Supplemental Material

RNA-DNA Sequence Differences in *Saccharomyces cerevisiae*

I. X. Wang*, C. Grunseich, Y. G. Chung, H. Kwak, G. Ramrattan, Z. Zhu, V. G. Cheung*

Supplemental Results

RNH1 deletion leads to petite phenotype

It was previously reported that deletion of *RNH1* leads to petite phenotype, suggesting RNase H1 is important for maintenance of mitochondria DNA (Bernardi 1979; El Hage et al. 2014). To examine frequency of petite colonies, ~500 cells from wild type and mutant strain were plated onto YPD plates, incubated at 25°C for 3 days and colonies were quantified for petite phenotype. *rnh1*$^{-}$ mutant showed an increase in petite colonies (18%) compared to wild type control (5%), consistent with previous report (El Hage et al. 2014).

RDDs are not sequencing or mapping errors

One of the motivations to study RDD in yeast is to take advantage of its smaller and less complex genome (fewer repetitive sequences; <5% genes contain introns) compared to human. However, even in yeast, analysis of the sequencing results is not trivial. Each read from deep sequencing is only about 100 nucleotides in length. To align the 40 million DNA-seq reads and 10 million RNA-seq reads to the 12 Mb genome is complex. Many alignment algorithms have been developed to analyze deep sequencing data. They have different strengths and weaknesses. For our initial analysis, we used GSNAP (Wu and Nacu, 2010). To determine how different alignment programs may affect our RDD counts, we repeated the analyses using TopHat2 and STAR (Djebali et al. 2012; Kim et al. 2013). We also tested different thresholds in GSNAP by varying the number of mismatches allowed in the alignment. The comparison of the different alignment programs showed that GSNAP is the most conservative in that it yielded the fewest number of RDDs (759 for S288C), while TopHat2 identified the largest number of RDDs (3,120 for S288C) (Supplemental Table S4). It is important to note that there are substantial overlaps between the alignments; hundreds of the RDDs are identified by multiple programs. In

Supplemental Figure S4, we showed an example of RDD in *IMG1* identified by GSNAP, STAR and TopHat2. These results show that most of the RDDs can be identified regardless of the alignment programs used in the sequence analysis.

In alignment of sequence reads, those that span splice junctions are more prone to misalignment, especially when those junctions fall in the ends of the sequence reads. The alignment programs have to decide whether to place the short overhangs in the RNA-seq reads to the intronic regions or to the next exons. This can lead to misalignment and false identification of differences in DNA and RNA sequences. In budding yeast, only 5% of genes contain introns, and the majority of these genes contain only one intron (less than 20 genes have two or more introns). Of the 759 RDDs in the S288C strain, only 7 RDDs mapped to introns. In our initial analysis, GSNAP is set as "splice-aware" mode where reads can be mapped to annotated and novel splice junctions. Some groups have suggested that the "splice-aware" method could generate misalignment. Given the paucity of introns in S. cerevisiae and the small number of RDDs found in intron-containing genes, we expect that splicing has little impact on our alignments and RDD calls. To confirm this, we re-aligned our RNA-seq reads using Bowtie2 local alignment mode and turned off splice-aware option of GSNAP; both of these will not allow alignment across exon/intron junctions. We compared the results from the GSNAP analyses with splice-aware option on and off, and from Bowtie2. The three methods showed highly similar results; most of the RDDs were identified by all three methods (Supplemental Fig. S5A); only 24 RDDs in S288C were identified by the splice-aware GSNAP alone (no overlap with the 7 intronic RDDs described above). This shows that splice junctions do not contribute to false discovery of RDDs in this project.

We believe RDD is unlikely due to sequencing errors. The error rate in deep sequencing is rather low (Illumina reports it to be <0.1%) and we required RDDs to be supported by at least two independent reads, therefore the likelihood of two independent errors occurring at same position is 1 per million. Regardless, to be sure that RDDs are not due to sequencing errors, we assessed our results in three different ways. First, we calculated error rates in PhiX libraries added as internal controls in our sequencing experiments. The PhiX sequence is known, which allowed us to calculate the error rate and we found it to be <0.1%, as one should expect from Illumina sequencing. We also compared the PhiX sequences we generated to its known DNA sequence using the same analysis parameters for RDD identification, and did not detect any discrepancies between the two (Supplemental Table S5). Second, we confirmed that the RDD frequencies in our samples are higher than predicted from simulated RNA-seq data generated using the yeast genome sequence and the Flux simulator (Griebel et al. 2012). The distribution of RDD types in the simulated data was different from what we observed in yeast samples (Supplemental Fig. S5B). Lastly, we estimated error rate using a probability-based test developed by Chepelev for identifying RNA editing events (Chepelev 2012), and found that our RDD sites are unlikely to result from sequencing ($P<0.05$) or mapping errors ($P<0.001$) (Supplemental Table S6).

Although we applied Phred score ≥20 as the cutoff in initial analysis, vast majority of sequenced bases have higher Phred scores. When we increase the threshold for Phred score, very similar numbers of RDDs were identified. Notably even more RDDs were identified when Phred score ≥35 was applied. This is due to a decrease of total number of RNA-seq reads; manual inspection confirmed that the reads that dropped out were those with lower phred scores at non-RDD bases, consequently there is an increase of RDD levels, and more RDDs (more that

pass the threshold of ≥5% level) (Supplemental Table S7). To assess whether distribution of RDDs within reads will bias our conclusion, we removed RDDs if they reside within 5 nt or 10 nt from the ends of reads. We still found all 12 types of RDDs at each threshold (Supplemental Fig. S5C)

Together, these analyses show that RDD sites are identified by different alignment programs, and the RDDs are not results of inaccuracies in sequencing technology.

RDDs are not caused by rare genomic mutations

To ensure that RDDs are detected at sites where there is no genomic mutation, only sites that are monomorphic in DNA are used for analysis. Since we sequenced the genomes to obtain high coverage, >96% of these sites are covered by 100 or more DNA-seq reads (minimum of 10 reads). This ensures that we are confident of the DNA sequences in our analysis. On average, each RDD site is supported by 180 DNA and 33 RNA reads; these sequence coverage allows us to be confident of the underlying sequences. Moreover, to assess the probability of detecting DNA mutation as RDDs, we split DNA sequencing data into two sets, and identified "DNA-DNA sequence difference (DDD)" using the same RDD identification methods. We found 125 DDD in over 10 million sites (10,026,631 sites), significantly fewer than RDDs ($\chi 2$, $P<0.0001$). None of these overlaps with the RDD sites. As expected, none of the sites with "DDD" are included in our analysis since we only included sites where DNA reads do not show alternative bases.

Identification of RDD-form peptides by mass spectrometry

We took three approached to ensure accurate identification of peptides encoded by RDDs. First, we applied stringent criteria in analysis of mass spectrometry data in MaxQuant. The peptide tolerance of 4.5ppm is much more stringent than required to detect small mass difference

between single amino acids encoded by DNA-form or RNA-form, respectively. Second, we searched sequence of each identified RDD-encoded peptides against yeast protein databases using BLAST to ensure they are unique peptides. Lastly, we used the target-decoy strategy to estimate false positive peptides (FDR<0.01).

