RNA-DNA sequence differences in *Saccharomyces cerevisiae*

Isabel X. Wang, Christopher Grunseich, Youree G. Chung, et al.

Genome Res. published online September 16, 2016
Access the most recent version at doi:10.1101/gr.207878.116

RNA-DNA Sequence Differences in *Saccharomyces cerevisiae*

I. X. Wang[1]*, C. Grunseich[2], Y. G. Chung[3], H. Kwak[1,4], G. Ramrattan[1,4], Z. Zhu[1], V. G. Cheung[1,4,5]*

[1]Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109, USA

[2]Neurogenetics Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA.

[3]College of Engineering, University of Michigan, Ann Arbor, MI 48109, USA

[4]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA

[5]Departments of Pediatrics and Genetics, University of Michigan, Ann Arbor, MI 48109, USA

*ixwang@umich.edu; vgcheung@umich.edu.

Running title: RNA-DNA sequence differences in budding yeast

Keywords: RNA editing; RNA-DNA sequence differences; R-loops

ABSTRACT

Alterations of RNA sequences and structures, such as those from editing and alternative splicing, result in two or more RNA transcripts from a DNA template. It was thought that in yeast, RNA editing only occurs in tRNAs. Here, we found that *Saccharomyces cerevisiae* have all 12 types of RNA-DNA sequence differences (RDDs) in the mRNA. We showed these sequence differences are propagated to proteins, as we identified peptides encoded by the RNA sequences in addition to those by the DNA sequences at RDD sites. RDDs are significantly enriched at regions with R-loops. A screen of yeast mutants showed that RDD formation is affected by mutations in genes regulating R-loops. Loss-of-function mutations in ribonuclease H, senataxin and topoisomerase I that resolve RNA-DNA hybrids lead to increases in RDD frequency. Our results demonstrate that RDD is a conserved process that diversifies transcriptomes and proteomes, and provide a mechanistic link between R-loops and RDDs.

INTRODUCTION

DNA is the genetic blueprint that provides the code for synthesizing RNA and proteins; however, there is not always a direct one-to-one relationship between DNA, RNA and protein. Co-transcriptional and post-transcriptional processing such as splicing and RNA editing alters the sequences and structures of RNA. As a result, different transcripts are produced from the same DNA sequences.

RNA editing was first identified in the mitochondrial genome of kinetoplastids (Benne et al. 1986). Subsequently, different types of RNA editing were found in the mitochondrial and nuclear genomes of diverse organisms including plants (Gualberto et al. 1989; Hiesel et al. 1989), viruses (Volchkov et al. 2001), molluscs (Garrett and Rosenthal 2012; Albertin et al. 2015) and human (Tennyson et al. 1989; Teng et al. 1993). The mechanisms that mediate these events and the functions of the edited transcripts are mostly unknown. However, as sequencing technology improves, information about RNA editing is accumulating. Adenosine-to-inosine editing mediated by the human ADAR (Adenosine deaminases acting on RNA) proteins was once regarded as rare events. But recently, hundreds of thousands of ADAR-mediated editing sites have been identified in human cells (Kawahara et al. 2004; Ju et al. 2011; Alon et al. 2012; Chen et al. 2012; Peng et al. 2012; Silberberg et al. 2012; Vesely et al. 2012; Chen 2013; Wang et al. 2013; Bazak et al. 2014). Furthermore, we and others have found that there are other types of RNA-DNA sequence differences (RDDs) that are unlikely to be mediated by these deaminases (Li et al. 2011; Bahn et al. 2012; Bar-Yaacov et al. 2013; Rubio et al. 2013; Turner et al. 2015). These events are found in normal cells, and altered patterns of RDDs were found in neurologic diseases (van Leeuwen et al. 1998; Silberberg et al. 2012; Krestel et al. 2013; Wang et al. 2014) and in cancers (Klimek-Tomczak et al. 2006; Martinez et al. 2008; Lee et al. 2013;

Avesson and Barry 2014; Han et al. 2014; Niavarani et al. 2015). Even though differences between RNA and their corresponding DNA templates were known for many years, their discoveries in human beyond the editing events mediated by ADAR and APOBEC families of deaminases were debated. Some groups posited that the RDDs are not as widespread as reported, and/or they are the results of inaccurate sequencing and/or analyses of deep sequencing data (Kleinman and Majewski 2012; Lin et al. 2012; Pickrell et al. 2012). We have addressed many of these concerns (Li et al. 2012), and other groups have identified RDDs in human cells. The first published work we know of is a G-to-A site in *WT1*, wilms tumor 1 (Sharma et al. 1994). Subsequently RDD sites in brain tissues from Alzheimer's and Down syndrome patients (van Leeuwen et al, 1998) and in other human cells were identified (Ju et al. 2011; Bahn et al. 2012; Silberberg et al. 2012; Turner et al. 2015; Zhang et al. 2015). A recent paper suggests that APOBEC3A is the protein that mediates G-to-A RDD in *WT1* (Niavarani et al. 2015).

Despite advances in technologies for sequencing and analysis of RNA transcripts, it is still challenging to survey enough human cells comprehensively; especially for those RDDs that are present at lower frequencies. To obtain such a comprehensive view, here we studied the DNA and RNA of budding yeast *Saccharomyces cerevisiae* and present data from 18 yeast strains, including wild-type and mutants.

RESULTS

RDDs in wild-type yeast strains

We sequenced the DNA and RNA of 6 wild-type budding yeast strains commonly used in research laboratories, including S288C, the reference strain of the yeast genome project (Engel et al. 2013). For each strain, we obtained about 40 million DNA-seq reads (>250× genome

coverage), and sequenced the transcriptomes to an average of 10 million 100-nt reads. To identify RDDs for each yeast strain, we compared the RNA sequences to the corresponding DNA sequences (Methods; Supplemental Results and Fig. S1). We found in S288C more than 750 RDD sites, corresponding to a frequency of 1.5 RDD per 10,000 nucleotides in the transcriptome (Supplemental Table S1). These include all 12 types of RDDs. Figure 1A shows the distribution by RDD type; T-to-C difference is the most common form, and transitions represent the 4 most common types (Fig. 1A). Similarly RDDs were also found in the other 5 wild-type strains and frequencies range from 1.3 to 1.8 per 10,000 nucleotides (Fig. 1B, Supplemental Fig. S2A). RDDs are significantly enriched in coding exons (Fisher's exact test, $P<0.05$) (Supplemental Fig. S2B).

We examined how various inclusion criteria affect RDD identification (Supplemental Results). All 12 types of RDDs were found even when we applied more stringent thresholds for sequencing depth and RDD levels (Supplemental Fig. S2C, S2D). We also ensured that RDDs are detected by different alignment programs and parameters (Supplemental Results). We re-aligned the sequence reads with different programs and tested parameters for handling splice sites. We assessed error rates with internal standards, simulated RNA-seq data and a probability-based method for estimating errors. All the results support our conclusion that RDDs are found in yeast; and they are not results of errors in sequencing or analysis of the sequence data. In addition, we used an independent method, droplet digital PCR to validate the RDDs. Among the 20 RDD sites examined, 18 were also found by this method in yeast grown in a replicate culture. Figure 2 shows a C-to-G RDD in *ADR1* (a*lcohol dehydrogenase II synthesis regulator 1*) that was identified by sequencing and droplet digital PCR.

There are 334 RDD sites that are found in 2 or more of the 6 wild-type strains; among them 234 are in coding exons, 8 are in introns, one is in snoRNA and 91 in "intergenic regions" (including unannotated transcripts and UTRs beyond the annotated genic region) (Supplemental Table S2). RDD-containing genes include *MTR4* that encodes an RNA helicase in the TRAMP complex, and *DIP2* that encodes components of U3 snoRNP that processes pre-18S rRNA (Table 1). Of the 216 RDDs in coding sequences with annotated open reading frame (excluding dubious ORF), there seems to be no bias against nonsynonymous changes since 161 (74%) are nonsynonymous and 55 (26%) are synonymous changes.

RNA forms of RDD-containing transcripts are translated into proteins

Next, we ask whether RDD-containing transcripts are translated to proteins. We analyzed the yeast proteome by mass spectrometry. Previously, proteomic studies identified peptides by prediction based on DNA sequences; thus peptides encoded by RNA forms at RDD sites would have been missed. Here, we embarked on finding peptides that result from RDDs. However, finding the RNA forms of proteins using shotgun mass spectrometry is challenging due to the "missing value" problem (Karpievitch et al. 2012; Nagaraj et al. 2012; Tyanova et al. 2014). Not all expressed proteins are identified by mass spectrometers; and even when a protein is sequenced, only some of its peptides are captured. In addition, since most of the RDD levels are not very high, assuming one-to-one translation of RNA to proteins, the majority of the peptides at an RDD site will be the DNA form. To look for peptides that are encoded by RDDs, we first analyzed raw data generated by Mann and colleagues who carried out a comprehensive study of the yeast proteome (Nagaraj et al. 2012). We translated RNA sequences at RDD sites into amino acid sequences and incorporated them into our peptide search. Using the same proteomics

software, MaxQuant (Nagaraj et al. 2012) as Mann and colleagues, we found peptides that are encoded by RNA rather than DNA sequences at RDD sites (Table 2). Encouraged by these results, we carried out our own mass spectrometry analysis with stringent thresholds sufficient for detecting mass difference between single amino acids (Methods). Just as with data from Mann and colleagues, we found peptides encoded by both DNA and RNA sequences at RDD sites (Table 2). For instance, a G-to-A RDD in *RPN1*, encoding a subunit of 26S proteasome, is translated to DNA- and RNA-forms of the protein (V221I), as depicted by sequencing reads and mass spectra in Figure 3. RNA-forms of Gip3 (S435L), Hsp90 (Q178E) and Tef1 (G34A) were found by mass spectrometry in both studies.

After finding the RDD-encoded peptides by whole proteome mass spectrometry, we attempted to enrich for protein isoforms resulting from RDDs. We began with one protein, Tup1, in a pilot study. There are two isoforms of Tup1, a DNA-encoded and an RNA-encoded form with alanine and valine as amino acid 459, respectively. For the enrichment of peptides for Tup1, we carried out immunoprecipitation with antibody against Tup1, then excised the protein band corresponding to the expected molecular weight of Tup1, and performed mass spectrometry. However, that did not lead to an enrichment, rather we got very few peptides corresponding to Tup1, and no valine-form of the protein. Most likely this is because Tup1 is a highly modified protein and it works in a large protein complex. The post-translational modifications of Tup1 led to different electrophoretic pattern(s) than its unmodified form. In addition, since Tup1 is part of a large protein complex, the immunoprecipitation likely pulls down its interacting partners with similar sizes which reduces our chance of detecting Tup1 itself (Krogan et al. 2006). Indeed, Cdc48, a known interacting partner of Tup1 with similar molecular weight, was detected in the immunoprecipitant. This suggests that unless we have ways to

enrich specifically for a protein, mass spectrometry of whole proteome may be a more practical approach for identifying protein isoforms from RDDs.

