by 100 plate-to-plate passages of single colonies (Fares et al. 2013). In the final generation, a total of 1003 base substitutions were detected in the genomes, of which 691 affected protein-coding regions. The identified mutations in the MA lines were previously mapped to the *S. cerevisiae* genome in Ensembl version 59 (Cunningham et al. 2015) by Fares and colleagues (Fares et al. 2013). The upstream 17 nt, the mutation itself, and the downstream 17 nt were extracted from Ensembl and remapped to the S288C genome at *Saccharomyces* Genome Database (SGD, February 2011) using SOAP (Li et al. 2008) with perfect matches. The mutation rate of a gene was estimated by the total number of mutations identified in the gene divided by the gene length.

Mutation rates in humans

We used HapMap release 27 (The International HapMap 3 Consortium et al. 2010), corresponding to Ensembl human genome version 54, to identify intronic SNPs. To ensure the data quality and selective neutrality of the genomic regions considered, we applied the following filters: First, SNPs should be in an intron sandwiched by two constitutive exons (based on the presence of the exons in all transcripts of the gene annotated in Ensembl), and should not be in any overlapping region of multiple genes; second, SNPs should not be within 200 nt from any other SNP; third, SNPs should not be within any 33-nt sequence whose genomic origin cannot be unambiguously determined by SOAP (Li et al. 2008) (i.e., repeats); and fourth, genes with less than 10 qualified SNPs were not considered. In the end, 7852 genes with 542,575 SNPs passed the preceding filters and were used in the analysis.

Distances between pauses of RNAP II

We used yeast native elongating transcript (NET-seq) data (Churchman and Weissman 2011) and human global run-on sequencing (GRO-seq) data (Core et al. 2008) to infer distances between pauses of RNAP II. Details of the computational analysis are provided in Supplemental Methods.

R-loop scores

Yeast R-loop scores were estimated using DNA–RNA immunoprecipitation tiling microarray (DRIP-chip) data (Chan et al. 2014) and R-loop data based on RNase H targets (El Hage et al. 2014). Human R-loop scores were estimated using human DNA–RNA immunoprecipitation sequencing (DRIP-seq) data (Ginno et al. 2013). Detailed computational analysis is provided in Supplemental Methods.

Nucleosome occupancy

We used the DNA micrococcal-nuclease-digested sequencing (MNase-seq) data (Weiner et al. 2010) to estimate nucleosome occupancy at each nucleotide in yeast. The protocol of our computational analysis is provided in Supplemental Methods.

Replication timing

DNA replication timing data from yeast (Koren et al. 2010) were analyzed. Details of the computational analysis are provided in Supplemental Methods.

Experimental mRNA folding strength data

The in vitro mRNA folding strength data (Kertesz et al. 2010) generated from *S. cerevisiae* strain S288C in YPD at 30°C were downloaded from NCBI (accession numbers: SRR066400–SRR066405, SRR066398–SRR066399, and SRR063372–SRR063374). The procedure of genome masking and short read alignment was the same as in the analysis of the NET-seq data. After a read is mapped to a transcript, the site in the transcript that is 1 nt upstream of the site aligned to the 5′ most nucleotide of the read is considered to have received a hit. Let the number of hits received by a nucleotide under RNase V1 treatment be N_V and the corresponding number under RNase S1 treatment be N_S. According to the original report (Kertesz et al. 2010), the mRNA folding strength of a site was defined by PARS = $\log_2(N_V/N_S)$. The higher the PARS value, the stronger the mRNA folding for the site. The mRNA folding strength of a gene was defined by the mean PARS of all of its nucleotides. Overlapping regions between multiple genes were excluded.

Human in vitro mRNA folding data (Wan et al. 2014) generated from GM12878, GM12891, and GM12892 cell lines were download from NCBI (accession number: GSE50676). Let the total number of reads from these cell lines mapped to a nucleotide be N_V and N_S under RNase V1 and RNase S1 treatments, respectively. The mRNA folding strength of a site was defined by PARS = $(N_V + 1)/(N_S + 1)$, because some sites have $N_S = 0$. The site was ignored if $N_V + N_S = 0$. The overall folding strength for the mRNA is $\log_2$(mean PARS), where mean PARS is the average PARS of all considered sites. In each gene, only the longest transcript was considered.

