CHARACTERIZATION OF BIOLOGICAL SIGNATURES OF RIBONUCLEOTIDES INCORPORATED INTO DNA USING THE RIBOSE-MAP BIOINFORMATICS TOOLKIT

A Dissertation Presented to
The Academic Faculty

By

Alli L. Gombolay

In Partial Fulfillment
of the Requirements for the Degree Doctor of Philosophy in Bioinformatics
School of Biological Sciences

Georgia Institute of Technology

August 2022

CHARACTERIZATION OF BIOLOGICAL SIGNATURES OF RIBONUCLEOTIDES INCORPORATED INTO DNA USING THE RIBOSE-MAP BIOINFORMATICS TOOLKIT

Approved by:

Dr. Francesca Storici, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. I. King Jordan
School of Biological Sciences
Georgia Institute of Technology

Dr. Mark Borodovsky
School of Biomedical Engineering and
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Fredrik Vannberg
Department of Biology
Georgia State University

Dr. Soojin Yi
Department of Ecology, Evolution, and Marine Biology
University of California, Santa Barbara

Date Approved: June 23, 2022

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Francesca Storici for allowing me the opportunity to be a part of her lab and her support throughout graduate school. I would also like to thank my committee members, Dr. King Jordan, Dr. Mark Borodovsky, Dr. Fred Vannberg, and Dr. Soojin Yi, for their advice and support. In addition, I would like to thank my family for their love and encouragement throughout my education. Most importantly, I would like to thank my Lord and Savior, Jesus Christ. “But those who hope in the Lord will renew their strength. They will soar on wings like eagles; they will run and not grow weary, they will walk and not be faint.” Isaiah 40:31

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

Alk-HydEn-seq	Alkali Hydrolytic End Sequencing
AtRNL	*Arabidopsis thaliana* tRNA Ligase
BAM	binary alignment map
DNA	Deoxyribonucleic Acid
DSB	Double-Strand Break
emRiboSeq	Embedded Ribonucleotide Sequencing
NCBI	National Center for Biotechnology Information
OH	Hydroxyl
Pu-seq	Polymerase Usage Sequencing
RED	Ribonucleotide Excision Defective
RER	Ribonucleotide-Excision Repair
RHII-HydEn-seq	RHII Hydrolytic End Sequencing
ribose-seq	Ribose Sequencing
RNA	Ribonucleic Acid
RNase H	Ribonuclease H
rNMPs	Ribonucleoside Monophosphates
SRA	Sequence Read Archive
SSB	Single-Strand Break
UMI	Unique Molecular Identifier

SUMMARY

The incorporation of ribonucleoside monophosphates (rNMPs) into DNA is one of the most frequently occurring errors during DNA synthesis. To maintain genome integrity, the ribonuclease (RNase) H enzymes efficiently remove rNMPs that are mistakenly incorporated into DNA during DNA replication or repair. However, if these enzymes fail to remove rNMPs from DNA, the 2'-hydroxyl (OH) group of the ribose sugar of rNMPs can attack the double-helix backbone of DNA, resulting in several types of genome instability, including SSBs, DSBs, short deletion mutations, replication stress, cell cycle checkpoint activation, aberrant recombination, formation of protein-DNA crosslinks, and alterations in the structural properties of DNA. Recently, five high-throughput rNMP sequencing (rNMP-seq) techniques have been developed (ribose-seq, emRiboSeq, Alk-HydEn-seq, RHII-HydEn-seq, and Pu-seq) to map the locations of rNMPs in DNA to single-nucleotide resolution. Since the development of rNMP-seq is recent, the biological signatures of rNMPs have yet to be thoroughly characterized. In addition, a standardized bioinformatics toolkit to characterize the biological signatures of rNMPs is needed. To address this, I created the Ribose-Map bioinformatics toolkit. In addition, I applied Ribose-Map to characterize the biological signatures of rNMPs in the DNA of different species, strains, and RNase H genotypes (wild type and mutant) of the yeast, *S. cerevisiae*, *S. pombe*, and *S. paradoxus*. This work serves as a foundational resource for the emerging field of rNMP mapping, leading to an improved understanding of the role of rNMPs in genome stability.

CHAPTER 1. INTRODUCTION

1.1 Incorporation of ribonucleotides into genomic DNA

Genetic information is stored in DNA rather than RNA partly due its greater stability (1). In contrast to deoxyribonucleotide monophosphates (dNMPs), the subunits of DNA, ribonucleoside monophosphates (rNMPs), the subunits of RNA, contain a reactive 2'-hydroxyl (OH) group (1) (**Figure 1**). During DNA synthesis, DNA polymerases (Pol) select nucleotides into the growing DNA strand that contain both the correct sugar and base, preferring dNMPs to rNMPs. However, selection errors can occur, resulting in the incorporation of rNMPs into DNA. In fact, the incorporation of rNMPs into DNA is one of the most frequently occurring errors during DNA synthesis (2). In yeast, Pols α, δ and ε have been shown to incorporate more than 13,000 rNMPs into DNA during each round of replication (3). In humans, Pols ε and δ have been shown to incorporate up to 3 million rNMPs during each round of replication (4,5). The incorporation of rNMPs into DNA is present across many different species, including *Escherichia coli* (*E. coli*), *Schizosaccharomyces pombe* (*S. pombe*), *Saccharomyces cerevisiae* (*S. cerevisiae*), *Chlamydomonas reinhardtii* (*C. reinhardtii*), mice, and humans (1,6,7).

To maintain genome stability, three pathways are responsible for removing rNMPs from DNA with varying degrees of efficiency- 1) ribonuclease (RNase H) enzymes (RNase H1 and RNase H2), 2) exonucleolytic processing by Pol δ and Pol ε, and 3) topoisomerase I (Top1) (1). The RNase H enzymes efficiently remove rNMPs from DNA. RNase H1 is a monomeric enzyme that cleaves stretches of ≥ 4 rNMPs in DNA, while RNase H2 is a heterotrimeric enzyme (RNH201, RNH202, and RNH203 subunits in yeast) cleaves both single and stretches of rNMPs in DNA (**Figure 2**) (8). The RNase H2 enzyme initiates ribonucleotide excision repair (RER) (1). During RER, RNase H2 enzyme recognizes rNMPs incorporated into DNA and cleaves at the 5'-end of

rNMPs (1). Then, the nicked strand is displaced, allowing DNA Pol δ or Pol ε in complex with proliferating cell nuclear antigen (PCNA) to bind and fill the gap (1). Although not essential for viability in yeast, RNase H2 is essential in mice and loss of this enzyme results in embryonic lethality (9). In addition, mutations in the genes encoding the subunits of RNase H2 in humans are associated with the neurodegenerative disorder, Aicardi-Goutières Syndrome (AGS) (10), and the autoimmune disorder, systemic lupus erythematosus (SLE) (11). The 3'-5' exonuclease activity of Pol δ and Pol ε can remove rNMPs from DNA but inefficiently (4,5,12,13). In the absence of RER, Top1 can remove rNMPs from DNA but aberrantly (14,15). If these pathways fail to remove rNMPs from DNA, the 2'-OH group of unrepaired rNMPs can attack the double-helix backbone of DNA, resulting in several types of genome instability, including SSBs, DSBs, short deletion mutations, replication stress, cell cycle checkpoint activation, aberrant recombination, formation of protein-DNA crosslinks, and alterations in the structural properties of DNA (16,17).

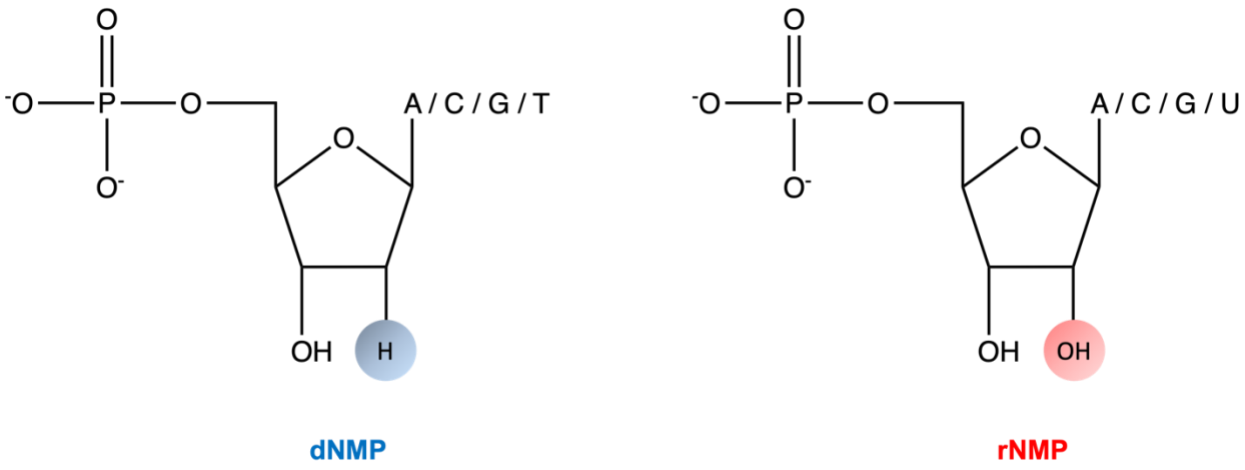


Figure 1. Chemical structure of dNMP vs. rNMP. dNMP consists of a deoxyribose sugar, a phosphate group, and a nitrogenous base (adenine monophosphate (dAMP), cytosine monophosphate (dCMP), guanine monophosphate (dGMP), or thymine monophosphate (dTMP)). rNMP consists of a ribose sugar, a phosphate group, and a nitrogenous base (adenine

monophosphate (rAMP), cytosine monophosphate (rCMP), guanine monophosphate (rGMP), or uracil monophosphate (rUMP)) In contrast to dNMPs, rNMPs contain a reactive 2'- OH group.

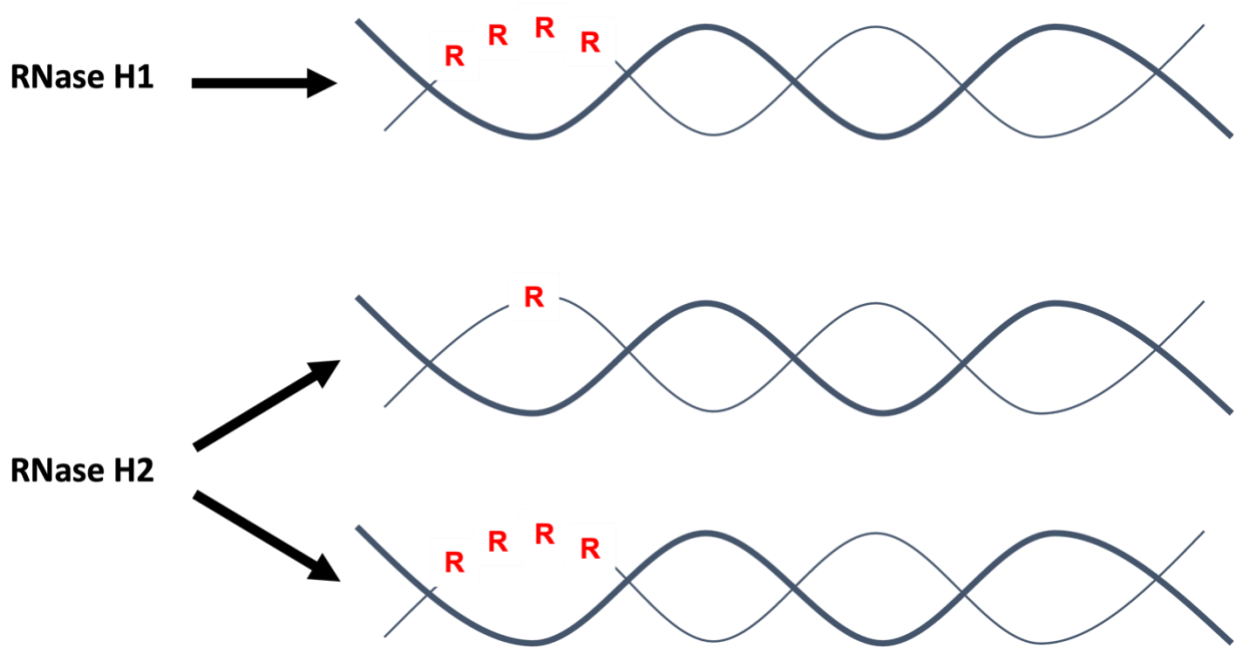


Figure 2. Cleavage specificity of RNase H1 and RNase H2. rNMPs are in red as 'R'. Double-helix DNA is in blue. RNase H1 cleaves only stretches of ≥ 4 rNMPs in DNA, while RNase H2 cleaves both single rNMPs in DNA and stretches of rNMPs (Modified from (8)).

1.2 High-throughput rNMP-seq techniques

To better understand the impact of rNMPs on genome stability, it is critical to map the locations of rNMPs in DNA to single-nucleotide resolution. In 2015, four high-throughput rNMP-seq techniques were developed to map the locations of rNMPs. These techniques include ribose sequencing (ribose-seq) developed by the Storici Lab at Georgia Tech (18), embedded ribonucleotide sequencing (emRiboSeq) developed by the Jackson and Taylor Labs at the University of Edinburgh (19), hydrolytic end sequencing (HydEn-seq) by the Kunkel Lab at the National Institute of Environmental Health Sciences (20), and polymerase usage (Pu-seq) by the

Carr Lab at the University of Sussex (21). Since the initial development of these four techniques, an improved version of HydEn-seq has been developed, RHII-HydEn-seq, and HydEn-seq has been renamed to alkali hydrolytic end sequencing (Alk-HydEn-seq) to distinguish the two techniques. The main challenge in mapping rNMPs is tagging the rNMPs relative to the 5' end of the sequencing read (tagged nucleotide), and each technique uses a unique approach. emRiboSeq and RHII-HydEn-seq use RNase H2 to generate nicks at the 5'-end of rNMPs, while Alk-HydEn-seq, Pu-seq, and ribose-seq use alkali to cleave at the 3'-end of rNMPs. In contrast to the other techniques, ribose-seq also takes advantage of *Arabidopsis thaliana* tRNA ligase (AtRNL) to directly capture sites of rNMPs, excluding Okazaki fragments and DNA abasic sites and thus minimizing background noise in the data. **Figure 3** describes the scheme of ribose-seq.

Once the data have been sequenced and aligned to the reference genome, emRiboSeq rNMPs are located one nucleotide downstream from the reverse complements of the tagged nucleotide. RHII-HydEn-seq rNMPs are located at the same position as the tagged nucleotide. Alk-HydEn-seq and Pu-seq rNMPs are located one nucleotide upstream from the tagged nucleotide. ribose-seq rNMPs are the reverse complements of the tagged nucleotide. **Figure 4** compares the positions of rNMPs in genomic DNA relative to the tagged nucleotide for each rNMP-seq technique.

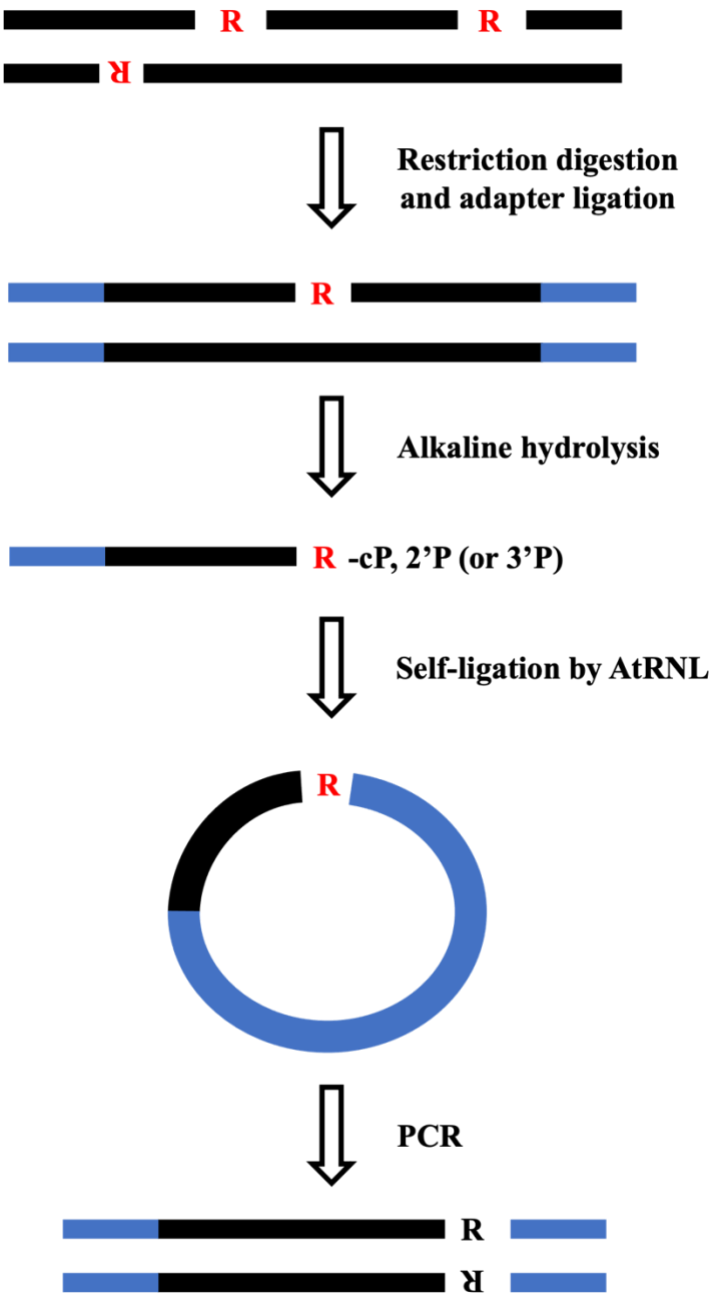


Figure 3. Scheme of ribose-seq. Genomic DNA is fragmented with restriction enzymes and ligated to a sequencing adaptor. Alkali denatures the DNA and cleaves at rNMPs, exposing 2',3'-cyclic phosphate (P) and 2'-P termini, which are self-ligated to 5'-phosphate ends by AtRNL. The rNMP-containing circular DNA molecules are PCR amplified and sequenced using Illumina. R in red represents the rNMPs and R in black represents the rNMP converted to a dNMP during PCR.

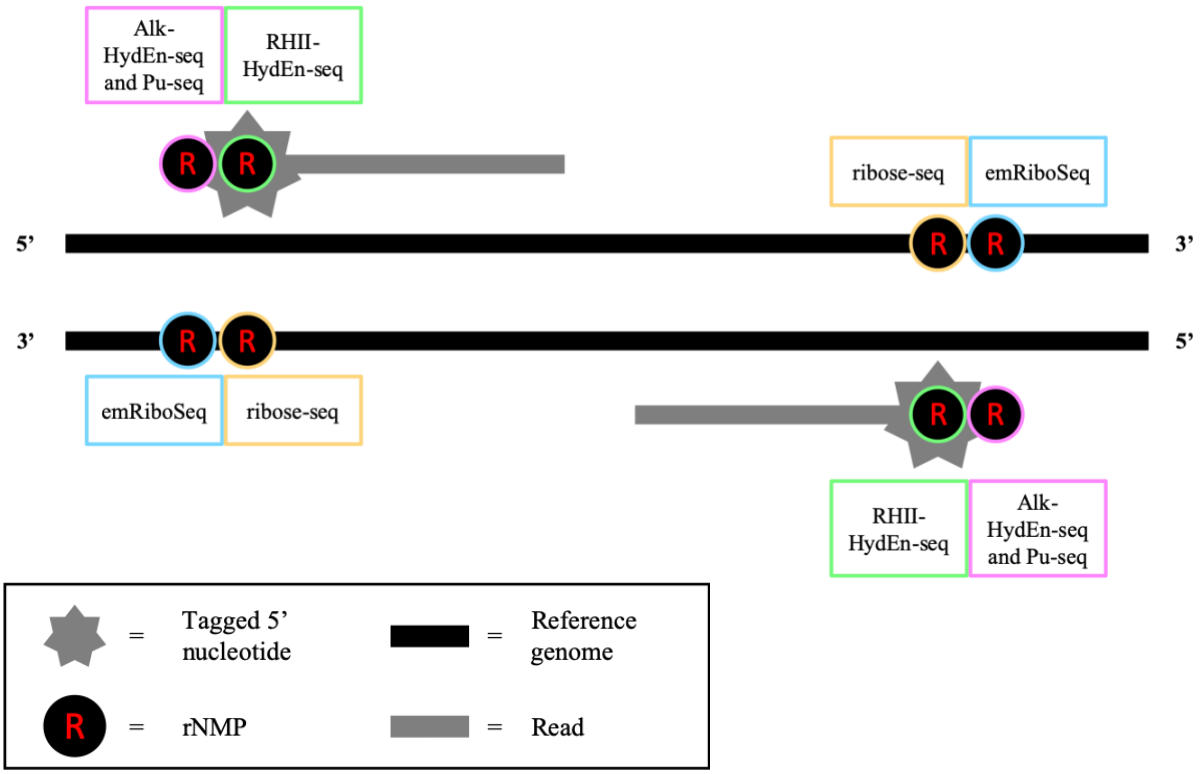


Figure 4. Positions of rNMPs in genomic DNA relative to 5' side of sequencing reads. emRiboSeq rNMPs are located one nucleotide downstream from the reverse complements of the 5' nucleotide, Alk-HydEn-seq and Pu-seq rNMPs are located one nucleotide upstream from the 5' nucleotide, ribose-seq rNMPs are the reverse complements of the 5' nucleotide, and RHII-HydEn-seq rNMPs are located at the same position as the 5' nucleotide.

1.3 Limitations of rNMP-seq

rNMP-seq can only capture one rNMP in a stretch of ≥ 2 rNMPs. For example, following treatment with alkali during ribose-seq, stretches of ≥ 2 rNMPs are reduced to single rNMPs and are subsequently removed by purification. If current rNMP-seq techniques were modified to capture each rNMP in a stretch of rNMPs or a new technique is developed, Ribose-Map's Coordinate Module could be readily updated to accommodate any change in the position of the rNMP relative to the tagged nucleotide. In addition, rNMP-seq techniques currently utilize short read sequencing technology. However, short read sequencing technology is unable to span

repetitive sequences of DNA, leading to errors when aligning reads originating from these regions. Since long reads often span repetitive sequences of DNA, rNMP-seq techniques could be updated to utilize long read sequencing technology (e.g., Oxford Nanopore). If current rNMP-seq techniques were modified to utilize long read sequencing technology or a new technique is developed, an option to align reads with a long read aligner (e.g., BLASR (22), GraphMap (23), or Kart (24)) could be added to Ribose-Map's Alignment Module. Once this option is added, the user could then specify short or long reads in the configuration file.

1.4 Alternative sequencing approaches

In addition to ribose-seq, emRiboSeq, RNHII-HydEn-seq, Alk-HydEn-seq, and Pu-seq, two additional sequencing techniques have been developed that are relevant to the field of rNMP mapping- Rare Damage and Repair sequencing (RADAR-seq) (25) and Ribonucleotide Scanning Quantification sequencing (RiSQ-seq) (26). RADAR-seq detects rare DNA damage events, such as rNMPs, but uses the third-generation sequencing technology, PacBio SMRT sequencing. RiSQ-seq quantifies rNMPs in DNA using background normalization and standard adjustment.

1.5 rNMP Mapping Software

To date, no standardized bioinformatics toolkit for the comprehensive analysis of rNMP-seq data exists. Current rNMP mapping software, Modmap (18), emRiboSeq Processor (19), and Puseq_app (21), are highly customized, limited in terms of scope of analysis, depend on proprietary software, and/or provide limited documentation. Modmap (https://github.com/hesselberthlab/modmap) is customized to analyze only ribose-seq data, does not provide online documentation, depends on proprietary software, and does not output a file containing the genomic

coordinates of rNMPs. emRiboSeq Processor (https://github.com/taylorLab/LaggingStrand) is customized to analyze only emRiboSeq data, does not provide online documentation, does not output a file containing the genomic coordinates of rNMPs (only counts of rNMPs), and does not characterize the biological signatures of rNMPs. The Puseq_app (https://github.com/AndreaKeszthelyi/Puseq_app) is customized to analyze only Pu-seq data, does not provide online documentation, does not output a file containing the genomic coordinates of rNMPs, and tracks the division of labor of DNA polymerases rather than characterizes the biological signatures of rNMPs. Since rNMP-seq techniques can generate large, complex datasets containing millions of rNMPs, software that can map the genomic coordinates of rNMPs to single-nucleotide resolution and characterize the biological signatures of rNMPs for data produced using any rNMP-seq technique is urgently needed. The bioinformatics workflow for rNMP-seq data can be divided into four main tasks: (i) trimming reads based on quality and adaptors, barcodes, and/or unique molecular identifiers (UMI's) (ii) aligning sequencing reads to the reference genome, (iii) mapping the genomic coordinates of rNMPs to single-nucleotide resolution, and (iv) characterizing the biological signatures of rNMPs (e.g., nucleotide composition of rNMPs, nucleotide sequence context of rNMPs, genome-wide distribution of rNMPs, and hotspot motifs of rNMPs).

1.6 Characterization of rNMPs in DNA

Since the development of rNMP-seq is recent, the biological signatures of rNMPs in DNA have yet to be thoroughly characterized in different strains, genotypes, and species. To date, one study characterized the biological signatures of rNMPs in the DNA of yeast cells but was limited to only one species- *S. cerevisiae*. By applying ribose-seq, Koh *et al.* showed widespread rNMP incorporation in RNase H2 mutant *S. cerevisiae* cells, with strong preference for rC and rG in both

the nucleus and mitochondria of these cells. In the mitochondria, rC and rG were present in (G+C)-rich regions. The remaining studies applied rNMP-seq to track the division of labor of DNA polymerases in yeast (19-21,27) rather than characterize biological signatures of rNMPs.

1.7 Research Goals

This project addresses a fundamental question in the field of molecular biology- what are the biological signatures of rNMP incorporation in DNA (e.g., nucleotide composition of rNMPs, nucleotide sequence context of rNMPs, genome-wide distribution of rNMPs, and hotspot motifs of rNMPs)? To address this question, we created and validated Ribose-Map, a standardized bioinformatics toolkit for the comprehensive analysis of rNMP-seq data, and applied this toolkit to characterize the biological signatures of rNMPs in wild type and mutant RNase H strains of the budding yeast, *S. cerevisiae* and *S. paradoxus*, and the fission yeast, *S. pombe*.

1.7.1 Create and validate standardized bioinformatics toolkit for rNMP-seq data

a) Create Ribose-Map, a bioinformatics toolkit to characterize the biological signatures of rNMPs in DNA captured using any of the five currently available rNMP-seq techniques

b) Validate Ribose-Map against current rNMP mapping software

Current rNMP mapping software are highly customized, limited in terms of scope of analysis, depend on proprietary software, and/or provide limited documentation. To address this, I created the Ribose-Map bioinformatics toolkit, the first known standardized bioinformatics

toolkit for the comprehensive analysis of rNMP-seq data. Through a series of modules, Ribose-Map calculates the genomic coordinates of rNMPs to single-nucleotide resolution and characterizes the biological signatures of rNMPs for data generated using any rNMP-seq technique. Then, I validated Ribose-Map against current rNMP mapping software.