Effect of RDD on protein structure and function

Next we examined whether the RDDs that lead to non-synonymous changes in amino acids affect protein structures. The amino acid changes that result from RDDs are found throughout the proteins, including key active sites. For instance, a G-to-A RDD results in two forms of Rsc4 that have different sequences in the bromodomain. To characterize the impact of RDDs, we compared the predicted structures of DNA- and RNA-forms of the proteins (Zhang 2008). We focused on 10 RDD-encoded proteins that were detected by mass spectrometry (Table 2). The analysis showed some of the amino acids are exposed on protein surfaces that interact with ligands and/or other proteins, and others affect tertiary structures (Fig. 4). For example, modeling shows that the I49 (DNA-form) and V49 (RNA-form) of Arf2 have different orientations in the GTP/GDP binding domain. The RNA-form (V49) of Arf2 is predicted to be more flexible for ligand binding than the DNA form, thus the two protein isoforms likely have different substrate affinities (Cheng et al. 2006). In other proteins, RDDs affect tertiary structures. An example is Vph1, a subunit of the V-type H^+-ATPase that is required for assembly of the ATPase complex and maintenance of pH in yeast vacuoles. Figure 4 shows that the DNA (I319) and RNA (N319) forms of Vph1 have distinct folding properties (TM-score = 0.34, in contrast to >0.5 for control proteins with similar structures; $P<10^{-6}$). Third, several RDDs are predicted to affect protein stabilities (Worth et al. 2011) (Table 3). For instance, for the transcription repressor Tup1, the RNA-form (V459) is predicted to be more stable ($\Delta\Delta G=1.02$) than the DNA-form (A459) (Worth et al. 2011).

Next, we experimentally validated results from computational prediction and determined the functional consequences of RDDs by focusing on Tup1. We cloned the DNA- and RNA-forms of *TUP1* ORF into plasmids under *GAL1* promoter, and induced their expression in yeast *tup1* knockout strain. By western blot, we showed that the protein level of the RNA-form (V459) of Tup1 is higher than that of the DNA-form (A459), while mRNA levels of both forms are similar (Fig. 5A, 5B). Cycloheximide chase study showed that the RNA-form of Tup1 has a longer half-life than that of the DNA-form (Fig. 5C). Together, these results are consistent with the prediction that the C-to-U RDD makes the Tup1 protein more stable. Next, we asked if this difference in Tup1 protein stability affects its function. Since the main role of Tup1 is transcriptional regulation, to assess the effect of the RDD on function, we carried out RNA sequencing of yeast cells with or without *TUP1* expression, and identified 292 genes ($P<0.0001$, *t*-test) whose expression levels are influenced by Tup1. Previously microarrays were used to identify Tup1-regulated genes; we compared our results to the 335 Tup1-regulated genes identified by Green and Johnson, among them 243 are also expressed at significantly different levels ($P<0.01$, *t*-test) in yeast with and without Tup1 in this study. Among these Tup1-regulated genes are *MAL11*, *HXT2* and *HSP12* that were previously found to be repressed by Tup1 (Green and Johnson 2004), and genes such as *HIS4* and *TAT2* that were induced by Tup1 (Tanaka and Mukai 2015). Tup1-Cyc8 does not directly binds to DNA. As a co-repressor complex, it affects gene expression by recruiting DNA-binding proteins mainly through the WD40 domain where the amino acid affected by the RDD resides (Sprague et al. 2000; Malavé and Dent 2006). It is likely that DNA- and RNA-forms of Tup1 interact differently with some DNA-binding proteins, and therefore have differential effects on expression of some of its target genes. To examine this possibility, we compared the expression levels of the 292 genes in the yeast expressing either the

DNA- or the RNA-form of Tup1. The results showed that 51 genes are differentially expressed (Fig. 5D); among them 42 had higher expression (>50%) and 9 genes had lower expression (<50%) in yeast expressing the RNA-form of Tup1. The 42 genes with higher expression in the RNA-form of Tup1 were repressed by Tup1; thus, the RNA-form conferred a lesser repressive effect on its target genes compared to the DNA-form. These 42 genes include several stress responsive genes such as *RCK1* and *HUG1*. Of the 9 genes that showed lower expression, 6 genes including *NDJ1* that encodes a telomere associated protein that promotes meiotic recombination (Wu and Burgess 2006) were induced by Tup1 but the RNA-form of Tup1 induced these target genes to a lesser extent. Our results therefore show that while the RNA-form of Tup1 is more stable, it has a lesser effect on the expression levels of its target genes than the DNA-form. This paradox where a more stable protein has lower activity was described by Matthews and colleagues for lysozyme (Shoichet et al. 1995) and found for other proteins including PTEN (Vazquez et al. 2000) and β-lactamase (Beadle et al. 1999). Lastly, to assess the cellular effect of RDD, we compared the stress response of yeast expressing the DNA and RNA-form of Tup1. We found that yeast that expresses the RNA-form of Tup1 is more sensitive to hygromycin B than the DNA-form (Fig. 5E); this could be due to the reduced activity of the valine-bearing form of Tup1. These results suggest that like other co-transcriptional processes, RDD produces transcripts with different functions and affects cellular phenotypes.

Deaminases do not play a key role in yeast RDD formation

After characterizing the RDDs, next we took advantage of yeast mutants to search for genes that are involved in RDD formation. The two known RNA editing enzymes in humans, ADAR and APOBEC, are deaminases that convert adenosine to inosine (then read by the translation machinery as guanosine), and cytidine to uridine. Knockdown of ADAR proteins in human B-

cells resulted in a greater than 90% decrease in A-to-G editing (Wang et al. 2013). To look for equivalent pathways in yeast, we sequenced deaminase mutants and assessed for RDDs. We studied mutant strains of tRNA adenosine deaminases, *tad1*$^{-}$, *tad2*^{ts} and *tad3*^{ts}, and other deaminases such as *aah1*$^{-}$ (adenine deaminase), *amd1*$^{-}$ (AMP deaminase) and *fcy1*$^{-}$ (cytosine deaminase). To confirm the mutant phenotypes, we first looked at the known editing sites in tRNA; as expected, we found a decrease in editing levels in the mutants. At non-permissive temperature, the temperature sensitive mutant of the essential *TAD2* gene (*tad2*^{ts}) has a lower A-to-G editing level at position A34 of tRNA-serine (Supplemental Fig. S3A). Then, we asked whether the deaminases that target tRNA also affect mRNA. Unlike in tRNA, there was no difference in the mRNA editing patterns between mutants and wild-type strains. Other types of RDDs were also unchanged in the deaminase mutants. The relative abundance of all 12 types is highly similar between the mutants and wild-type controls (Supplemental Fig S3B). For example, the average level of A-to-G RDD was 11% in wild-type yeast, and those among the mutants were 9 to 12%; and for C-to-T, the level in wild-type strains was 9% and those in mutants were 7 to 10%. These results suggest that Tad1, Tad2, Tad3, Aah1, Amd1 and Fcy1 do not contribute to the majority of RDDs in mRNAs of yeast. A few other putative deaminase genes, including *CDD1*, *DCD1* and *RIB2*, are essential for yeast survival and null mutants are not available. These deaminases remain to be ruled out as enzymes contributing to RDDs.

RDDs are dependent on R-loop formation

Next, we assessed whether R-loops play a role in RDD formation. Previously, we showed RDDs do not arise during RNA synthesis by RNA Polymerase II nor as a direct consequence of incorporation of modified bases; rather we showed that RDDs begin to occur in nascent RNA

chains ~55 nucleotides from the RNA Polymerase II active site (Wang et al. 2014). Given that RDDs are found so soon after transcription, we posited that R-loops which often occur outside of the polymerase complex may be involved in RDD formation. Results from studying human cells with a gain-of-function mutation in senataxin (L389S) that leads to juvenile amyotrophic lateral sclerosis showed lower RDD frequency, consistent with the involvement of R-loops (Wang et al. 2014).

Yeast provides us with an opportunity to examine the connection between R-loops and RDD more deeply using several approaches. We asked whether R-loops and RDDs co-localize, and if yeast mutants with defective R-loops have altered RDD profiles. First, we mapped R-loops by DNA-RNA hybrid immunoprecipitation and deep sequencing (DRIP-seq) using the S9.6 antibody (Boguslawski et al. 1986). We identified more than 1,500 R-loop peaks in BY4741. Then we asked whether RDDs co-localize with these R-loops, and found that RDDs are significantly enriched ($P<0.0001$) in R-loop regions (Supplemental Results; Fig 6A). A zoomed in figure of a C-to-T RDD and R-loops in *BUG1* is shown in (Supplemental Fig S6).

Next we search for genetic evidence that perturbation of R-loops affects RDDs. We took advantage of the different mechanisms that yeast has developed to resolve R-loops from directly unwinding the R-loops with the helicase, senataxin, to degrading the RNA with ribonuclease H and altering the supercoils in DNA with topoisomerase enzyme. First, we studied senataxin homolog (Sen1) in yeast. Proudfoot and colleagues showed that the yeast senataxin homolog (Sen1) plays a role in resolving R-loops and there is an accumulation of R-loops in the *sen1-1* temperature-sensitive mutant (Mischo et al. 2011). Here we quantified RDDs in the same senataxin mutant (Fig. 6B). By shifting the *sen1-1* mutant to a non-permissive temperature, we found significantly more R-loops ($P<0.02$, χ^2) and a genome-wide increase in RDD frequency

(Fig. 6C). At the higher temperature with a loss of senataxin function, we identified more RDD sites, and a trend for higher RDD levels. Examples include G-to-A sites in *HSE1* and *SEC26* (Fig. 6D). The higher RDD levels at the non-permissive temperature allowed more RDD sites to be identified; some sites were found only in the non-permissive temperature, while others that were present at 25°C but below our 5% level inclusion criterion had higher levels at 34°C, and therefore are included as RDD sites. Sen1 plays a role in pathways other than resolving R-loops. To ensure that the difference in RDD frequency in yeast senataxin mutant is due to R-loop defects and not from other functions of senataxin, we examined additional yeast mutants that are defective in resolving RNA-DNA hybrids. We studied yeast deletion mutants of ribonuclease H1 (RNaseH1, *RNH1*) and topoisomerase 1 (*TOP1*). RNase H1 resolves R-loops by hydrolysis of the RNA component of RNA-DNA hybrids (Stein and Hausen 1969; Keller and Crouch 1972; Wyers et al. 1973; Karwan et al. 1984). Topoisomerase 1 minimizes transcription-dependent R-loops by relaxing the supercoils that follow RNA polymerases (Drolet et al. 1995; Tuduri et al. 2009; El Hage et al. 2010; Williams et al. 2013). As expected, using the S9.6 antibody that recognizes RNA-DNA hybrids, we confirmed that deletion of *RNH1* or *TOP1* leads to more R-loops, as in *sen1-1* mutant (Fig. 6B). Then, we sequenced and studied RDDs in *rnh1* and *top1* mutants in BY4741 and SNM8 strains of *S. cerevisiae*. The results showed that genome-wide the *rnh1* mutant has more RDDs (≥25%) than wild-type controls in both strains (Fig. 6E). Similarly, loss of *TOP1* leads to about 10% more RDDs. As in the senataxin mutants, the RDD levels are significantly higher in both mutants ($P \leq 0.006$, χ^2). Figure 6F shows examples of RDD sites where the *rnh1* and *top1* mutants have higher levels compared to the wild-type yeast. Together these findings show that yeasts that have more R-loops, from different mechanisms

ranging from deficiencies in senataxin, to the loss of ribonuclease H1 and topoisomerase 1 activities, have more RDDs.