Moreover, we performed immunoprecipitation with Tup1 antibody followed by gel electrophoresis, then we carried out mass spectrometric analysis on the gel-purified protein band of the expected molecular weight for Tup1, in the hope that we would enrich for peptides corresponding to Tup1. In the first attempt, we did not get any peptide that corresponds to the RDD site. Then after we scaled up the experiment, we found four Tup1 peptides that span the RDD site but these four peptides corresponded to the DNA form and not the RDD-encoded peptide. We reasoned that we found so few peptides that span the RDD site because multiple post-translational modifications of Tup1 led to different electrophoretic pattern(s) than its unmodified form. Tup1 is modified by acetylation, phosphorylation and ubiquitination at over 19 residues (Albuquerque et al. 2008; Soulard et al. 2010; Swaney et al. 2013; Weinert et al. 2013). Since mass spectrometry analysis of the excised band yielded only a few peptides corresponding to Tup1, a large fraction of the protein most likely migrated differently from that of the unmodified protein. In addition, Tup1 is part of a large protein complex thus 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.

Co-localization of R-loops and RDDs

We mapped R-loops by DNA-RNA immunoprecipitation with S9.6 antibody. We identified 1,505 R-loop peaks in BY4741 that span 7% of the genome. We asked whether RDDs

co-localize with these R-loops, and found that RDDs are significantly enriched ($P<0.0001$) in R-loop regions. Among the 829 RDDs found in the same strain, 96 were found within the R-loop regions. When we lowered the thresholds for read depth and fold enrichment in R-loop peak calling (DRIP sequencing read depth = 5RPM, fold enrichment = 1.2), 346 (42%) RDDs were found in regions covered by DRIP-seq reads.

Supplemental Methods

Yeast cultures. Each strain was cultured on YPAD plates. Single colonies were inoculated into 2 ml start cultures of YPAD medium, and kept in a shaking incubator at 25°C, 250 rpm overnight. Cells were then counted and diluted into 200 ml fresh YPAD medium to 5 X 10^5 cells/ml and incubated at 25°C, 250 rpm until reaching early log phase (2~4 X 10^6 cells/ml). Yeast cells were harvested by centrifugation at 1,000 g for 10 min, and washed twice in PBS. For temperature-sensitive mutants, yeast cells were grown in YPAD at 25°C to early log phase and then shifted to 34°C and cultured for 4 hours.

DNA sequencing and RNA sequencing. 5 X 10^6 yeast cells harvested from the same cultures were divided for purification of DNA or RNA. DNA and RNA were extracted using MasterPure Yeast DNA or RNA Purification Kit (Epicentre). Sequencing libraries were prepared from 1μg of genomic DNA or total RNA using TruSeq Nano DNA LT Sample Preparation Kit or TruSeq Stranded RNA LT kit with Ribo zero gold (Illumina), respectively. Then they were sequenced on a HiSeq 2500 instrument to 100-nt read lengths. On average 40 million reads were obtained from each DNA-seq sample and 10 million reads from each RNA-seq sample.

Functional analysis of RNA-form of Tup1. The pBY011-TUP1 plasmid containing *GAL1* promoter was obtained from the Harvard PlasmID repository (Cat# ScCD00095253). RNA-form of *TUP1* was generated using the QuikChange II XL Site-Directed Mutagenesis Kit (Stratagene) following the manufacture's instruction. Primers used to mutagenize *TUP1* (A459V) are listed in Supplemental Table S8. Plasmids were sequenced to confirm no other mutations had been introduced. Yeast cells were transformed with either pBY011-TUP1 A459 (DNA form), pBY011-TUP1-V459 (RNA form), or empty vector using the lithium acetate method. Yeast transformants were cultured in synthetic dropout medium without uracil (SD URA-) containing

2% glucose to mid-log phase at 25°C, 250rpm. Cells were pelleted and washed 3 times with PBS. Cells were then re-suspended in SD URA- medium containing 2% galactose and 2% sucrose, and cultured for 18 hours. Total RNA was extracted using MasterPure Yeast RNA Purification Kit (Epicentre) with the following modifications. Reverse transcription using 1μg of RNA was done using the TaqMan Reverse Transcription kit (Life Technologies) following the manufacture's instruction. Quantitative PCR was preformed using SYBR Green PCR Master mix (Life Technologies). Tup1 antibody (Abcam, Cat#24313) and anti-GAPDH (Thermo Scientific, # MA5-15738) were used for western blot analysis. Hygromycin-B (Sigma-Aldrich) or DMSO control was added to a final concentration of 100μg/ml to SD URA- plates supplemented with either 2% glucose or 2% galactose and 2% sucrose. Cells were cultured overnight in SD URA- liquid medium containing 2% glucose and were then washed and re-suspended in equal volume of PBS. 3μl of ten-fold serial dilution of the cultures were spotted on the appropriate plates. The plates were covered with aluminum foil and incubated at 25°C. Pictures were taken 3-5 days after spotting.

For cycloheximide chase assay, yeast transformants were grown in SD URA- medium containing 2% glucose to mid-log phase. Cells were then pelleted and washed three times with PBS. Expression was then induced by re-suspending in SD URA- medium containing 2% galactose and 2%sucrose for 18 hours. Cells were collected by centrifugation and washed with PBS once before being resuspended in an equal volume of SD URA- containing 2% glucose to repress new transcription of *TUP1*. Cycloheximide (Sigma-Aldrich) was added to a final concentration of 0.75ug/ml. Immediately after addition of cycloheximide, 2 X 10^6 cells were harvested as samples at baseline. Cells were harvested in this manner for each subsequent time point. Whole cell lysates were prepared using denaturing conditions as described and analyzed

by western blot (Kushnirov 2000). The intensity of each band were quantified using ImageJ and normalized to that from baseline to determine the fraction of protein remaining.

Experimental validation of RDD using droplet digital PCR. We picked a few sites of each RDD type that are suitable for primer and ddPCR assay design. DNA probes specific to the DNA and RNA alleles at RDD sites were synthesized and labeled by VIC and FAM, respectively and custom Taqman assays were designed (ABI Biosystems). PCR reaction was prepared by mixing genomic DNA or cDNA from same yeast strains, Taqman assay reagents containing VIC- and FAM- probes, and ddPCR Supermix (Bio-Rad). Emulsion PCR was carried out on a Bio-Rad thermocycler using following cycles: 95°C 10 min, (94°C 30 sec, 58~61°C 1 min) X 40 cycles, 98°C 10 min (Bio-Rad). Fluorescent signal representing each variant was quantified by QuantaLife Droplet Reader and analyzed using manufacturer's software (Bio-Rad). Primers and probes are listed in Supplemental Table S9.

Additional RDD Filtering. First, we use BLAT to ensure the RDD-containing reads are correctly mapped and that they cannot be attributed to sequences in other parts of the genome. We extracted genomic sequences 25 bp, 50 bp, and 75 bp upstream and downstream of each site, and aligned each of the 6 sequences to the reference genome using BLAT (v. 34x11) (Kent 2002) with parameters '-stepSize=5' and 'repMatch=2253'. RDD sites were removed if any of the 6 surrounding sequences aligned to another genomic location with ≤3 mismatches and with sequences that explain the RDD call (that is, if the mapped genomic sequences match the RDD allele). Second, we removed all the sites that reside in repetitive genome regions annotated by RepeatMasker (version 3.2.7). Third, we removed all the sites that are less than 5nt away from splicing junctions and the sites where RNA-seq reads were spliced at non-annotated splicing sites.