1.7.2 Characterize the biological signatures of rNMPs in the DNA of wild type and mutant RNase H strains of *S. cerevisiae, S. paradoxus*, and *S. pombe* cells

a) Apply Ribose-Map to wild type strains of *S. cerevisiae*, *S. paradoxus*, and *S. pombe* cells

b) Apply Ribose-Map to mutant RNase H strains of *S. cerevisiae*, *S. paradoxus*, and *S. pombe* cells

Since the development of rNMP-seq is recent, the biological signatures of rNMPs in DNA have yet to be thoroughly characterized in different strains, genotypes, and species. To address this, I applied the Ribose-Map bioinformatics toolkit to ribose-seq libraries of wild type, mutant RNase H1, and mutant RNase H2 strains of *S. cerevisiae*, *S. paradoxus*, and *S. pombe* cells and characterized the biological signatures of rNMPs in the DNA of these cells.

CHAPTER 2. Ribose-Map Bioinformatics Toolkit

The work presented in this chapter consists of the research projects published in 1) Gombolay, A.L., Vannberg, F.O., and Storici, F. (2019) Ribose-Map: a bioinformatics toolkit to map ribonucleotides embedded in genomic DNA. *Nucleic Acids Research*, 47: e5; 2) Gombolay, A.L. and Storici, F. (2021) Mapping ribonucleotides embedded in genomic DNA to single-nucleotide resolution using Ribose-Map. *Nature Protocols*, 16: 3625-3638; and 3) Gombolay, A.L. and Storici, F. (2021) Ribose-Map: A bioinformatics toolkit for ribonucleotide sequencing experiments. *Software Impacts*, 10: 100149.

2.1 Abstract

Recently, five high-throughput rNMP-seq techniques, 1) emRiboSeq, 2) RHII-HydEn-seq, 3) Alk-HydEn-seq, 4) Pu-seq, and 5) ribose-seq, have been developed to map sites of rNMPs in DNA. These techniques can capture potentially millions of rNMPs, generating large, complex datasets that require software that can map the genomic coordinates of rNMPs to single-nucleotide resolution and characterize the biological signatures of rNMPs for data generated using any rNMP-seq technique. However, current rNMP mapping software are highly customized, limited in terms of scope of analysis, depend on proprietary software, and/or provide limited documentation. To address this, I developed the Ribose-Map bioinformatics toolkit for the comprehensive analysis of rNMP-seq data. Through a series of modules, Ribose-Map calculates the genomic coordinates of rNMPs to single-nucleotide resolution and characterizes the biological signatures of rNMPs for data generated using any rNMP-seq technique. The Alignment Module aligns rNMP-seq data to the reference genome of interest (and de-multiplexes and/or de-duplicates data if needed), the Coordinate Module calculates the genomic coordinates of rNMPs to single-nucleotide resolution, the Composition Module calculates and plots the nucleotide composition of rNMPs, the Sequence Module calculates and plots the nucleotide sequence context of rNMPs, the Distribution Module calculates and plots the per-nucleotide counts of rNMPs, and the Hotspot Module calculates the most abundant sites of rNMPs and plots their consensus sequences.

When tested against current rNMP mapping software, Ribose-Map is the only software that analyzes data generated from any rNMP-seq technique, analyzes data from any organism with a sequenced reference genome, normalizes per-nucleotide counts of rNMP to account for sequencing depth, outputs a file containing the single-nucleotide genomic coordinates of rNMPs, and depends on only open-source software. By accommodating data from any rNMP-seq

technique, Ribose-Map standardizes the analysis of rNMP-seq experiments and facilitates direct comparisons of the results from these experiments. Ribose-Map is documented, maintained, and available for download at https://github.com/agombolay/ribose-map.

2.2 Materials and Methods

2.2.1 Create Ribose-Map bioinformatics toolkit

To create a standardized bioinformatics toolkit for the comprehensive analysis of rNMP-seq data, I created the Ribose-Map bioinformatics toolkit. Ribose-Map consists of six modules that calculate the genomic coordinates of rNMPs to single-nucleotide resolution and characterize the biological signatures of rNMPs (e.g., nucleotide composition of rNMPs, nucleotide sequence context of rNMPs, genome-wide distribution of rNMPs, and hotspot motifs of rNMPs). These modules include the 1) Alignment Module, 2) Coordinate Module, 3) Composition Module, 4) Sequence Module, 5) Distribution Module, and 6) Hotspot Module (28-30). Ribose-Map consists of custom Bash and R scripts that incorporate Linux commands and standard bioinformatics tools, such as Bowtie2 (31), SAMtools (32), UMI-tools (33), and the MEME Suite (34). Ribose-Map can be run via the Unix/Linux command line along with a configuration file described below. Ribose-Map is documented, maintained, and available for download at https://github.com/agombolay/ribose-map. **Figure 5** shows a graphical abstract of Ribose-Map.

To demonstrate the potential of Ribose-Map, we consider two rNMP-seq datasets available on the National Center for Biotechnology Information's (NCBI) Short Read Archive (SRA)- 1) ribose-seq data from mutant RNase H2 strains (*rnh201*, strain E134) of *S. cerevisiae* (accession number SRR11364933) (18) and 2) emRiboSeq data from mutant RNase H2 strains (*rnh201*, strain (Δl(−2)l-7BYUNI300) of *S. cerevisiae* (accession number SRR1734967) (19).

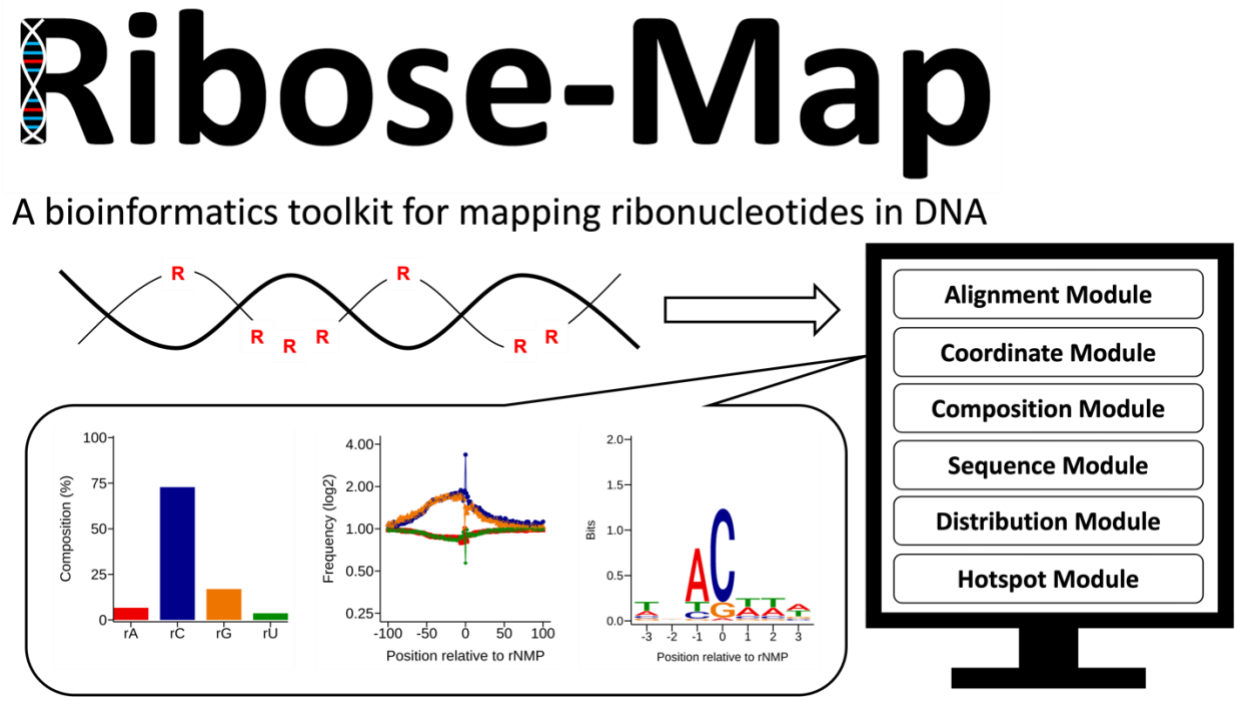


Figure 5. Ribose-Map Graphical Abstract. The Ribose-Map bioinformatics toolkit is a toolkit for the comprehensive analysis of rNMP-seq data. Through a series of modules, Ribose-Map transforms rNMP-seq data generated from ribose-seq, emRiboSeq, RHII-HydEn-seq, Alk-HydEn-seq, and PU-seq into summary datasets and publication-ready visualizations of results.

2.2.1.1 Expertise needed to run Ribose-Map

This protocol assumes experience with the Unix/Linux command-line interface. Users should be able to run programs from the command line and edit text files in the Unix/Linux environment.

2.2.1.2 Applications of Ribose-Map

Ribose-Map can be applied to rNMP-seq data generated from any organism (provided a sequenced reference genome is available) using any rNMP-seq technique. In addition to rNMPs, Ribose-Map can also be used to characterize the biological signatures of any single-nucleotide genomic coordinates of interest (e.g., single-nucleotide polymorphisms). To perform such an analysis, the user should bypass Ribose-Map's Alignment Module and Coordinate Module and

save a browser extensible data (BED) file of the coordinates to a folder according to Ribose-Map's output directory structure ('path/to/output/results/sample/coordinate') and then run Ribose-Map's Composition Module, Sequence Module, Distribution Module, and Hotspot Module.

2.2.1.3 Limitations of the protocol

Since Ribose-Map's Alignment Module uses the short-read aligner, Bowtie2 (31), long reads generated using third-generation sequencing technology (i.e., Oxford Nanopore), should not be input into Ribose-Map's Alignment Module. However, after aligning long reads to the reference genome of interest using a long-read aligner (e.g., BLASR (22), GraphMap (23), or Kart (24)), the user can input the aligned reads into Ribose-Map to calculate the single-nucleotide genomic coordinates of rNMPs and characterize the biological signatures of rNMPs. To perform such an analysis, the user should bypass Ribose-Map's Alignment Module and save a binary alignment map (BAM) file of the aligned reads to a folder according to Ribose-Map's output directory structure ('path/to/output/results/sample/alignment') and then run Ribose-Map's Coordinate Module, Composition Module, Sequence Module, Distribution Module, and Hotspot Module.

2.2.1.4 Equipment

- mamba (https://github.com/mamba-org/mamba)
- Git version-control system (https://git-scm.com/)
- Ribose-Map (https://github.com/agombolay/ribose-map)
- Hardware (64-bit computer running Linux or Mac OS X)

2.2.1.5 Installing software

Ribose-Map is available for download at https://github.com/agombolay/ribose-map. To download Ribose-Map, the user should install git on their local computer (https://git-scm.com/). To create a software environment in which to run Ribose-Map, the user should install mamba (https://github.com/mamba-org/mamba) on their local computer. The YAML file required to create the software environment is provided in the lib folder of the Ribose-Map GitHub repository. All commands shown below assume a Bash Shell and should be run via the Unix/Linux command-line (indicated by '$'). The user should replace all instances of 'path/to/' with their own file path.

Download Ribose-Map:

$ git clone https://github.com/agombolay/ribose-map.git

Create software environment:

Prefix should be set to the full path of the user-preferred location

$ mamba env create -p PREFIX -f path/to/ribose-map/lib/ribosemap.yaml

Activate software environment:

$ conda activate PREFIX

2.2.1.6 Required Data

Ribose-Map requires a FASTQ file of single- or paired-end rNMP-seq data, a FASTA file of the reference genome nucleotide sequence, a file containing the reference genome chromosome

sizes, Bowtie2 indexes for the reference genome, and a configuration file. Please see instructions below for creating a file containing chromosome sizes, Bowtie2 indexes, and a configuration file.

2.2.1.7 Preparing required reference genome data files

As an example, the data files required for the yeast sacCer2 genome are used.

Download FASTA file from UCSC genome browser for sacCer2 reference genome:

$ wget http://hgdownload.soe.ucsc.edu/goldenPath/sacCer2/bigZips/sacCer2.fa.gz

Download chrom.sizes file from UCSC genome browser for sacCer2 reference genome:

$ wget http://hgdownload.soe.ucsc.edu/goldenPath/sacCer2/bigZips/sacCer2.chrom.sizes

Unzip FASTA file:

$ gunzip path/to/sacCer2.fa.gz

Activate the software environment:

$ conda activate PREFIX

Download FASTQ files of rNMP-seq data from NCBI's Sequence Read Archive:

$ fastq-dump SRR11364933

$ fastq-dump SRR1734967

Create Bowtie2 indexes for reference genome:

$ bowtie2-build path/to/sacCer2.fa sacCer2

2.2.1.8 Create file of the reference genome chromosome sizes

If the chrom.sizes file is not readily available for download for the genome of interest, this file can be created from the FASTA file of the reference genome nucleotide sequence using the following commands. As an example, the FASTA file for the yeast sacCer2 genome is used.

$ samtools faidx sacCer2.fa > sacCer2.fa.fai

$ cut -f1,2 sacCer2.fa.fai > sacCer2.chrom.sizes

2.2.1.9 Create Configuration File

The configuration file should contain the following parameters: ‘sample’, ‘technique’, ‘fasta’, ‘basename’, ‘repository’, ‘read1’, ‘read2’ if paired-end, ‘mismatches’, ‘quality’, ‘threads’ and ‘percentile’. The ‘sample’ parameter should be specified as the sample name. The ‘technique’ parameter should be specified as ‘ribose-seq’, ‘emRiboseSeq’, ‘RHII-HydEn-seq’, ‘Alk-HydEn-seq’, or ‘Pu-seq’. The ‘fasta’ parameter should be specified as the FASTA filepath. The ‘basename’ parameter should be specified as the Bowtie2 index filepath excluding the bt2 extension. The ‘repository’ parameter should be specified as the output filepath. The ‘read1’ parameter should be specified as the read 1 FASTQ filepath. If the input data is paired-end, then an additional parameter, the ‘read2’ parameter, should be specified as the read 2 FASTQ filepath. The ‘mismatches’ parameter should be specified as the number of mismatches allowed in a seed alignment during multi-seed alignment (0 or 1). The ‘quality’ parameter should be specified as the

minimum alignment Phred quality score threshold for the aligned reads. The ‘threads’ parameter should be specified as the number of parallel computing threads to be used during alignment. The ‘percentile’ parameter should be specified as the percentile for rNMPs to be used by the Hotspot Module (e.g., 0.99 for 99th percentile). The ‘units’ parameter should be specified as genomic unit(s) that should be analyzed separately (e.g., chr1). If applicable, the ‘barcode’ parameter should be specified as the nucleotide sequence of the barcode. If applicable, the ‘pattern’ parameter should be specified as the nucleotide pattern of the UMI (e.g., ‘NNNNNNXXXNN’).

2.2.1.10 Trim reads based on quality, length, and adaptor content

The 5’ end of the ribose-seq reads contain an 8 nucleotide UMI and 3 nucleotide barcode sequence (‘NNNNNNXXXNN’) plus the tagged nucleotide. Thus, the minimum read length was selected as 62 nucleotides (12 nucleotides of UMI/barcode sequence + tagged nucleotide + 50 nucleotides of genomic DNA for alignment). The 3’ end of the ribose-seq reads contain a custom adaptor sequence (‘AGTTGCGACACGGATCTATCA’) that should be trimmed from the reads prior to alignment to the reference genome. The emRiboSeq reads do not contain a UMI or barcode sequence, so the minimum read length was selected as 51 nucleotides (tagged nucleotide + 50 nucleotides of genomic DNA for alignment). The user should replace ‘path/to/’ with their own filepaths. ‘Output’ should be set to the user-preferred location of the output files.

Trim the ribose-seq reads with the following command:

```
$ trim_galore -a AGTTGCGACACGGATCTATCA -q 15 --length 62 path/to/SRR11364933.fastq -o ‘output’
```

Trim the emRiboSeq reads with the following command:

```
$ trim_galore --illumina -q 15 --length 50 path/to/SRR1734967.fastq -o ‘output’
```

2.2.1.11 Running Ribose-Map

Ribose-Map should be run via the Unix/Linux command-line (indicated below by '$'). **Table 1** shows troubleshooting advice for possible errors when running Ribose-Map.

Table 1. Troubleshooting advice for Ribose-Map

Problem	Possible Reason	Solution
Coordinate Module outputs empty BED file	'Technique' parameter is not specified correctly in config	Check 'technique' is spelled correctly
Composition, Sequence, Distribution, and/or Hotspots Modules do not output results	BED file is empty or missing	Run the Coordinate Module to obtain file of genomic coordinates of rNMPs
Software cannot be located	ribosemap_env not activated	Activate ribosemap_env environemnt

Run Alignment Module:

$ path/to/ribose-map/modules/ribosemap alignment path/to/SRR11364933.config

$ path/to/ribose-map/modules/ribosemap alignment path/to/SRR1734967.config

Run Coordinate Module:

$ path/to/ribose-map/modules/ribosemap coordinate path/to/SRR11364933.config

$ path/to/ribose-map/modules/ribosemap coordinate path/to/SRR1734967.config

Run Sequence Module:

$ path/to/ribose-map/modules/ribosemap sequence path/to/SRR11364933.config

$ path/to/ribose-map/modules/ribosemap sequence path/to/SRR1734967.config

Run Distribution Module:

$ path/to/ribose-map/modules/ribosemap coordinate path/to/SRR11364933.config

$ path/to/ribose-map/modules/ribosemap coordinate path/to/SRR1734967.config

Run Hotspot Module:

$ path/to/ribose-map/modules/ribosemap hotspot path/to/SRR11364933.config

$ path/to/ribose-map/modules/ribosemap hotspot path/to/SRR1734967.config

2.2.1.12 Timing

Running this protocol on the example data provided will take about 30 minutes of hands-on time and 3 hours of computing time. This protocol was run using a high-performance computer cluster with two nodes, four processors per node and 5 GB of memory per core; however, a more standard set-up (e.g., personal laptop with Mac operating system) can be used to run this protocol, and timing would be comparable. For other datasets, the timing may take longer depending on the volume of rNMP-seq data, the size of the reference genome, and the computer used.

2.2.2 Test Ribose-Map against current rNMP mapping software

To validate Ribose-Map against current rNMP mapping software, I applied the following criteria: 1) analyzes data generated from any rNMP-seq technique, 2) analyzes data from any

organism with a reference genome, 3) normalizes per-nucleotide rNMP counts to account for sequencing depth among datasets, 4) outputs a file containing the genomic coordinates of rNMPs that can be used for downstream analyses, and 5) depends on only open-source software.

2.3 Results and Discussion

2.3.1 Create Ribose-Map bioinformatics toolkit

The Ribose-Map bioinformatics is the first known standardized bioinformatics toolkit for the comprehensive analysis of rNMP-seq data. Through a series of modules, Ribose-Map calculates the genomic coordinates of rNMPs to single-nucleotide resolution and characterizes the biological signatures of rNMPs (e.g., nucleotide composition of rNMPs, nucleotide sequence context of rNMPs, genome-wide distribution of rNMPs, and hotspot motifs of rNMPs). **Figure 6** shows an overview of Ribose-Map, including input/output. Ribose-Map is documented, maintained, and available for download at https://github.com/agombolay/ribose-map.

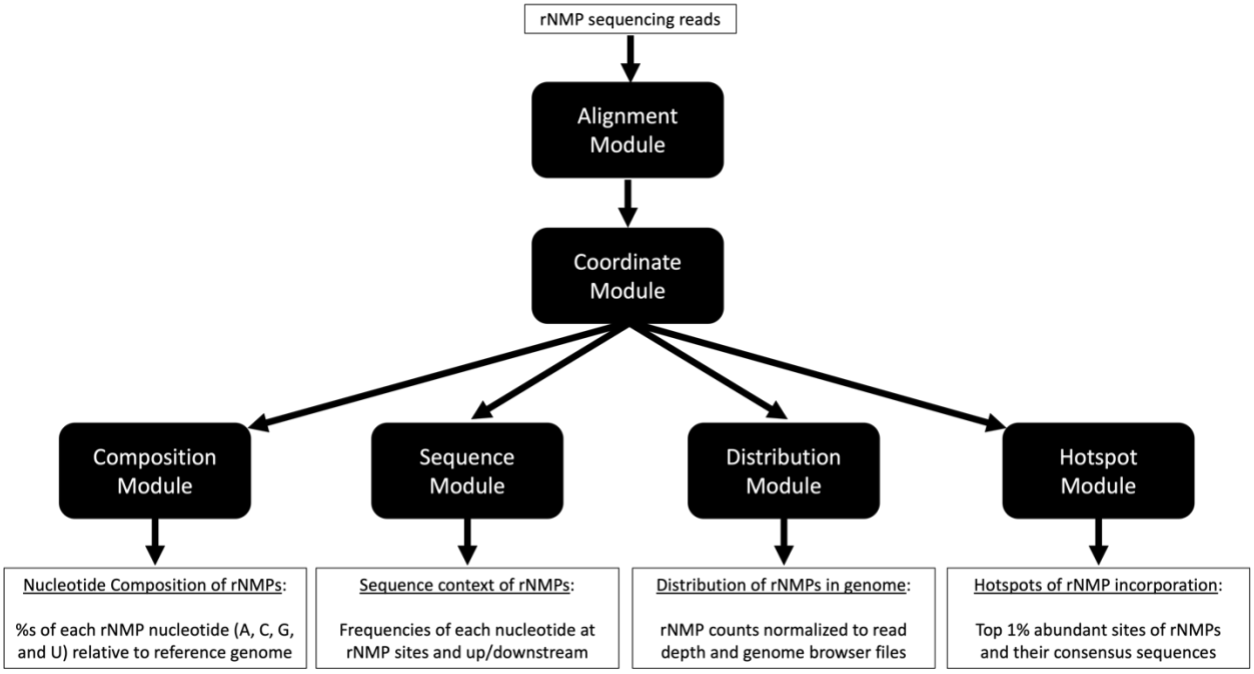


Figure 6. Overview of Ribose-Map. The Alignment Module first aligns input rNMP sequencing reads to the reference genome (de-duplicates/de-multiplexes reads if needed). Then, based on the aligned reads, the Coordinate Module calculates the genomic coordinates of rNMPs to single-nucleotide resolution and per-nucleotide rNMP counts for the user-specified rNMP sequencing technique. Next, the Composition, Sequence, Distribution, and Hotspot Modules use these coordinates to produce data tables, genome browser annotation files, and visualizations of results.

2.3.1.1 Read Alignment with the Alignment Module

The Alignment Module aligns single- or paired-end ribose-seq, emRiboSeq, RHII-HydEn-seq, Alk-HydEn-seq, and Pu-seq reads to the reference genome using Bowtie2 (31). Prior to alignment, the reads should be processed using data cleaning software (e.g., Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)) to remove any low-quality base calls and/or sequencing adaptors. If needed, the Alignment Module also de-duplicates the reads using UMI-tools (33) and de-multiplexes the reads based on barcodes using seqtk (https://github.com/lh3/seqtk). The Alignment Module outputs a BAM file of the aligned reads and a log file stating the overall alignment rate. If the reads contain a UMI and/or barcode, the log

file will also include the percentage of reads remaining after de-duplication and the percentage of reads containing the barcode. If the number of reads present in the BAM file is substantially less than the number of input reads, the user should check the log file to assess if this discrepancy is due to low alignment rate, low barcode rate, and/or high duplication rate. About 90% or more of the reads should align to the reference genome (35). Lower alignment rates could indicate low quality reads and/or the presence of contaminants (e.g., sequencing adaptors) in the reads.

2.3.1.2 Locating genomic coordinates of rNMPs with the Coordinate Module

Once the reads are processed and aligned to the reference genome, the Coordinate Module outputs a BED file containing the single-nucleotide genomic coordinates of rNMPs (chromosome, 0-based start coordinate, 0-based end coordinate, read name, and DNA strand (+/−)) and a tab-delimited file containing the per-nucleotide counts of rNMPs. **Table 2** outlines the arithmetic used by the Coordinate Module to calculate the genomic coordinates of rNMPs for ribose-seq, emRiboSeq, RHII-HydEn-seq, Alk-HydEn-seq, and Pu-seq. The Coordinate Module also screens the genomic coordinates of rNMPs for biological relevance. Although likely rare, contamination or inefficiencies during rNMP-seq library preparation could generate reads that align to the 5'-most ends of a given chromosome. Since rNMPs captured by emRiboSeq, Alk-HydEn-seq, and Pu-seq are located either upstream from the tagged nucleotides or downstream from the reverse complements of the tagged nucleotides, reads that align to the 5'-most ends of the chromosome for any of these techniques would result in genomic coordinates that are located beyond the ends of the chromosome. If such biologically meaningless coordinates (e.g., −1) were input into Ribose-Map's Sequence Module, Distribution Module, or Hotspot Module or any other downstream analytical tools (e.g., UCSC genome browser), these programs would error out.

Table 2. Arithmetic to calculate the genomic coordinates of rNMP sites. Start represents the 0-based start coordinate of rNMP; End represents the 0-based end coordinate of rNMP; S represents the 0-based start coordinate of a read; E represents the 0-based end coordinate of a read; Strand represents the Strand of rNMP; + represents the Watson strand and – represents the Crick strand.

	Reads aligned to + DNA Strand			**Reads aligned to – DNA strand**		
	Start	**End**	**Strand**	**Start**	**End**	**Strand**
ribose-seq	S	S + 1	–	E – 1	E	+
emRiboSeq	S – 1	S	–	E	E + 1	+
RHII-HydEn-seq	S	S + 1	+	E – 1	E	–
Alk-HydEn-seq and Pu-seq	S – 1	S	+	E	E + 1	–

2.3.1.3 Characterizing biological signatures of rNMPs with Composition, Sequence, Distribution, and Hotspot Modules

Based on the genomic coordinates of rNMPs calculated by the Coordinate Module, the Composition Module, Sequence Module, Distribution Module, and Hotspot Module can be used to characterize the biological signatures of rNMPs, including nucleotide composition of rNMPs, nucleotide sequence context of rNMPs, and hotspot motifs of rNMPs. The Composition Module calculates and plots the percentage of each type of rNMP (r[A, C, G, U]) for each genomic unit (e.g., mitochondria) normalized to the nucleotide composition of the reference genome. **Figure 7** shows an example of the plots output by the Composition Module. The Sequence Module calculates and plots the frequencies of each type of rNMP and the dNMPs up/downstream from rNMPs for each genomic unit normalized to the nucleotide composition of the reference genome. The Distribution Module outputs tab-delimited files containing the per-nucleotide coverage of rNMPs for each genomic unit normalized to account for sequencing depth and plots the per-nucleotide coverage of rNMPs for each genomic unit separated by DNA strand. **Figure 8** shows

an example of the plots output by the Sequence Module. The Distribution Module also outputs BedGraph files containing per-nucleotide coverage of rNMPs that can be directly uploaded to a genome browser, such as the UCSC Genome Browser (http://genome.ucsc.edu/), as custom annotation tracks. **Figure 9** shows an example of the plots output by the Distribution Module. The Hotspot Module calculates hotspots of rNMPs and plots their consensus sequences using the MEME Suite (34). **Figure 10** shows an example of the plots output by the Hotspot Module.