DISCUSSION

Nascent RNAs are highly modified through splicing and alternative use of start and termination sites, which generate mature RNAs and proteins of different structures. In some cases, the different transcripts from the same DNA templates even have opposing functions, such as BCL-x where one splice form (BCL-xS) promotes apoptosis and another form (BCL-xL) inhibits apoptosis (Boise et al. 1993). Besides splicing, another way to diversify transcripts is through changes in RNA sequences by RNA editing such as those mediated by the ADAR proteins. Previously, RNA editing was reported in organisms from kinetoplastids to plants and human cells, but not in yeast. Here, we show that RNA-DNA sequence differences occur at about the same frequency in yeast (1/10,000) as in human cells. All 12 types of RDDs were found. Our mutant screens showed that RDD formation is associated with RNA-DNA hybrids. Comparison of DNA and RNA sequences of mutants known to have more R-loops (Wyers et al. 1973; Tuduri et al. 2009) reveals that all the mutants have higher RDD frequencies than wild-type strains. This effect is specific since consistent changes in RDD frequency and level were found among yeasts with R-loop defects from loss-of-function mutations in senataxin, ribonuclease H1 or topoisomerase 1 but not in other mutants such as those with defective deaminases. The increase in RDD is not just at a few sites but rather there is an increase genome-wide.

Previously, our results in human cells show that RDDs are formed in nascent RNA soon after RNA synthesis and suggest that their formation is coupled to R-loops. Here in yeast, we

provide molecular and genetic evidence that RDDs are coupled to R-loops. We posit that the R-loops promote RNA structures that facilitate RDD formation. Multi-step processes must be involved in RDD formation; in particular, for the transversion events. For the sites reported here, the RDD types are highly consistent; for example, at an A-to-C site, all RDD-containing transcripts contain a C, instead of multiple alleles. To us, this suggests that RNA modification is not likely to explain the RDDs, since reverse transcriptases mis-incorporate nucleotides when they encounter modified bases. At RDD sites, we do not see the multi-allelic patterns that are hallmarks of RNA modifications, rather the same types of nucleotides are observed. Thus, RNA modifications do not explain most of the RDDs. Rather, we favor enzymatic steps that convert one nucleotide to another for the transition events, and multi-step processes that may include abasic intermediates for the transversions. It remains unknown what confers the specificity in RDD formation. Organisms such as trypanosomes use short RNAs as guides for editing RNA (reviewed by Benne 1992). Guide RNAs have not been identified in yeast nor in human; however since many small RNAs remain to be characterized, we cannot exclude this possibility.

To begin to explore the functional consequences of RDDs, we studied a C-to-U site that results in a valine substitution for alanine in the WD40 domain of Tup1. This protein in conjunction with Cyc8 regulates gene expression by recruiting DNA-binding proteins to target genes. The WD40 domain is the platform for these protein interactions. We showed the RNA-form (V459) of Tup1 is more stable than the DNA form yet its regulatory activity on over 50 target genes is lower. Matthews and colleagues suggested that to maximize their activities, proteins can adopt a conformation that is less stable. They first showed supporting data in lysozyme (Shoichet et al. 1995); since then they and others have provided many additional

examples. Here the RDD in *TUP1* appeared to make the protein more stable yet less active. This could be a mechanism to fine tune the kinetics and the expression levels of its target genes.

To conclude, since the initial identification of RDDs in human cells (Li et al, 2011), additional studies have enabled us to characterize further this co-transcriptional RNA processing step. By identifying RDDs in yeast, we show that they are conserved. Characterization of transcripts with RDDs and the resulting protein isoforms shows that RDDs, like alternate splicing, diversify transcriptome and proteome by providing additional RNA and protein isoforms with different functions and/or regulation. In addition, we show that RDDs are coupled to R-loops. This information should lead us and others closer to uncovering mechanistic details that underlie RDD formation and determining their functional roles.

METHODS

Yeast strains. Yeast strains used in this study are listed in Supplemental Table S3. Yeast cultures and deep sequencing of DNA and RNA are described in Supplemental Methods.

DNA-RNA immunoprecipitation and deep sequencing (DRIP-seq). Immunoprecipitation procedure was adapted from previous studies (Skourti-Stathaki et al. 2011; Ginno et al. 2012). 1 X 10^8 cells yeast cells were used for each immunoprecipitation. Genomic DNA containing R-loops was purified using MasterPure Yeast DNA Purification Kit (Epicentre). DNA was fragmented with a cocktail of five restriction enzymes (HindIII, EcoRI, BsrGI, XbaI and SspI) (New England Biolabs). 5μg of S9.6 monoclonal antibody (gift from Dr. Stephen H. Leppla at NIH) or non-specific mouse IgG was used for each immunoprecipitation. Input and precipitated DNA was made into sequencing libraries using DNA SMART ChIP-Seq Kit (Clontech).

Functional analysis of RNA-form of Tup1. The pBY011-TUP1 plasmid (Cat# ScCD00095253, Harvard PlasmID repository) contains *GAL1* promoter and DNA-form of *TUP1*. RNA-form of *TUP1* was generated using the QuikChange II XL Site-Directed Mutagenesis Kit (Stratagene). Yeast cells were transformed with either pBY011-TUP1 A459 (DNA form), pBY011-TUP1-V459 (RNA form), or empty vector using the lithium acetate method. Gene induction, drug sensitivity and cycloheximide chase assay are described in detail in supplemental methods.

R-loop dot blot using S9.6 antibody. 250 ng of DNA containing R-loops from each strain was incubated with 1 unit of RNase H (#M0297, NEB) or mock for no RNase H digestion control, in 1X RNase H digestion buffer at 37°C for 12 hours. DNA was phenol extracted and ethanol precipitated and reconstituted in 10 ul TE buffer. 5 ul of 1X or 0.2X titration of DNA solution was loaded to nitrocellulose membrane, air dried for 30 min and baked at 80°C for 2 hours. The

membrane was blocked in 5% BSA for one hour and incubated with 1:1000 S9.6 antibody (gift from Proudfoot lab) (Boguslawski et al. 1986) overnight at 4°C in order to detect RNA-DNA hybrids.

Analysis of sequencing data. Sequencing reads were pre-processed as follows: adapter sequences from the end of reads were trimmed using the program fastx_clipper from FASTX-Toolkit (Hannon Lab). Low-quality sequences at ends of reads represented by a stretch of "#" in the quality score string in FASTQ file were also removed. PolyA (>5 consecutive As) at end of reads were removed. Reads that were >35nt after trimming were included for further analyses. Sequencing reads were then aligned to sacSer3 reference using GSNAP (Version 2013-10-28) (Wu and Nacu 2010) with the following parameters and bam files generated: Mismatches ≤[(read length+2)/12-2]; Mapping score ≥20; Soft-clipping on (-trim-mismatch-score=-3). Sequencing reads with identical sequences were counted as one read in order to remove possible artefacts from PCR amplification.

Expression levels of RNA transcripts were analyzed using Cufflinks (version 2.1.1.) at default parameters (Trapnell et al. 2010) and RPKM (reads per kilobase per million mapped reads) value is reported as expression value. Genes with RPKM >1 in at least 75% samples were retained for downstream analysis. To identify genes whose expression levels are influenced by *TUP1* expression, we compared gene expression in yeast with (galactose-induced) and without (empty vector and glucose-repressed) *TUP1* expression by Student's *t*-test. And among the 292 genes with nominal P<0.0001, we determined fold difference of expression levels in yeast expressing valine- (RNA-form) versus alanine-containing (DNA-form) of TUP1.

To identify R-loop peaks from sequencing data, read depth at each nucleotide site was calculated using BamTools (Barnett et al. 2011). Only genomic sites covered by sequencing

reads in input samples are retained for further analysis. An R-loop peak is called in DRIP samples when contiguous sequencing reads cover a genomic region with sequencing depth ≥ 20 RPM (reads per nucleotide per million of uniquely aligned reads), and fold enrichment of S9.6 antibody over input ≥2.5 fold.

RDD identification. RNA-DNA sequence differences were identified as previously described with some modifications (Wang et al. 2013, 2014; Li et al. 2011). In DNA-seq data, we required a minimum coverage of 10 reads at a given site and the DNA sequence at this site to be 100% concordant (no reads containing alternate alleles). For each site that passed DNA-seq coverage cutoff, we compared RNA-seq read sequence to its corresponding DNA sequence from the same strain, requiring 4 criteria for RDD identification: 1) a minimum of 10 RNA-seq reads covering this site; 2) RDD level ≥5% (level = [# of RNA-seq reads containing non-DNA allele]/[# all RNA-seq reads covering a given site]); 3) a minimum of two non-identical RNA-seq reads containing RDD; 4) Phred score ≥ 20.

We applied additional filtering steps and statistical analyses to ensure correct RDD identification (Supplemental Methods & Results). We picked 20 sites of each type that are suitable for primer and assay design for droplet digital PCR and experimentally validated them (Supplemental Methods).

Mass spectrometry analysis. For protein encoded by RDDs to be identified by mass spectrometry, RDD sequence information needs to be incorporated into protein database. For each non-synonymous RDDs identified in six wild-type strains, we translated the RNA sequence into amino acid sequence. Thus we constructed a protein database containing both yeast protein sequences coded by reference sequence (sacCer3) and those coded by RDDs. This database

includes 5,887 proteins, of which 2,336 proteins are represented by both DNA-form and RNA-form sequences.

Yeast proteomic data generated by the Mann group using a Single-shot Ultra HPLC Runs on a Bench Top Orbitrap were downloaded from Tranche proteome repository (Nagaraj et al. 2012). In our own LC/MS-MS analysis, 5 X 10^7 yeast cells of S288C and W303-1A collected from exponential cultures were frozen in liquid nitrogen and ground using Retsch ball mill homogenizer (Retsch). Homogenized cells were then dissolved in 1ml lysis buffer (20 mM Tris HCl pH 8, 137 mM NaCl, 10% glycerol 1% Nonidet P-40, 2 mM EDTA). 60 µg of total protein was pre-fractionated and digested as previously described (Li et al. 2011).

MaxQuant typically performs an initial search, mass recalibration and then carry out the main searches in order to achieve high accuracy. We applied peptide tolerance (4.5ppm) that is much more stringent than required to detect small mass difference between single amino acids. RAW files were analyzed using MaxQuant vesion 1.5.1.2 (Cox and Mann 2008) and searched against the above RDD-containing protein database. The following search parameters were applied: initial search with peptide tolerance of 20 ppm, main search after mass recalibration with peptide tolerance of 4.5ppm, parent ion mass tolerance of 0.3 Da, MS/MS tolerance 0.5 Da, complete carbaminomethyl modification of cysteine and variable N-terminal acetylation and methionine oxidation, allowing up to one trypsin miscleavage, protein false discovery rate <1%, peptide false discovery rate <1% (determined by decoy database search). To confirm uniqueness of peptides, we searched sequence of each identified RDD-encoded peptides against yeast protein databases using BLAST. None of the RDD-encoded peptides matched known peptides, confirming these peptides are not encoded by other regions of the genome. This ensures that each RDD-form peptide is truly unique.

To assess the false discovery rate of peptide search, we generated a negative control database. We used the target-decoy strategy, a method widely used to estimate false positive peptides, developed by Gygi's group (Elias and Gygi 2010). In this method, a control protein database is generated by reversing amino acid sequence of each protein, and number of peptides detected from decoy search is used to determine FDR. Choosing a randomized control for proteomic analysis is challenging. A simple scrambled amino acid sequence will differ from a biological sample in amino acid composition, location of tryptic sites, length of tryptic peptides, mass accuracy and probability to be detected. The decoy method minimizes such drawbacks. The FDR of peptide in our analysis is 1%, as determined by this method.