To examine the relationship between mRNA folding strength and nascent RNA folding strength, we correlated mRNA folding strength with F_{RNA} for yeast genes as well as human exons.

Yeast and human transcriptome data

Yeast RNA sequencing (RNA-seq) data (Nagalakshmi et al. 2008) were generated from *S. cerevisiae* strain BY4741 in YPD at 30°C. Human RNA-seq data (Mortazavi et al. 2008) were generated from the testis. In these two RNA-seq data sets, gene expression is measured in reads per kilobase of exon model per million mapped reads (RPKM).

Odds ratios

We defined and calculated two odds ratios. To calculate OR_1, a 2 × 2 table was constructed for each gene by respectively classifying its SNPs and non-SNP control sites into two categories based on whether their F_{RNA} values are higher or lower than the mean F_{RNA} of all intronic SNPs and non-SNP control sites of the gene. Let the numbers of sites that fall into the four categories be: a_1 (SNPs with F_{RNA} > mean); b_1 (SNPs with $F_{RNA} \leq$ mean); c_1 (non-SNP control sites with F_{RNA} > mean); and d_1 (non-SNP control sites with $F_{RNA} \leq$ mean), respectively. $OR_1 = (a_1d_1)/(b_1c_1)$; thus, $OR_1 < 1$ if strong nascent RNA folding reduces mutation rate.

To calculate OR_2, a 2 × 2 table was constructed for each gene by respectively classifying its SNPs and non-SNP control sites into two categories based on whether their R-loop scores are higher or lower than the mean R-loop score of all SNPs and non-SNP control sites of the gene. Let the numbers of sites that fall in the four categories be: a_2 (SNPs with R-loop scores > mean); b_2 (SNPs with R-loop scores ≤ mean); c_2 (non-SNP control sites with R-loop scores > mean); and d_2 (non-SNP control sites with R-loop scores ≤ mean), respectively. $OR_2 = (a_2d_2)/(b_2c_2)$; thus, $OR_2 > 1$ if weak R-loops decrease mutation rate.

Selective advantage of nascent RNA folding for reducing mutational load

Because nascent RNA folding affects the mutation rate only locally, it is reasonable to assume no recombination between an antimutator and its target (i.e., sites where the mutation rate is reduced). Under no recombination, the fitness advantage (k) conferred by an antimutator approximates the reduction in deleterious mutation rate of its target ($\Delta\mu_d$) (Kimura 1967; Lynch 2011). The per-generation mutation rate in yeast is on average 3.3×10^{-10} per nucleotide (Lynch et al. 2008). Assuming that (1) a nucleotide change that increases local nascent RNA folding can reduce the mutation rate of a target of 10 sites by 80%; (2) the fraction of mutation that is deleterious at the target is 75%; and (3) the mean mutation rate at the target is 10 times the genomic average, we can estimate that $\Delta\mu_d = 3.3 \times 10^{-10} \times 10 \times 10 \times 0.75 \times 0.8 = 2.0 \times 10^{-8}$. Note that because the assumptions made here are quite extreme, the actual benefit is likely to be smaller. Because $\Delta\mu_d$ is much smaller than 10^{-7}, the inverse of the effective population size of yeast (Wagner 2005), the benefit is too small to be detectable by natural selection (Kimura 1983). For humans, the mutation rate is on average 1.3×10^{-8} per nucleotide per generation (Lynch 2010b). Using the same assumptions made for yeast, we can estimate that $\Delta\mu_d = 1.3 \times 10^{-8} \times 10 \times 10 \times 0.75 \times 0.8 = 7.8 \times 10^{-7}$. This tiny benefit is much smaller than 10^{-4}, the inverse of the human effective population size (Takahata 1993). These analyses suggest that it is unlikely for natural selection to enhance nascent RNA folding for the benefit of reducing mutational load. Thus, selection for reduced mutational load is unable to generate differential nascent RNA folding strengths among genes with different expression levels.