Simulation of yeast RNA-seq data. Yeast reference genome (sacCer3) was used to simulate RNA-seq data by Flux simulator (Griebel et al. 2012) using default parameters. The built-in Illumina error profile was used to simulate sequencing errors. The simulated data were then analyzed for RNA-DNA sequence differences using the same methods used for yeast samples.

Probability-based error rate estimation. The statistical analysis is adapted from Chepelev 2012 (Chepelev 2012). At a given nucleotide position, the base error probability p was computed using Phred base quality score in RNA-seq data as $Q=-10\log_{10}p$. Assuming there is no RDD, the observation of alternate allele in RNA-seq reads that differ from the DNA allele is due to base-calling error (null hypothesis H_0). Let k be the subset of RNA-seq reads supporting DNA allele S_0, and m be the subset of reads supporting alternate allele S_1 at the same position, under null hypothesis H_0, the probability of observing S_1 is $P(D|H_0)=(\prod_{m\in S1}p_m)(\prod_{k\in S0}(1-p_k)$, where p is the base error probability. In RNA-seq data, the frequency of alternate allele S_1 is $f=n(S_1)/(n(S_0)+n(S_1)$, where $n(S_0)$ and $n(S_1)$ are numbers of reads representing S_0 and S_1. The probability that the observation of S_1 is a true RDD with frequency of f is $P(D|H_1)=f^{n(S1)}(1-f)^{n(S0)}$ (alternative hypothesis H_1). According to the Wilks's theorem, the test statistic of likelihood ratio $-2\log(P(D|H_0)/P(D|H_1)$ follows a chi-square distribution with the degree of freedom =1, and we derived p-values and False Discovery Rate adjusted p-values for each RDD sites using the R statistics package. Similarly, the mapping error probability was computed using mapping scores (from GSNAP) for the reads mapped to the position.

Metagene Analysis of R-loops and RDD. Each gene annotated in serSac is divided into 3 regions, "5'UTR" as transcription start site to 500 bp upstream, "3'UTR" as end of last exon to 500bp downstream, and "gene body: as region between transcription start site and end of last exon. Each region is divided into 100 bins. For R-loop metagene plot, read count from aligned

bam files from DRIP-seq and input samples in each bin is normalized to bin size and total number of uniquely mapped reads (number of reads per nucleotide per million of uniquely aligned reads; RPM). Only genes with 2.5 fold enrichment are included. Average read counts from each bin are plotted. For RDD metagene plot, number of RDD event in each bin is similarly normalized to bin size and total number of reads. Sum of RDD event in each bin is plotted.

Supplemental Figure legends

Figure S1 Summary of analysis steps for identifying RNA-DNA differences. Numbers of RDDs that were filtered out in each step are noted.

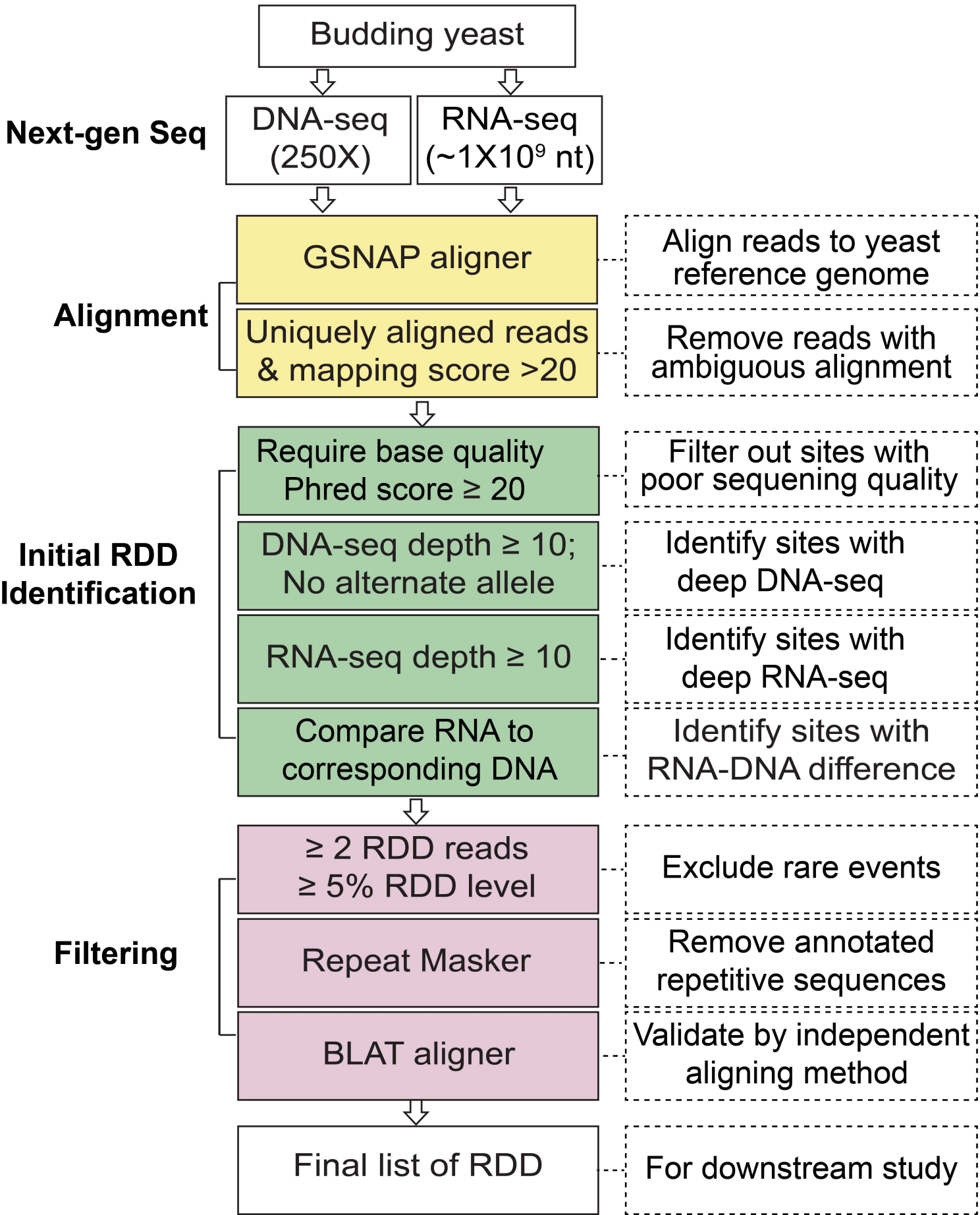
Figure S2 RDDs were identified with more stringent thresholds. (A) Distribution of RDDs by type is similar among the wild-type strains. (B) Distribution of RDDs in different genomic regions is similar among the wild-type strains. Compared to genome background, RDDs are significantly enriched in coding exons (Fisher exact test, $P<0.05$). (C) Majority of RDD sites (>98%) are covered by more than 10 RNA-seq reads. (D) 12 types of RDDs are detected using more stringent thresholds of sequencing depth and RDD level.