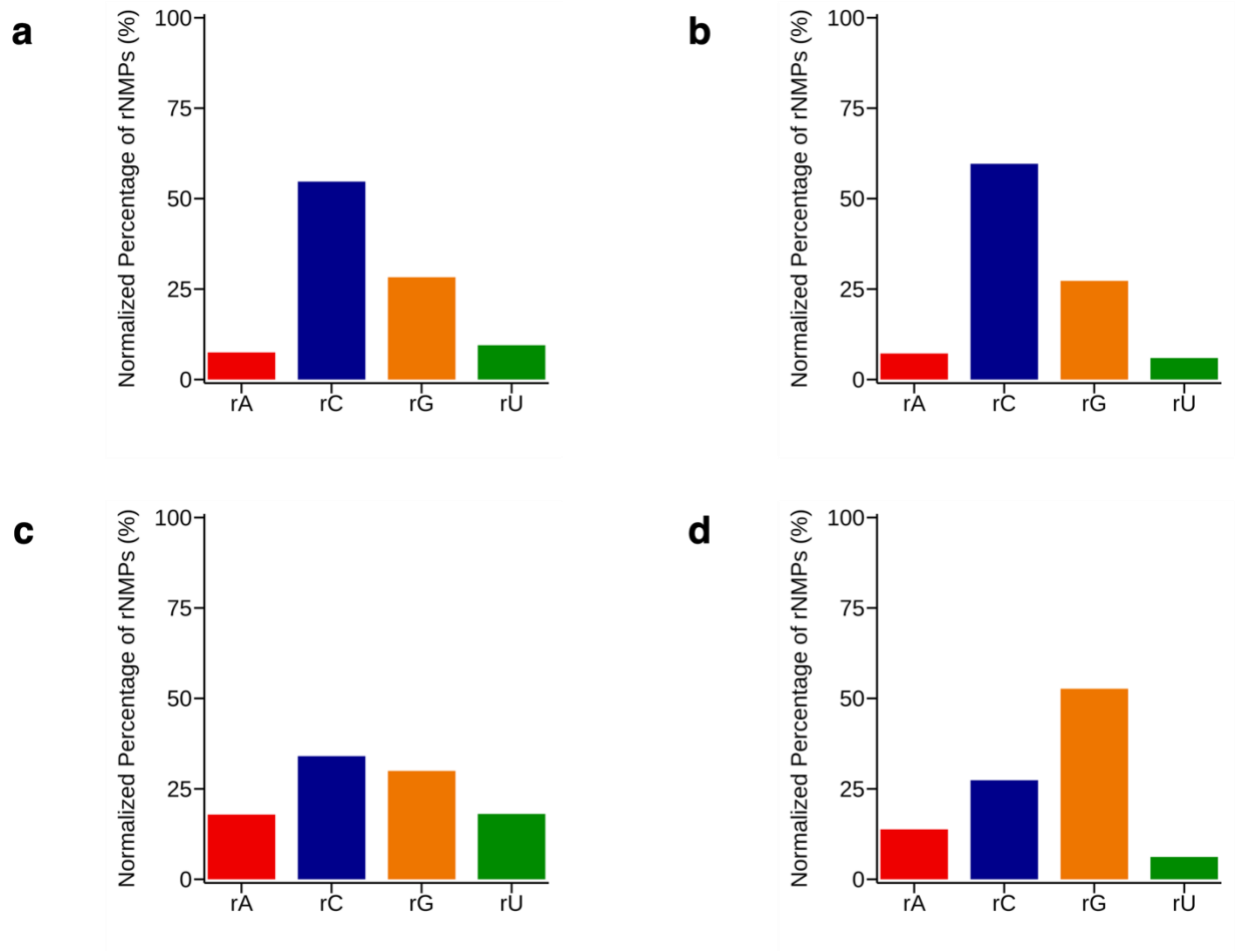


Figure 7. Nucleotide composition of rNMPs in *rnh201 S. cerevisiae* DNA. Ribose-seq strain E134 data for (a) nuclear DNA and (b) mitochondrial DNA; emRiboSeq strain Δl(-2)l-7BYUNI300 data for (c) nuclear DNA and (d) mitochondrial DNA. Percentages were normalized to the nucleotide composition of the sacCer2 reference genome and plotted using the Composition Module.

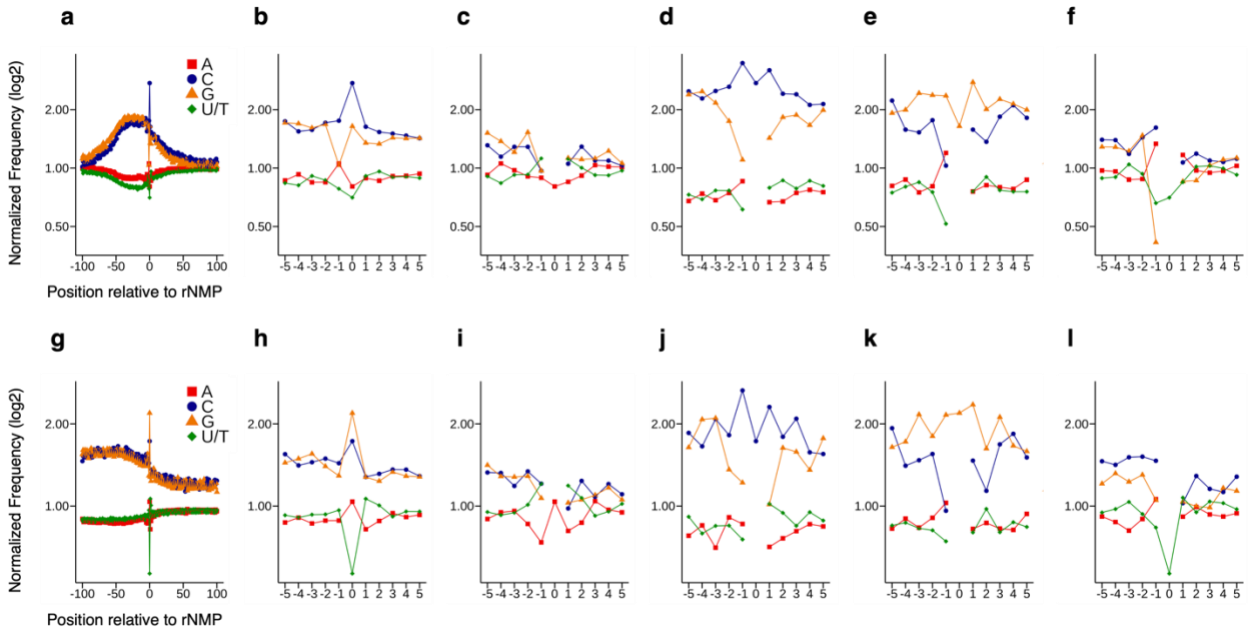


Figure 8. Nucleotide sequence context of rNMPs in *rnh201 S. cerevisiae* mitochondrial DNA. (a) Ribose-seq strain E134 data zoomed-out for r[A, C, G, U], (b) zoomed-in for r[A, C, G, U], (c) rA, (d) rC, (e) rG, (f) and rU. EmRiboSeq strain Δl(-2)l-7BYUNI300 data zoomed-out for (g) r[A, C, G, U], (h) zoomed-in for r[A, C, G, U], (i) rA, (j) rC, (k) rG, and (l) rU. Nucleotide frequencies were normalized to nucleotide composition of sacCer2 and plotted using the Sequence Module. Positions –100 to -1 on the x-axis represent nucleotides upstream from rNMPs, position 0 on the x-axis represents rNMPs, and positions 1 to 100 on the x-axis represent nucleotides downstream from rNMPs. The y-axis represents normalized frequency (log2) of rNMPs.

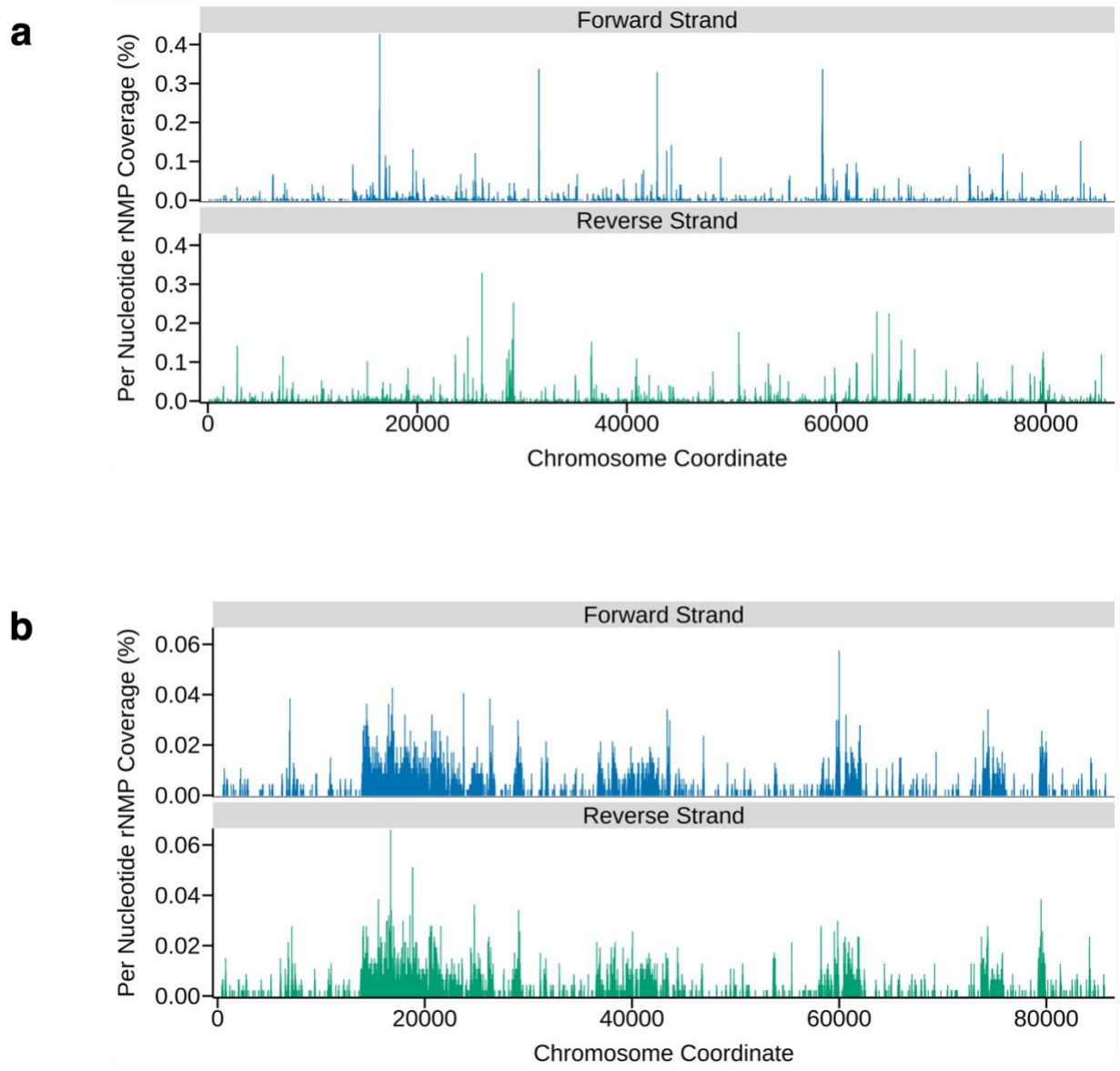


Figure 9. Per-nucleotide coverage of rNMPs in *rnh201 S. cerevisiae* mitochondrial DNA. (a) Ribose-seq strain E134 data and (b) emRiboSeq strain Δl(−2)l-7BYUNI300 data. Per-nucleotide coverage was normalized for sequencing depth and plotted using the Distribution Module.

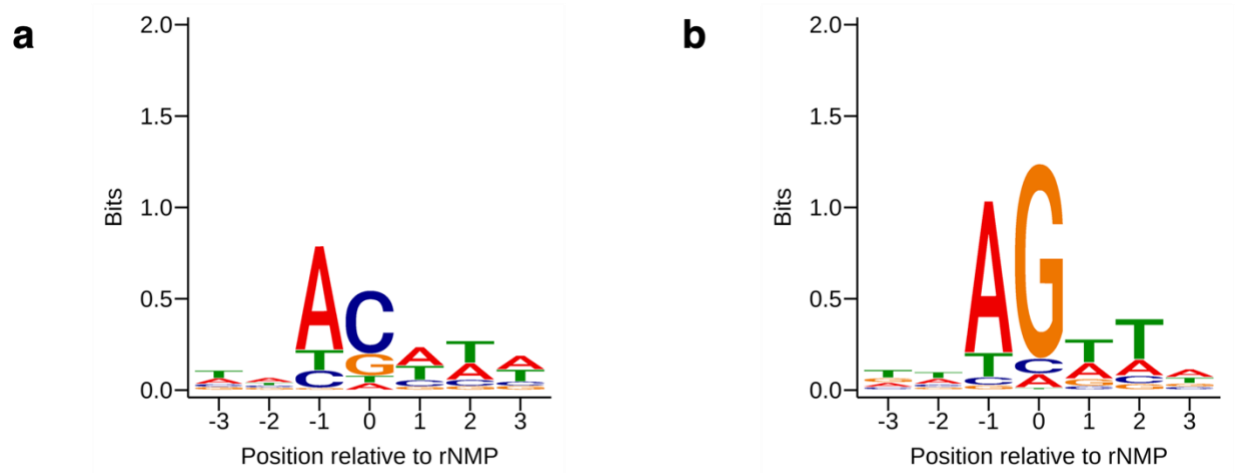


Figure 10. Hotspot motifs of rNMP hotspots in *rnh201 S. cerevisiae* mitochondrial DNA. (a) Ribose-seq strain E134 data (b) emRiboSeq strain Δl(−2)l-7BYUNI300 data. Top 1% most abundant sites of rNMPs were calculated and plotted using the Hotspot Module. Positions -3 to -1on the x-axis represent nucleotides upstream from rNMP sites, position 0 on the x-axis represents rNMPs, and positions 1 to 3 on the x-axis represent nucleotides downstream from rNMPs. The y-axis represents conservation of nucleotide sequence measured in bits.

2.3.2 Test Ribose-Map against current rNMP mapping software

When tested against emRiboSeqProcessor, Modmap, and Puseq_app, Ribose-Map is the only rNMP mapping software that meets all of the evaluation criteria (**Table 3**). Ribose-Map analyzes data generated from any rNMP-seq technique, analyzes data from any organism with a sequenced reference genome, normalizes per-nucleotide counts of rNMPs to account for sequencing depth, outputs a file containing the single-nucleotide genomic coordinates of rNMPs that can be used for downstream analyses, and depends on only open-source software. emRiboSeqProcessor (19), Modmap (18), and Puseq_app (21) are highly customized, limited in terms of scope of analysis, depend on proprietary software, and/or provide limited documentation. Modmap is customized to analyze only ribose-seq data, does not provide online documentation, depends on proprietary software, and does not output a file containing the genomic coordinates of rNMPs. emRiboSeq Processor is customized to analyze only emRiboSeq data, does not provide

online documentation, does not output a file containing the genomic coordinates of rNMPs (only counts of rNMPs), and does not characterize the biological signatures of rNMPs. The Puseq_app is customized to analyze only Pu-seq data, does not provide online documentation, does not output a file containing the genomic coordinates of rNMPs, and tracks the division of labor of DNA polymerases rather than characterizes the biological signatures of rNMPs.

Table 3. Comparison of Ribose-Map and current rNMP mapping software.

	Ribose-Map	**Modmap**	**emRiboSeq Processor**	**Puseq _app**
1) Analyzes data from any rNMP sequencing technique	✓	✗	✗	✗
2) Analyzes data from any organism with reference genome	✓	✗	✗	✓
3) Normalizes rNMP counts to account for sequencing depth	✓	✗	✓	✗
4) Outputs file containing genomic coordinates of rNMPs	✓	✗	✓	✓
5) Depends on only open-source software	✓	✗	✓	✓

2.4 Conclusion

Current rNMP mapping software are highly customized, limited in terms of scope of analysis, depend on proprietary software, and/or provide limited documentation. The Ribose-Map bioinformatics toolkit is the first standardized bioinformatics toolkit for the comprehensive analysis of rNMP-seq data. Through a series of modules, Ribose-Map calculates the genomic coordinates of rNMPs to single-nucleotide resolution and characterizes the biological signatures of rNMPs (e.g., nucleotide composition of rNMPs, nucleotide sequence context of rNMPs, genome-wide distribution of rNMPs, and hotspot motifs of rNMPs). By accommodating data from

any rNMP-seq technique, Ribose-Map standardizes the analysis of rNMP-seq experiments and facilitates direct comparisons of the results from these experiments.

2.5 Acknowledgments

This work was supported by the National Institutes of Health (grant R01ES026243-01 to F.S.), the Parker H. Petit Institute for Bioengineering and Bioscience at Georgia Institute of Technology (grant 12456H2 to F.S.), and the Howard Hughes Medical Institute (grant 55108574 to F.S.).

CHAPTER 3. Characterization of the biological signatures of rNMP incorporation in different species, strains, and ribonuclease H genotypes of yeast cells using Ribose-Map

The work presented in this chapter consists of the research project published in Balachander, S.*, Gombolay, A.L.*, Yang, T.*, Xu, P.*, Newnam, G., Keskin, H., El-Sayed, W.M.M., Bryskin, A.V., Tao, S., Bowen, N.E., Schinazi, R.F., Kim, B., Koh, K.D., Vannberg, F.O., and Storici, F. (2020) Ribonucleotide incorporation in yeast genomic DNA shows preference for cytosine and guanosine preceded by deoxyadenosine. *Nature Communications*, 11: 2447. *co-first authors

3.1 Abstract

Since the development of rNMP-seq techniques is recent, the biological signatures of rNMP incorporation in DNA have yet to be thoroughly characterized. Here, we built ribose-seq sequencing libraries derived from mitochondrial and nuclear DNA of *S. cerevisiae*, *S. paradoxus*, and *S. pombe* cells and applied the Ribose-Map bioinformatics toolkit to characterize the biological signatures of rNMP incorporation of these libraries. We uncovered distinct characteristics of rNMP incorporation among yeast species and between wild type and different RNase H-mutant genotypes, suggesting non-random incorporation of rNMPs in DNA. In the mitochondrial DNA of all three yeast species, rC and rG are consistently the most abundant type of rNMP present in both wild type and mutant RNase H cells. In the nuclear DNA, rC and rG are consistently the most abundant type of rNMP present in only *S. cerevisiae* and *S. paradoxus rnh201* cells. In addition, rC and rG were found in C/G-rich regions of *S. cerevisiae* and *S. paradoxus* mitochondrial DNA but not in *S. pombe* or in the nucleus of any of the yeast species. Furthermore, deoxyadenosine is found directly upstream from the most abundant genomic rC's and rG's in all three yeast species.

3.2 Materials and Methods

3.2.1 Yeast Strains

All the yeast strains used in this study are presented in **Supplementary Table 1**. We used haploid *S. cerevisiae* strains from different backgrounds: E134, BY4741, BY4742, YFP17, W303, and S288C. In addition to *S. cerevisiae*, we also used *S. paradoxus* (DG2204) and *S. pombe* (JZ105). Standard genetic and molecular biology methods were used for yeast growth, gene disruption, isolation of mutants, yeast marker selection, yeast genome engineering, yeast colony PCR, and sequence analysis of yeast DNA (36-38). All RNH201 deletion strains were made by

replacing RNH201 via transformation with a PCR product containing hygMX4 or kanMX4 cassette flanked by 50 nucleotides of sequence homologous to regions upstream and downstream of RNH201 ORF. KK-172 was made from KK-44 by replacement of RNH1 with the kanMX4 cassette. Yeast strains SB-285 and SB-286 were derived from KK-44 by using the delitto perfetto method (38) and then by popping out the CORE cassette by a pair of oligonucleotides, primers 202PIP.F and 202PIP.R to mutate the PCNA-interacting peptide box, which is present in Rnh202 to make *rnh202*-FF346,347A18,34. SB-311 was derived from KK-2 by using the delitto perfetto method to generate *rnh201*-P45D and *rnh201*-Y219A by using primers RNH P45D.60 and RNH Y219A.60 to result in an RNase H2 ribonucleotide excision defective (RED) (39) mutant. All mutations were confirmed by sequence analysis of PCR products obtained from amplification of a DNA region surrounding the specific mutation.

3.2.2 Properties of ribose-seq

The ribose-seq technique allows to build libraries of rNMP sites that are present in any DNA source of interest (18,40). The key enzyme of ribose-seq is AtRNL. AtRNL recognizes the 2′,3′ cyclic phosphate end of a single-stranded (ss) genomic DNA fragment terminated with an rNMP, generated upon treatment of double-stranded DNA by alkali. AtRNL directly ligates the rNMP-terminated ssDNA to the 5′-phosphate end of the same ssDNA fragment, to which an adaptor sequence has been attached before the alkali treatment (18,40). Such ssDNA circles, each containing one rNMP next to the adaptor, constitute the rNMP library for a given DNA sample. Ribose-seq cannot capture primers of Okazaki fragments because these do not have an adaptor ligated at their 5′ end and are degraded upon alkali treatment. Moreover, thanks to the specificity of the enzyme activity, AtRNL cannot capture abasic sites or DNA sequences upstream or

downstream from DNA breaks (18). Thus, cell cycle stage and/or integrity of DNA do not generate false positives in the preparation of the ribose-seq libraries from the DNA samples of choice.

3.2.3 Choice of yeast backgrounds for ribose-seq library preparation

We built 34 ribose-seq libraries of yeast mitochondrial DNA and 25 libraries of yeast nDNA (**Tables 4-9)**. These include 25 mitochondrial and 18 nuclear libraries derived from *S. cerevisiae* of six different strain backgrounds E134, BY4741, BY4742, YFP17, W303, and S288C, four mitochondrial and three nuclear from *S. paradoxus* of strain DG2204, and five mitochondrial and four nuclear from *S. pombe* of strain JZ105 (**Supplementary Table 1)**. For *S. cerevisiae*, we utilized some of the most commonly used haploid yeast laboratory strains. Strain S288C was used in the systematic sequencing project of *S. cerevisiae* (41); strains BY4741 and BY4742 both are derivatives of S288C, were used in the *S. cerevisiae* gene disruption project, and have opposite mating type (42); W303 is another common yeast laboratory strain more distantly related to S288C47; YFP17 is a derivative of DBY745 (MATα ura3-52 leu2-3 leu2-112 adel-100) (43); and E134 is a derivative of CG379 (44). The ribose-seq libraries derive from wild-type RNase H2 or *rnh201* cells of all three yeast species, and also from *S. cerevisiae* cells containing the RED mutations of *rnh201* (P45D-Y219A), which block RNase H2 activity on single rNMPs in DNA but allow cleavage at long RNA/DNA hybrids (39), the pip mutation in *rnh202* (FF346,347AA), which impedes interaction with PCNA (45), or a *rnh1* allele (**Supplementary Table 1)**. For wild-type cells in the different species, and for RNase H-mutant cells in *S. cerevisiae*, we constructed two or more ribose-seq libraries using two or three different sets of restriction enzymes (RE1: DraI, EcoRV, and SspI; RE2: AluI, DraI, EcoRV, and SspI; and RE3: RsaI and HaeIII). This strategy allowed us (i) to verify that the conclusions taken from our analyses of ribose-seq data are

not influenced by a particular set of restriction enzymes used to fragment the DNA and (ii) to further confirm reproducibility of the results. All ribose-seq libraries have a specific barcode within the UMI to distinguish the libraries from each other in the sequencing run and eliminate PCR duplicates.

Table 4. Ribose-seq *S. cerevisiae* mitochondrial libraries

Species	Genotype	Strain	SRR
S. cerevisiae	WT	E134	SRR11364942
			SRR11364941
		BY4741	SRR11364930
		BY4742	SRR11364919
			SRR11364914
			SRR11364913
		YFP17	SRR11364912
			SRR11364911
			SRR11364910
		W303	SRR11364909
		S288C	SRR11364940
	rnh201	E134	SRR11364935
			SRR11364934
			SRR11364933
			SRR11364932
			SRR11364931
			SRR11364929
		BY4742	SRR11364928
		YFP17	SRR11364927
	pip	E134	SRR11364939
		BY4742	SRR11364938
	rnh1	E134	SRR11364937
			SRR11364936
	RED	BY4742	SRR11364926
			SRR11364925

Table 5. Ribose-seq *S. cerevisiae* nuclear libraries

Species	Genotype	Strain	SRR
S. cerevisiae	WT	E134	SRR11364942
			SRR11364941
		BY4741	SRR11364930
		BY4742	SRR11364919
			SRR11364913
		YFP17	SRR11364912
	rnh201	E134	SRR11364935
			SRR11364934
			SRR11364933
			SRR11364932
			SRR11364931
			SRR11364929
		BY4742	SRR11364928
		YFP17	SRR11364927
	pip	BY4742	SRR11364938
	rnh1	E134	SRR11364936
	RED	BY4742	SRR11364926
			SRR11364925

Table 6. Ribose-seq *S. paradoxus* mitochondrial libraries

Species	Genotype	Strain	SRR
S. paradoxus	WT	DG2204	SRR11364924
			SRR11364923
			SRR11364922
	rnh201		SRR11364921

Table 7. Ribose-seq *S. paradoxus* nuclear libraries

Species	Genotype	Strain	SRR
S. paradoxus	WT	DG2204	SRR11364924
			SRR11364922
	rnh201		SRR11364921

Table 8. Ribose-seq *S. pombe* mitochondrial libraries

Species	Genotype	Strain	SRR
S. pombe	WT	JZ105	SRR11364920
			SRR11364918
			SRR11364917
	rnh201		SRR11364916
			SRR11364915

Table 9. Ribose-seq *S. pombe* nuclear libraries

Species	Genotype	Strain	SRR
S. pombe	WT	JZ105	SRR11364920
			SRR11364918
	rnh201		SRR11364916
			SRR11364915

3.2.4 Ribose-seq library preparation

Libraries were prepared using the ribose-seq method with some modifications (18,40). Specifically, we optimized the ribose-seq protocol by (i) redesigning the molecular barcode-containing adaptor, making it shorter and removing overlapping sequences, (ii) fragmenting the genome of interest in smaller fragments (~450 bp); (iii) performing two rounds of PCR and overall reducing the PCR cycle number; (iv) cutting and purifying a specific size range of the ribose-seq library from the non-denaturing gel to eliminate any primer dimers formed during PCR and any long products that are not proficient for sequencing (40). All the commercial enzymes utilized in the ribose-seq protocol were used according to the manufacturer's instructions. *S. cerevisiae* cells were cultured in liquid rich medium (150 mL of a 250 mL glass flask) containing yeast extract, peptone, and 2% (wt/vol) dextrose (YPD) for 2 days at 30 °C with shaking to reach stationary phase with a density of ~108 cells/mL. Genomic DNA was extracted using Qiagen Genomic DNA protocol "Preparation of Yeast Samples." Successively, 40 μg of yeast genomic DNA were fragmented using restriction enzymes to produce blunt-ended fragments with an average size of 450 base pairs (bp) in length. Multiple sets of restriction enzymes were used for different library preparation. The different combinations used were (i) RE1: DraI, EcoRV, and SspI; (ii) RE2: AluI, DraI, EcoRV, and SspI; and (iii) RE3: RsaI and HaeIII. Following restriction digestion, the

fragmented DNA was purified by spin column (Qiagen). The fragments were tailed with dATP (Sigma Aldrich) by using Klenow Fragment (3′→5′ exo-) (NEB) for 30 min at 37 °C and purified by using spin column. Following dA-tailing and purification, the DNA fragments were annealed with a partially double-stranded adaptor (Adaptor. L1 or Adaptor.L2 with Adaptor.S) by using T4 DNA ligase (NEB) incubating overnight at 15 °C. Following overnight ligation, the products were purified using RNA Clean XP beads (Beckman Coulter). The annealed fragments were treated with 0.3M NaOH for 2 h at 55 °C to denature the DNA strands, and to cleave at the rNMP sites resulting in 2′,3′-cyclic phosphate and 2′-phosphate termini. This was followed with neutralization using 0.3M HCl and purification using RNA Clean XP beads. All the successive purification steps were performed using RNA Clean XP beads. The single-stranded ssDNA fragments were incubated with 1 μM *Arabidopsis thaliana* tRNA ligase (AtRNL), 50mM Tris-HCl pH 7.5, 40mM NaCl, 5mM MgCl2, 1mM DTT, and 300 μM ATP in a volume of 20 μL for 1 h at 30 °C, followed by purification. AtRNL aids in ligating the 2′-phosphate ends of rNMP-terminated ssDNA fragment to its opposite 5′-phosphate end, which results in a circular ssDNA. Due to the efficient removal of rNMPs by RNase H2, nuclear libraries of wild-type RNH201, *rnh202*-pip, and *rnh1* cells generally had a much lower number of reads compared to the mitochondrial libraries of the same cells, and thus had a higher number of background reads that originated from the capture of restriction enzyme ends likely by residual activity of T4 DNA ligase. These background reads were identified computationally and were found to constitute less than 5% of the total reads in the mitochondrial and less than 7.5% nuclear *rnh201* and *rnh201*-RED libraries. To determine whether the background reads could influence our results and conclusions, we subtracted the background reads from the total reads. When the background reads were <12% of the total reads, the results after background subtraction were found to be the same as those without subtraction. Therefore,

for our analyses, we selected only those nuclear libraries for which the background was <12%, so that our results and conclusions are not biased. The fragments were then treated with T5 Exonuclease (NEB) 50 units in 50 μL volume for 1 h 30 min at 37 °C to degrade the unligated ssDNA fragments. After purification, the circular fragments were incubated with 1 μM 2′-phosphotransferase (Tpt1), 20mM Tris-HCl pH 7.5, 5 mM MgCl2, 0.1mM DTT, 0.4% Triton X-100, and 10mM NAD+ in a volume of 40 μL for 1 h at 30°C to remove the 2′-phosphate present at the ligation junction. After Tpt1 treatment and purification, the circular fragments were PCR-amplified using two rounds of amplifications to result in ribose-seq library: both PCR rounds begin with an initial denaturation at 98°C for 30 s. Then denaturation at 98°C for 10 s, primer annealing at 65 °C for 30 s, and DNA extension at 72 °C for 30 s are performed. These three steps are repeated for 6–15 cycles in the first PCR round, and for 7–13 cycles in the second PCR round depending on the concentration of the circular ssDNAs containing the rNMPs. Successively, there is a final extension reaction at 72 °C for 2 min for both PCRs. A first round of PCR was performed to amplify and introduce the sequences of Illumina TruSeq CD Index primers. The primers (PCR.1 and PCR.2) used for the first round were the same for all libraries. A second round of PCR was performed to attach specific indexes i7 and i5 for each library. PCR round 1 and 2 were performed using Q5-High Fidelity polymerase (NEB) for 10 and 7 cycles, respectively (unless specified otherwise). Following the PCR cycles, the ribose-seq library was loaded on a 6% non-denaturing polyacrylamide gel and stained using 1× SYBR Gold (Life Technologies) for 40–45 min. As shown in Koh et al. (18), in control experiments for the optimized ribose-seq protocol, we found that exclusion of either AtRNL or alkali treatment prevented library formation. Fragments between 200 and 700 bp were cut and gel purified using the crush and soak method (46). The resulting

ribose-seq libraries were mixed at equimolar concentrations and normalized to 1.5 nM. The libraries were sequenced using Illumina in the Core Facility at the Georgia Institute of Technology.