Protein Structure Prediction. 3-dimensional structures of proteins were predicted using I-TASSER without restraints and templates(Zhang 2008). TM-score (value 0-1 with 1 indicating the perfect match) was used to measure structural similarity between DNA-form and RNA-form models. Protein stability is predicted as previously published (Worth et al. 2011).

Data Access. The DNA and RNA sequencing and DRIP-seq data from this study have been submitted to the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena/) under study accession number PRJEB8021. The mass spectrometry data have been submitted to the PeptideAtlas database with the identifier PASS00687.

ACKNOWLEDGEMENTS

We thank Drs. Susana Cerritelli, Robert Crouch, Daniel Klionsky, Thomas Kunkel and Rolf Sternglanz for providing yeast strains, insightful suggestions and discussions. We thank Drs Nicholas Proudfoot, Konstantina Skourti-Stathaki and Stephen Leppla for providing S9.6 antibody and discussions. We are grateful for the many discussions and encouragements from Dr. John Lis. We thank Dr. Hsin-Yao Tang at Wistar Proteomic Center for technical support and

discussion on proteomic analysis. We thank Mr. Hongjiu Zhang in Dr. Yang Zhang's lab for help and suggestions on protein structure prediction. We thank Ms. Jennifer Fox and Ms. Yaojuan Liu for technical support.

Author Contributions:

I.X.W. and V.G.C. conceived and supervised the project and designed experiments. I.X.W. and G.R. carried out experiments. Z.Z. and C.G. processed next-generation sequencing data and analyzed RDD. I.X.W., Y.G.C., Z.Z. and G.R. analyzed proteomic data. H.K. performed statistical analyses. I.X.W. and V.G.C. wrote the manuscript with contributions from all authors.

REFERENCES

Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, Brenner S, Ragsdale CW, Rokhsar DS. 2015. The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* **524**: 220–224.

Alon S, Mor E, Vigneault F, Church GM, Locatelli F, Galeano F, Gallo A, Shomron N, Eisenberg E. 2012. Systematic identification of edited microRNAs in the human brain. *Genome Res* **22**: 1533–1540.

Avesson L, Barry G. 2014. The emerging role of RNA and DNA editing in cancer. *Biochim Biophys Acta BBA - Rev Cancer* **1845**: 308–316.

Bahn JH, Lee J-H, Li G, Greer C, Peng G, Xiao X. 2012. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res* **22**: 142–150.

Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. 2011. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinforma Oxf Engl* **27**: 1691–1692.

Bar-Yaacov D, Avital G, Levin L, Richards AL, Hachen N, Rebolledo Jaramillo B, Nekrutenko A, Zarivach R, Mishmar D. 2013. RNA-DNA differences in human mitochondria restore ancestral form of 16S ribosomal RNA. *Genome Res* **23**: 1789–1796.

Bazak L, Haviv A, Barak M, Jacob-Hirsch J, Deng P, Zhang R, Isaacs FJ, Rechavi G, Li JB, Eisenberg E, et al. 2014. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res* **24**: 365–376.

Beadle BM, McGovern SL, Patera A, Shoichet BK. 1999. Functional analyses of AmpC beta-lactamase through differential stability. *Protein Sci Publ Protein Soc* **8**: 1816–1824.

Benne R. 1992. RNA editing in trypanosomes. The us(e) of guide RNAs. *Mol Biol Rep* **16**: 217–227.

Benne R, Van Den Burg J, Brakenhoff JPJ, Sloof P, Van Boom JH, Tromp MC. 1986. Major transcript of the frameshifted coxll gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* **46**: 819–826.

Boguslawski SJ, Smith DE, Michalak MA, Mickelson KE, Yehle CO, Patterson WL, Carrico RJ. 1986. Characterization of monoclonal antibody to DNA · RNA and its application to immunodetection of hybrids. *J Immunol Methods* **89**: 123–130.

Boise LH, González-García M, Postema CE, Ding L, Lindsten T, Turka LA, Mao X, Nuñez G, Thompson CB. 1993. bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. *Cell* **74**: 597–608.

Chen L. 2013. Characterization and comparison of human nuclear and cytosolic editomes. *Proc Natl Acad Sci* **110**: E2741–E2747.

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, et al. 2012. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148**: 1293–1307.

Cheng X, Lu B, Grant B, Law RJ, McCammon JA. 2006. Channel opening motion of alpha7 nicotinic acetylcholine receptor as suggested by normal mode analysis. *J Mol Biol* **355**: 310–324.

Drolet M, Phoenix P, Menzel R, Massé E, Liu LF, Crouch RJ. 1995. Overexpression of RNase H partially complements the growth defect of an Escherichia coli delta topA mutant: R-loop formation is a major problem in the absence of DNA topoisomerase I. *Proc Natl Acad Sci* **92**: 3526–3530.

El Hage A, French SL, Beyer AL, Tollervey D. 2010. Loss of Topoisomerase I leads to R-loop-mediated transcriptional blocks during ribosomal RNA synthesis. *Genes Dev* **24**: 1546–1558.

Elias JE, Gygi SP. 2010. Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. In *Proteome Bioinformatics* (eds. S.J. Hubbard and A.R. Jones), Vol. 604 of, pp. 55–71, Humana Press, Totowa, NJ http://link.springer.com/10.1007/978-1-60761-444-9_5 (Accessed July 25, 2015).

Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, Dwight SS, Hitz BC, Karra K, Nash RS, et al. 2013. The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now. *G3 GenesGenomesGenetics* **4**: 389–398.

Garrett S, Rosenthal JJC. 2012. RNA Editing Underlies Temperature Adaptation in K+ Channels from Polar Octopuses. *Science* **335**: 848–851.

Ginno PA, Lott PL, Christensen HC, Korf I, Chédin F. 2012. R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters. *Mol Cell* **45**: 814–825.

Green SR, Johnson AD. 2004. Promoter-dependent Roles for the Srb10 Cyclin-dependent Kinase and the Hda1 Deacetylase in Tup1-mediated Repression in Saccharomyces cerevisiae. *Mol Biol Cell* **15**: 4191–4202.

Gualberto JM, Lamattina L, Bonnard G, Weil JH, Grienenberger JM. 1989. RNA editing in wheat mitochondria results in the conservation of protein sequences. *Nature* **341**: 660–662.

Han S-W, Kim H-P, Shin J-Y, Jeong E-G, Lee W-C, Kim KY, Park SY, Lee D-W, Won J-K, Jeong S-Y, et al. 2014. RNA editing in RHOQ promotes invasion potential in colorectal cancer. *J Exp Med* **211**: 613–621.

Hiesel R, Wissinger B, Schuster W, Brennicke A. 1989. RNA editing in plant mitochondria. *Science* **246**: 1632–1634.

Ju YS, Kim J-I, Kim S, Hong D, Park H, Shin J-Y, Lee S, Lee W-C, Kim S, Yu S-B, et al. 2011. Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nat Genet* **43**: 745–752.

Karpievitch YV, Dabney AR, Smith RD. 2012. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics* **13 Suppl 16**: S5.

Karwan R, Blutsch H, Wintersberger U. 1984. A ribonuclease H from yeast stimulates DNA polymerase in vitro. *Adv Exp Med Biol* **179**: 513–518.

Kawahara Y, Ito K, Sun H, Aizawa H, Kanazawa I, Kwak S. 2004. Glutamate receptors: RNA editing and death of motor neurons. *Nature* **427**: 801.

Keller W, Crouch R. 1972. Degradation of DNA RNA hybrids by ribonuclease H and DNA polymerases of cellular and viral origin. *Proc Natl Acad Sci U S A* **69**: 3360–3364.

Kleinman CL, Majewski J. 2012. Comment on “Widespread RNA and DNA sequence differences in the human transcriptome.” *Science* **335**: 1302; author reply 1302.

Klimek-Tomczak K, Mikula M, Dzwonek A, Paziewska A, Karczmarski J, Hennig E, Bujnicki JM, Bragoszewski P, Denisenko O, Bomsztyk K, et al. 2006. Editing of hnRNP K protein mRNA in colorectal adenocarcinoma and surrounding mucosa. *Br J Cancer* **94**: 586–592.

Krestel H, Raffel S, von Lehe M, Jagella C, Moskau-Hartmann S, Becker A, Elger CE, Seeburg PH, Nirkko A. 2013. Differences between RNA and DNA due to RNA editing in temporal lobe epilepsy. *Neurobiol Dis* **56**: 66–73.

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, et al. 2006. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**: 637–643.

Lee RD-W, Song M-Y, Lee J-K. 2013. Large-scale profiling and identification of potential regulatory mechanisms for allelic gene expression in colorectal cancer cells. *Gene* **512**: 16–22.

Li M, Wang IX, Cheung VG. 2012. Response to Comments on "Widespread RNA and DNA Sequence Differences in the Human Transcriptome." *Science* **335**: 1302–1302.

Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. 2011. Widespread RNA and DNA Sequence Differences in the Human Transcriptome. *Science* **333**: 53–58.

Lin W, Piskol R, Tan MH, Li JB. 2012. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome." *Science* **335**: 1302; author reply 1302.

Malavé TM, Dent SYR. 2006. Transcriptional repression by Tup1-Ssn6. *Biochem Cell Biol Biochim Biol Cell* **84**: 437–443.

Martinez HD, Jasavala RJ, Hinkson I, Fitzgerald LD, Trimmer JS, Kung H-J, Wright ME. 2008. RNA Editing of Androgen Receptor Gene Transcripts in Prostate Cancer Cells. *J Biol Chem* **283**: 29938–29949.

Mischo HE, Gómez-González B, Grzechnik P, Rondón AG, Wei W, Steinmetz L, Aguilera A, Proudfoot NJ. 2011. Yeast Sen1 Helicase Protects the Genome from Transcription-Associated Instability. *Mol Cell* **41**: 21–32.

Nagaraj N, Kulak NA, Cox J, Neuhauser N, Mayr K, Hoerning O, Vorm O, Mann M. 2012. System-wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-shot Ultra HPLC Runs on a Bench Top Orbitrap. *Mol Cell Proteomics* **11**: M111.013722.

Niavarani A, Currie E, Reyal Y, Anjos-Afonso F, Horswell S, Griessinger E, Luis Sardina J, Bonnet D. 2015. APOBEC3A Is Implicated in a Novel Class of G-to-A mRNA Editing in WT1 Transcripts ed. D.T. Starczynowski. *PLOS ONE* **10**: e0120089.

Peng Z, Cheng Y, Tan BC-M, Kang L, Tian Z, Zhu Y, Zhang W, Liang Y, Hu X, Tan X, et al. 2012. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat Biotechnol* **30**: 253–260.

Pickrell JK, Gilad Y, Pritchard JK. 2012. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome." *Science* **335**: 1302; author reply 1302.

Rubio MAT, Paris Z, Gaston KW, Fleming IMC, Sample P, Trotta CR, Alfonzo JD. 2013. Unusual Noncanonical Intron Editing Is Important for tRNA Splicing in Trypanosoma brucei. *Mol Cell* **52**: 184–192.