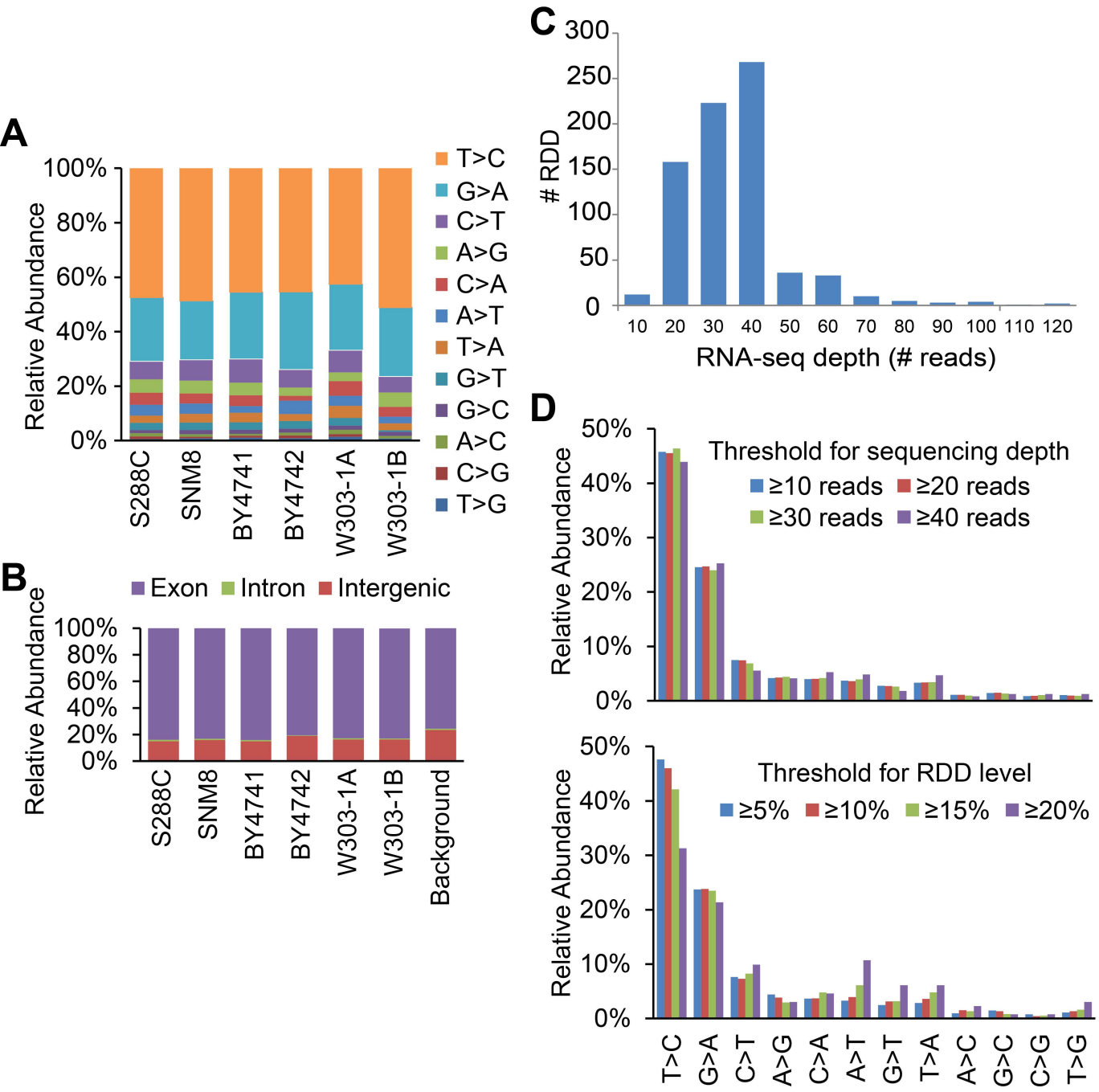
Figure S3 RDD frequencies of deaminase mutants are similar to those in wild-type strains. (A) A-to-G editing level at A34 of tRNA-Ser is reduced in *tad2*^{ts} mutant at non-permissive temperature. Editing level was measured using droplet digital PCR. Error bar: SEM of duplicate PCR. (B) RDD levels in wild type and deaminase mutants measured in RNA-seq data.

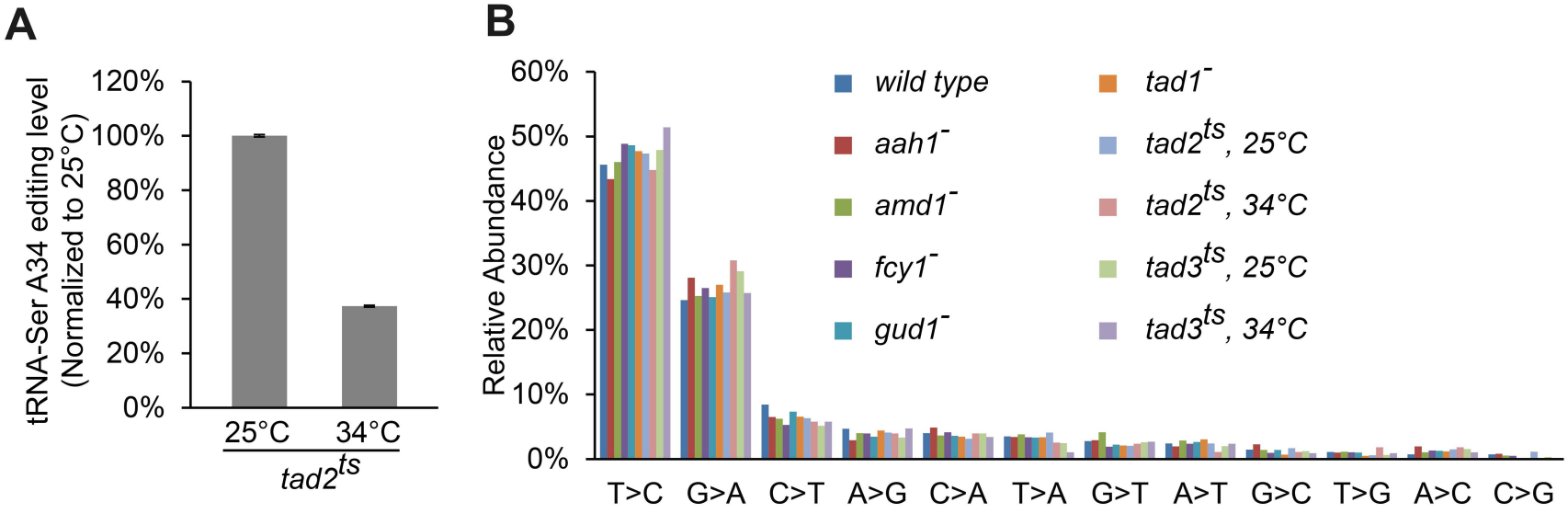
Figure S4 The A-to-G in *IMG1* at chrIII:210075 was identified by 4 different alignment algorithms. Screenshots from the Integrative Genomics Viewer are shown. *IMG1* is on negative strand and reverse complementary sequence of reads is shown.

Figure S5 Analysis of RDDs using different alignments or simulated data. (A) Three aligning methods that handle spliced reads differently were used to identify RDDs. DNA-seq data were aligned using Bowtie2-local mode, and RNA-seq data were aligned using Bowtie2-local mode, GSNAP with or without "splice-aware" mode. (B) Sequence differences between the simulated RNA-seq data and DNA sequences do not resemble characteristics of RDDs identified from yeast cells. RNA-seq data were simulated using Flux simulator and sequence differences were identified using the same algorithm for RDD identification. (C) 12 types of RDDs were identified after RDDs within 5 nt or 10 nt from both ends of sequencing reads were removed. Only RDDs in S288C are shown.