3.2.5 dNTP and rNTP measurements

Yeast cell lysate were appropriately prepared to extract the dNTPs and rNTPs using an established protocol (47) with some modifications. Yeast cells were grown as described above in YPD medium for 2d at 30 °C with shaking to reach the stationary phase with a density of ~108 cells/mL. Cells were then harvested, washed with DI water, and resuspended in a solution of 1M sorbitol, 100 mM EDTA, 14 mM B-mercaptoethanol, and 1 mg of Zymolase and incubated at 37°C for 2 h. The mixture was then spun down and the pellet of cells were washed two times with a Phosphate Buffer Saline solution. The pellet was then resuspended in 65% methanol and mixed by pipetting. The mixture was heated at 95 °C for 3min and then placed on ice for 1 min. The cells were spun at 16,000 × g for 3 min. The supernatant was transferred to a 30 kDa column where it is spun for 30 min at 18,000 × g. The flow through was lyophilized and stored at −80°C. To quantify the intracellular dNTPs and rNTPs, an ion pair chromatography-tandem mass spectrometry method (48) was applied, with modifications. Chromatographic separation and detection were performed on a Vanquish Flex system (Thermo Scientific) coupled with a TSQ Quantiva triple quadrupole mass spectrometer (Thermo Scientific). Analytes were separated using a Kinetex EVO-C18 column (100 × 2.1 mm, 2.6 μm) (Phenomenex) at a flow rate of 250 μL/min. The mobile phase A consisted of 2 mM of ammonium phosphate monobasic and 3 mM of hexylamine in water and the mobile phase B consisted of acetonitrile. The LC gradient increased from 10% to 35% of mobile phase B in 5 min, and then returned to the initial condition. Selected

reaction monitoring in both positive and negative modes (spray voltage: 3200 V (pos) or 2500 V (neg); sheath gas: 35 Arb; auxiliary gas: 20 Arb; ion transfer tube temperature: 350°C; vaporizer temperature: 380°C) was used to detect the targets: dATP (492→136, pos), dGTP (508→152, pos), dCTP (466→158.9, neg), TTP (481→158.9, neg), ATP (508→136, pos), GTP (524→152, pos), CTP (482→158.9, neg), UTP (483→158.9, neg). Extracted samples were reconstituted in 100 µL of mobile phase A. After centrifuging at 13,800 × g for 10 min, 40 µL of supernatant was mixed with 10 µL of 13C and 15N labeled dNTPs and rNTPs as internal standards, and then subjected to analysis. Data were collected and processed by Thermo Xcalibur 3.0 software. Calibration curves were generated from standards by serial dilutions in mobile phase A (dATP and dGTP 0.1–400 nM, dCTP and TTP 0.2–400 nM, rNTPs 1–4000 nM). The calibration curves had r2 value greater than 0.99. All the chemicals and standards are analytical grade or higher and were obtained commercially from Sigma Aldrich. Nucleotides were at least 98% pure.

3.2.6 Processing and alignment of sequencing reads

For the ribose-seq libraries, the sequencing reads consist of an eight-nucleotide UMI, a three-nucleotide molecular barcode, the tagged nucleotide (the nucleotide tagged during ribose-seq from which the position of the rNMP is determined), and the sequence directly downstream from the tagged nucleotide. The UMI corresponds to sequencing cycles 1–6 and 10–11, the molecular barcode corresponds to cycles 7–9, the tagged nucleotide corresponds to cycle 12, and the tagged nucleotide's downstream sequence corresponds to cycles 13+ of the raw FASTQ sequences. The rNMP is the reverse complement of the tagged nucleotide. Before aligning the sequencing reads to the reference genome, the reads were trimmed based on sequencing quality and custom ribose-seq adaptor sequence using cutadapt 1.16 (-q 15 -m 62 -a

"AGTTGCGACACGGATCTATCA"). In addition, to ensure accurate alignment to the reference genome, reads containing fewer than 50 bases of genomic DNA (those bases located downstream from the tagged nucleotide) after trimming were discarded. Following quality control, the Alignment and Coordinate Modules of the Ribose-Map toolkit were used to process and analyze the reads30. The Alignment Module de-multiplexed the trimmed reads by the appropriate molecular barcode, aligned the reads to the reference genome using Bowtie 2 (31), and de-duplicated the aligned reads using UMI-tools. Based on the alignment results, the Coordinate Module filtered the reads to retain only those with a mapping quality score of at least 30 (probability of misalignment <0.001) and calculated the genomic coordinates and per-nucleotide counts of rNMPs. All ribose-seq libraries were then checked for background noise of restriction enzyme reads. We counted the number of reads ending with a restriction enzyme cut site, which is expected not to be generated by ribonucleotides incorporation. Some reads captured the dAMP, which is added by dA-tailing at the restriction cut site. We summed up such background reads and calculated the percentage of background noise. All mitochondrial libraries (34/34) had very low background (0.04–4.85%). Majority of the nuclear libraries (25/34) had background <12% (0.02–11.74%), and these were studied. To allow comparison between sequencing libraries of different read depth, the per-nucleotide coverage was calculated by normalizing raw rNMP counts to counts per hundred. For the emRiboSeq libraries, we downloaded libraries SRR1734967, SRR1734969, SRR1734972, SRR1734980, and SRR1734982 from NCBI's SRA using the SRA toolkit and obtained the genomic coordinates of rNMP sites using the Alignment and Coordinate Module of Ribose-Map. The FASTQ files and configuration files used as input into the Alignment and Coordinate Modules are available as Supplementary Material at NCB online.

3.2.7 Nucleotide sequence context of rNMPs

Using the Sequence Module of Ribose-Map, the frequencies of the nucleotides at rNMP sites and 100 nucleotides upstream and downstream from those sites were calculated for the nuclear and mitochondrial genomes. The Sequence Module normalizes the nucleotide frequencies to the frequencies of the corresponding reference genome. To normalize the nucleotide frequencies, the number of each type of nucleotide (A, C, G, U/T) present in the ribose-seq data was counted and divided by the total number of nucleotides at a given position to yield the raw proportion. In addition, the number of each type of nucleotide (A, C, G, T) present in the reference genome was counted and divided by the total number of nucleotides in the reference genome to yield the reference proportion. Then, the raw proportions were divided by the corresponding reference proportions to yield the normalized nucleotide frequencies.

3.2.8 Data presentation

Bar graphs representing the percentage of rNMPs were made using GraphPad Prism 8 (GraphPad Software). The nucleotide sequence context plots were created using the ggplot2 R package. Consensus sequences around rNMP sites were identified using Multiple Em for Motif Elicitation (MEME) (34) and plotted using ggseqlogo (49).

3.2.9 Statistical analysis for heatmaps

To compare frequency results of heatmap data obtained for each rNMP, or dinucleotide pair containing an rNMP, with those obtained with all other rNMPs, or dinucleotide pairs containing an rNMP, within a specific genotype of mitochondrial or nuclear libraries for a given yeast species, we used the two-sided Mann–Whitney U test. Each pair of mononucleotides, or

dinucleotide pairs containing the same rNMP, was tested to determine whether its frequency was significantly greater or smaller than that of the other samples.

3.2.10 Genome browser and hotspots

BedGraph files were generated using the Distribution Module of Ribose-Map and then visualized using the JBrowse genome browser (50). Top 1% and top 100 most abundant rNMP sites for the ribose-seq and emRiboSeq libraries were calculated based on the BED files created by the Coordinate Module of Ribose-Map. For the analysis of short-nucleotide repeat tracts, we only considered reads that were longer than the repeated regions to ensure accuracy of our findings.

3.2.11 Data availability

The ribose-seq datasets generated during the current study are available in NCBI's SRA via BioProject "PRJNA613920". The emRiboSeq datasets analyzed during the current study are available in NCBI's Gene Expression Omnibus via accession number "GSE64521."

3.2.12 Code Availability

Ribose-Map is available for download at GitHub (https://github.com/agombolay/ribose-map). Customized python3 scripts for background subtraction are available on GitHub under GPLv3.0 license (https://github.com/xph9876/ArtificialRiboseDetection). Scripts for heatmaps are available at GitHub (https://github.com/xph9876/RibosePreferenceAnalysis).

3.3 Results

3.3.1 Biased rC and rG pattern in mitochondrial DNA of wild-type *S. cerevisiae*

We built and analyzed 11 ribose-seq libraries with mitochondrial DNA of wild-type RNase H2 cells from six commonly utilized haploid yeast laboratory strains using three sets of restriction enzymes (RE1, RE2, and RE3). The percentages of r[A, C, G, and U] among these libraries were similar, regardless of the RE set used; however, we found some variation among the strains (**Figures 11 and 12**). rA, followed by rC and then rG, is the most abundant rNMP found in almost all libraries examined. Interestingly, the two ribose-seq libraries prepared from strain E134 had dominant rC and lower incorporation of rG, close to the dG count of 8.55%, compared to the libraries derived from all other strains. Consistently, rU was rarely incorporated in all libraries (**Figures 11 and 12**). Normalization of single rNMP frequencies to the nucleotide content of *S. cerevisiae* mitochondrial DNA revealed bias for rC and/or rG in all libraries over rA and especially over rU (**Figure 12**). These data in part reflect the nucleotide pool imbalance, as proposed before (18) and corroborated recently (51). By measuring rNTPs/dNTPs ratios for strains E134 and BY4742, we found that the rGTP/dGTP ratio was significantly lower in E134 compared to BY4742 ($P = 0.0015$), possibly accounting for the lower rG incorporation in E134. Furthermore, while proportions of rC and rG were similar in most libraries, rC was incorporated at a larger number of different sites than rG (**Figure 11 and Supplementary Figure 1**).

We normalized the frequency by which dNMPs upstream or downstream of the rNMPs are found next to each rNMP to the A/T (41.45%) or C/G (8.55%) content. We found that in all 11 libraries rC and in part rG are located within C/G-rich areas of mitochondrial DNA, with rC being in a dC-rich, and rG in a dG-rich area (**Figure 13**). This is a feature of the *S. cerevisiae* mitochondrial genome, in which dCMPs and dGMPs cluster together (52) (sacCer2 genome

database). Thus, despite the low percentage of dCMPs and dGMPs in mitochondrial DNA, rCs and rGs were often surrounded by dCMPs or dGMPs, respectively.

We then studied whether rA, rC, rG, and rU were randomly incorporated in mitochondrial DNA. We reasoned that, if for example rA was randomly incorporated, the frequency by which the dNMP A, C, G, or T was found at position −1 and +1 relative to rA should reflect the frequency of the dinucleotides AA, CA, GA, TA, AC, AG, and AT obtained from the sequence of *S. cerevisiae* mitochondrial DNA (standard frequency). The frequency of CrA (**Figure 14**) was above the standard value for this pair and above the frequency of the other nucleotide pairs. For rC, rG and rU, ArC, ArG, and CrU were the highest, respectively (**Figure 14**). Much less prominent difference was found among pair combinations for the dNMPs at position +1 (**Supplementary Figure 2**). We also examined the dNMPs at positions −2, +2, −3, +3, −4, and +4. Less pronouncedly than for position −1, for −2 position, we found predominant occurrences of C-rA and T-rG; for +2, rG-T; for −3, T-rG; and for −4, G-rA (**Supplementary Figure 2**). As a control, for dNMPs at position −100 or +100, the observed frequencies of dinucleotides with an rNMP matched well with those obtained from the sacCer2 mitochondrial genome (**Supplementary Figure 2**). Overall, these results, being also conserved among all of the 11 mitochondrial wild-type libraries, demonstrate that the dNMPs upstream from the rNMP, particularly the ones at position −1, have the most impact on the incorporation of a specific rNMP type in a given genomic position of yeast mitochondrial DNA.

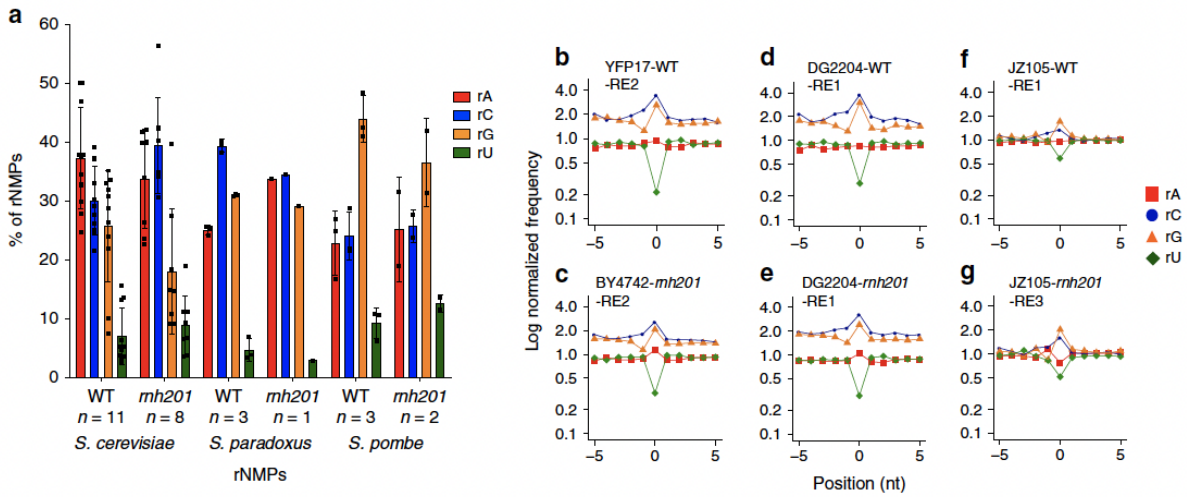
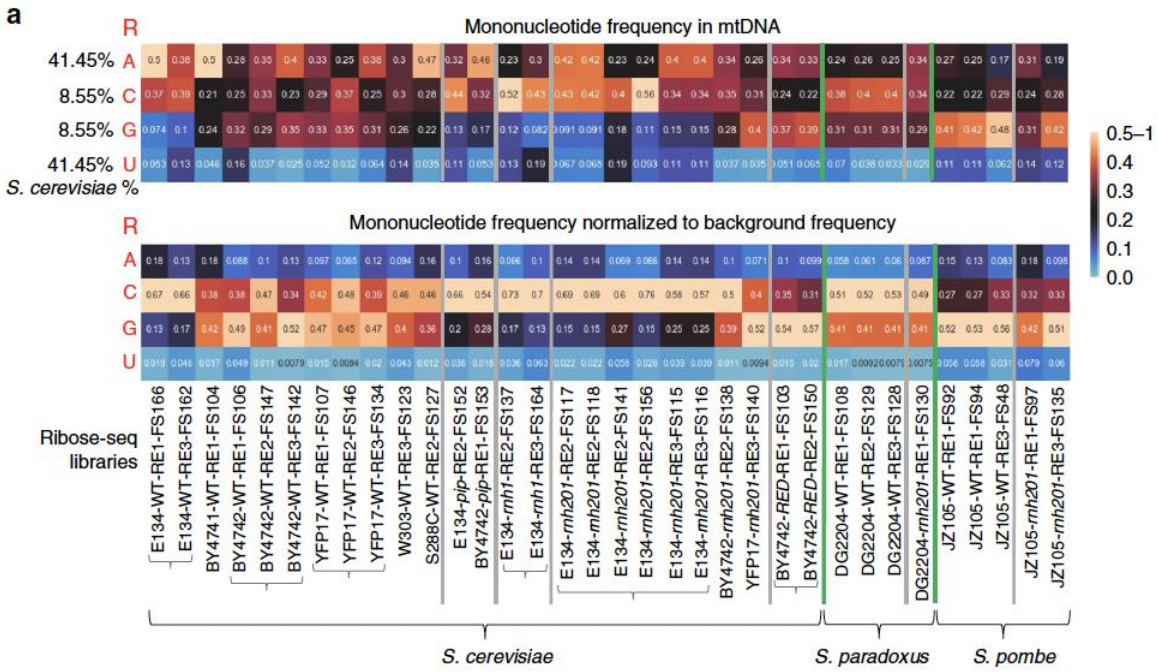


Figure 11 Identity and frequency of rNMP types in the mitochondrial yeast genome. (a) Bar graph with corresponding data points showing percentage of r[A, C, G, and U] found in mitochondrial DNA of WT or *rnh201* strains of *S. cerevisiae*, *S. paradoxus*, and *S. pombe*. Mean, standard deviation, and number of libraries (n) analyzed for each genotype are shown. (b–g) Zoomed-in plots of normalized nucleotide frequencies relative to mapped positions of sequences from mitochondrial ribose-seq libraries. Plots derived from example libraries are shown for *S. cerevisiae* (b) WT (YFP17-WT-RE2-FS146) and (c) *rnh201* (BY4742-*rnh201*-RE2-FS138), *S. paradoxus* (d) WT (DG2204-WT-RE1-FS108) and (e) *rnh201* (DG2204-*rnh201*-RE1-FS130), and *S. pombe* (f) WT (JZ105-WT-RE1-FS94) and (g) *rnh201* (JZ105-*rnh201*-RE3-FS135). Position 0 is the rNMP, − and + positions are upstream and downstream dNMPs, respectively, normalized to the A, C, G, and T content in the corresponding yeast species.



b

$$R_{N_{R,\text{raw}}} = \frac{R_N}{R_A + R_C + R_G + R_U}$$

$$P_{RN} = \frac{R_N}{N_N} \qquad R_{N_{N,\text{norm}}} = \frac{P_{R_N}}{P_{R_A} + P_{R_C} + P_{R_G} + P_{R_U}}$$

Figure 12 Heatmap analyses of rNMPs in yeast mitochondrial DNA. (a) Heatmap analyses with (top) frequency of each type of rNMP (rA, rC, rG, and rU), and (bottom) frequency of each type of rNMP normalized to the nucleotide frequencies of the corresponding reference genome for all the mitochondrial riboseseq libraries of this study. The corresponding formulas used are shown in b and explained in Methods. Each column of the heatmap shows results of a specific ribose-seq library. Each library name is indicated underneath each column of the heatmap with its corresponding strain name, genotype, and restriction enzyme (RE) set used. The yeast species of the ribose-seq libraries are also indicated. *S. cerevisiae* libraries derived from the same strains are grouped together by curly brackets. Thick, vertical, green lines separate data from the different yeast species. Vertical gray lines separate data obtained from different RNase H genotypes within each species. Each row shows results obtained for an rNMP (R in red) of base A, C, G, or U for each library. The actual percentages of A, C, G, and T bases present in mitochondrial DNA of *S. cerevisiae* are shown to the left of the top heatmap. The bar to the right shows how different frequency values are represented as different colors: black for 0.25; black to yellow for 0.25 to 0.5–1, and black to light blue for 0.25 to 0. (b) Formulas used to calculate the frequency and the normalized frequency values of the mitochondrial heatmaps in (a).

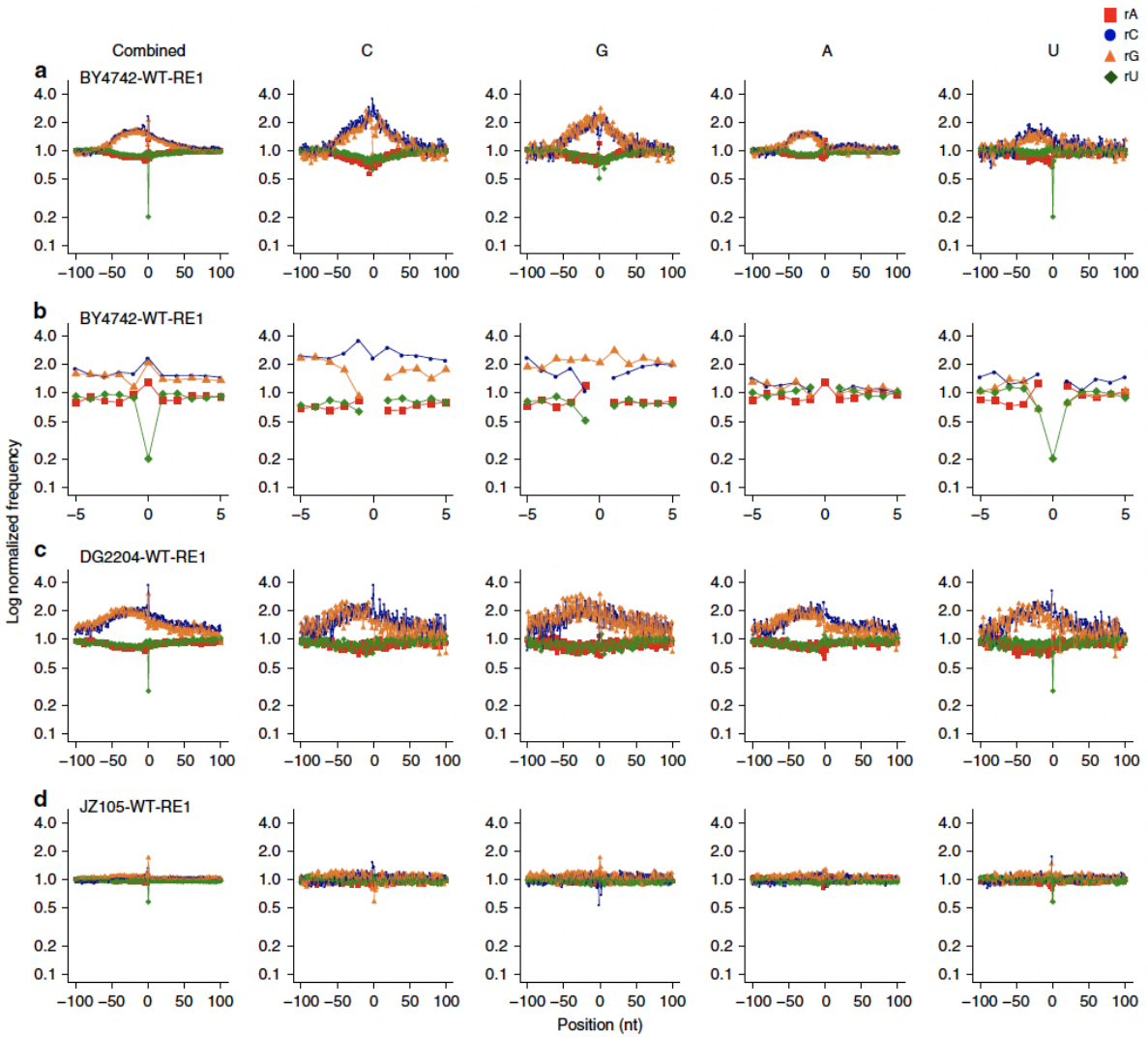


Figure 13 Sequence context of rNMPs in wild-type yeast mitochondrial DNA. Shown are examples of combined and single rNMP plots of normalized nucleotide frequencies relative to mapped positions of rNMPs from examples of mitochondrial ribose-seq libraries. (a) Zoom-out and (b) zoom-in plots for mitochondrial library BY4742-WT-RE1-FS104 of *S. cerevisiae* WT cells. (c) Zoom-out plots for mitochondrial library DG2204-WT-RE1-FS108 of *S. paradoxus* WT cells. (d) Zoom-out plots for mitochondrial library JZ105-WT-RE1-FS94 of S. pombe WT cells. Position 0 on the x-axis represents the site of rNMP incorporation, − and + positions represent upstream and downstream dNMPs, respectively. The y-axis shows the frequency of each type of nucleotide present in the ribose-seq data normalized to the frequency of the corresponding nucleotide present in the reference genome of the indicted yeast species.