Sharma PM, Bowman M, Madden SL, Rauscher FJ 3rd, Sukumar S. 1994. RNA editing in the Wilms' tumor susceptibility gene, WT1. *Genes Dev* **8**: 720–731.

Shoichet BK, Baase WA, Kuroki R, Matthews BW. 1995. A relationship between protein stability and protein function. *Proc Natl Acad Sci U S A* **92**: 452–456.

Silberberg G, Lundin D, Navon R, Öhman M. 2012. Deregulation of the A-to-I RNA editing mechanism in psychiatric disorders. *Hum Mol Genet* **21**: 311–321.

Skourti-Stathaki K, Proudfoot NJ, Gromak N. 2011. Human Senataxin Resolves RNA/DNA Hybrids Formed at Transcriptional Pause Sites to Promote Xrn2-Dependent Termination. *Mol Cell* **42**: 794–805.

Stein H, Hausen P. 1969. Enzyme from calf thymus degrading the RNA moiety of DNA-RNA Hybrids: effect on DNA-dependent RNA polymerase. *Science* **166**: 393–395.

Tanaka N, Mukai Y. 2015. Yeast Cyc8p and Tup1p proteins function as coactivators for transcription of Stp1/2p-dependent amino acid transporter genes. *Biochem Biophys Res Commun* **468**: 32–38.

Teng B, Burant CF, Davidson NO. 1993. Molecular cloning of an apolipoprotein B messenger RNA editing protein. *Science* **260**: 1816–1819.

Tennyson GE, Sabatos CA, Higuchi K, Meglin N, Brewer HB. 1989. Expression of apolipoprotein B mRNAs encoding higher- and lower-molecular weight isoproteins in rat liver and intestine. *Proc Natl Acad Sci U S A* **86**: 500–504.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–515.

Tuduri S, Crabbé L, Conti C, Tourrière H, Holtgreve-Grez H, Jauch A, Pantesco V, De Vos J, Thomas A, Theillet C, et al. 2009. Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. *Nat Cell Biol* **11**: 1315–1324.

Turner AJ, Aggarwal P, Miller HE, Waukau J, Routes JM, Broeckel U, Robinson RT. 2015. The introduction of RNA-DNA differences underlies interindividual variation in the human IL12RB1 mRNA repertoire. *Proc Natl Acad Sci* **112**: 15414–15419.

Tyanova S, Mann M, Cox J. 2014. MaxQuant for in-depth analysis of large SILAC datasets. *Methods Mol Biol Clifton NJ* **1188**: 351–364.

van Dijk EL, Chen CL, d'Aubenton-Carafa Y, Gourvennec S, Kwapisz M, Roche V, Bertrand C, Silvain M, Legoix-Né P, Loeillet S, et al. 2011. XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature* **475**: 114–117.

van Leeuwen FW, de Kleijn DPV, van den Hurk HH, Neubauer A, Sonnemans MAF, Sluijs JA, Köycü S, Ramdjielal RDJ, Salehi A, Martens GJM, et al. 1998. Frameshift Mutants of β Amyloid Precursor Protein and Ubiquitin-B in Alzheimer's and Down Patients. *Science* **279**: 242–247.

Vazquez F, Ramaswamy S, Nakamura N, Sellers WR. 2000. Phosphorylation of the PTEN tail regulates protein stability and function. *Mol Cell Biol* **20**: 5010–5018.

Vesely C, Tauber S, Sedlazeck FJ, von Haeseler A, Jantsch MF. 2012. Adenosine deaminases that act on RNA induce reproducible changes in abundance and sequence of embryonic miRNAs. *Genome Res* **22**: 1468–1476.

Volchkov VE, Volchkova VA, Muhlberger E, Kolesnikova LV, Weik M, Dolnik O, Klenk HD. 2001. Recovery of infectious Ebola virus from complementary DNA: RNA editing of the GP gene and viral cytotoxicity. *Science* **291**: 1965–1969.

Wang IX, Core LJ, Kwak H, Brady L, Bruzel A, McDaniel L, Richards AL, Wu M, Grunseich C, Lis JT, et al. 2014. RNA-DNA Differences Are Generated in Human Cells within Seconds after RNA Exits Polymerase II. *Cell Rep* **6**: 906–915.

Wang IX, So E, Devlin JL, Zhao Y, Wu M, Cheung VG. 2013. ADAR Regulates RNA Editing, Transcript Stability, and Gene Expression. *Cell Rep* **5**: 849–860.

Williams JS, Smith DJ, Marjavaara L, Lujan SA, Chabes A, Kunkel TA. 2013. Topoisomerase 1-Mediated Removal of Ribonucleotides from Nascent Leading-Strand DNA. *Mol Cell* **49**: 1010–1015.

Worth CL, Preissner R, Blundell TL. 2011. SDM--a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res* **39**: W215-222.

Wu H-Y, Burgess SM. 2006. Ndj1, a telomere-associated protein, promotes meiotic recombination in budding yeast. *Mol Cell Biol* **26**: 3683–3694.

Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinforma Oxf Engl* **26**: 873–881.

Wyers F, Sentenac A, Fromageot P. 1973. Role of DNA-RNA hybrids in eukaryotes. Ribonuclease H in yeast. *Eur J Biochem FEBS* **35**: 270–281.

Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**: 40.

Zhang Z, Li C, Wu F, Ma R, Luan J, Yang F, Liu W, Wang L, Zhang S, Liu Y, et al. 2015. Genomic variations of the mevalonate pathway in porokeratosis. *eLife* **4**: e06322.

Table 1. Examples of RDDs in genes that play a role in ribosome biogenesis.*

Gene	Chr	Position	RDD	Amino Acid	RDD levels (%)					
					BY4741	BY4742	S288C	SNM8	W303-1A	W303-1B
NOP6	IV	77382	C>A	syn	-	-	-	10.5	-	14.3
NOP14	IV	189726	T>A	D>E	9.5	-	-	6.5	14	-
RRP1	IV	617886	T>C	W>R	5.8	-	5.6	-	-	-
ESF1	IV	1205918	T>C	W>R	-	-	5.9	-	-	8.3
ENP2	VII	783759	T>C	S>P	-	-	-	5.7	5	-
SDA1	VII	980392	C>A	N>K	7.5	-	-	7.1	-	-
CIC1	VIII	211854	T>C	I>T	11.5	-	-	-	5.3	-
MTR4	X	344335	T>C	L>S	10.5	-	-	-	8.7	-
DIP2	XII	400595	T>A	syn	-	-	12	-	9.8	-
ERB1	XIII	368739	C>T	syn	5.8	-	5	6.3	-	-
ECM16	XIII	525831	A>T	E>D	5.1	-	5.9	-	-	-
RRP5	XIII	727288	G>A	G>D	-	-	8.3	7.1	6.9	-
KRI1	XIV	55287	G>A	G>E	5.6	-	-	-	-	6.1
LSM7	XIV	351083	G>A	syn	-	8.7	-	-	6.6	11.8
HRR25	XVI	164682	T>C	M>T	-	-	5.3	-	5.9	-
NOP4	XVI	471281	T>C	V>A	-	-	-	9.1	5.3	-

*RDD levels for 6 wild-type yeast strains are shown. "-" are RDD with levels below 5% or where no RDD was detected. syn = synonymous

Table 2. Peptides from RNA- and DNA-bases at RDD sites detected by mass spectrometry.

Protein	Peptide Sequence	Encoded by	RDD position	Codon Change	AA change	Source
Abp1	IGNPLPGMHIEADNEEEPEENDDDWDDDEDEAAQPPLPSR	DNA	chrIII:266393	GAT>GAG	D442E	This study
	IGNPLPGMHIEADNEEEPEENDDDWDDEEDEAAQPPLPSR	RNA				
Arf2	LGEVITTIPTIGFNVETVQYK	DNA	chrIV:216673	ATT>GTT	I49V	This study
	LGEVITTIPTVGFNVETVQYK	RNA				
Ded1	GLHEILTEANQEVPSFLK	DNA	chrXV:724489	TCA>CCA	S345P	This study
	GLHEILTEANQEVPPFLK	RNA				
Fra1	DAVCLVQYFAWLEQQLAGR	RNA	chrXII:82920	GTG>GCG	V487A	This study
Gcr1	LIENYDWRR	RNA	chrXVI:415019	GGC>GAC	G724D	This study
Gip3	LAENLLK	RNA	chrXVI:295344	TCA>TTA	S435L	Both
Hsp90	EEVQELEELNK	DNA	chrXIII:633135	CAA>GAA	Q178E	Both
	EEVEELEELNK	RNA				
Kcs1	NSFCNSSSPILTATNSR	RNA	chrIV:480262	TTA>TCA	L669S	This study
Krs1	LAMFLTDSNTIR	DNA	chrIV:527109	CTG>CGG	L403R	This study
	RAMFLTDSNTIR	RNA				
Ksp1	LSTEQKFK	RNA	chrVIII:269066	ATG>ACG	M697T	This study
Mdm34	ISLNLDPKK	RNA	chrVII:82991	TCC>CCC	S251P	Nagaraj et
Noc2	SEQMELEK	DNA	chrXV:728017	ATG>ACG	M60T	This study
	SEQTELEK	RNA				
Nup159	TVTFFEK	RNA	chrIX:148342	TCT>TTT	S493F	Nagaraj et
Pop1	LNADQFISSR	RNA	chrXIV:233520	GTG>GCG	V59A	This study
Psk1	EGDEFEQSLR	DNA	chrI:120439	TTC>TCC	F72S	This study
	EGDESEQSL	RNA				
Rho3	VALTAGPVATEVK	DNA	chrIX:140376	GTT>CTT	V209L	Nagaraj et al
	LALTAGPVATEVK	RNA				
Rpn1	HNGEEDAVDLLLEIESIDK	DNA	chrVIII:164051	GTA>ATA	V221I	This study
	HNGEEDAIDLLLEIESIDK	RNA				
Rsa4	TVRVWDINSQGR	DNA	chrIII:241443	TGG>CGG	W170R	Nagaraj et al
	TVRVRDINSQGR	RNA				
Tef1	SVEMHHEQLEQGVPGDNVGFNVK	DNA	chrXVI:701513	GGT>GCT	G34A	Both
	SVEMHHEQLEQGVPGDNVAFNVK	RNA				
Tup1	FLATGAEDR	DNA	chrIII:261077	GCA>GTA	A459V	This study
	FLVTGAEDR	RNA				
Vph1	AIFEILNK	DNA	chrXV:829619	ATT>AAT	I196N	This study
	ANFEILNK	RNA				
Vps5	NGMEISLEEAIESQK	RNA	chrXV:455731	GCG>GAG	A292E	Nagaraj et
YIL108W	TFPFVEEFTWDTLFER	DNA	chrIX:161376	TGG>CGG	W164R	This study
	TFPFVEEFTR	RNA				
Ypk1	GTINPSNSSVVPVR	DNA	chrXI:205902	GTC>ATC	V66I	This study
	GTINPSNSSVIPVR	RNA				
Ypt10	DANIALIVFESGDVSSLQCAK	RNA	chrII:738095	TTG>TCG	L111S	This study
Yta12	SMVKVMLNDNEK	RNA	chrXIII:447368	GGA>GAA	G240E	Nagaraj et

* Amino acids from RDDs are underscored. For each RDD-containing peptide, the corresponding DNA-form counterpart is also shown if detected by mass spectrometry.

.