Figure S6 A C-to-T RDD was found in R-loops of *BUG1*. R-loop peak was identified at *BUG1* by DRIP-seq. Arrow indicates the RDD site. RPM = number of reads per million of uniquely mapped reads. The RNA-seq and DNA-seq data are displayed using the Integrated Genomics Viewer.

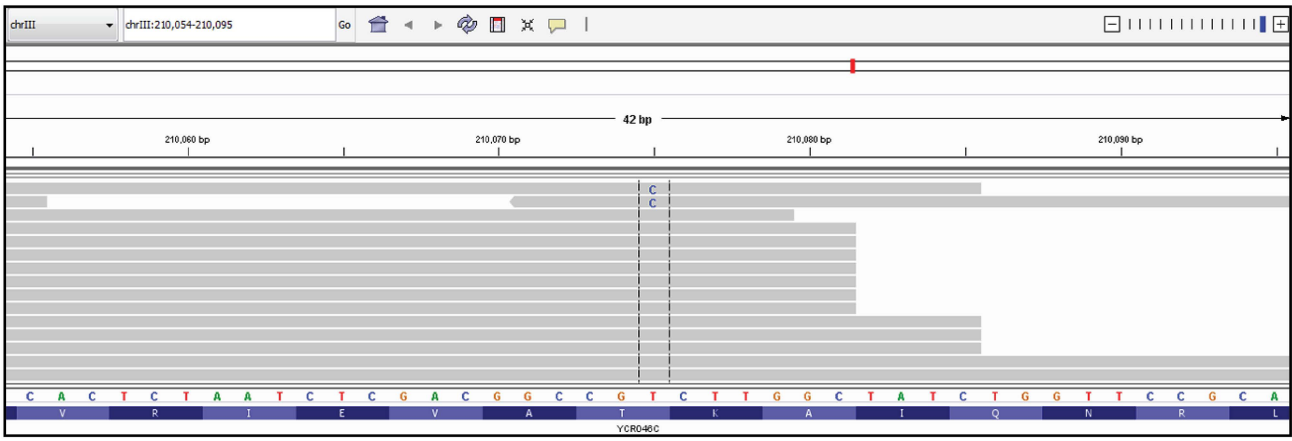

Budding yeast
Next-gen Seq
DNA-seq (250X)
RNA-seq (~1X10⁹ nt)
Alignment
GSNAP aligner
Align reads to yeast reference genome
Uniquely aligned reads & mapping score >20
Remove reads with ambiguous alignment
Initial RDD Identification
Require base quality Phred score ≥ 20
Filter out sites with poor sequening quality
DNA-seq depth ≥ 10; No alternate allele
Identify sites with deep DNA-seq
RNA-seq depth ≥ 10
Identify sites with deep RNA-seq
Compare RNA to corresponding DNA
Identify sites with RNA-DNA difference
Filtering
≥ 2 RDD reads ≥ 5% RDD level
Exclude rare events
Repeat Masker
Remove annotated repetitive sequences
BLAT aligner
Validate by independent aligning method
Final list of RDD
For downstream study


A
100%
80%
60%
40%
20%
0%
Relative Abundance
S288C
SNM8
BY4741
BY4742
W303-1A
W303-1B
T>C
G>A
C>T
A>G
C>A
A>T
T>A
G>T
G>C
A>C
C>G
T>G
B
Exon
Intron
Intergenic
100%
80%
60%
40%
20%
0%
Relative Abundance
S288C
SNM8
BY4741
BY4742
W303-1A
W303-1B
Background
C
300
250
200
150
100
50
0
RDD
10 20 30 40 50 60 70 80 90 100 110 120
RNA-seq depth (# reads)
D
50%
40%
30%
20%
10%
0%
Relative Abundance
Threshold for sequencing depth
≥10 reads
≥20 reads
≥30 reads
≥40 reads
Threshold for RDD level
≥5%
≥10%
≥15%
≥20%
T>C
G>A
C>T
A>G
C>A
A>T
G>T
T>A
A>C
G>C
C>G
T>G


A
tRNA-Ser A34 editing level
(Normalized to 25°C)
120%
100%
80%
60%
40%
20%
0%
25°C
34°C
tad2ts
B
Relative Abundance
60%
50%
40%
30%
20%
10%
0%
wild type
aah1-
amd1-
fcy1-
gud1-
tad1-
tad2ts, 25°C
tad2ts, 34°C
tad3ts, 25°C
tad3ts, 34°C
T>C
G>A
C>T
A>G
C>A
T>A
G>T
A>T
G>C
T>G
A>C
C>G

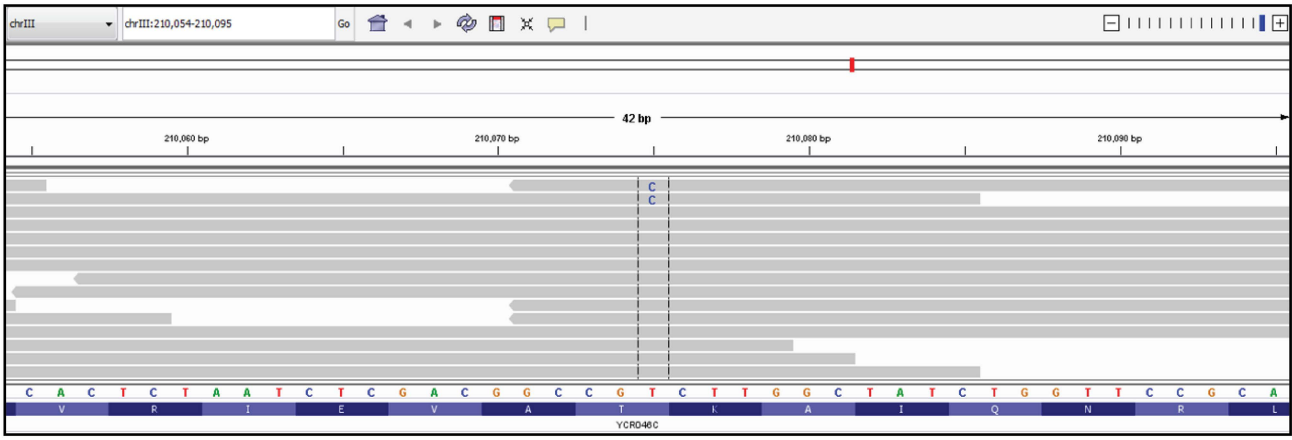
IMG1, chrIII:210075 , A>G

GSNAP (#mismatch<7)