Data access

The DNA sequences from this study have been submitted to NCBI GenBank (http://www.ncbi.nlm.nih.gov/genbank/) under accession numbers KT852337–KT852338.

Acknowledgments

The *RNASEH1* overexpression plasmid was kindly provided by the Douglas Koshland laboratory. We thank Calum Maclean, Wenfeng Qian, and three anonymous reviewers for constructive comments. This work was supported in part by a US National Institutes of Health grant R01GM103232 and a US National Science Foundation grant MCB-1329578 to J.Z.

References

Aguilera A, García-Muse T. 2012. R loops: from transcription byproducts to threats to genome stability. *Mol Cell* **46:** 115–124.

Braberg H, Jin H, Moehle EA, Chan YA, Wang S, Shales M, Benschop JJ, Morris JH, Qiu C, Hu F, et al. 2013. From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II. *Cell* **154:** 775–788.

Bürger R, Willensdorfer M, Nowak MA. 2006. Why are phenotypic mutation rates much higher than genotypic mutation rates? *Genetics* **172:** 197–206.

Chan YA, Aristizabal MJ, Lu PY, Luo Z, Hamza A, Kobor MS, Stirling PC, Hieter P. 2014. Genome-wide profiling of yeast DNA:RNA hybrid prone sites with DRIP-chip. *PLoS Genet* **10:** e1004288.

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

Chen X, Zhang J. 2014. Yeast mutation accumulation experiment supports elevated mutation rates at highly transcribed sites. *Proc Natl Acad Sci* **111:** E4062.

Chen X, Chen Z, Chen H, Su Z, Yang J, Lin F, Shi S, He X. 2012. Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science* **335:** 1235–1238.

Churchman LS, Weissman JS. 2011. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469:** 368–373.

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322:** 1845–1848.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2015. Ensembl 2015. *Nucleic Acids Res* **43(**Database issue**):** D662–D669.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134:** 341–352.

El Hage A, Webb S, Kerr A, Tollervey D. 2014. Genome-wide distribution of RNA-DNA hybrids identifies RNase H targets in tRNA genes, retrotransposons and mitochondria. *PLoS Genet* **10:** e1004716.

Fares MA, Keane OM, Toft C, Carretero-Paulet L, Jones GW. 2013. The roles of whole-genome and small-scale duplications in the functional specialization of *Saccharomyces cerevisiae* genes. *PLoS Genet* **9:** e1003176.

Ginno PA, Lim YW, Lott PL, Korf I, Chédin F. 2013. GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Res* **23:** 1590–1600.

Gong J. 2008. "A systematic screen of the *Saccharomyces cerevisiae* deletion mutant collection for novel genes required for DNA damage-induced mutagenesis." PhD thesis, University of North Texas Health Science Center, Fort Worth.

Goodman MF, Woodgate R. 2013. Translesion DNA polymerases. *Cold Spring Harb Perspect Biol* **5:** a010363.

Hanawalt PC, Spivak G. 2008. Transcription-coupled DNA repair: two decades of progress and surprises. *Nat Rev Mol Cell Biol* **9:** 958–970.

Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12:** 756–766.

Hofacker IL. 2009. RNA secondary structure analysis using the Vienna RNA package. *Curr Protoc Bioinformatics* **Chapter 12:** Unit12.2.

The International HapMap 3 Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467:** 52–58.

Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E. 2010. Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467:** 103–107.

Kim N, Jinks-Robertson S. 2012. Transcription as a source of genome instability. *Nat Rev Genet* **13:** 204–214.

Kimura M. 1967. On the evolutionary adjustment of spontaneous mutation rates. *Genet Res* **9:** 23–34.

Kimura M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, UK.

Koren A, Soifer I, Barkai N. 2010. *MRC1*-dependent scaling of the budding yeast DNA replication timing program. *Genome Res* **20:** 781–790.

Lang GI, Murray AW. 2011. Mutation rates across budding yeast Chromosome VI are correlated with replication timing. *Genome Biol Evol* **3:** 799–811.

Lange SJ, Maticzka D, Möhl M, Gagnon JN, Brown CM, Backofen R. 2012. Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res* **40:** 5215–5226.