Table 3. Differences in protein structure and stability between RNA-form and DNA-form of proteins.

Protein	RDD	Amino Acid Change	ΔΔG*	Stability Change
Arf2	A>G	I49V	-0.41	Neutral
Ded1	T>C	S527P	0.32	Neutral
Krs1	T>G	L557R	-3.25	Destabilized
Noc2	T>C	M169T	-0.13	Neutral
Psk1	T>C	F72S	-2.06	Destabilized
Rpn1	G>A	V221I	-0.1	Neutral
Rsa4	T>C	W304R	-0.79	Destabilized
Tup1	C>T	A459V	1.02	Stabilized
Vph1	T>A	I319N	-5.16	Destabilized
YIL108W	T>C	W164R	-2.83	Destabilized

* $-0.5 \leq \Delta\Delta G \leq 0.5$ is considered neutral.

Figure Legends

Figure 1. RNA-DNA sequence differences in *S. cerevisiae*. (A) 12 types of RDDs are found in S288C. (B) RDD frequencies are similar in six wild-type yeast (1.3~1.8 per 10,000 nucleotides).

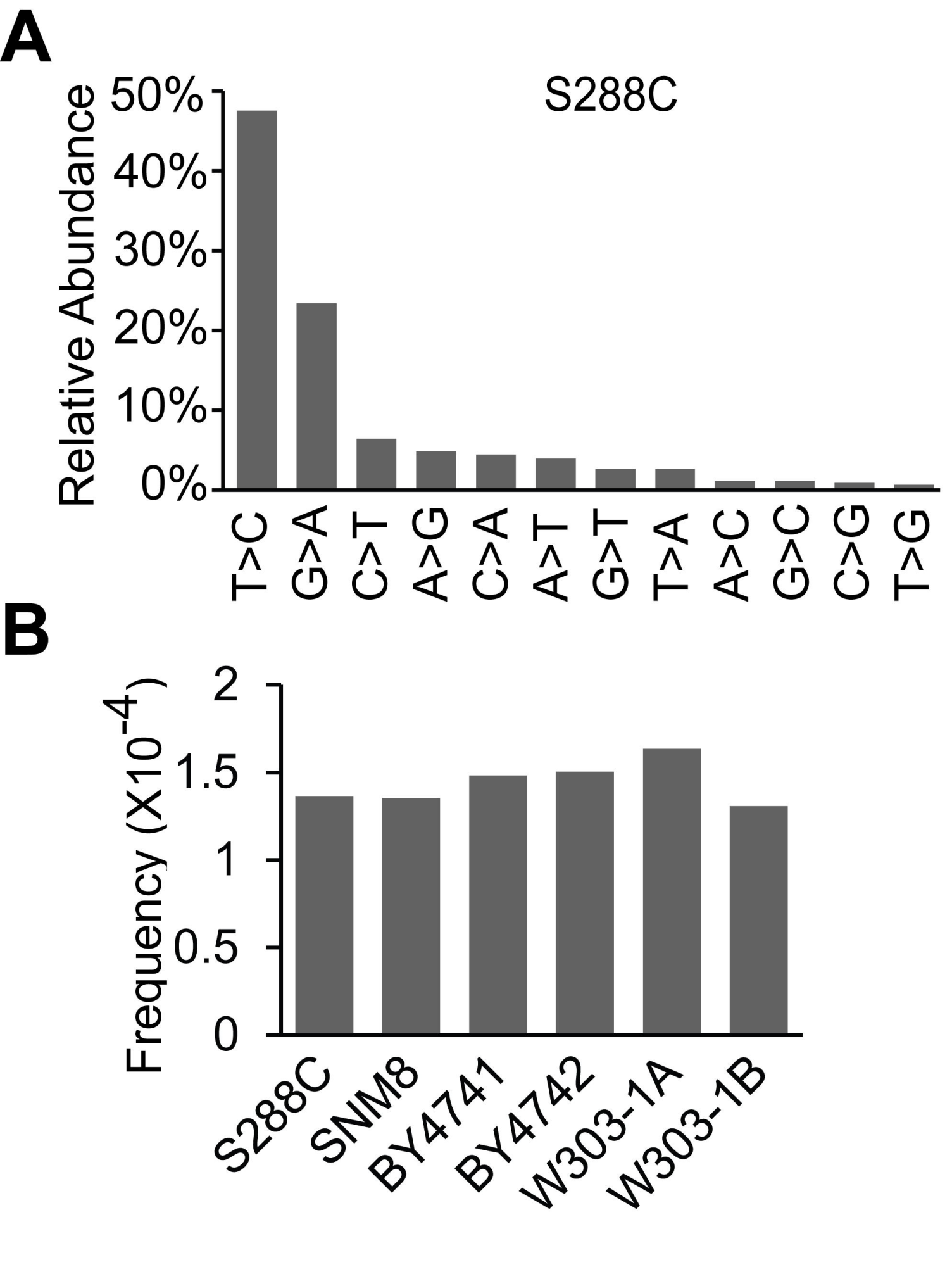
Figure 2. Representative example of an RDD identified by deep sequencing and droplet digital PCR. A C-to-G RDD in coding exon of *ADR1* shown by (A) RNA-seq and DNA-seq data on the Integrated Genomics Viewer, (B) and by droplet digital PCR.

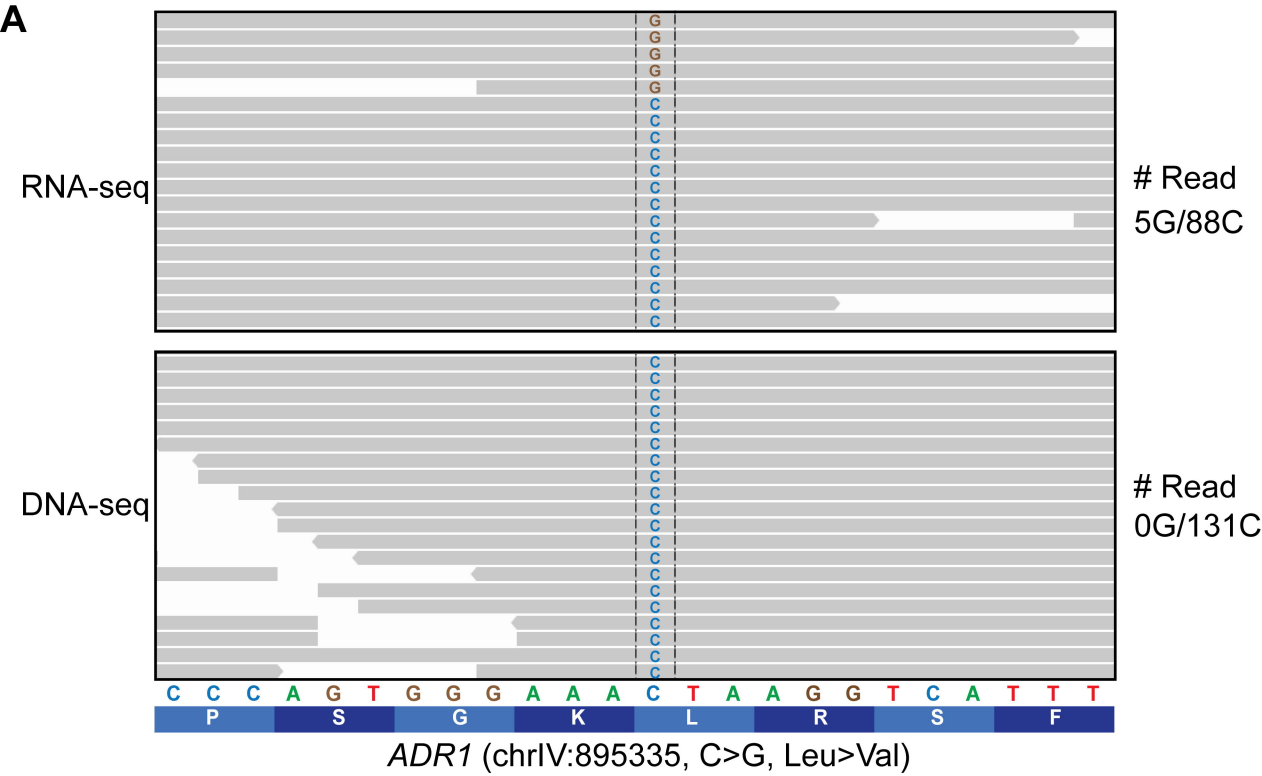
Figure 3. LC/MS-MS identified DNA-encoded isoleucine and RNA-encoded valine forms of Rpn1. (A) G-to-A RDD in the coding exon of *RPN1* shown by RNA-seq and DNA-seq data on the Integrated Genomics Viewer. (B) Spectra from the MaxQuant Program for the resulting Val- and Ile-bearing peptides of Rpn1.

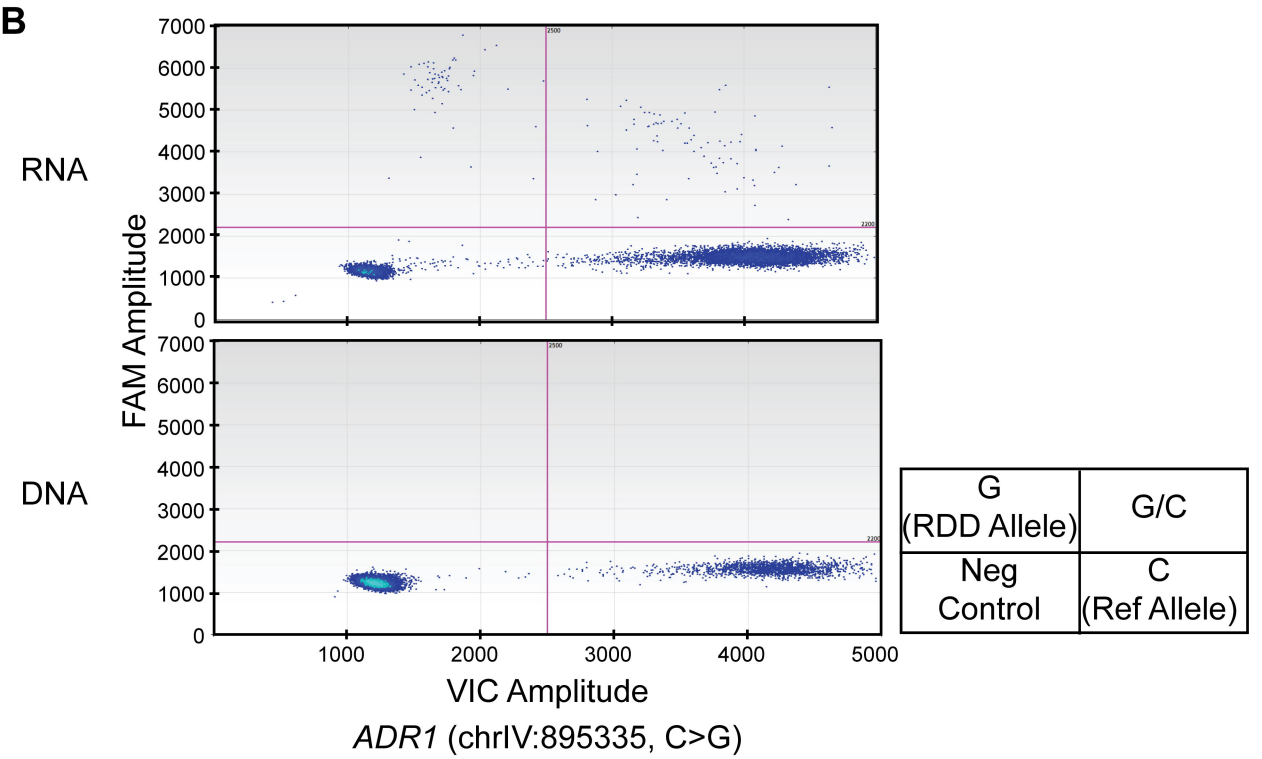
Figure 4. 3D-structures of DNA and RNA-forms of proteins. Arf2, Vph1 and Tup1 are shown as examples of DNA and RNA forms of the proteins with distinct features as predicted using I-TASSER.