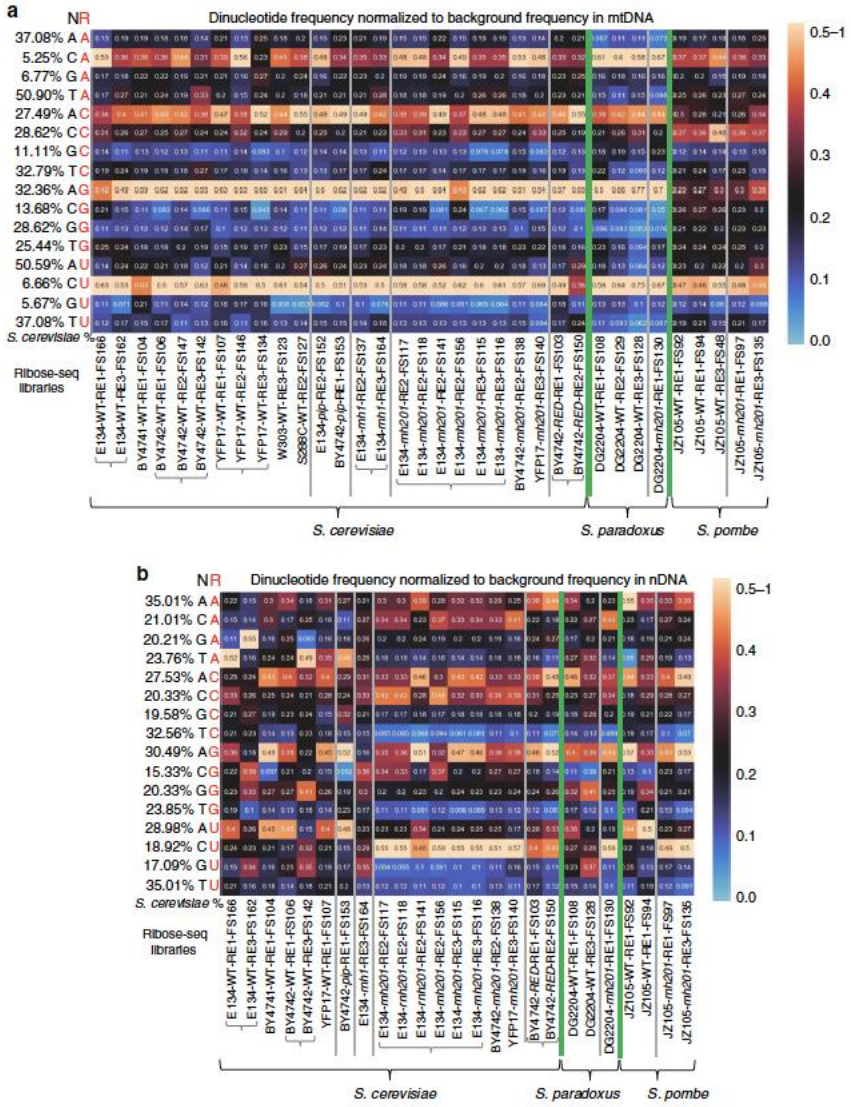


Figure 14 dNMP directly upstream from rNMP affects frequency of rNMP incorporation. Heatmap analyses with normalized frequency of a mitochondrial and b nuclear NR dinucleotides (rA, rC, rG, and rU with the upstream deoxyribonucleotide with base) for all the ribose-seq libraries of this study. Each column of the heatmap shows results of a specific library. Each library name is indicated underneath each column of the heatmap with its corresponding strain name, genotype, and RE set used. The yeast species are also indicated. *S. cerevisiae* libraries derived from the same strains are grouped together by brackets. Thick, vertical, green lines separate data from the different yeast species. Vertical gray lines separate data obtained from different RNase H genotypes within each species. Each row shows results obtained for a dinucleotide NR (R in red) of fixed rNMPs A, C, G, or U for each library. The actual percentages of dinucleotides of fixed base A, C, G, or T for the indicated base combinations present in mitochondrial and nuclear DNA of *S. cerevisiae* are shown to the left of the heatmaps. The observed % of dinucleotides with rNMPs A, C, G, or U were divided by the actual % of each dinucleotide with fixed base A, C, G, or T in mitochondrial or nuclear DNA of the corresponding species.

3.3.2 rNMP patterns in mitochondrial DNA of *S. cerevisiae* RNase H-mutant cells

Recent records revealed lack of activity for RNase H2 in yeast mitochondrial DNA (51,53,54). Here, we constructed a null mutation of the catalytic subunit of RNase H2, and the ribonucleotide-excision defective (RED) mutations in the same subunit, impairing RER, but not RNase H2 cleavage at long RNA-DNA hybrids (39). We built eight ribose-seq libraries of *S. cerevisiae* mitochondrial DNA derived from *rnh201* cells of strains E134, BY4742, and YFP17, and two libraries with the RED mutations derived from BY4742. As expected, the mitochondrial *rnh201* libraries had rNMP content similar to the mitochondrial wild-type libraries for the same strains (**Figures 11 and 12; Supplementary Figure 1**). The two mitochondrial libraries derived from the RED cells were also similar to wild type and *rnh201* libraries except for slightly increased level of rG (**Figure 12**). We further deleted the RNH1 gene in E134 and constructed two independent ribose-seq libraries. Data analysis from the *rnh1* libraries showed rNMP content similar to that of mitochondrial wild type and *rnh201* cells of E134 (**Figure 12 and Supplementary Figure 1**). Consistently with our results in wild-type RNase H2 cells, we found that CrA, ArC, ArG, and CrU were significantly above the standard count for the respective dinucleotides in mitochondrial DNA and significantly above the other dinucleotide pair combinations in all *rnh201*, RED, and *rnh1* cells (**Figure 14 and Supplementary Figure 2**). Moreover, to ensure that the results of the dinucleotides were not artifacts of the ribose-seq technique, using Ribose-Map (28), we analyzed data from five libraries of *S. cerevisiae rnh201* cells prepared with the emRiboSeq technique, which does not employ restriction enzymes, alkali, and AtRNL, but sonication, human recombinant RNase H2, and T4-quick ligase, respectively (19). We generated heatmaps for the mononucleotides and examined the dNMPs at positions −1, +1, −2, and +2 relative to the rNMPs in these libraries (**Supplementary Figure 3**). The similarities

between these emRiboSeq and the ribose-seq results are remarkable. This comparison further supports our findings that the dNMP upstream of the rNMP has the most impact on the incorporation of a specific rNMP type in yeast mitochondrial DNA.

3.3.3 Patterns of rNMPs in *S. paradoxus* and *S. pombe* mitochondrial DNA

To determine whether the patterns of rNMP incorporation in mitochondrial DNA vary in different yeast species, we built ribose-seq libraries from mitochondrial DNA of wild type and *rnh201 S. paradoxus*. We analyzed three ribose-seq libraries with mitochondrial DNA of wild type and one from *rnh201* S. paradoxus of strain DG2204, as well as three ribose-seq libraries with mitochondrial DNA of wild type and two from *rnh201 S. pombe* of strain JZ105 (**Supplementary Table 1)**. The rNMP patterns in mitochondrial DNA of *S. paradoxus* were comparable to that of *S. cerevisiae*, being quite similar to strains YFP17, W303, and S288C, displaying a preference for rC and low incorporation of rU (**Figures 11 and 12**). *S. pombe* mitochondrial DNA had higher presence of rG, both in wild type and *rnh201* cells, while still displaying low rU (**Figures 11 and 12**). Like in *S. cerevisiae* cells, the rUTP/dTTP ratio was the lowest among rNTPs/dNTPs ratios in *S. pombe* strain JZ105 (Supplementary Fig. 1). We did not detect major differences in the rNMP frequencies between wild type and *rnh201* mitochondrial libraries in either *S. paradoxus* or *S. pombe* (**Figure 12**).

While rNMPs were located within C/G-rich areas of *S. paradoxus* mitochondrial DNA, similarly to *S. cerevisiae*, this was not the case for rNMPs in *S. pombe* mitochondrial DNA (**Figure 13 and Supplementary Figure 1**), highlighting a unique feature of mitochondrial rNMPs of budding yeasts. Markedly, like in *S. cerevisiae*, we found that the dNMP upstream from the rNMP had the most impact on rNMP incorporation both in *S. paradoxus* and *S. pombe*. For *S. paradoxus*,

we found that, as in *S. cerevisiae*, CrA, ArC, ArG, and CrU were the most abundant pairs. *S. pombe* mitochondrial DNA showed preference for CrA and CrU, and no preference for any dNMP following the rNMP, like *S. cerevisiae*. Differently from budding yeasts, in *S. pombe* wild-type libraries, rC showed preference for CrC (**Figure 14 and Supplementary Figure 2**).

3.3.4 Wild-type *S. cerevisiae* nDNA has low rG and high rC

We built six ribose-seq libraries of nuclear DNA from wild-type RNase H2 cells of strains E134, BY4741, BY4742, and YFP17, eight libraries from nuclear DNA of *rnh201* cells from strains E134, BY4742, and YFP17, two libraries from RED cells of strain BY4742, one from *rnh1* cells of strain E134, and one from *rnh202*-pip cells of strain BY4742. The rNMP content and distribution in nuclear DNA were different from those in mitochondrial DNA of the same yeast strains. In wild-type cells, rG was consistently the least abundant rNMP in all libraries. Normalization of single rNMP frequencies to the nucleotide-base content of *S. cerevisiae* nuclear DNA revealed bias for rC (**Figures 15 and 16**). While rA and rU were incorporated proportionally to the abundance of dA and dT in *S. cerevisiae* nuclear DNA, in all of these eight libraries, rC occurred above the 19.00% C count, and its normalized content was significantly greater than that of the other rNMPs (**Figures 15 and 16; Supplementary Figure 4**).

Due to the activity of RNase H2 on nuclear DNA, the nuclear rNMP content in *rnh201* cells was distinct from that in wild-type cells. In nuclear DNA of rnh201 cells, rC was by far the most abundant rNMP. On average, over 50% of rNMPs were rCs. The rA fraction dropped substantially, and rU was the least abundant (**Figures 15 and 16; Supplementary Figure 4**). The data were consistent among the different strains. The two RED libraries had increased level of rG (**Figure 16; Supplementary Figure 4**).

In vitro studies have shown that the interaction with PCNA, via the PCNA interacting peptide domain (PIP-box), enhances cleavage of misincorporated rNMPs by the archaeal RNase HII but not the yeast or human RNase H2 (39,55). In line with these results, we found that in *rnh202*-pip mutant cells of *S. cerevisiae*, the nuclear rNMP content remained similar to that obtained in wildtype cells (**Figure 16 and Supplementary Figure 4**). Moreover, as observed in mitochondria, in *rnh1* cells, the rNMP content in nuclear DNA was also similar to that in wildtype cells (**Figure 16 and Supplementary Figure 4**). These results suggest that Rnh1 does not have a strong impact on rNMP removal from nuclear DNA.

We then examined whether rA, rC, rG, and rU were randomly incorporated in nuclear DNA of the wild type and mutant RNase H strains by determining the frequency of the dNMPs preceding or following each rNMP and comparing this frequency with the given frequency of each dinucleotide in *S. cerevisiae* nuclear DNA. For rNMPs in nuclear DNA of wild type, *rnh202*-pip, and *rnh1* cells, the dinucleotide pattern was less evident (**Figure 14**) than that observed in mitochondrial DNA (**Figure 14**), likely due to the lower number of rNMPs detected. Nevertheless, we found that ArC, ArG, and ArU pairs were dominant, with ArC being significantly different from GrC and TrC (**Figure 14**). This effect is more evident when results obtained for the −1 are compared with those for the +1 position, and with those for the −2, +2, −3, +3, −4, +4, −100, and +100 positions, in which no particular pair was dominant across these different libraries (**Supplementary Figure 5**). Nuclear *rnh201* and RED ribose-seq libraries showed frequencies of ArG and CrU greater than the other dinucleotide pairs for rG and rU, respectively (**Figure 14**). However, differently from mitochondrial DNA, we found low frequency of dT upstream of any rNMP (**Figure 14**). Moreover, while in mitochondrial DNA CrA and ArC also stood out among the dNMPs upstream of rA and rC, respectively, these two pairs were abundant but not dominant

in *rnh201* and RED libraries, showing frequencies similar to ArA and CrC, respectively (**Figure 14**). These results show that rNMP incorporation in nuclear DNA of mutant RNase H cells is not random. We did not detect major preference for the +1, −2, +2, −3, +3, −4, and +4 dNMP in *rnh201* and RED libraries (**Supplementary Figure 5**). Analysis of the five *rnh201* emRiboSeq libraries, which are all from the same strain background (Δl(−2)l-7BYUNI300), revealed common trends in the sequence context of rNMPs observed in the ribose-seq libraries of the same genotype. In particular, the strongest biases were seen for the −1 dNMP (**Supplementary Figure 3**). Therefore, more strictly than what we found in mitochondrial libraries, the dNMP immediately upstream from the rNMP rather than the one downstream or further upstream or downstream has the most impact on the incorporation of a specific rNMP type in the nuclear DNA of both wild type and RNase H-mutant *S. cerevisiae* cells.

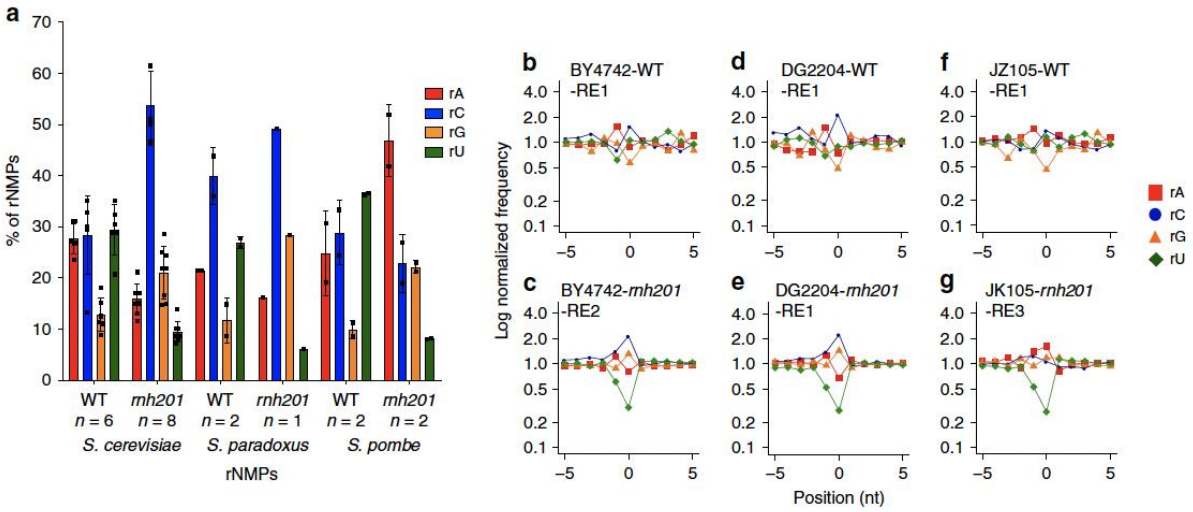
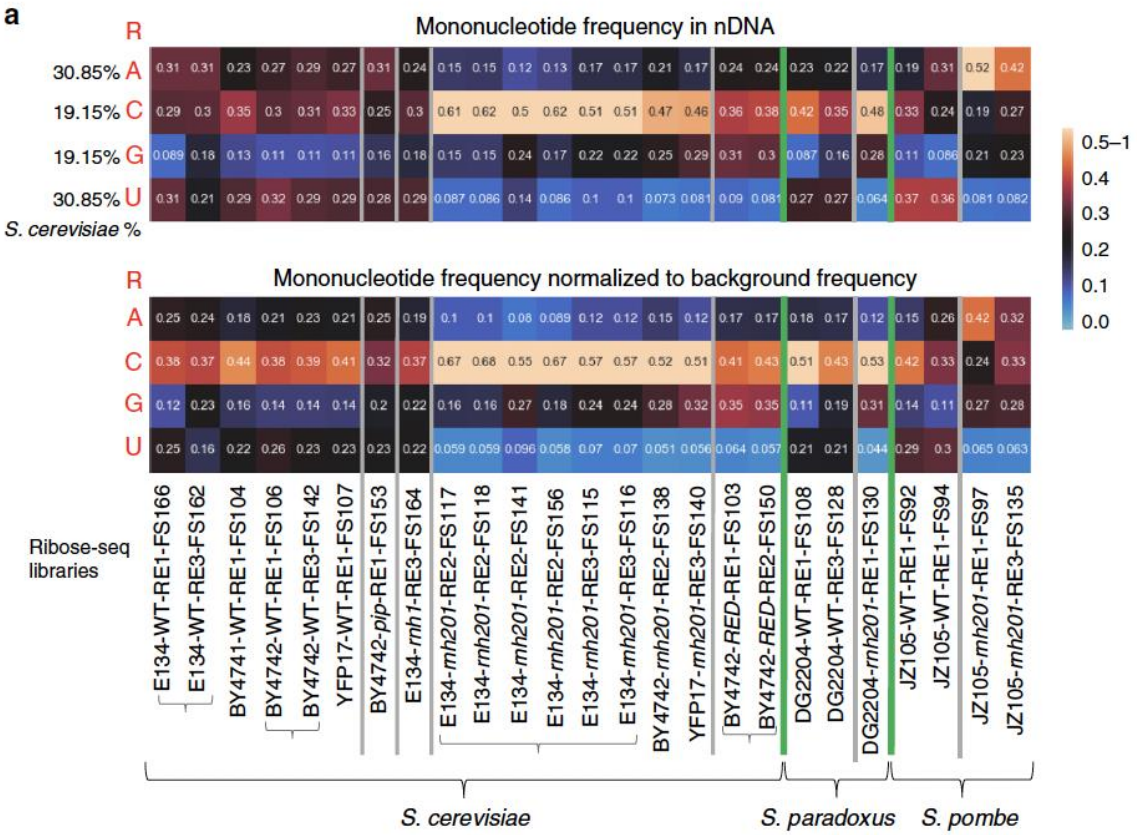


Figure 15 Identity and frequency of rNMP types in the nuclear yeast genome. (a) Bar graph with corresponding data points showing percentage of rA, rC, rG, and rU found in nDNA of WT or *rnh201* strains of *S. cerevisiae*, *S. paradoxus*, and *S. pombe*. Mean, standard deviation, and number of libraries analyzed for each genotype are shown. (b–g) Zoomed-in plots of normalized nucleotide frequencies relative to mapped positions of sequences from nuclear ribose-seq libraries. Plots derived from example libraries are shown for *S. cerevisiae* (b) WT (BY4742-WT-RE1-FS106) and (c) *rnh201*Δ (BY4742-*rnh201*-RE2-FS138), *S. paradoxus* (d) WT (DG2204-WT-RE1-FS108) and (e) *rnh201* (DG2204-*rnh201*-RE1-FS130), and *S. pombe* (f) WT (JZ105-WT-RE1-FS94) and (g) *rnh201* (JZ105-*rnh201*-RE3-FS135). Position 0 on the x-axis represents the site of rNMP incorporation, − and + positions represent upstream and downstream dNMPs, respectively. The y-axis shows the frequency of each type of nucleotide present in the data normalized to the frequency of the corresponding nucleotide present in the reference genome.



b

$$R_{N_{R,\text{raw}}} = \frac{R_N}{R_A + R_C + R_G + R_U}$$

$$P_{R_N} = \frac{R_N}{N_N} \qquad R_{R_N,\text{norm}} = \frac{P_{R_N}}{P_{R_A} + P_{R_C} + P_{R_G} + P_{R_U}}$$

Figure 16 Heatmap analyses of rNMPs in yeast nuclear DNA. (a) Heatmap analyses with (top) frequency of each type of rNMP (rA, rC, rG, and rU), and (bottom) frequency of each type of rNMP normalized to the nucleotide frequencies of the corresponding reference genome for all the nucleus ribose-seq libraries of this study. The corresponding formulas used are shown in (b). Each column of the heatmap shows results of a specific ribose-seq library. Each library name is indicated underneath each column of the heatmap with its corresponding strain name, genotype, and RE set used. The yeast species of the ribose-seq libraries are also indicated. *S. cerevisiae* libraries derived from the same strains are grouped together by curly brackets. Thick, vertical, green lines separate data from the different yeast species. Vertical gray lines separate data obtained from different RNase H genotypes within each species. Each row shows results obtained for an rNMP (R in red) of base A, C, G, or U for each library. The actual percentages of A, C, G, and T bases present in nDNA of *S. cerevisiae* are shown to the left of the top heatmap. (b) Formulas used to calculate the frequency and the normalized frequency values of the nuclear heatmaps in (a).

3.3.5 Patterns of rNMPs in *S. paradoxus* and *S. pombe* nuclear DNA

We analyzed two ribose-seq libraries from nuclear DNA of wild type and one of *rnh201 S. paradoxus* cells of strain DG2204, as well as two ribose-seq libraries from nuclear DNA of wild type and two of *rnh201 S. pombe* cells of strain JZ105. The patterns of rNMP incorporation in nuclear DNA of wild type and *rnh201 S. paradoxus* were similar to those of the corresponding genotypes of *S. cerevisiae*, with rG the lowest and rC the dominant rNMP, similar frequency of rU and rA in wild type, and high rC, low rA, and very low rU in *rnh201* cells. For wild-type *S. pombe* cells, like for *S. cerevisiae* and *S. paradoxus*, rC was the highest and rG the lowest. In *rnh201* cells of *S. pombe*, while rU was still the lowest, rA was the most frequent rNMP (**Figures 15 and 16; Supplementary Figure 4**). Interestingly, the rATP/dATP ratio was high in *S. pombe* compared to *S. cerevisiae* cells, possibly supporting elevated incorporation of rA in the absence of RNase H2 function in nuclear DNA.

Data analysis of nuclear DNA libraries of *S. paradoxus*, generally in line with results from *S. cerevisiae*, showed ArC and ArG as higher pairs in wild type, and mainly CrA, ArC, CrC, ArG, and CrU above the standard frequency in *rnh201* cells (**Figure 14**). For *S. pombe*, together with ArC, ArG, and ArU, also ArA was above the standard value in wild-type cells, while in *rnh201* cells, ArA, ArC, ArG, and CrU were the highest (**Figure 14**). These findings show that rNMPs are not randomly incorporated in *S. paradoxus* and *S. pombe* nuclear DNA and that factors beyond variation in nucleotide pools affect distribution and patterns of rNMP incorporation in nuclear DNA. Although there is some variability for wild-type libraries, likely due to the lower number of detected rNMPs, the frequencies of dNMPs at positions +1, −2, +2, −3, +3, −4, and +4 for nuclear DNA of wild type and *rnh201* cells of *S. paradoxus* and *S. pombe* did not deviate much from the dinucleotide frequencies found in the nuclear DNA of these yeast species (Supplementary Figure

5). Overall, the results reveal substantial similarity in the dinucleotide patterns for nuclear DNA among yeast *S. cerevisiae*, *S. paradoxus*, and *S. pombe*.

3.3.6 Hotspots of rNMPs occur at ArC or ArG sites in yeast DNA

The ribose-seq data analysis revealed rNMP sites that were in common among libraries and preferred rNMP sites in each library. We generated a list of rNMP sites that were shared among all the different mitochondrial or nuclear libraries of wild type or *rnh201 S. cerevisiae* cells. Interestingly, the most abundant and shared rNMP sites in the mitochondrial libraries were found on the Crick (−) strand, which in most cases corresponds to the template strand for transcription. Moreover, the majority of these common sites were ArC or ArG in mitochondrial wild type (8/11 shared sites) and *rnh201* (15 of the top 25) libraries. No shared sites were found among nuclear libraries of wild-type *S. cerevisiae*. Among the shared sites in the nuclear DNA of *rnh201* cells, rC was dominant and, in the majority of cases, preceded by dA (17 of the top 25).

To determine whether there were overlapping features and specific signatures among rNMPs that were most frequently incorporated in mitochondrial DNA and nuclear DNA, we selected the top 1% most abundant rNMP sites from each mitochondrial and nuclear library and analyzed them using MEME (34). Because the number of rNMPs at a particular site can vary depending on sequencing coverage, calculating the top 1% allowed us to compare the frequency of rNMPs at each site among each library independently of the sequencing coverage of the libraries. For comparison, we also selected the top 100 most abundant rNMP sites. The results of both analyses revealed specific consensus motifs of rNMP incorporation for these hotspot sites in mitochondrial DNA (**Figure 17 and Supplementary Figure 6**) and in nuclear DNA for *rnh201* libraries (**Figure 17 and Supplementary Figure 6**). We found rG followed by rC to be the most

prevalent in all mitochondrial libraries except for those from strain E134, which had rC followed by rA. Mitochondrial libraries from strains with *rnh202*-pip, *rnh1*, *rnh201*, and *rnh201*-RED mutations had similar rNMP preference as wild-type cells of the same strains. For hotspots in mitochondrial libraries of *S. paradoxus*, all had rC/rG as most abundant rNMP, while for *S. pombe*, rG was dominant followed by rC. Strikingly, dA was dominant at the −1 position in all hotspot motifs for all mitochondrial libraries of *S. cerevisiae*, *S. paradoxus*, and *S. pombe*, as well as in the *S. cerevisiae* emRiboSeq libraries (**Supplementary Figure 7**). The MEME analysis covered three nucleotides upstream and three downstream from the rNMP site. Although less pronounced, dT was conserved at position −3 and +2 in most of the mitochondrial *S. cerevisiae* libraries, including the emRiboSeq libraries, as well as in those of *S. paradoxus* and *S. pombe*. The consensus motif that emerged from all the hotspot sites of the yeast mitochondrial libraries is TNArSWTW (**Figure 17**). Analysis of nuclear DNA data from *rnh201* libraries revealed ArC as the dominant motif in all *rnh201* libraries of *S. cerevisiae*, including the S. paradoxus *rnh201* library. *S. pombe* showed ArA as the dominant motif in *rnh201* cells. No particular nucleotide further upstream or downstream from rC was conserved in the nuclear DNA of these hotspot sites (**Figure 17**).

3.3.7 Patterns of rNMPs in short-nucleotide repeats

Concentration of rNMPs at short-nucleotide repeated tracts is a feature that emerged by browsing the genomic data of the ribose-seq libraries. We found a series of trinucleotide-repeated sequences that contain abundant rNMPs with a specific pattern. These are not artifacts of PCR because all reads in a given short-nucleotide repeated tract have different UMI, not just those reads that are at the same position within the repeated tract. In *S. cerevisiae*, for example, region chrXI:576134..576175 on the reverse strand for nuclear library FS115 (E134 *rnh201*) displayed

multiple hotspots of rNMP incorporation at the C-nucleotide in the triplet GAC (**Figure 17**). The same pattern was reproducible at the same locus in FS116 (E134 *rnh201*) and FS140 (YFP17 *rnh201*). Another interesting example is locus chrM:63583..63651 in FS156, also seen in FS141, FS138, FS162, and FS166, that has rNMPs at the G-nucleotide of TAAGTA-repeated sequence on the forward strand and at the C-nucleotide on the reverse strand in TACTTA-repeated sequence (**Figure 17**). Interestingly, this pattern was also evident in emRiboSeq libraries SRR1734980 (**Supplementary Figure 7**) and SRR173982. A series of similar patterns with ArC in trinucleotide, dinucleotide, or short-nucleotide repeat tracts were found in other loci in the *S. cerevisiae* genome (**Supplementary Figure 8**).