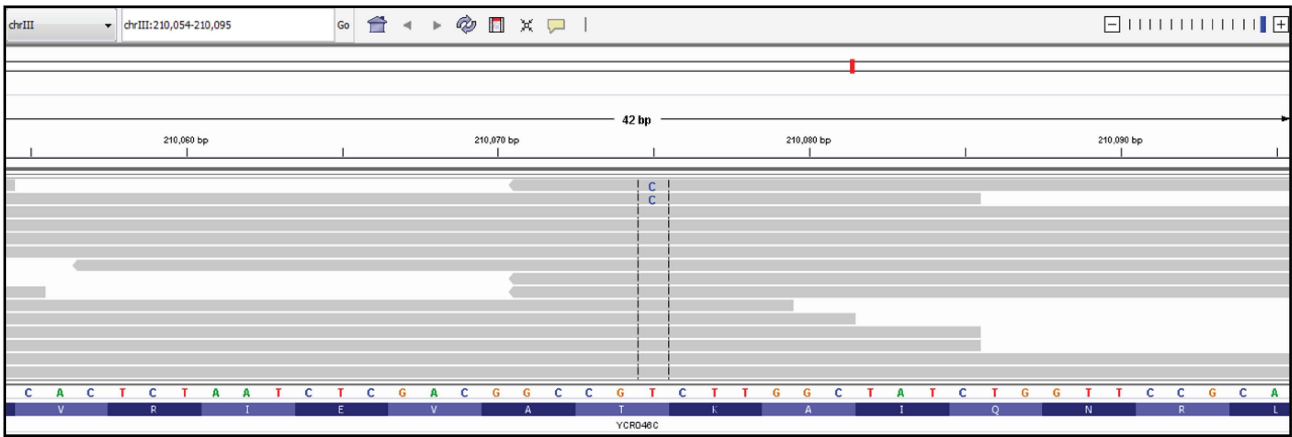


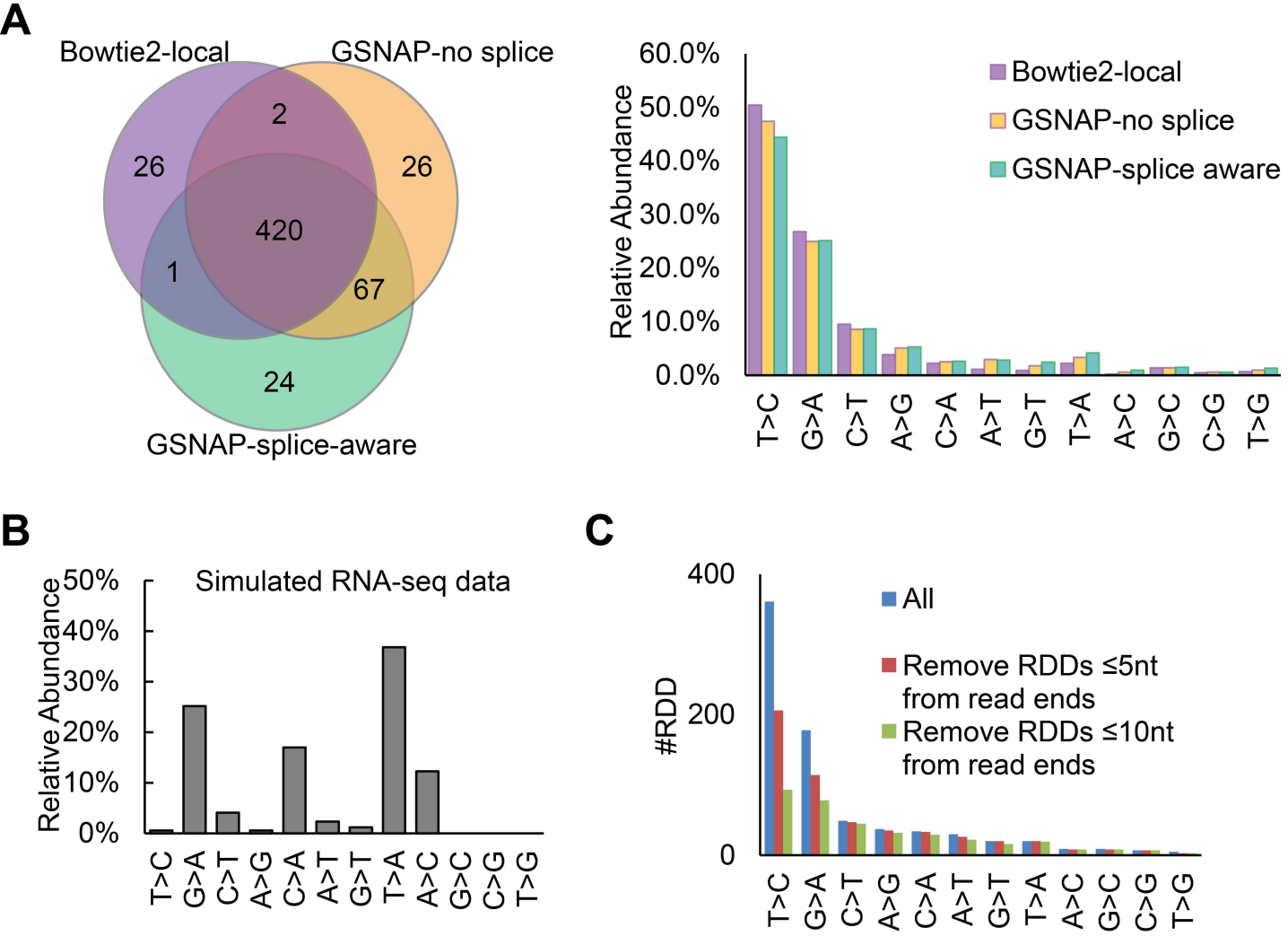
GSNAP (#mismatch≤3)

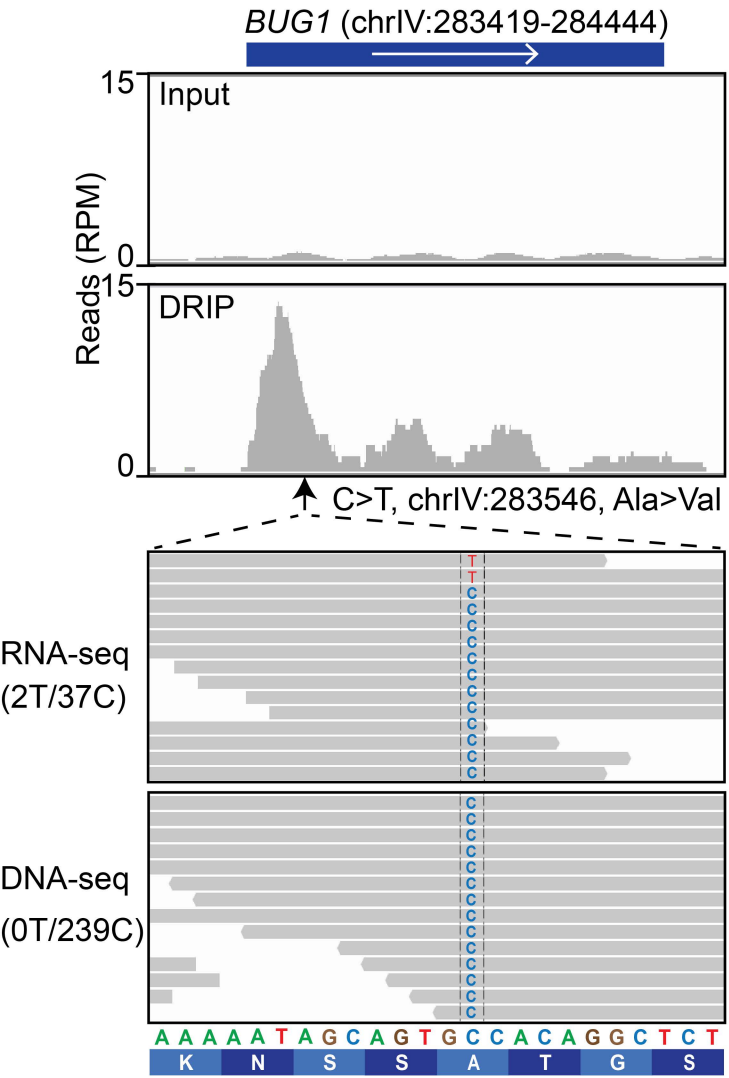
STAR



TopHat2




A
Bowtie2-local
GSNAP-no splice
GSNAP-splice-aware
26
2
26
420
1
67
24
Relative Abundance
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
Bowtie2-local
GSNAP-no splice
GSNAP-splice aware
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
Simulated RNA-seq data
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
C
#RDD
400
200
0
All
Remove RDDs ≤5nt from read ends
Remove RDDs ≤10nt from read ends
T>C
G>A
C>T
A>G
C>A
A>T
G>T
T>A
A>C
G>C
C>G
T>G


BUG1 (chrIV:283419-284444)
15
Input
0
Reads (RPM)
15
DRIP
0
C>T, chrIV:283546, Ala>Val
RNA-seq
(2T/37C)
DNA-seq
(0T/239C)
A A A A A T A G C A G T G C C A C A G G C T C T
K N S S A T G S

Supplemental Table S1. RDDs identified from six common wild-type strains.