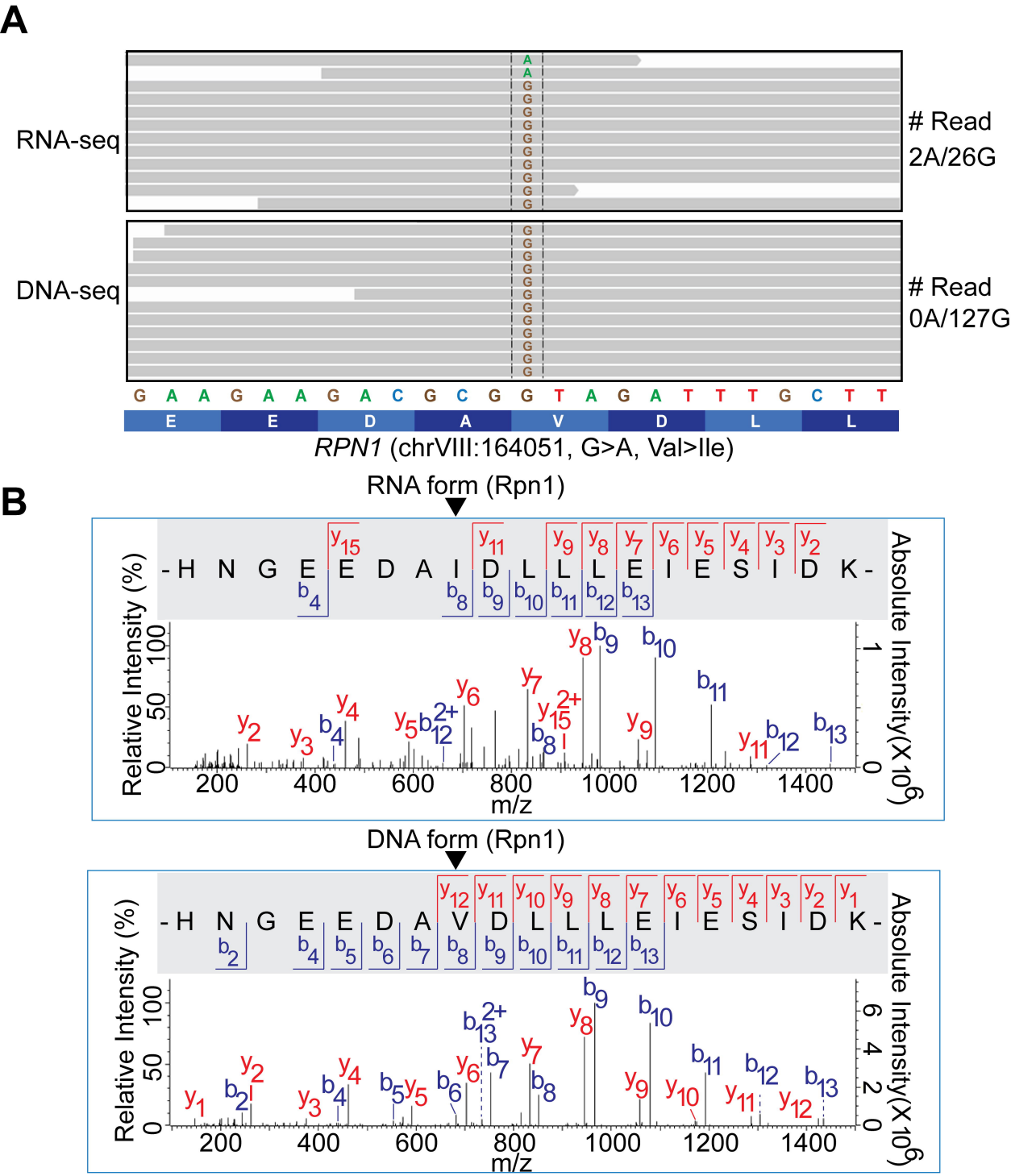
Figure 5. Effects of an RDD in *TUP1* on protein stability and function. (A) DNA-form and RNA-form of *TUP1* were cloned into yeast plasmids under *GAL1* promoter, and expression of each form was induced by galactose. The two forms had similar mRNA levels, as determined by real-time RT-PCR. Yeast colonies (n=3) transformed with empty vector, DNA-form and RNA-form of *TUP1* are shown. Error bars represent standard deviation of three colonies. (B) Protein level of the RNA-form of Tup1 is higher than that of DNA-form. Whole cell extracts of the same colonies in (A) were analyzed by western blot. Intensity of each band of Tup1 is quantified using ImageJ and normalized to that of GAPDH, and their values are shown under each band. (C) Cycloheximide chase assay showed that RNA-form of Tup1 is more stable than the DNA-form. Act1 level from each transformant was measured as control. (D) Genes that are differentially regulated by DNA- and RNA-form of Tup1. Gene expression levels were measured by RNA-seq of yeast cells expressing either form of Tup1. (E) Yeast expressing the RNA-form of Tup1 is more sensitive to Hygromycin-B. Yeast cells were cultured in liquid medium; 10X serial dilution were spotted onto plates containing Hygromycin-B or DMSO control to measure cell growth. Tup1 expression is induced by galactose in plates. Cells were also spotted on plates containing glucose as negative control.

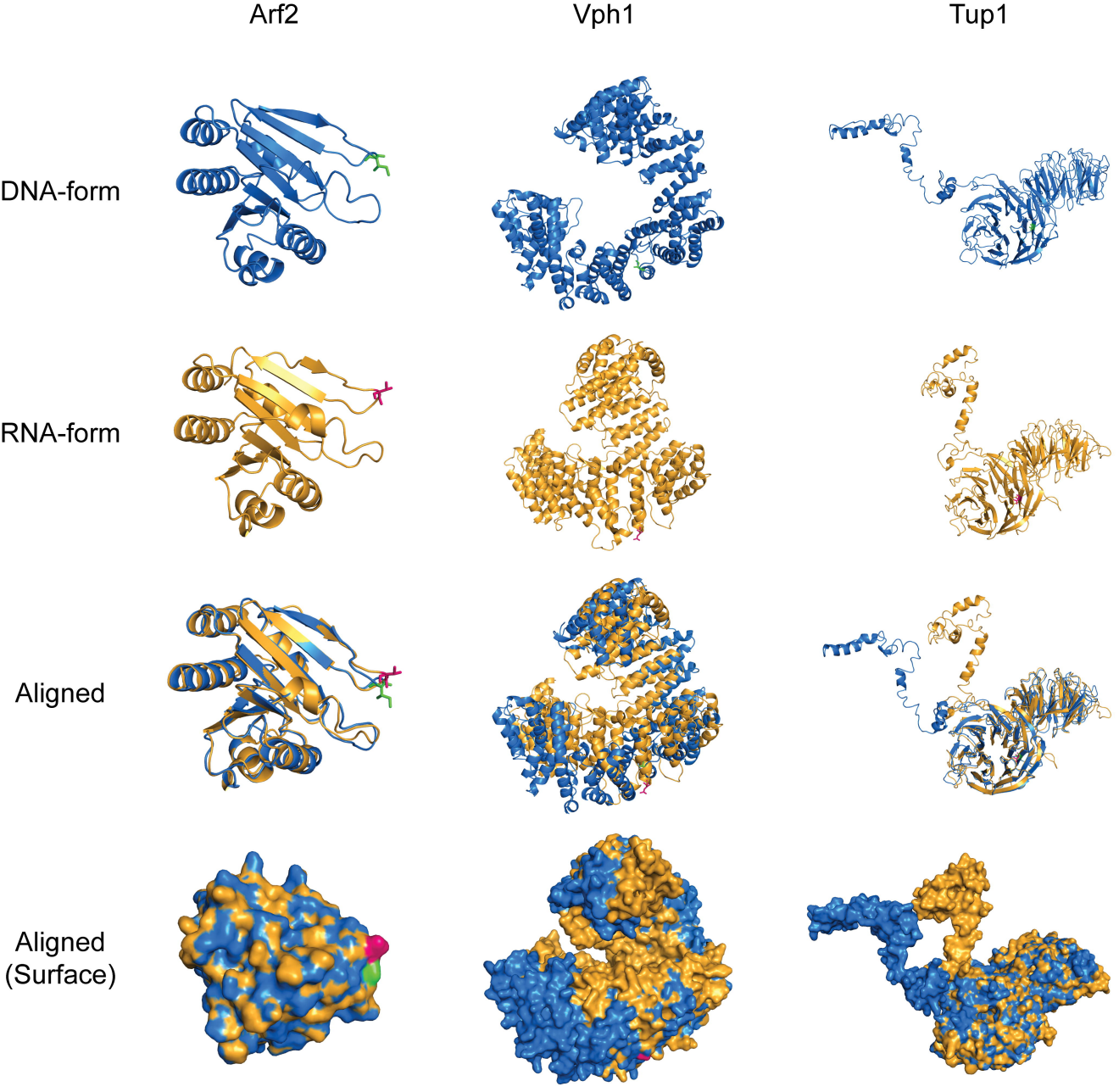
Figure 6. RDDs and R-loops are coupled. (A) Genome-wide distribution of R-loops and RDDs. Metagene plots show overlap between R-loops and RDDs. TSS: transcription start site. CPS: cleavage and polyadenylation signal. Gene body: annotated coding sequence in reference genome. (B) Nucleic acid blots probed with S9.6 antibody show more RNA-DNA hybrids in *rnh1*$^{-}$, *top1*$^{-}$ and *sen1-1* mutants, compared to wild-type control. In the temperature sensitive mutant of *sen1-1*, at non-permissive temperature, RDD frequency (C) and levels (D) are higher. (E) In two different strain background, RDD frequencies are higher in *rnh1*$^{-}$ and *top1*$^{-}$ mutants than in wild-type control. (F) RDD levels are higher in *rnh1*$^{-}$ and *top1*$^{-}$ mutants than in wild-type control. Two Xrn1-sensitive unstable transcripts (XUTs) are annotated according to XUT track in SGD genome browser (van Dijk et al. 2011).