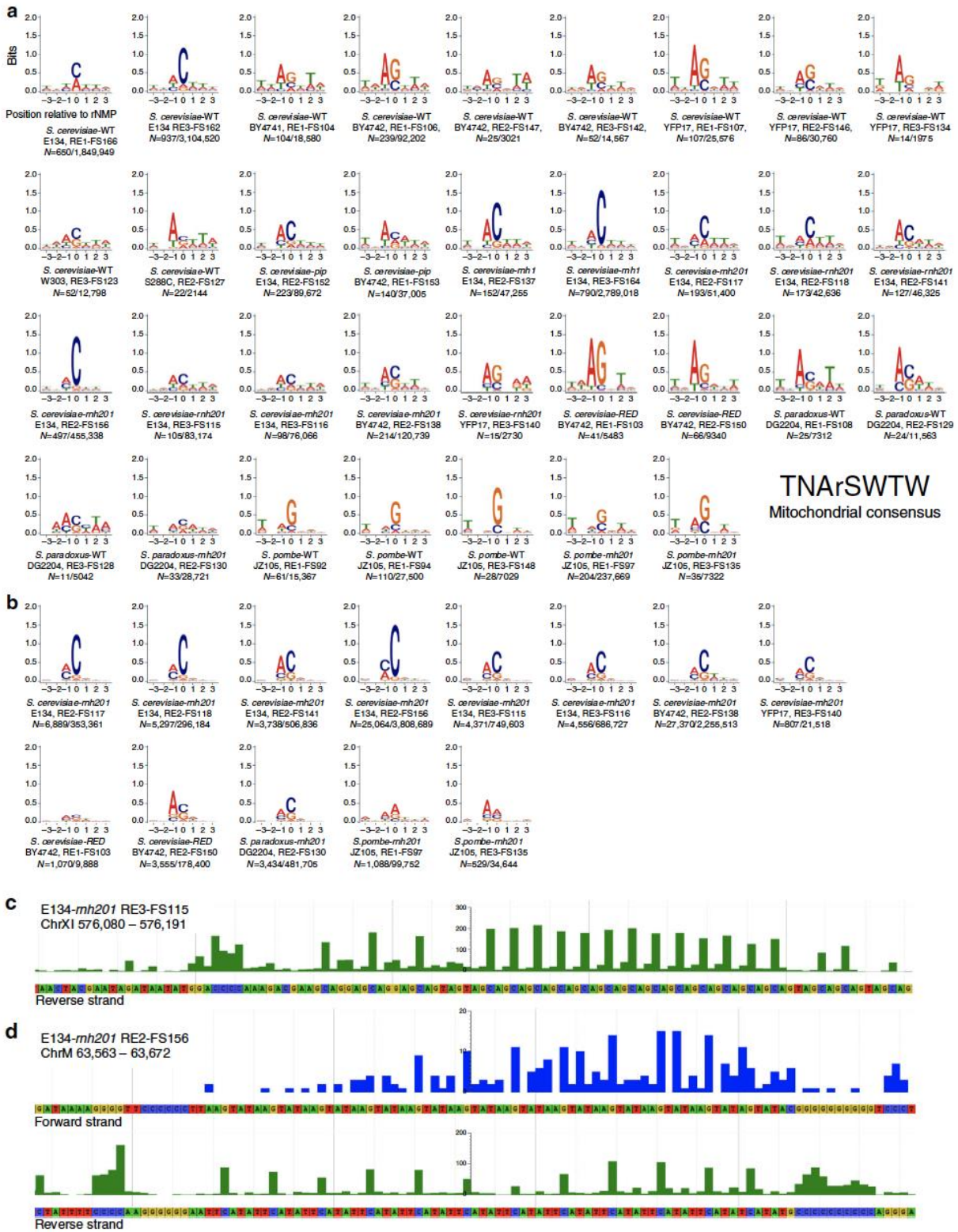

a
Bits
Position relative to rNMP
S. cerevisiae-WT
E134, RE1-FS166
N=650/1,849,949
S. cerevisiae-WT
E134 RE3-FS162
N=937/3,104,520
S. cerevisiae-WT
BY4741, RE1-FS104
N=104/18,580
S. cerevisiae-WT
BY4742, RE1-FS106,
N=239/92,202
S. cerevisiae-WT
BY4742, RE2-FS147,
N=25/3021
S. cerevisiae-WT
BY4742, RE3-FS142,
N=52/14,567
S. cerevisiae-WT
YFP17, RE1-FS107,
N=107/25,576
S. cerevisiae-WT
YFP17, RE2-FS146,
N=86/30,760
S. cerevisiae-WT
YFP17, RE3-FS134
N=14/1975
S. cerevisiae-WT
W303, RE3-FS123
N=52/12,798
S. cerevisiae-WT
S288C, RE2-FS127
N=22/2144
S. cerevisiae-pip
E134, RE2-FS152
N=223/89,672
S. cerevisiae-pip
BY4742, RE1-FS153
N=140/37,005
S. cerevisiae-mh1
E134, RE2-FS137
N=152/47,255
S. cerevisiae-mh1
E134, RE3-FS164
N=790/2,789,018
S. cerevisiae-mh201
E134, RE2-FS117
N=193/51,400
S. cerevisiae-rnh201
E134, RE2-FS118
N=173/42,636
S. cerevisiae-rnh201
E134, RE2-FS141
N=127/46,325
S. cerevisiae-mh201
E134, RE2-FS156
N=497/455,338
S. cerevisiae-rnh201
E134, RE3-FS115
N=105/83,174
S. cerevisiae-mh201
E134, RE3-FS116
N=98/76,066
S. cerevisiae-mh201
BY4742, RE2-FS138
N=214/120,739
S. cerevisiae-rnh201
YFP17, RE3-FS140
N=15/2730
S. cerevisiae-RED
BY4742, RE1-FS103
N=41/5483
S. cerevisiae-RED
BY4742, RE2-FS150
N=66/9340
S. paradoxus-WT
DG2204, RE1-FS108
N=25/7312
S. paradoxus-WT
DG2204, RE2-FS129
N=24/11,563
S. paradoxus-WT
DG2204, RE3-FS128
N=11/5042
S. paradoxus-mh201
DG2204, RE2-FS130
N=33/28,721
S. pombe-WT
JZ105, RE1-FS92
N=61/15,367
S. pombe-WT
JZ105, RE1-FS94
N=110/27,500
S. pombe-WT
JZ105, RE3-FS148
N=28/7029
S. pombe-mh201
JZ105, RE1-FS97
N=204/237,669
S. pombe-mh201
JZ105, RE3-FS135
N=35/7322
TNArSWTW
Mitochondrial consensus
b
S. cerevisiae-mh201
E134, RE2-FS117
N=6,889/353,361
S. cerevisiae-mh201
E134, RE2-FS118
N=5,297/296,184
S. cerevisiae-mh201
E134, RE2-FS141
N=3,738/506,836
S. cerevisiae-mh201
E134, RE2-FS156
N=25,064/3,808,689
S. cerevisiae-mh201
E134, RE3-FS115
N=4,371/749,603
S. cerevisiae-mh201
E134, RE3-FS116
N=4,556/686,727
S. cerevisiae-mh201
BY4742, RE2-FS138
N=27,370/2,255,513
S. cerevisiae-mh201
YFP17, RE3-FS140
N=807/21,518
S. cerevisiae-RED
BY4742, RE1-FS103
N=1,070/9,888
S. cerevisiae-RED
BY4742, RE2-FS150
N=3,555/178,400
S. paradoxus-mh201
DG2204, RE2-FS130
N=3,434/481,705
S.pombe-mh201
JZ105, RE1-FS97
N=1,088/99,752
S.pombe-mh201
JZ105, RE3-FS135
N=529/34,644
c
E134-mh201 RE3-FS115
ChrXI 576,080 – 576,191
Reverse strand
d
E134-mh201 RE2-FS156
ChrM 63,563 – 63,672
Forward strand
Reverse strand

Figure 17 Hotspot motifs with rNMPs in mitochondrial DNA and nuclear DNA. (a) Sequence motif plots for mitochondrial and (b) nuclear *rnh201* hotspots (top 1% of rNMP sites). Position 0 on the x-axis represents the site of rNMP incorporation, − and + positions represent upstream and downstream dNMPs, respectively. The y-axis shows the level of sequence conservation, represented in bits. The species, genotype, strain, restriction enzyme (RE) set, library name, and number of rNMP sites are included below each plot. The consensus sequence for mitochondrial hotspots is shown: N any nucleotide, W weak (A or T), and S strong (G or C). (c) Genome browser snapshot showing an rC hotspot within the GAC-repeated sequence in *S. cerevisiae* nDNA at the locus chrXI:576134..576175 for ribose-seq library FS115 (*rnh201*). (d) Genome browser snapshot showing an rG hotspot within the TAAGTA-repeated sequence on the forward strand, and an rC hotspot in the complementary TACTTA-repeated sequence on the reverse strand in *S. cerevisiae* mitochondrial DNA at the locus chrM:63583..63651 for ribose-seq library FS156 (*rnh201*). All reads shown here have distinct UMI, meaning they do not represent PCR duplicates caused by slippage of DNA polymerase. Similarly, reads for other hotspot sites within short-nucleotide repeats are shown in **Supplementary Figure 8**.

3.4 Discussion

We report a genome-wide analysis of rNMP sites in mitochondrial DNA and nuclear DNA of wild type and RNase H mutants of different strains of *S. cerevisiae*, *S. paradoxus*, and *S. pombe*. Our results, observed consistently across multiple genotypes and/or strains, show that rNMPs are found at preferential sites in DNA, rather than being randomly distributed along the mitochondrial and nuclear genomes. A remarkable feature across *S. cerevisiae* and *S. paradoxus* and in most part conserved in *S. pombe* is that the deoxyribonucleotide immediately upstream from the rNMPs embedded in DNA is found to have the strongest impact on rNMP distribution in DNA, compared to the nucleotide immediately downstream or those further up or downstream. This feature may reflect an accommodation mechanism of yeast replicative DNA polymerases that is favored by specific sequence contexts. Exonucleolytic proofreading activity on rNMPs by DNA polymerases (Pols) is absent or weak. No proofreading was detected for human mitochondrial Pol γ (56) or yeast Pol δ, while only limited proofreading activity was shown for Pol ε on rU (4,12,18). Studies using Xray crystal structures of human DNA polymerase μ showed that the enzyme can

accommodate rNMPs almost as well as dNMPs in its active site, with no alteration of protein structure (57). While yeast nuclear Pol δ and ε, and mitochondrial Pol γ have stronger sugar discrimination capacity than Pol μ (3,58), it is possible that specific sequence contexts provide for sufficient accommodation of the rNMPs in the polymerase active site, allowing rNMP incorporation in DNA by these replicative enzymes.

We found that the overall distribution and dinucleotide patterns of rNMPs in *S. cerevisiae* and *S. paradoxus* are similar. The rNMP patterns in *S. pombe* have more similarity with the budding yeasts at the nuclear than at the mitochondrial level. Mitochondrial rC and rG were found in C/G-rich regions in the budding yeasts but not in *S. pombe* (**Figure 13**). rC and rG were often found downstream of dA in nuclear DNA of all the three species, while this was evident only in mitochondrial DNA of the budding yeasts. In fact, the dinucleotide patterns show clear similarities at the nuclear level across the three yeast species (**Figure 14**). A major common feature across the three yeast species studied here is the strong similarity of the rNMP patterns of mitochondrial DNA between wild type and *rnh201* cells of the same species and strains. Not only did we confirm the lack of activity of RNase H2 in *S. cerevisiae* mitochondria, but we also found that this applies to mitochondrial DNA of *S. paradoxus* and *S. pombe*. Interestingly, despite many conserved features of rNMP patterns across the yeast species, we revealed variation among yeast strains of the same species. Particularly, within the six different strains of *S. cerevisiae* studied, strain E134 displayed marked preference for rC vs. rG in mitochondrial DNA (**Figure 12**). Although different yeast strains may have DNA sequence polymorphisms, it is unlikely these prominently alter the A, C, G, and T content of the genome. Lower ratio for rGTP/dGTP in E134 compared to BY4742 strain may account for such preference. We found more consistent rNMP distribution in nuclear DNA among all the strains for wild-type RNase H2 cells of *S. cerevisiae*, and similarly for *rnh201*

cells (**Figure 16**). Another common feature among the three yeast species is the underrepresentation of rU in mitochondrial DNA of any genotype, and in nuclear DNA of *rnh201* cells. While the low level of rU can be explained by general high concentration of dTTP in the nucleotide pools (4,51) with rUTP/dTTP being the lowest rNTP/dNTP ratio, which we observed both in *S. cerevisiae* and *S. pombe* strains, potential activity of topoisomerase I on sequences with rU (16,59), and some proofreading activity for Pol ε on rU (18), its incorporation in mitochondrial DNA and nuclear DNA of *rnh201* cells is not random. In fact, we show that rU is found in most cases after dC, and this feature is highly conserved across *S. cerevisiae*, *S. paradoxus*, and *S. pombe* (**Figure 14**).

Wild-type RNase H2 cells have active RER to remove incorporated rNMPs from nuclear DNA. Therefore, we expected to detect transiently incorporated rNMPs, not yet removed by RER, or rNMPs that escaped RER removal. Interestingly, we found rG to be present below the proportion of dGMP in all nuclear wild-type libraries of *S. cerevisiae*, *S. paradoxus*, and *S. pombe*. This was also noted for *S. cerevisiae* strain W1588-4C (51), derivative of W303. It was shown that lack of RNase H2 functionality does not alter nucleotide pools in *S. cerevisiae* cells (51). Therefore, the fact that wild-type RNase H2 cells have low rG and equal absolute amount of rA, rC, and rU in nuclear DNA, with an overall rNMP distribution that is significantly different from that observed in *rnh201* cells (**Figures 12 and 16**), may indicate that yeast RNase H2 cleaves rNMPs in yeast DNA with differential efficiency and may have preference for rG, as preliminary data showed, using protein extracts from yeast, HeLa cells, and *E. coli* cells (60). A more recent study, exploiting microarray analysis of thousands of rNMP-containing DNA hairpins of different sequence, demonstrated clear cleavage preference of *E. coli* RNase HII at and around the cleavage site (61). It is also possible that other mechanisms to remove rNMPs from DNA may be more

active in *rnh201* than in wild-type cells, like topoisomerase I (16,59), which may have preferred genomic targets, or both cases may apply.

The presence of rG was also not random. Like rC, rG was most often preceded by dA both in *S. cerevisiae* and *S. paradoxus* mitochondrial libraries and in all nuclear libraries across all the three yeast species for all strains and genotypes examined in this study. Strikingly, ArC and ArG were dominant motifs in all hotspots, as well as in trinucleotide, dinucleotide, and other shortnucleotide repeats across nuclear and mitochondrial DNA. Interestingly, we found enrichment of rNMPs in short-nucleotide repeated sequences. The observed patterns of rNMPs demonstrate a strong propensity for rNMP incorporation when nucleotides with base C or G are in di, tri, or other short-nucleotide repeats. Remarkably, in all repeats that we analyzed except one (9/10), the rCs and rGs were preceded by dA (**Figure 17 and Supplementary Figures 7 and 8**). Thus, if the same dinucleotide-sequence context with AC or AG is repeated within a given region of the genome, like in trinucleotide or dinucleotide repeated tracts containing such dinucleotides, any AC or AG site within the repeated tracts may be equally prone to rNMP incorporation, explaining the observed patterns. These rNMP sites at short-nucleotide repeats may contribute to the instability of these regions, possibly driving mechanisms of repeat expansion/contraction.

In conclusion, via mapping and genome-wide analysis of many ribose-seq libraries across three yeast species with wild type and mutant RNase H2 genotypes, which consistently generated reproducible results, we reveal new biological features and a fundamental rule shaping the patterns of rNMP incorporation. Not only is genomic rNMP incorporation far from random, but it is also driven by the sequence context, particularly by the sequence of the nucleotide immediately upstream of the site of incorporation. Potentially, such sequence context-driven mechanism of rNMP incorporation reflects a physiological function of rNMPs in DNA. More broadly, our work

provides robust tools and a model framework to further study biological, chemical, and structural aspects of genomic rNMP incorporation from yeast throughout all kingdoms of life.

3.5 Acknowledgments

We thank Yury Chernoff for yeast strains BY4741, W303, and S288C; Shweta Biliya and Naima Djeddar for high-throughput sequencing; Troy Hilley for help setting up the genome browser, the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology for their research cyberinfrastructure resources and services; Markus W. Germann for discussion of results; Patrick McGrath and Stefania Marsili for critically reading the manuscript; and all members of the Storici laboratory for assistance and feedback on this study. We acknowledge funding from the National Institutes of Health, NIH AI136581 (to Baek Kim), AI150451 (to Baek Kim), MH116695 (to Raymond F. Schinazi), R01ES026243 (to Francesca Storici), the Parker H. Petit Institute for Bioengineering and Bioscience at Georgia Institute of Technology 12456H2 (to Francesca Storici), and the Howard Hughes Medical Institute Faculty Scholar grant 55108574 (to Francesca Storici) for supporting this work.

3.6 Author Contributions

Francesca Storici, Sathya Balachander, Alli L. Gombolay, Taehwan Yang, Penghao Xu, and Kyung Duk Koh formulated the project and designed experiments. Sathya Balachander and Taehwan Yang built the ribose-seq libraries. Alli L. Gombolay and Penghao Xu performed the bioinformatics analysis of the sequencing data with guidance from Fredrik Vannberg.

CHAPTER 4

CONCLUSION

4.1 Incorporation of ribonucleotides into genomic DNA

Genetic information is stored in DNA rather than RNA partly due its greater stability (1). In contrast to dNMPs, the subunits of DNA, rNMPs, the subunits of RNA, contain a reactive 2'-OH group (1). During DNA synthesis, DNA Pols select nucleotides into the growing DNA strand that contain both the correct sugar and base, preferring dNMPs to rNMPs. However, selection errors can occur, resulting in the incorporation of rNMPs into DNA. In fact, the incorporation of rNMPs into DNA is one of the most frequently occurring errors during DNA synthesis (2).

To maintain genome stability, three pathways are responsible for removing rNMPs from DNA with varying degrees of efficiency- 1) RNase H enzymes (RNase H1 and RNase H2), 2) exonucleolytic processing by Pol δ and Pol ε, and 3) Top1 (1). The RNase H enzymes efficiently remove rNMPs from DNA. RNase H1 cleaves stretches of ≥ 4 rNMPs in DNA, while RNase H2 cleaves both single and stretches of rNMPs in DNA (8). The RNase H2 enzyme initiates RER (1). Although not essential for viability in yeast, RNase H2 is essential in mice and loss of this enzyme results in embryonic lethality (9). In addition, mutations in the genes encoding the subunits of RNase H2 in humans are associated with AGS (10) and SLE (11). The 3'-5' exonuclease activity of Pol δ and Pol ε can remove rNMPs from DNA but inefficiently (4,5,12,13). In the absence of RER, Top1 can remove rNMPs from DNA but aberrantly (14,15). If these pathways fail to remove rNMPs from DNA, the 2'-OH group of unrepaired rNMPs can attack the double-helix backbone of DNA, resulting in several types of genome instability, including SSBs, DSBs, short deletion

mutations, replication stress, cell cycle checkpoint activation, aberrant recombination, formation of protein-DNA crosslinks, and alterations in the structural properties of DNA (16,17).

The recent development of rNMP-seq techniques (ribose-seq, emRiboSeq, RHII-HydEn-seq, Alk-HydEn-seq, and Pu-seq), has enabled us to map the locations of rNMPs to single-nucleotide resolution. However, since the development of rNMP-seq techniques is recent, the biological signatures of rNMPs had yet to be thoroughly characterized. Previously, Koh *et al.* applied ribose-seq to characterize the biological signatures of rNMPs in the DNA of yeast cells but was limited to only one strain of the species- *S. cerevisiae* (18). The remaining studies applied rNMP-seq to track the division of labor of DNA polymerases in yeast (19-21,27) rather than characterize biological signatures of rNMPs. In addition, since current rNMP mapping software are highly customized, limited in terms of scope of analysis, depend on proprietary software, and/or provide limited documentation, a standardized bioinformatics toolkit to map the genomic coordinates of rNMPs to single-nucleotide resolution and characterize the biological signatures of rNMPs for data produced using any rNMP-seq technique was urgently needed.

4.2 Ribose-Map and Its Application to Yeast Ribose-seq Data

To address this, I created the Ribose-Map bioinformatics toolkit, the first known standardized bioinformatics toolkit for the comprehensive analysis of rNMP-seq data. Through a series of modules, Ribose-Map calculates the single-nucleotide genomic coordinates of rNMPs and characterizes the biological signatures of rNMPs for data generated using any rNMP-seq technique. The Alignment Module aligns rNMP-seq data to the reference genome of interest (and de-multiplexes and/or de-duplicates data if needed), the Coordinate Module calculates the genomic coordinates of rNMPs to single-nucleotide resolution, the Composition Module calculates and

plots the nucleotide composition of rNMPs, the Sequence Module calculates and plots the nucleotide sequence context of rNMPs, the Distribution Module calculates and plots the per-nucleotide counts of rNMPs, and the Hotspot Module calculates the most abundant sites of rNMPs and plots their consensus sequences. Then, I validated Ribose-Map against current rNMP mapping software. When compared to current rNMP mapping software (Modmap, emRiboSeqProcessor, and Puseq_app), Ribose-Map is the only software that analyzes data generated from any rNMP-seq technique, analyzes data from any organism with a sequenced reference genome, normalizes per-nucleotide counts of rNMPs to account for sequencing depth, outputs a file containing the single-nucleotide genomic coordinates of rNMPs, and depends on only open-source software.

We then built nuclear and mitochondrial ribose-seq libraries from wild type, RNase H1 mutant, and RNase H2 mutant strains of *S. cerevisiae*, *S. paradoxus*, and *S. pombe* and applied Ribose-Map to characterize the biological signatures of rNMPs in these libraries. We found deoxyadenosine upstream from the most abundant rCs and rGs, suggesting an accommodation of rNMPs by DNA polymerases. We also found rC and rG in C/G-rich regions of *S. cerevisiae* and *S. paradoxus* mitochondrial DNA, suggesting a physiological function of rNMPs in DNA.

In addition to applying Ribose-Map to characterize the biological signatures of rNMPs in ribose-seq yeast data, we also applied Ribose-Map to compare the results of our yeast ribose-seq study to the previously published yeast emRiboSeq study to assess the agreement between these two techniques. To compare ribose-seq to emRiboSeq, we applied Ribose-Map to *rnh201 S. cerevisiae* nuclear and mitochondrial data generated using ribose-seq and *rnh201 S. cerevisiae* nuclear and mitochondrial data generated using emRiboSeq. For both ribose-seq and emRiboSeq, we found deoxyadenosine upstream from the most abundant rCs and rGs (**Figure 10)**. We also found rC and rG in C/G-rich regions of *S. cerevisiae* mitochondrial DNA (**Figure 8)**. This

comparison further supports our findings that the dNMP upstream of rNMPs has the most impact on the incorporation of a specific rNMP type in yeast mitochondrial DNA.

We use Ribose-Map to analyze the data for all rNMP mapping projects in the Storici Lab. In addition to yeast, I also applied Ribose-Map to characterize the biological signatures of rNMPs in the unicellular green algae, *C. reinhardtii* for the first time (7). We found increased incorporation of rA in the mitochondrial and chloroplast DNA of *C. reinhardtii* compared to the nuclear DNA relative to the nucleotide content of the corresponding reference genomes (7).

4.3 Limitations of rNMP-seq

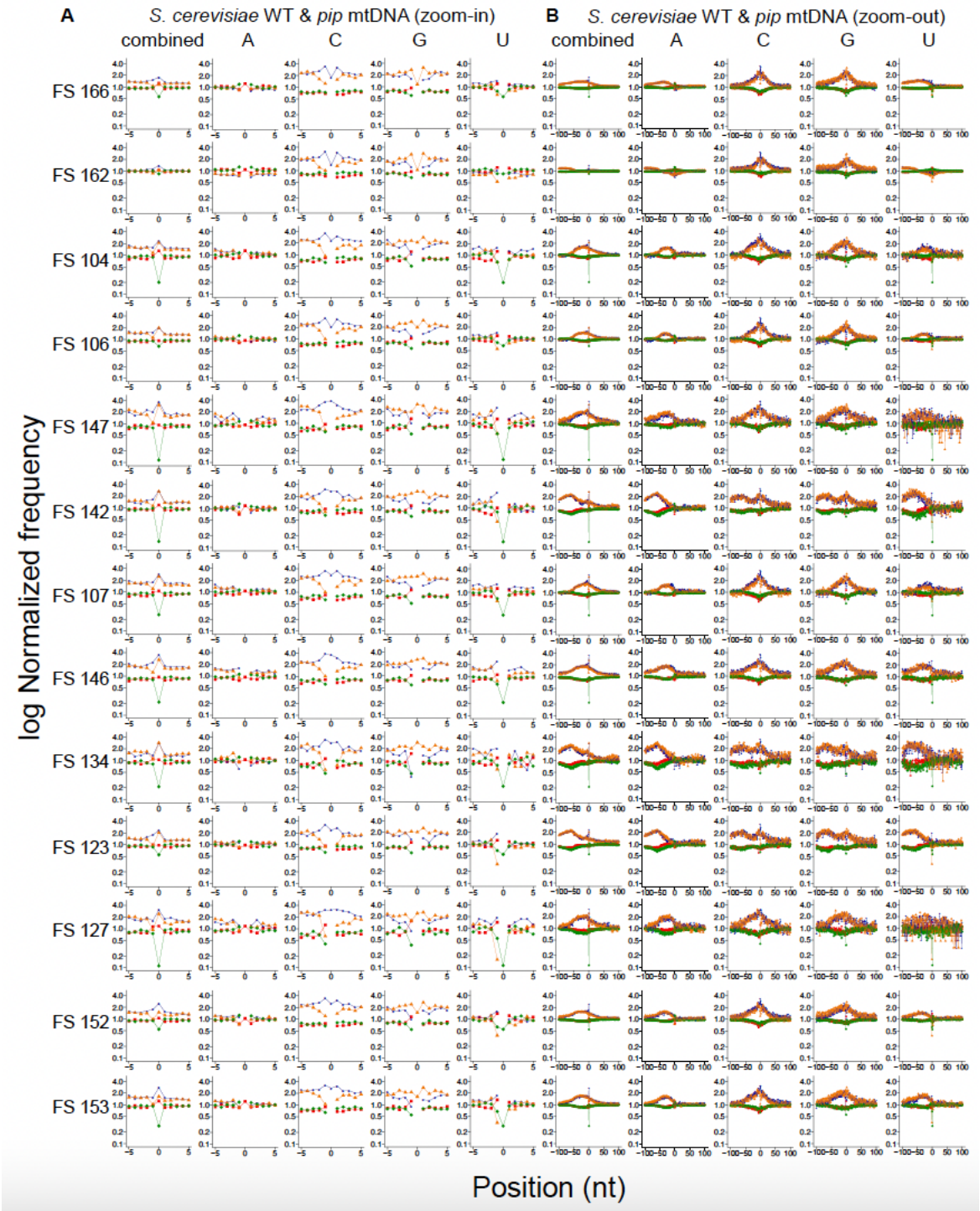
rNMP-seq can only capture one rNMP in a stretch of ≥ 2 rNMPs. For example, following treatment with alkali during ribose-seq, stretches of ≥ 2 rNMPs are reduced to single rNMPs and are subsequently removed by purification. If current rNMP-seq techniques were modified to capture each rNMP in a stretch of rNMPs or a new technique is developed, Ribose-Map's Coordinate Module could be readily updated to accommodate any change in the position of the rNMP relative to the tagged nucleotide. In addition, rNMP-seq techniques currently utilize short read sequencing technology. However, short read sequencing technology is unable to span repetitive sequences of DNA, leading to errors when aligning reads originating from these regions. Since long reads often span repetitive sequences of DNA, rNMP-seq techniques could be updated to utilize long read sequencing technology (e.g., Oxford Nanopore). If current rNMP-seq techniques were modified to utilize long read sequencing technology or a new technique is developed, an option to align reads with a long read aligner (e.g., BLASR (22), GraphMap (23), or Kart (24)) could be added to Ribose-Map's Alignment Module. Once this option is added, the user could then specify short or long reads in the configuration file.