Supplemental Table S2. DNA-seq and RNA-seq depth at RDD sites shared by multiple wild type strains.

Supplemental Table S3. *Saccharomyces cerevisiae* strains used in this study.

Strain ID	Genotype	Source
S288C	*MATα SUC2 mal mel gal2 CUP1 flo1 flo8-1 hap1*	ATCC
BY4741	*MATa his3Δ0 leu2Δ0 met15Δ0 ura3Δ0*	R Crouch
BY4742	*MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*	D Klionsky
W303-1A	*MATa ade2-1 ura3-1 his3-11 trp1-1 leu2-3 leu2-112 can1-100*	ATCC
W303-1B	*MATα leu2-3 leu2-112 trp1-1 ura3-1 his3-11 his3-15 ade2-1 can1-100*	ATCC
SNM8	*MATa CAN1 his7-2 leu2-Δ::kanMX ura3-Δ trp1-289 ade2-1 lys2-ΔGG2899-2900 agp1::URA3-OR1*	T Kunkel (Williams et al. 2013)
YTAK015	*SNM8, top1::natMX4*	T Kunkel (Williams et al. 2013)
YTAK030	*SNM8, rnh1::natMX4*	T Kunkel (Williams et al. 2013)
VCY201	*BY4741, rnh1::kanMX4*	Yeast Deletion Project(Kelly et al. 2001)
VCY203	*BY4741, top1::kanMX4*	Yeast Deletion Project(Kelly et al. 2001)
VCY224	*BY4741, tad1::kanMX4*	Yeast Deletion Project(Kelly et al. 2001)
VCY225	*BY4741, aah1::kanMX4*	Yeast Deletion Project(Kelly et al. 2001)
VCY226	*BY4741, amd1::kanMX4*	Yeast Deletion Project(Kelly et al. 2001)
VCY227	*BY4741, fcy1::kanMX4*	Yeast Deletion Project(Kelly et al. 2001)
VCY228	*BY4741, gud1::kanMX4*	Yeast Deletion Project(Kelly et al. 2001)
VCY251	*BY4741, tup1::kanMX4*	Yeast Deletion Project(Kelly et al. 2001)
VCY216	*BY4741, sen1-1*	D Klionsky (Mischo et al. 2011, 1)
VCY229	*BY4741, tad2^{ts}*	D Klionsky (Winey and Culbertson 1988)
VCY230	*BY4741, tad3^{ts}*	D Klionsky (Winey and Culbertson 1988)

Supplemental Table S4. RDDs are identified using various aligning algorithms.

	Aligner*	S288C	BY4741	BY4742	W303-1A	W303-1B	SNM8
Individual aligner	GSNAP(7MM)	741	817	645	1004	392	846
	GSNAP(3MM	529	549	424	763	309	720
	STAR	906	957	724	1334	547	1328
	TopHat2	3120	3340	2529	4355	1830	4478
Overlap between 2 aligners	GSNAP(7MM) & GSNAP(3MM)	501	519	391	710	289	650
	GSNAP(7MM) & STAR	429	442	337	592	249	554
	GSNAP(7MM) & TopHat2	366	368	276	464	210	438
	GSNAP(3MM) & STAR	440	456	350	612	258	590
	GSNAP(3MM) & TopHat2	378	380	292	484	217	473
	STAR & TopHat2	623	652	506	869	403	858
Overlap between 3 aligners	GSNAP(7MM) & GSNAP(3MM) & STAR	425	439	335	587	249	550
	GSNAP(7MM) & GSNAP(3MM) & TopHat2	363	365	276	459	209	437
	GSNAP(3MM) & STAR & TopHat2	364	371	281	470	208	457
	GSNAP(7MM) & STAR & TopHat2	352	358	268	452	201	425
Overlap between 3 aligners	GSNAP(7MM) & GSNAP(3MM) & STAR & TopHat2	351	358	268	451	201	425

* GSNAP(7MM): GSNAP with default parameter for mismatches (< 7 mismatches for 100-nt reads); GSNAP(3MM): GSNAP with ≤3 mismatches for each read.

Supplemental Table S5. Sequencing errors detected in PhiX control spiked in RNA-seq samples.

Sequencing Sample	#sites ≥ 10 reads	# sites with errors	# sites with ≥ 2 reads containing same errors
Sample 1	5051	32	0
Sample 2	5013	29	0
Sample 3	5098	51	0
Sample 4	3724	20	0
Sample 5	5019	30	0
Sample 6	5139	47	0

Supplemental Table S6. Probability-based error rate estimation

P-value (Base Error)	S288C	SNM8	BY4741	BY4742	W303-1A	W303-1B
<0.01	739	845	814	641	1000	392
0.01-0.05	2	1	1	3	3	0
0.05-0.1	0	0	1	1	1	0
0.1-1	0	0	1	0	0	0

P-value (Mapping Error)	S288C	SNM8	BY4741	BY4742	W303-1A	W303-1B
<0.01	741	846	816	645	1004	392
0.01-0.05	0	0	1	0	0	0
0.05-0.1	0	0	0	0	0	0
0.1-1	0	0	0	0	0	0

Supplemental Table S7. Phred score threshold has minimal effect on RDD identification.

Phred Score Cutoff	S288C	BY4741	BY4742	W303-1A	W303-1B	SNM8
Phred≥20 (initial)	759	829	666	1023	394	867
Phred≥25	744	812	648	999	392	852
Phred≥30	744	806	643	988	373	854
Phred≥35	790	827	667	1082	381	891

Supplemental Table S8. Primers and probes used in this study.

Gene	Experiments	Forward Primer	Reverse Primer
TUP1	Cloning	5'-CCAGATGGGAAATTTTTGGTAACAGGTGCTGAAGACAG-3'	5'-CTGTCTTCAGCACCTGTTACCAAAAATTTCCCATCTGG-3'.
TUP1	Real-time PCR	5'-CATCGGCCTTCCCAGTACAA-3'	5'-ACAGGCAAAGTGGTGGTAGG-3'
RPL15A	Real-time PCR	5'-TGTGACCCAGTTCACAAGCAC-3'	5'-GTATCTCCACAAGGACAAAGTG-3'

Supplemental Table S9. Sequence of probes and primers in droplet digital PCR.