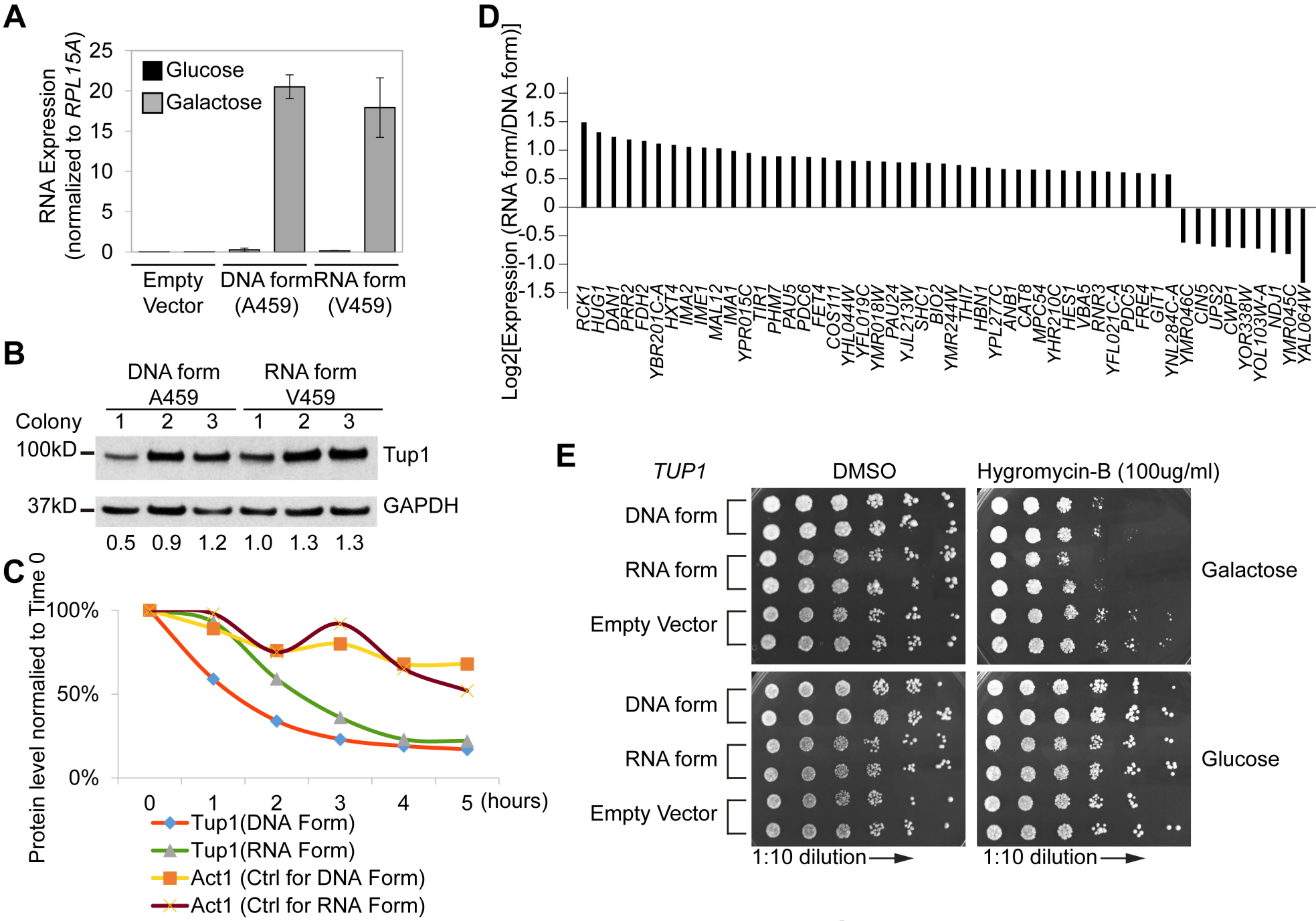

A
S288C
Relative Abundance
50%
40%
30%
20%
10%
0%
T>C
G>A
C>T
A>G
C>A
A>T
G>T
T>A
A>C
G>C
C>G
T>G
B
Frequency (X10-4)
2
1.5
1
0.5
0
S288C
SNM8
BY4741
BY4742
W303-1A
W303-1B

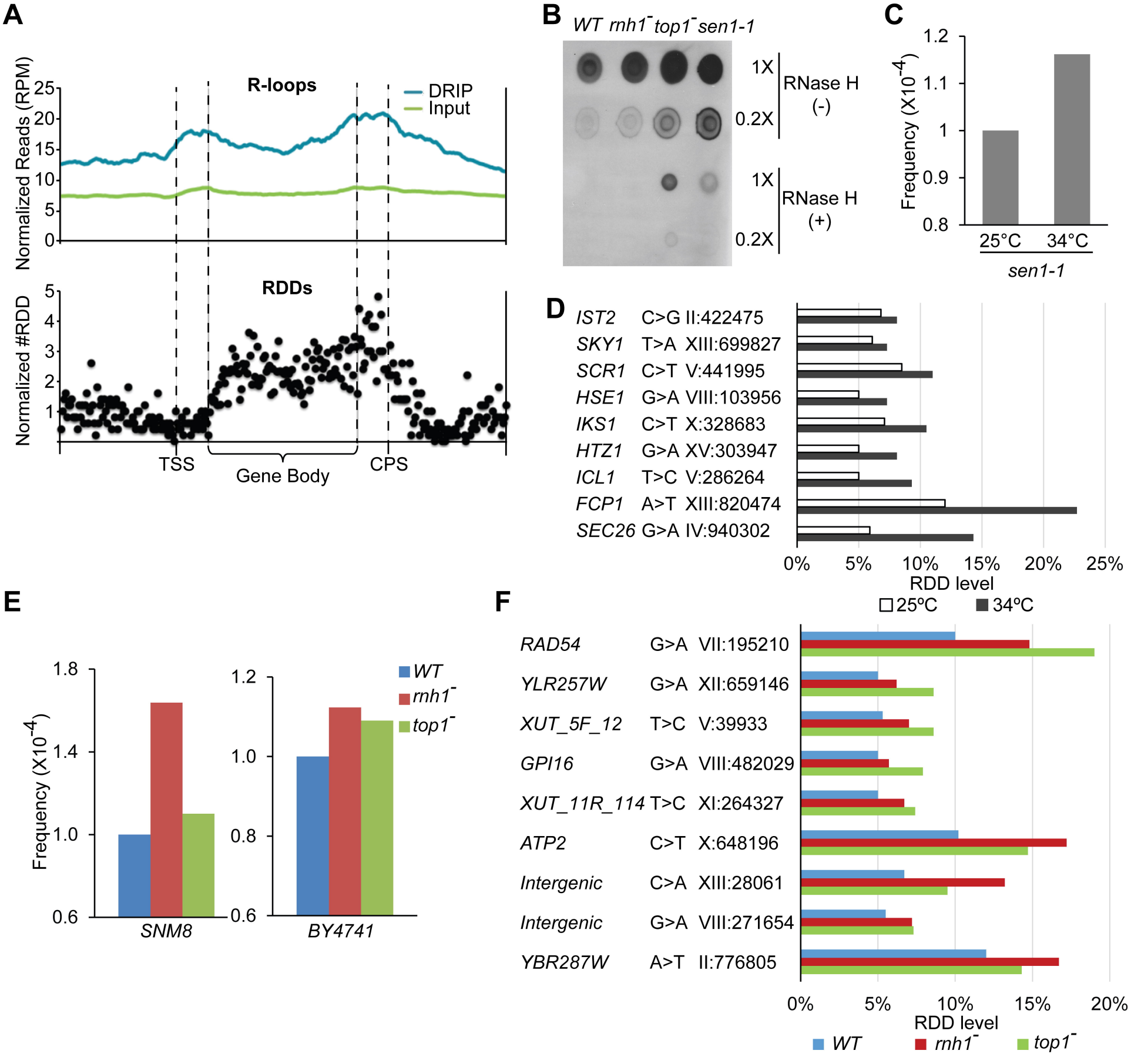

A
RNA-seq
Read
5G/88C
DNA-seq
Read
0G/131C
C C C A G T G G G A A A C T A A G G T C A T T T
P S G K L R S F
ADR1 (chrIV:895335, C>G, Leu>Val)


B
RNA
DNA
FAM Amplitude
VIC Amplitude
7000
6000
5000
4000
3000
2000
1000
0
1000
2000
3000
4000
5000
G (RDD Allele)
G/C
Neg Control
C (Ref Allele)
ADR1 (chrIV:895335, C>G)


A
RNA-seq
Read
2A/26G
DNA-seq
Read
0A/127G
G A A G A A G A C G C G G T A G A T T T G C T T
E E D A V D L L
RPN1 (chrVIII:164051, G>A, Val>Ile)
B
RNA form (Rpn1)
-H N G E E D A I D L L L E I E S I D K-
Relative Intensity (%)
Absolute Intensity(X 10^6)
m/z
DNA form (Rpn1)
-H N G E E D A V D L L L E I E S I D K-
Relative Intensity (%)
Absolute Intensity(X 10^6)
m/z


Arf2
Vph1
Tup1
DNA-form
RNA-form
Aligned
Aligned
(Surface)


A
RNA Expression (normalized to RPL15A)
25
20
15
10
5
0
Glucose
Galactose
Empty Vector
DNA form (A459)
RNA form (V459)
B
DNA form A459
RNA form V459
Colony 1 2 3 1 2 3
100kD
Tup1
37kD
GAPDH
0.5 0.9 1.2 1.0 1.3 1.3
C
Protein level normalied to Time 0
100%
50%
0%
0 1 2 3 4 5 (hours)
Tup1(DNA Form)
Tup1(RNA Form)
Act1 (Ctrl for DNA Form)
Act1 (Ctrl for RNA Form)
D
Log2[Expression (RNA form/DNA form)]
2.0
1.5
1.0
0.5
0
-0.5
-1.0
-1.5
RCK1
HUG1
DAN1
PRR2
FDH2
YBR201C-A
HXT4
IMA2
IME1
MAL12
IMA1
YPR015C
TIR1
PHM7
PAU5
PDC6
FET4
COS111
YHL044W
YFL019C
YMR018W
PAU24
YJL213W
SHC1
BIO2
YMR244W
THI7
HBN1
YPL277C
ANB1
CAT8
MPC54
YHR210C
HES1
VBA5
RNR3
YFL021C-A
PDC5
FRE4
GIT1
YNL284C-A
YMR046C
CIN5
UPS2
CWP1
YOR338W
YOL103W-A
NDJ1
YMR045C
YAL064W
E
TUP1
DMSO
Hygromycin-B (100ug/ml)
DNA form
RNA form
Empty Vector
Galactose
DNA form
RNA form
Empty Vector
Glucose
1:10 dilution
1:10 dilution


A
R-loops
DRIP
Input
Normalized Reads (RPM)
RDDs
Normalized #RDD
TSS
Gene Body
CPS
B
WT rnh1- top1- sen1-1
1X
0.2X
RNase H (-)
RNase H (+)
C
Frequency (X10-4)
25°C
34°C
sen1-1
D
IST2 C>G II:422475
SKY1 T>A XIII:699827
SCR1 C>T V:441995
HSE1 G>A VIII:103956
IKS1 C>T X:328683
HTZ1 G>A XV:303947
ICL1 T>C V:286264
FCP1 A>T XIII:820474
SEC26 G>A IV:940302
RDD level
25ºC
34ºC
E
Frequency (X10-4)
WT
rnh1-
top1-
SNM8
BY4741
F
RAD54 G>A VII:195210
YLR257W G>A XII:659146
XUT_5F_12 T>C V:39933
GPI16 G>A VIII:482029
XUT_11R_114 T>C XI:264327
ATP2 C>T X:648196
Intergenic C>A XIII:28061
Intergenic G>A VIII:271654
YBR287W A>T II:776805
RDD level
WT
rnh1-
top1-