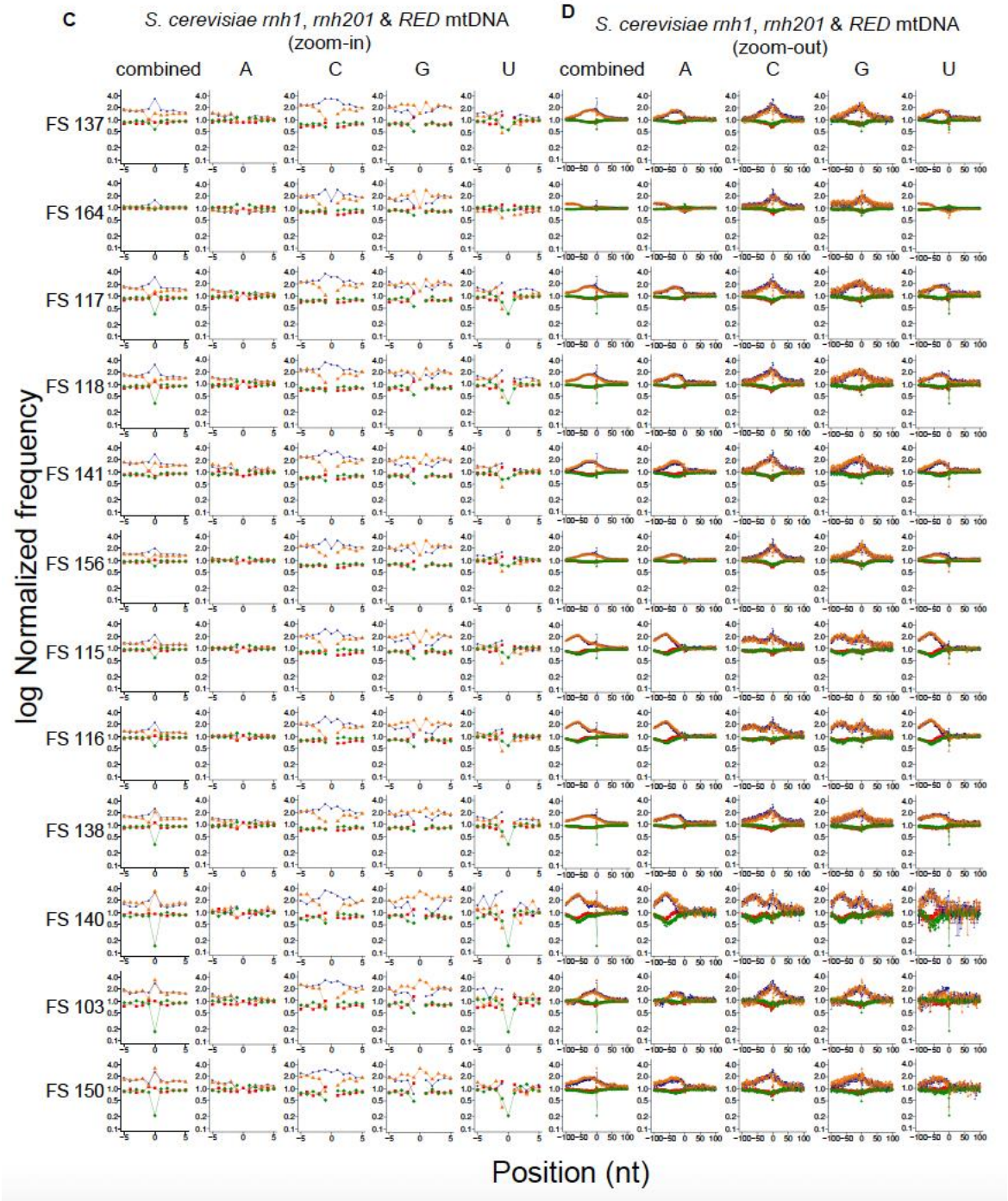
4.4 Future Directions

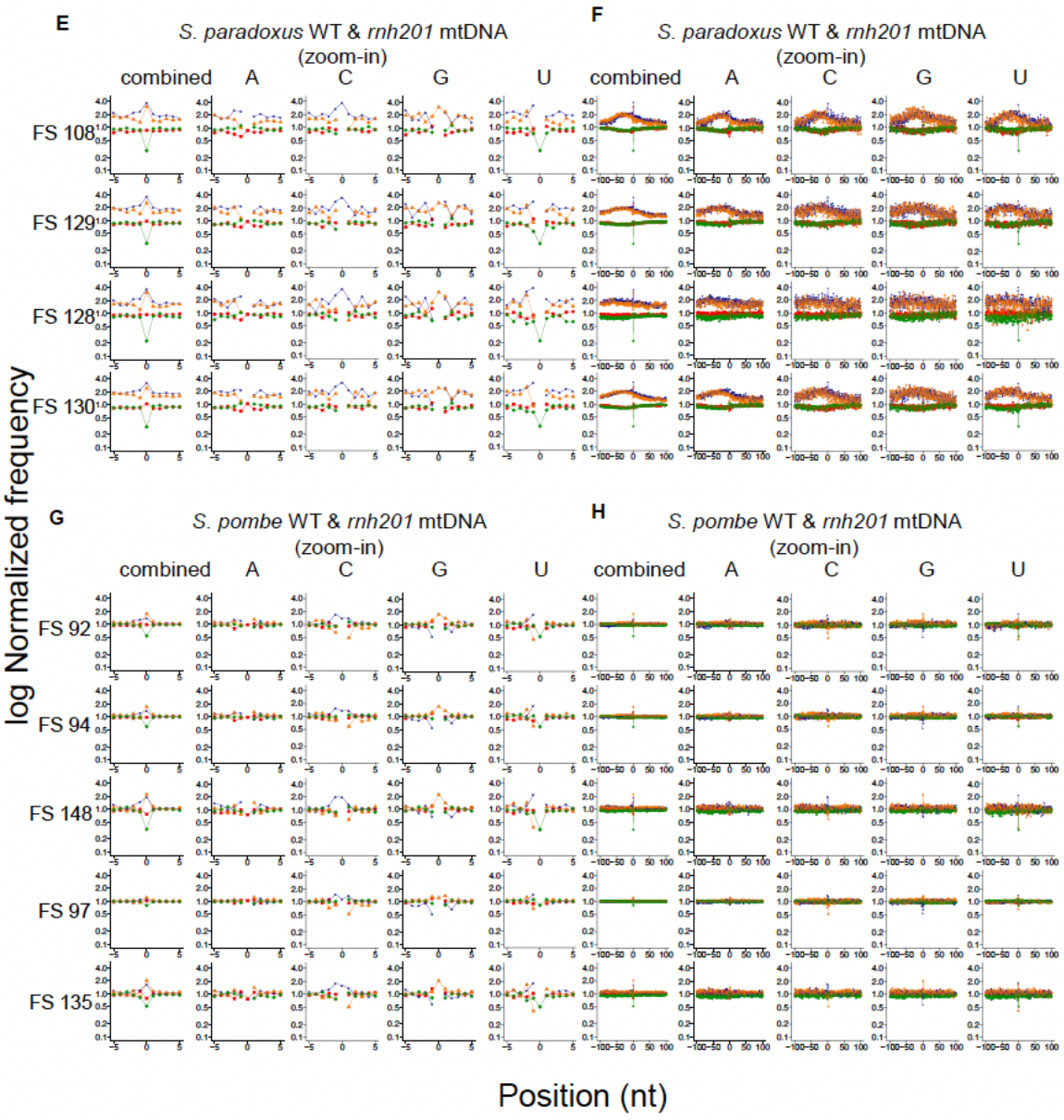
As the field of rNMP mapping grows and gains more interest from the broader scientific community, several improvements to rNMP-seq techniques and Ribose-Map would be helpful. First, rNMP-seq techniques should be improved to capture both single rNMPs and stretches of rNMPs to allow us to gain an even better understanding of the locations and biological signatures of rNMPs in DNA. Second, rNMP-seq techniques should be modified or a new technique should be developed to utilize long read sequencing technology and Ribose-Map should be updated accordingly. Third, it would be helpful to have a conda installation available for Ribose-Map to further increase ease of installation and use in the bioinformatics community. Lastly, it would be helpful to add new modules to Ribose-Map to further characterize the biological signatures of rNMPs, including a module to identify genomic features associated with rNMPs.

Moving forward, it will be important to perform a direct comparison of the results from all rNMP-seq techniques to assess the accuracy of each of these techniques in mapping rNMPs. It will also be important to characterize the biological signatures of rNMPs in a variety of strains, genotypes, and species to determine if the biological signatures of rNMPs that we found in *S. cerevisiae*, *S. paradoxus*, and *S. pombe* also apply to other species. In conclusion, the development of Ribose-Map and its application to yeast serves as a foundational resource for the emerging field of rNMP mapping, leading to an improved understanding of the role of rNMPs in genome stability.

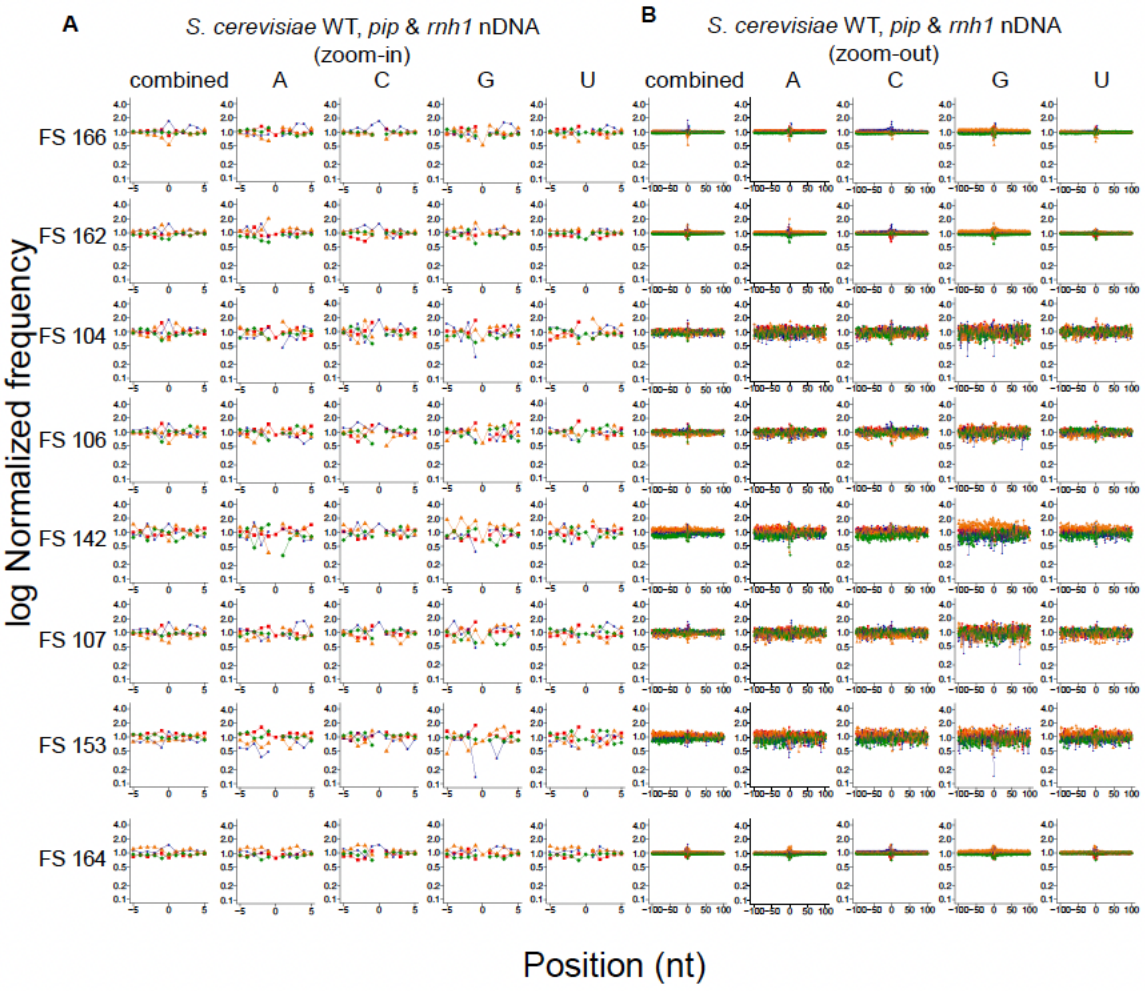
APPENDIX

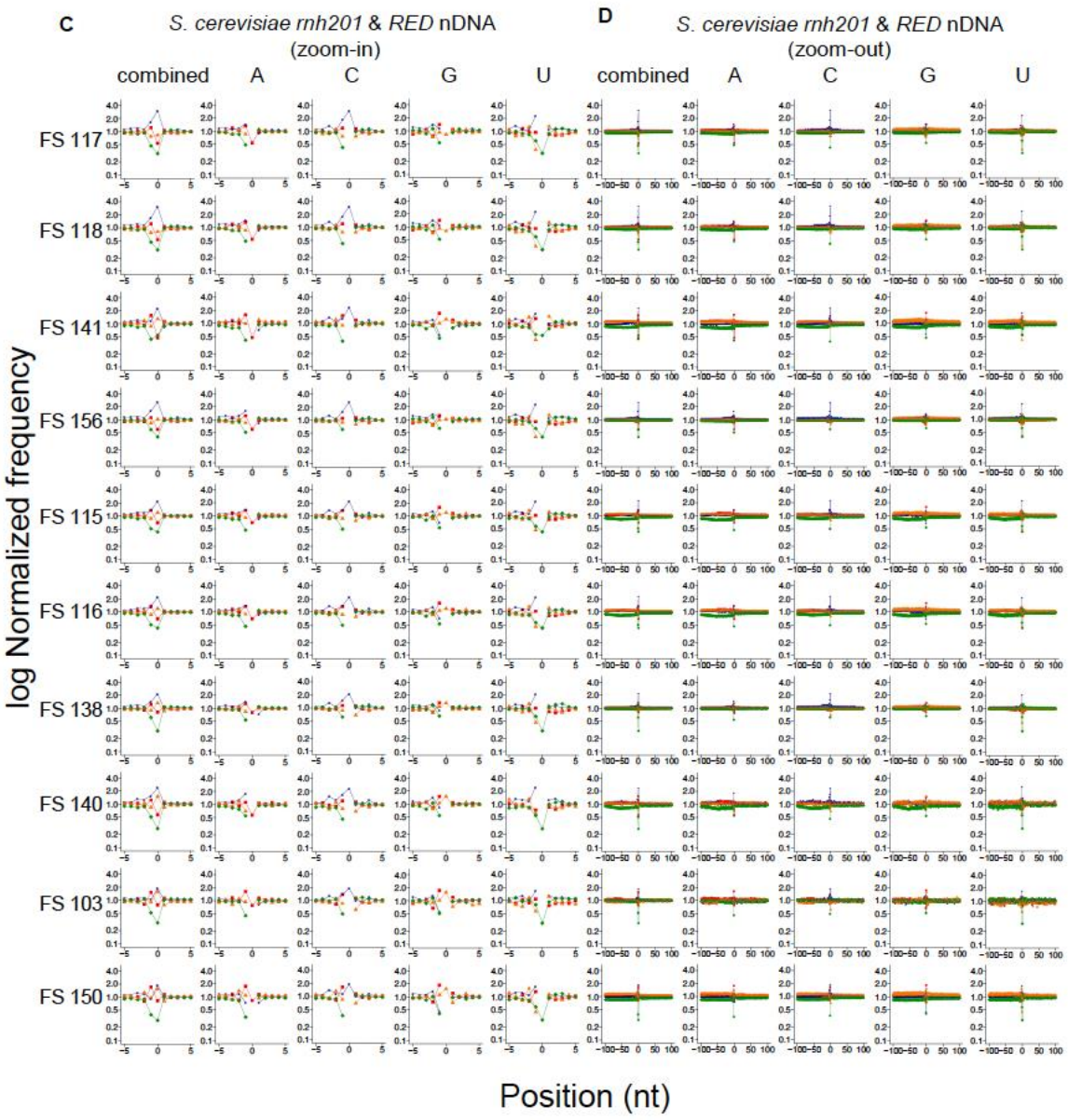

A
S. cerevisiae WT & pip mtDNA (zoom-in)
B
S. cerevisiae WT & pip mtDNA (zoom-out)
combined
A
C
G
U
FS 166
FS 162
FS 104
FS 106
FS 147
FS 142
FS 107
FS 146
FS 134
FS 123
FS 127
FS 152
FS 153
log Normalized frequency
Position (nt)

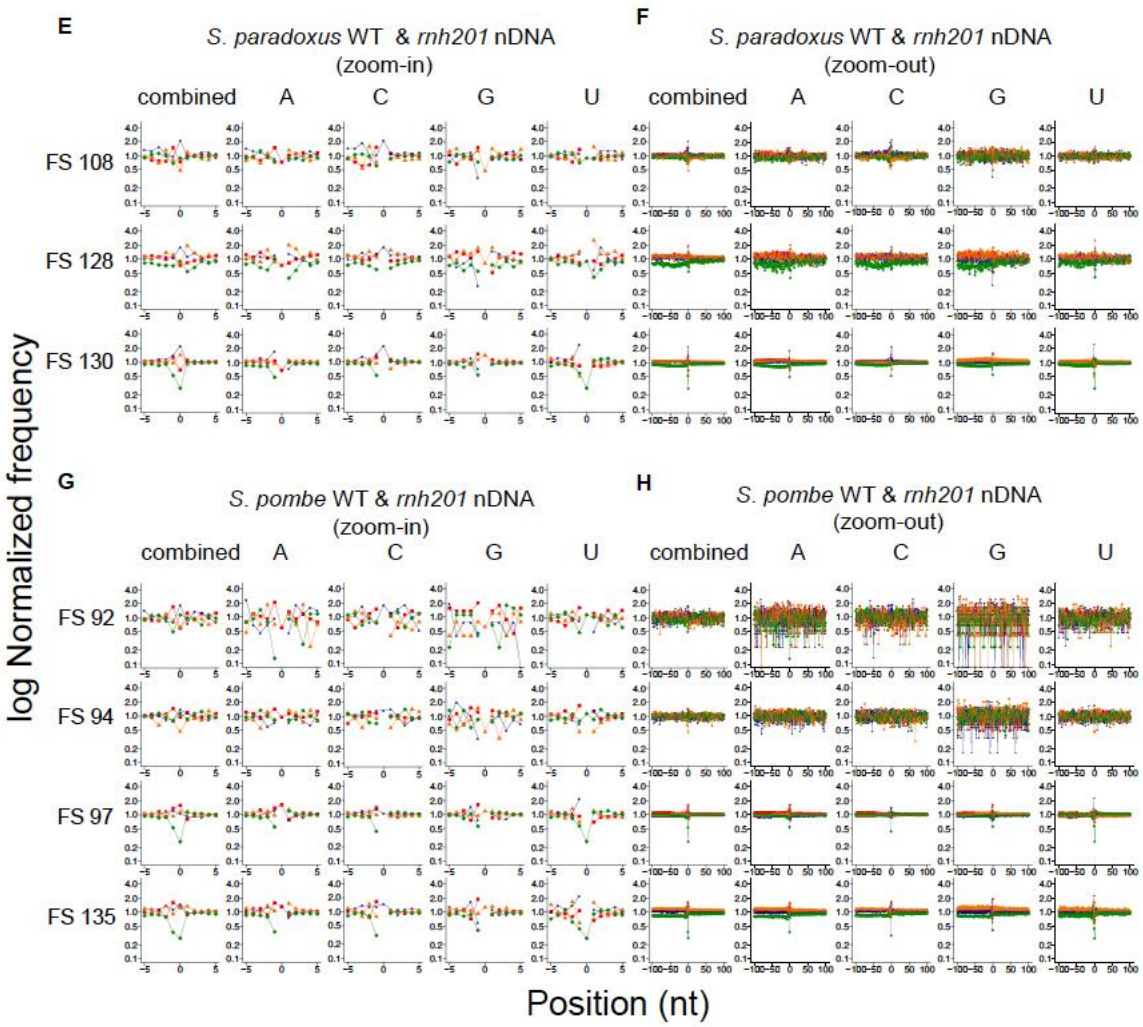

C
S. cerevisiae rnh1, rnh201 & RED mtDNA
(zoom-in)
D
S. cerevisiae rnh1, rnh201 & RED mtDNA
(zoom-out)
combined
A
C
G
U
FS 137
FS 164
FS 117
FS 118
FS 141
FS 156
FS 115
FS 116
FS 138
FS 140
FS 103
FS 150
log Normalized frequency
Position (nt)



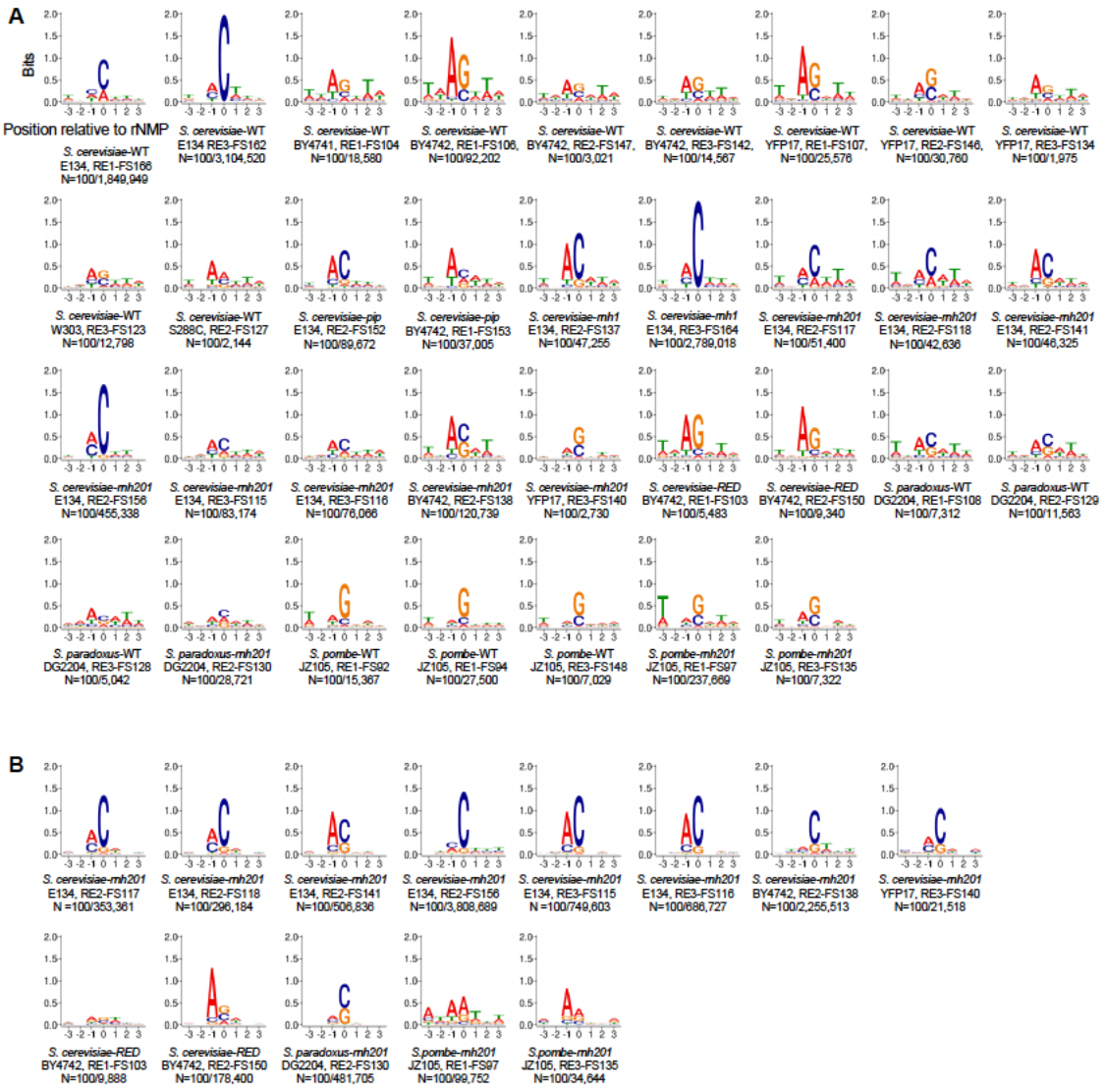
Supplementary Figure 1. Nucleotide plots of all mitochondrial libraries. (A-H) Plots of normalized nucleotide frequencies relative to mapped positions of sequences from the (A,B) 11 wild type and 2 *pip S. cerevisiae* mitochondrial libraries generated in this study, combined and single, (A) zoom in, and (B) zoomed out plots (C,D) 2 *rnh1*, 8 *rnh201*, and 2 RED *S. cerevisiae* mitochondrial libraries combined and single, (C) zoom in, and D zoomed out plots (E,F) 3 S paradoxus wild type and 1 *rnh201* mitochondrial libraries combined and single, (E) zoom in, and (F) zoomed out plots (G,H) 3 *S. pombe* wild type and 2 *rnh201* mitochondrial libraries combined and single, (G) zoom in, and (H) zoomed out plots. Position 0 on the x-axis represents the site of rNMP incorporation, and positions represent upstream and downstream dNMPs respectively. The y-axis shows the frequency of each type of nucleotide present in the ribose-seq data normalized to the frequency of the corresponding nucleotide present in the reference genome of the indicted yeast species. Red square, A blue circle, C orange triangle, G and green rhombus, U.