Genomic Location	Gene Name	RDD type	Forward Primer	Reverse Primer	VIC Probe Sequence	FAM Probe Sequence
chrII:396658	*RPG1*	A>G	CCACAAGAAACT GAAGACGGTGAA	TGTTGTGGATGTA AGAATTGCGGAT A	CAGATTCTTCTTC CTTTTC	AGATTCTTCCTCC TTTTC
chrIV:454727	*RCR2*	G>A	CTCCGACTGTTGA ATCTTCTTCCTT	CGAAAAATCACTT ACTCACTTTCGCT	ATAACGCGCCGG CAAG	AATAACGCACCG GCAAG
chrIV:1086040	*TFB1*	G>C	GGTGACGTAATC ATTGACAGGTACT	GTCGTCCTGTATG TTACCATCTAAAT CTATAATTTT	CGACAGAAAAGA TGATGACAT	CGACAGAAAAGA TCATGACAT
chrIV:895335	*ADR1*	C>G	CAAGTTGCCCGA AAATTTAAGGCTT A	TCTCGTACAAACC TCGCAAACAA	ATGACCTTAGTTT CCC	ATGACCTTACTTT CCC
chrVII:146359	*CDC55*	A>T	ACCATAAACATA CATGAGCAATTG AAGGA	CACTTGAACTATC ACCACTAAAATTA ACTTCAAA	TGAGTGATACCTA TGAAAAC	AGTGATACCTTTG AAAAC
chrIX:220698	*RPN2*	G>C	GACCACCCAGAA TAGATTGAGTTCA A	GCTAAAAGAGGA GCAGCAGTCG	CCGGAAAAGATG TCTTTGA	CCGGAAAAGATC TCTTTGA
chrXI:103490	*FAS1*	C>T	AAGAGATTGGTG GAATTAATGTTCA TCAGA	GACGTAGGAAAT CACCAGTAAAGG T	CATGTGACGTCAA ACC	CATGTGACATCAA ACC
chrXVI:450033	*PDR12*	C>T	GTCTCGTCAATTG GAACAGGGAAT	GCAGCAGAGGCA TCAACAC	ATTCTGGTATTGC CTTTAAA	ATTCTGGTATTGC TTTTAAA
chrXVI:829933	*NCE102*	G>A	CTTCTCAAAGCAT ACCTAATAACAAT ATAATCCCA	GATTAAACCGAT GGAAATAACCAA AAATAGGAA	CTAGCCCTAGCTG ATAAC	TAGCCCTAGCTAA TAAC
chrXVI:23121	*SAM3*	G>A	GCATGTCATTCCA GAAGACCTTGAA	GCTGAGAAGCTA GTTTCCATTGGAT	AGACAGAGCAGG AAAA	ACAGAGCAGAAA AA
chrX:524060	*tRNA-Ser*	A>G	ACTTGGCCGAGTG GTTAAGG	TGCGCGGGCAAA GC	CGAAAGATTAGA AATC	CGAAAGATTGGA AATC
chrII:540135	*ARA1*	C>A	GACAGCAAATCC TCACGAAAAGTT A	GCCTGTATCCAGC TTTGATTGC	CGGCTTGTTTTGT TTCAG	CGGCTTGTTTTTT TTCAG
chrIV:33734	*YPD1*	T>C	CGACAGCTGGAC GGTGAAAA	CCTAATGCAGCA GAAGAACCCTTTA	AAATGGCCCAGA TTGT	AATGGCCCGGATT GT
chrVII:351377	*GUP1*	T>C	GCCGGCCCCATTA TAACATTCA	CAATAACGAATCT CACCGCATAGTA AAA	CAATCGAAACAT ACCTTGCCAT	ATCGAAACATAC CCTGCCAT
chrVII:447735	*TRP5*	T>A	ACGACGTTGCTAA GGAATATGTACA G	GCGGACTTCAAG GACTCTTTTTGA	TGGACAGAACCTT ATGCTTA	TGGACAGAACCTT TTGCTTA
chrVII:885558	*PDX1*	G>A	GTCTTGGAAATAT AAAGTTGGCGAA CC	GCTTCCACATCAA TTTGAGATTTATC TGTT	CAGCGCGGGCGA T	ATTCAGCACGGG CGAT
chrX:701486	*MGM101*	C>G	AGGCGAACAAGA CTATTTCAACGAA	CCGAGATCTTTGC AACACCTCATTA	CTGTAGCAGTTGG TATGC	CTGTAGCAGTTCG TATGC
chrXII:724769	*YHC1*	G>A	CCAAGAAGCGGA TGCATTCC	ATTTCCATACCGC CTCCTTTTATATC C	ACGGCATACGGA AAC	CGGCATACGAAA AC
chrXIV:134490	*BNI1*	G>A	CTCTTTCATCGGT AGGTACGTCAAC	GAGATGGCTGCTT TTTCAAACTCC	CTCACTAATTTTT TCCCCTTCG	CTCACTAATTTTT TTCCCTTCG
chrXV:545227	*RPT5*	G>A	GTCCCATGAAAA CAACGTTATGCT	CCACAAGGTACG GTAACTGTCTATT	AGATTAAGGACA ATAAGGAAAA	AAGGACAATAAG AAAAA
chrXVI:823056	*ASN1*	T>C	GAGACCCAATCG GTATTACGACAT	GTTCGGATGCAA AATAAACGGTCTT	AAGAGCGTCCCA TATATA	AAGAGCGTCCCG TATATA

Supplemental References:

Albuquerque CP, Smolka MB, Payne SH, Bafna V, Eng J, Zhou H. 2008. A multidimensional chromatography technology for in-depth phosphoproteome analysis. *Mol Cell Proteomics MCP* **7**: 1389–1396.

Bernardi G. 1979. The petite mutation in yeast. *Trends Biochem Sci* **4**: 197–201.

Chepelev I. 2012. Detection of RNA editing events in human cells using high-throughput sequencing. *Methods Mol Biol Clifton NJ* **815**: 91–102.

Djebali S, Lagarde J, Kapranov P, Lacroix V, Borel C, Mudge JM, Howald C, Foissac S, Ucla C, Chrast J, et al. 2012. Evidence for Transcript Networks Composed of Chimeric RNAs in Human Cells. *PLoS ONE* **7**: e28213.

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.

Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigó R, Sammeth M. 2012. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res* **40**: 10073–10083.

Kelly DE, Lamb DC, Kelly SL. 2001. Genome-wide generation of yeast gene deletion strains. *Comp Funct Genomics* **2**: 236–242.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656–664.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36.

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.

Kushnirov VV. 2000. Rapid and reliable protein extraction from yeast. *Yeast* **16**: 857–860.

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.

Soulard A, Cremonesi A, Moes S, Schütz F, Jenö P, Hall MN. 2010. The rapamycin-sensitive phosphoproteome reveals that TOR controls protein kinase A toward some but not all substrates. *Mol Biol Cell* **21**: 3475–3486.

Swaney DL, Beltrao P, Starita L, Guo A, Rush J, Fields S, Krogan NJ, Villén J. 2013. Global analysis of phosphorylation and ubiquitylation cross-talk in protein degradation. *Nat Methods* **10**: 676–682.

Weinert BT, Schölz C, Wagner SA, Iesmantavicius V, Su D, Daniel JA, Choudhary C. 2013. Lysine succinylation is a frequently occurring modification in prokaryotes and eukaryotes and extensively overlaps with acetylation. *Cell Rep* **4**: 842–851.

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.

Winey M, Culbertson MR. 1988. Mutations affecting the tRNA-splicing endonuclease activity of Saccharomyces cerevisiae. *Genetics* **118**: 609–617.