Lesnik EA, Freier SM. 1995. Relative thermodynamic stability of DNA, RNA, and DNA:RNA hybrid duplexes: relationship with base composition and structure. *Biochemistry* **34:** 10807–10815.

Li R, Li Y, Kristiansen K, Wang J. 2008. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24:** 713–714.

Lind PA, Andersson DI. 2008. Whole-genome mutational biases in bacteria. *Proc Natl Acad Sci* **105:** 17878–17883.

Lippert MJ, Kim N, Cho JE, Larson RP, Schoenly NE, O'Shea SH, Jinks-Robertson S. 2011. Role for topoisomerase 1 in transcription-associated mutagenesis in yeast. *Proc Natl Acad Sci* **108:** 698–703.

Lorenz R, Hofacker IL, Bernhart SH. 2012. Folding RNA/DNA hybrid duplexes. *Bioinformatics* **28:** 2530–2531.

Lynch M. 2010a. Evolution of the mutation rate. *Trends Genet* **26:** 345–352.

Lynch M. 2010b. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci* **107:** 961–968.
Lynch M. 2011. The lower bound to the evolution of mutation rates. *Genome Biol Evol* **3:** 1107–1118.
Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci* **105:** 9272–9277.
Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci* **101:** 7287–7292.
Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628.
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320:** 1344–1349.
Pan T, Sosnick T. 2006. RNA folding during transcription. *Annu Rev Biophys Biomol Struct* **35:** 161–175.
Park C, Qian W, Zhang J. 2012. Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO Rep* **13:** 1123–1129.
Park C, Chen X, Yang JR, Zhang J. 2013. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc Natl Acad Sci* **110:** E678–E686.
SantaLucia J Jr, Hicks D. 2004. The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct* **33:** 415–440.
Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet* **41:** 393–395.
Takahashi T, Burguiere-Slezak G, Van der Kemp PA, Boiteux S. 2011. Topoisomerase 1 provokes the formation of short deletions in repeated sequences upon high transcription in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci* **108:** 692–697.
Takahata N. 1993. Allelic genealogy and human evolution. *Mol Biol Evol* **10:** 2–22.
Turner DH, Mathews DH. 2010. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res* **38(**Database issue**):** D280–D282.
Wagner A. 2005. Energy constraints on the evolution of gene expression. *Mol Biol Evol* **22:** 1365–1374.
Wahba L, Amon JD, Koshland D, Vuica-Ross M. 2011. RNase H and multiple RNA biogenesis factors cooperate to prevent RNA:DNA hybrids from generating genome instability. *Mol Cell* **44:** 978–988.
Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, Ouyang Z, Zhang J, Spitale RC, Snyder MP, Segal E, et al. 2014. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505:** 706–709.
Weiner A, Hughes A, Yassour M, Rando OJ, Friedman N. 2010. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res* **20:** 90–100.
Xia X. 2010. The effect of probe length and GC% on microarray signal intensity: characterizing the functional relationship. *Int J Syst Synth Biol* **1:** 171–183.
Yang JR, Liao BY, Zhuang SM, Zhang J. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc Natl Acad Sci* **109:** E831–E840.
Yang JR, Chen X, Zhang J. 2014. Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS Biol* **12:** e1001910.
Zamft B, Bintu L, Ishibashi T, Bustamante C. 2012. Nascent RNA structure modulates the transcriptional dynamics of RNA polymerases. *Proc Natl Acad Sci* **109:** 8948–8953.
Zhang J, He X. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol Biol Evol* **22:** 1147–1155.
Zhang J, Yang JR. 2015. Determinants of the rate of protein sequence evolution. *Nat Rev Genet* **16:** 409–420.
Zur H, Tuller T. 2012. Strong association between mRNA folding strength and protein abundance in *S. cerevisiae*. *EMBO Rep* **13:** 272–277.

Received May 28, 2015; accepted in revised form October 14, 2015.

Nascent RNA folding mitigates transcription-associated mutagenesis

Xiaoshu Chen, Jian-Rong Yang and Jianzhi Zhang

Genome Res. 2016 26: 50-59 originally published online October 30, 2015
Access the most recent version at doi:10.1101/gr.195164.115