A
S. cerevisiae WT, pip & rnh1 nDNA
(zoom-in)
B
S. cerevisiae WT, pip & rnh1 nDNA
(zoom-out)
combined
A
C
G
U
FS 166
FS 162
FS 104
FS 106
FS 142
FS 107
FS 153
FS 164
log Normalized frequency
Position (nt)

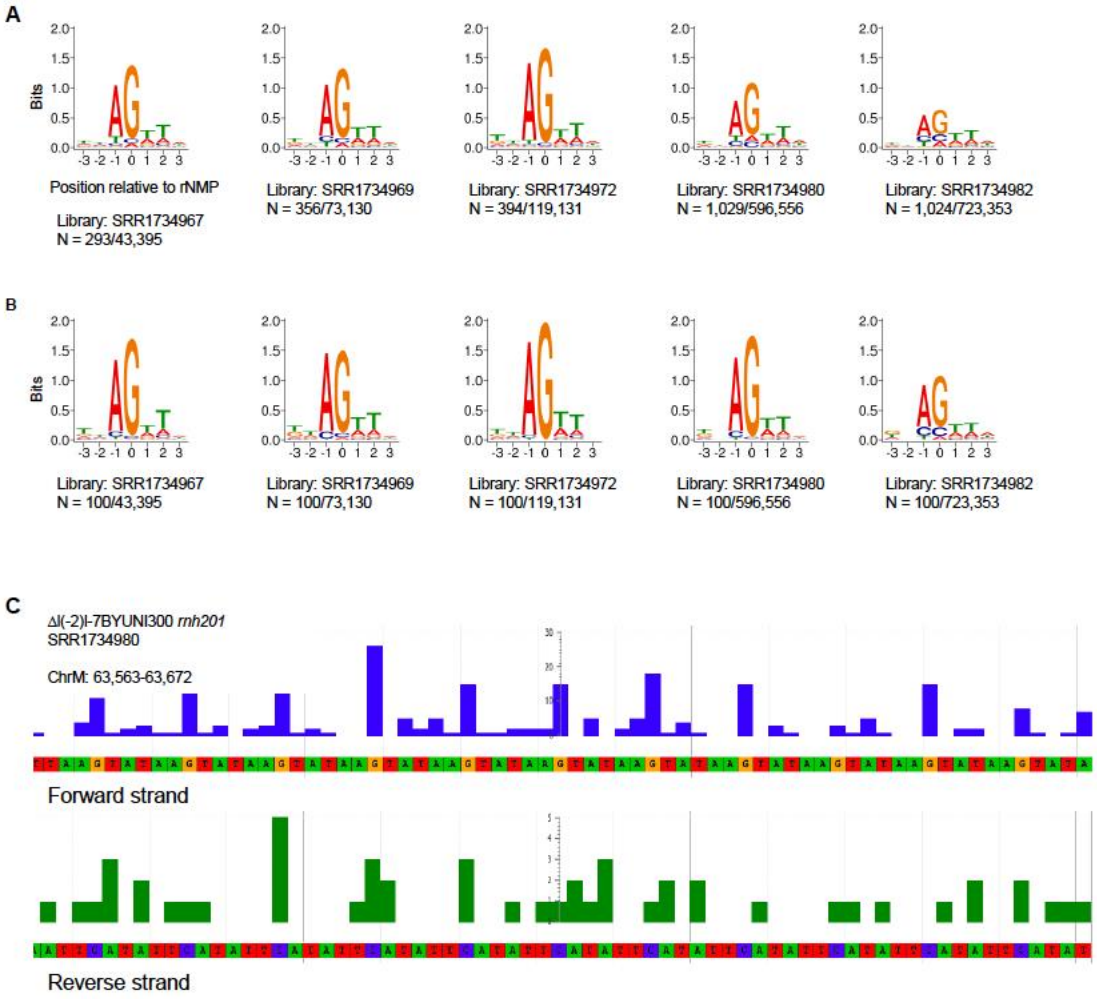

C
S. cerevisiae rnh201 & RED nDNA
(zoom-in)
combined
A
C
G
U
D
S. cerevisiae rnh201 & RED nDNA
(zoom-out)
combined
A
C
G
U
FS 117
FS 118
FS 141
FS 156
FS 115
FS 116
FS 138
FS 140
FS 103
FS 150
log Normalized frequency
Position (nt)



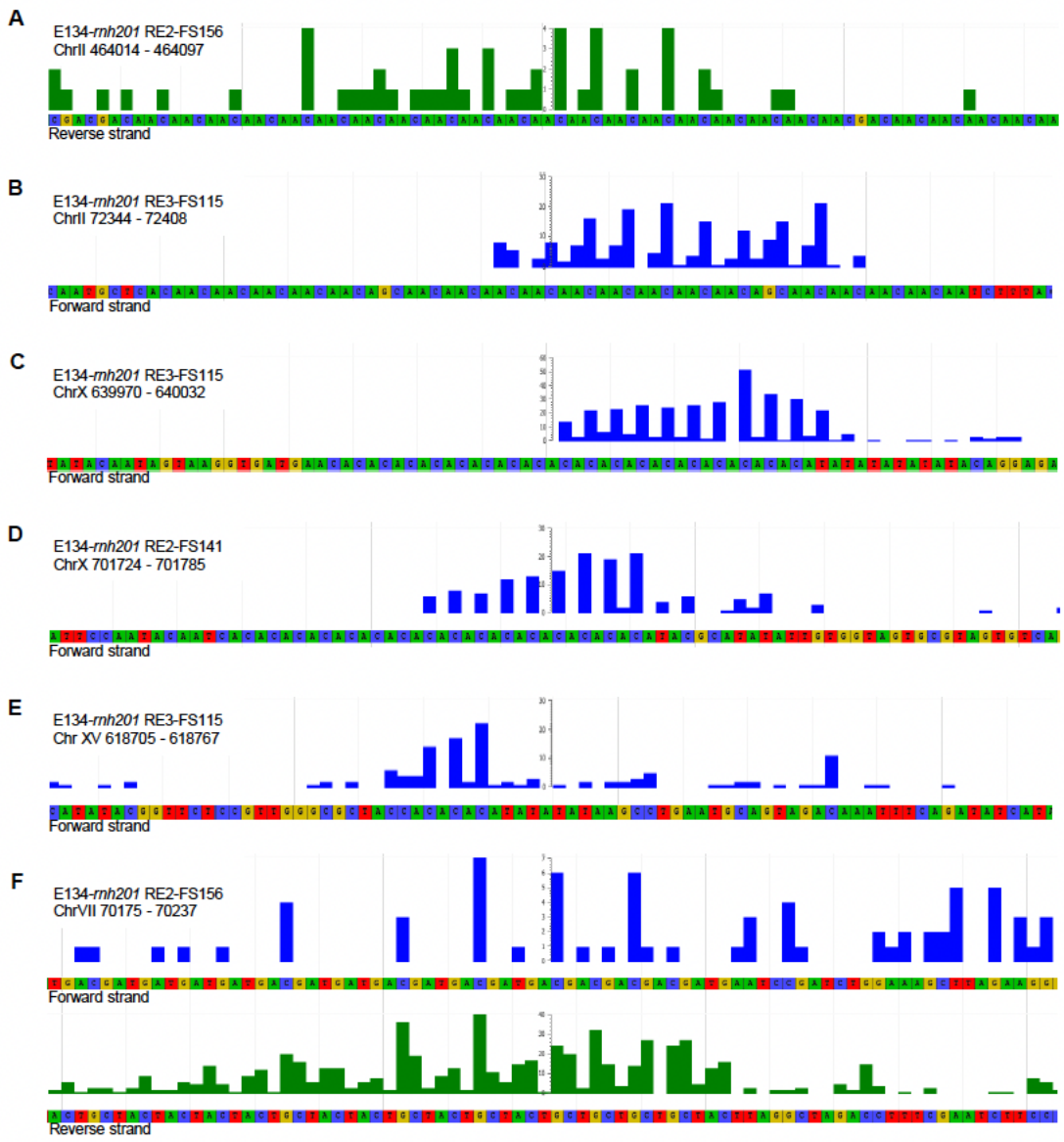
Supplementary Figure 2. Nucleotide plots from all nuclear libraries. (A-H) Plots of normalized nucleotide frequencies relative to mapped positions of sequences from the (A,B) 6 wild-type, 1 *pip* and 1 *rnh1 S. cerevisiae* nuclear libraries generated in this study, combined and single, (A) zoom-in, and (B) zoomed-out plots; (C,D) 8 *rnh201* and 2 RED *S. cerevisiae* nuclear libraries combined and single, (C) zoom-in, and (D) zoomed-out plots; (E,F) 2 *S. paradoxus* wild-type and 1 *rnh201* nuclear libraries combined and single, (E) zoom-in, and (F) zoomed-out plots; (G,H) 2 *S. pombe* wild-type and 2 *rnh201* nuclear libraries combined and single, (G) zoom-in, and (H) zoomed-out plots. Position 0 on the x-axis represents the site of rNMP incorporation, - and + positions represent upstream and downstream dNMPs, respectively. The y-axis shows the frequency of each type of nucleotide present in the ribose-seq data normalized to the frequency of the corresponding nucleotide present in the reference genome of the indicted yeast species. Red square, A; blue circle, C; orange triangle, G; and green rhombus, U.



Supplementary Figure 3. Hotspot motifs from mitochondrial and nuclear DNA from top 100 rNMP sites. (A) Sequence motif plots for mitochondrial and (B) nuclear hotspots (top 100 of rNMP sites). Position 0 on the x-axis represents the site of rNMP incorporation, and positions represent upstream and downstream dNMPs respectively. The y-axis shows the level of sequence conservation, represented in bits. The species, genotype, strain, restriction enzyme (RE) set, library name, and the number of rNMP sites are included below each plot.



Supplementary Figure 4. Hotspot motifs from mitochondrial and nuclear DNA of emRiboSeq libraries. (A) Sequence motif plots for mitochondrial hotspots, top 1 of rNMP sites, and (B) top 100 of rNMP sites of emRiboSeq libraries derived from *rnh201 S. cerevisiae* cells. Position 0 on the x-axis represents the site of rNMP incorporation, and positions represent upstream and downstream dNMPs respectively. The y-axis shows the level of sequence conservation, represented in bits (C) Genome browser snapshot showing an rG hotspot within the TAAGTA repeated sequence on the forward strand and an rC hotspot on the complementary strand in *S. cerevisiae* mitochondrial DNA at the locus chrM 63 563 63 672 for emRiboSeq library SRR 1734980 (*rnh 201*).



Supplementary Figure 5. Patterns of rNMPs in tri- and di-nucleotide repeat tracts. Genome browser snapshots of *S. cerevisiae* nDNA showing examples of tri- and di- nucleotide repeat tracts with a specific rNMP pattern. (A) rC in the repeated motif AAC at locus chrII 464014..464097 on the reverse strand for library FS156 (E134 *rnh201* RE 2 and similarly for FS141 (E134 *rnh201* RE 2 FS138 (BY4742 *rnh201* RE 2 FS150 (BY4742 *rnh201-RED* RE 2 FS115 and FS116 (E134 *rnh201* RE 3 FS117 and FS118 (E134 *rnh201* RE 2). (B) Pattern AArC in AAC repeated sequence at locus chrII 72344..72408 on the forward strand for library FS115 (E134 *rnh201* RE 3 and similarly for FS116, FS117, FS118, FS138, FS140, FS141, FS156 (E134 *rnh201*), and FS150 (BY4742 *rnh201-RED*). (C) rC hotspot within the AC repeated sequence at locus chrX 639970..640032 on the forward strand for library FS115 (E134 *rnh201* RE 3 and similarly for FS115 and FS116 (E134 *rnh201* RE 3 FS117 and FS118 (E134 *rnh201* RE 2 FS 138 (BY4742 *rnh201* RE 2 FS140 (YFP17 *rnh201* RE 3 FS141 (E134 *rnh201* RE 2 FS 150 (BY4742 *rnh201-RED* RE 2 and FS156 (E134 *rnh201* RE 2). (D) rC hotspot within the AC repeated sequence at

locus chrX 701724 701785 on the forward strand for library FS141 (E134 *rnh201* RE 2 and similarly for FS116 (E134 *rnh201* RE 3 FS117 and FS118 (E134 *rnh201* RE 2 FS138 (BY4742 *rnh201* RE 2 FS150 (BY4742 *rnh201-RED* RE 2 and FS156 (E134 *rnh201* RE 2). (E) rC hotspot within the AC repeated sequence at locus chrXV 618732..618740 on the forward strand for library FS 115 (E134 *rnh201* RE 3 and similarly for FS 116 (E134 *rnh201* RE 3 FS117 and FS 118 (E134 *rnh201* RE 2 FS 138 (BY4742 *rnh 201* RE 2 S 140 (YFP17 *rnh201* RE 3 FS141 (E134 *rnh201* RE 2 FS150 (BY4742 *rnh201-RED* RE 2 and FS156 (E134 *rnh201* RE 2). (F) rC hotspot within GArC repeated sequence at locus chrVII 70175..70237 on the forward strand for library FS156 (E134 *rnh201* RE 2 The same locus has also rG hotspot on the reverse complement site in TCrG of TGC repeated sequence, and similarly for FS117, FS118, FS138, FS141 (all *rnh201*), and FS150 (BY4742 *rnh201-RED* RE 2). This pattern with an rNMP hotspot on both strands in the same site was also seen in mitochondrial DNA of *S. cerevisiae* at locus chrM 63583..63651.

Supplementary Table 1. Yeast strains used in this study. Yeast strains used in this study for ribose-seq library construction and their corresponding genotype. The strains include *S. cerevisiae* strains, as well as strains of *S. paradoxus* and *S. pombe*, as indicated in parenthesis.

Strain	Relevant genotype	Source
E134 (KK-44)	*MATα ade5-1 lys2-14A trp1-289 his7-2 leu2-3,112 ura3-52*	Koh *et al.*, 2015 (18)
KK-100	KK-44 *rnh201*Δ::*hygMX4*	Koh *et al.*, 2015 (18)
KK-172	KK-44 rnh1Δ::kanMX4	this study
SB-286	KK-44 rnh202-FF346,347AA	this study
BY4742 (KK-2)	*MATα his3*Δ*1 leu2*Δ*0 lys2*Δ*0 ura3* Δ0	Storici *et al.*, 2001 (62)
SB-305	KK-2 *rnh201*Δ::*hygMX4*	this study
SB-285	KK-2 rnh202-FF346,347AA	this study
SB-311	KK-2 rnh201-P45D Y219A	this study
BY4741 (SB-292)	*MATa his3-1 leu2-0 met15-0 ura3-0*	Brachmann *et al.*, 1998 (42)
YFP17 (SB-288)	*ho*Δ *hml*Δ::*ADE1 mata*Δ::*hisG Δhmr*::*ADE1 ade1 leu2-3,1122 lys5 trp1*::*hisG ura3-52 ade3*::*GAL*::*HO*	Keskin *et al.*, 2014 (36)
SB-293	SB-293 *rnh201*Δ::*hygMX4*	this study
W303 (SB-316)	*MATa can1-100 his3-11,15 leu2-3,112 trp1-1 ura3-1 ade2-1*	Ralser *et al.*, 2012 (63)
S288C (SB-313)	*MATα SUC2 gal2 mal2 mel flo8-1 hap1 ho bio1 bio6*	Mortimer and Johnston (1986) (41)
DG2204 (HK-692)	*MATa LYS+ trp1 ADE+ LEU+ ura3 his3-Δ200 hisG Gal+ Spo+ Ty-less* (*S. paradoxus*)	Garfinkel *et* al., 2005 (64)
HK-705	HK-692 *rnh201*Δ::*hygMX4* (*S. paradoxus*)	this study
JZ105 (KK-154)	*Mat1M mat2,3*Δ::*LEU2 ade6-210 leu1-32 ura4-D18 his2* (*S. pombe*)	Vengrova *et al.*, 2004 (65)
HK-983	KK-154 *rnh201*Δ::*kanMX4* (*S. pombe*)	this study

REFERENCES

1. Williams, J.S., Lujan, S.A. and Kunkel, T.A. (2016) Processing ribonucleotides incorporated during eukaryotic DNA replication. *Nat Rev Mol Cell Biol*, **17**, 350-363.
2. Kellner, V. and Luke, B. (2020) Molecular and physiological consequences of faulty eukaryotic ribonucleotide excision repair. *EMBO J*, **39**, e102309.
3. McElhinny, S.A.N., Kumar, D., Clark, A.B., Watt, D.L., Watts, B.E., Lundstrom, E.B., Johansson, E., Chabes, A. and Kunkel, T.A. (2010) Genome instability due to ribonucleotide incorporation into DNA. *Nat Chem Biol*, **6**, 774-781.
4. Clausen, A.R., Zhang, S., Burgers, P.M., Lee, M.Y. and Kunkel, T.A. (2013) Ribonucleotide incorporation, proofreading and bypass by human DNA polymerase delta. *DNA Repair (Amst)*, **12**, 121-127.
5. Goksenin, A.Y., Zahurancik, W., LeCompte, K.G., Taggart, D.J., Suo, Z. and Pursell, Z.F. (2012) Human DNA polymerase epsilon is able to efficiently extend from multiple consecutive ribonucleotides. *J Biol Chem*, **287**, 42675-42684.
6. Wallace, B.D. and Williams, R.S. (2014) Ribonucleotide triggered DNA damage and RNA-DNA damage responses. *RNA Biol*, **11**, 1340-1346.
7. El-Sayed, W.M.M., Gombolay, A.L., Xu, P., Yang, T., Jeon, Y., Balachander, S., Newnam, G., Tao, S., Bowen, N.E., Bruna, T. *et al.* (2021) Disproportionate presence of adenosine in mitochondrial and chloroplast DNA of Chlamydomonas reinhardtii. *iScience*, **24**, 102005.
8. Cerritelli, S.M. and Crouch, R.J. (2009) Ribonuclease H: the enzymes in eukaryotes. *FEBS J*, **276**, 1494-1505.
9. Reijns, M.A., Rabe, B., Rigby, R.E., Mill, P., Astell, K.R., Lettice, L.A., Boyle, S., Leitch, A., Keighren, M., Kilanowski, F. *et al.* (2012) Enzymatic removal of ribonucleotides from DNA is essential for mammalian genome integrity and development. *Cell*, **149**, 1008-1022.
10. Crow, Y.J., Leitch, A., Hayward, B.E., Garner, A., Parmar, R., Griffith, E., Ali, M., Semple, C., Aicardi, J., Babul-Hirji, R. *et al.* (2006) Mutations in genes encoding ribonuclease H2 subunits cause Aicardi-Goutieres syndrome and mimic congenital viral brain infection. *Nat Genet*, **38**, 910-916.
11. Gunther, C., Kind, B., Reijns, M.A., Berndt, N., Martinez-Bueno, M., Wolf, C., Tungler, V., Chara, O., Lee, Y.A., Hubner, N. *et al.* (2015) Defective removal of ribonucleotides from DNA promotes systemic autoimmunity. *J Clin Invest*, **125**, 413-424.
12. Williams, J.S., Clausen, A.R., Nick McElhinny, S.A., Watts, B.E., Johansson, E. and Kunkel, T.A. (2012) Proofreading of ribonucleotides inserted into DNA by yeast DNA polymerase varepsilon. *DNA Repair (Amst)*, **11**, 649-656.
13. Williams, J.S. and Kunkel, T.A. (2014) Ribonucleotides in DNA: origins, repair and consequences. *DNA Repair (Amst)*, **19**, 27-37.
14. Kim, N., Huang, S.N., Williams, J.S., Li, Y.C., Clark, A.B., Cho, J.E., Kunkel, T.A., Pommier, Y. and Jinks-Robertson, S. (2011) Mutagenic processing of ribonucleotides in DNA by yeast topoisomerase I. *Science*, **332**, 1561-1564.
15. Sekiguchi, J. and Shuman, S. (1997) Site-specific ribonuclease activity of eukaryotic DNA topoisomerase I. *Mol Cell*, **1**, 89-97.
16. Klein, H.L. (2017) Genome instabilities arising from ribonucleotides in DNA. *DNA Repair (Amst)*, **56**, 26-32.

17. Chiu, H.C., Koh, K.D., Evich, M., Lesiak, A.L., Germann, M.W., Bongiorno, A., Riedo, E. and Storici, F. (2014) RNA intrusions change DNA elastic properties and structure. *Nanoscale*, **6**, 10009-10017.
18. Koh, K.D., Balachander, S., Hesselberth, J.R. and Storici, F. (2015) Ribose-seq: global mapping of ribonucleotides embedded in genomic DNA. *Nat Methods*, **12**, 251-257, 253 p following 257.
19. Reijns, M.A.M., Kemp, H., Ding, J., de Proce, S.M., Jackson, A.P. and Taylor, M.S. (2015) Lagging-strand replication shapes the mutational landscape of the genome. *Nature*, **518**, 502-506.
20. Clausen, A.R., Lujan, S.A., Burkholder, A.B., Orebaugh, C.D., Williams, J.S., Clausen, M.F., Malc, E.P., Mieczkowski, P.A., Fargo, D.C., Smith, D.J. *et al.* (2015) Tracking replication enzymology in vivo by genome-wide mapping of ribonucleotide incorporation. *Nat Struct Mol Biol*, **22**, 185-191.
21. Daigaku, Y., Keszthelyi, A., Muller, C.A., Miyabe, I., Brooks, T., Retkute, R., Hubank, M., Nieduszynski, C.A. and Carr, A.M. (2015) A global profile of replicative polymerase usage. *Nat Struct Mol Biol*, **22**, 192-198.
22. Chaisson, M.J. and Tesler, G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, **13**, 238.
23. Sovic, I., Sikic, M., Wilm, A., Fenlon, S.N., Chen, S. and Nagarajan, N. (2016) Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat Commun*, **7**, 11307.
24. Lin, H.N. and Hsu, W.L. (2017) Kart: a divide-and-conquer algorithm for NGS read alignment. *Bioinformatics*, **33**, 2281-2287.
25. Zatopek, K.M., Potapov, V., Maduzia, L.L., Alpaslan, E., Chen, L., Evans, T.C., Jr., Ong, J.L., Ettwiller, L.M. and Gardner, A.F. (2019) RADAR-seq: A RAre DAmage and Repair sequencing method for detecting DNA damage on a genome-wide scale. *DNA Repair (Amst)*, **80**, 36-44.
26. Iida, T., Iida, N., Sese, J. and Kobayashi, T. (2021) Evaluation of repair activity by quantification of ribonucleotides in the genome. *Genes Cells*, **26**, 555-569.
27. Zhou, Z.X., Lujan, S.A., Burkholder, A.B., Garbacz, M.A. and Kunkel, T.A. (2019) Roles for DNA polymerase delta in initiating and terminating leading strand DNA replication. *Nat Commun*, **10**, 3992.
28. Gombolay, A.L., Vannberg, F.O. and Storici, F. (2019) Ribose-Map: a bioinformatics toolkit to map ribonucleotides embedded in genomic DNA. *Nucleic Acids Res*, **47**, e5.
29. Gombolay, A.L. and Storici, F. (2021) Mapping ribonucleotides embedded in genomic DNA to single-nucleotide resolution using Ribose-Map. *Nat Protoc*, **16**, 3625-3638.
30. Gombolay, A.L. (2021) Ribose-Map: A bioinformatics toolkit for ribonucleotide sequencing experiments. *Software Impacts*.
31. Langdon, W.B. (2015) Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Min*, **8**, 1.
32. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M. *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**.

33. Smith, T., Heger, A. and Sudbery, I. (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res*, **27**, 491-499.
34. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W. and Noble, W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*, **37**, W202-208.
35. Thankaswamy-Kosalai, S., Sen, P. and Nookaew, I. (2017) Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics*, **109**, 186-191.
36. Keskin, H., Shen, Y., Huang, F., Patel, M., Yang, T., Ashley, K., Mazin, A.V. and Storici, F. (2014) Transcript-RNA-templated DNA recombination and repair. *Nature*, **515**, 436-439.
37. Storici, F., Bebenek, K., Kunkel, T.A., Gordenin, D.A. and Resnick, M.A. (2007) RNA-templated DNA repair. *Nature*, **447**, 338-341.
38. Stuckey, S., Mukherjee, K. and Storici, F. (2011) In vivo site-specific mutagenesis and gene collage using the delitto perfetto system in yeast Saccharomyces cerevisiae. *Methods Mol Biol*, **745**, 173-191.
39. Chon, H., Sparks, J.L., Rychlik, M., Nowotny, M., Burgers, P.M., Crouch, R.J. and Cerritelli, S.M. (2013) RNase H2 roles in genome integrity revealed by unlinking its activities. *Nucleic Acids Res*, **41**, 3130-3143.
40. Balachander, S., Yang, T., Newnam, G., El-Sayed, W.M.M., Koh, K.D. and Storici, F. (2019) Capture of Ribonucleotides in Yeast Genomic DNA Using Ribose-Seq. *Methods Mol Biol*, **2049**, 17-37.
41. Mortimer, R.K. and Johnston, J.R. (1986) Genealogy of principal strains of the yeast genetic stock center. *Genetics*, **113**, 35-43.
42. Brachmann, C.B., Davies, A., Cost, G.J., Caputo, E., Li, J., Hieter, P. and Boeke, J.D. (1998) Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast*, **14**, 115-132.
43. Rudin, N. and Haber, J.E. (1988) Efficient repair of HO-induced chromosomal breaks in Saccharomyces cerevisiae by recombination between flanking homologous sequences. *Mol Cell Biol*, **8**, 3918-3928.
44. Morrison, A., Bell, J.B., Kunkel, T.A. and Sugino, A. (1991) Eukaryotic DNA polymerase amino acid sequence required for 3'----5' exonuclease activity. *Proc Natl Acad Sci U S A*, **88**, 9473-9477.
45. Chon, H., Vassilev, A., DePamphilis, M.L., Zhao, Y., Zhang, J., Burgers, P.M., Crouch, R.J. and Cerritelli, S.M. (2009) Contributions of the two accessory subunits, RNASEH2B and RNASEH2C, to the activity and properties of the human RNase H2 complex. *Nucleic Acids Res*, **37**, 96-110.
46. Sambrook, J. and Russell, D.W. (2006) Isolation of DNA fragments from polyacrylamide gels by the crush and soak method. *CSH Protoc*, **2006**.
47. Diamond, T.L., Roshal, M., Jamburuthugoda, V.K., Reynolds, H.M., Merriam, A.R., Lee, K.Y., Balakrishnan, M., Bambara, R.A., Planelles, V., Dewhurst, S. *et al.* (2004) Macrophage tropism of HIV-1 depends on efficient cellular dNTP utilization by reverse transcriptase. *J Biol Chem*, **279**, 51545-51553.

48. Fromentin, E., Gavegnano, C., Obikhod, A. and Schinazi, R.F. (2010) Simultaneous quantification of intracellular natural and antiretroviral nucleosides and nucleotides by liquid chromatography-tandem mass spectrometry. *Anal Chem*, **82**, 1982-1989.
49. Wagih, O. (2017) ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*, **33**, 3645-3647.
50. Buels, R., Yao, E., Diesh, C.M., Hayes, R.D., Munoz-Torres, M., Helt, G., Goodstein, D.M., Elsik, C.G., Lewis, S.E., Stein, L. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol*, **17**, 66.
51. Wanrooij, P.H., Engqvist, M.K.M., Forslund, J.M.E., Navarrete, C., Nilsson, A.K., Sedman, J., Wanrooij, S., Clausen, A.R. and Chabes, A. (2017) Ribonucleotides incorporated by the yeast mitochondrial DNA polymerase are not repaired. *Proc Natl Acad Sci U S A*, **114**, 12466-12471.
52. de Zamaroczy, M. and Bernardi, G. (1986) The GC clusters of the mitochondrial genome of yeast and their evolutionary origin. *Gene*, **41**, 1-22.
53. Cerritelli, S.M., Frolova, E.G., Feng, C., Grinberg, A., Love, P.E. and Crouch, R.J. (2003) Failure to produce mitochondrial DNA results in embryonic lethality in Rnaseh1 null mice. *Mol Cell*, **11**, 807-815.
54. El Hage, A., Webb, S., Kerr, A. and Tollervey, D. (2014) Genome-wide distribution of RNA-DNA hybrids identifies RNase H targets in tRNA genes, retrotransposons and mitochondria. *PLoS Genet*, **10**, e1004716.
55. Bubeck, D., Reijns, M.A., Graham, S.C., Astell, K.R., Jones, E.Y. and Jackson, A.P. (2011) PCNA directs type 2 RNase H activity on DNA replication and repair substrates. *Nucleic Acids Res*, **39**, 3652-3666.
56. Berglund, A.K., Navarrete, C., Engqvist, M.K., Hoberg, E., Szilagyi, Z., Taylor, R.W., Gustafsson, C.M., Falkenberg, M. and Clausen, A.R. (2017) Nucleotide pools dictate the identity and frequency of ribonucleotide incorporation in mitochondrial DNA. *PLoS Genet*, **13**, e1006628.
57. Moon, A.F., Pryor, J.M., Ramsden, D.A., Kunkel, T.A., Bebenek, K. and Pedersen, L.C. (2017) Structural accommodation of ribonucleotide incorporation by the DNA repair enzyme polymerase Mu. *Nucleic Acids Res*, **45**, 9138-9148.
58. DeLucia, A.M., Grindley, N.D. and Joyce, C.M. (2003) An error-prone family Y DNA polymerase (DinB homolog from Sulfolobus solfataricus) uses a 'steric gate' residue for discrimination against ribonucleotides. *Nucleic Acids Res*, **31**, 4129-4137.
59. Cho, J.E. and Jinks-Robertson, S. (2018) Topoisomerase I and Genome Stability: The Good and the Bad. *Methods Mol Biol*, **1703**, 21-45.
60. Rydberg, B. and Game, J. (2002) Excision of misincorporated ribonucleotides in DNA by RNase H (type 2) and FEN-1 in cell-free extracts. *Proc Natl Acad Sci U S A*, **99**, 16654-16659.
61. Lietard, J., Damha, M.J. and Somoza, M.M. (2019) Large-Scale Photolithographic Synthesis of Chimeric DNA/RNA Hairpin Microarrays To Explore Sequence Specificity Landscapes of RNase HII Cleavage. *Biochemistry*, **58**, 4389-4397.
62. Storici, F., Lewis, L.K. and Resnick, M.A. (2001) In vivo site-directed mutagenesis using oligonucleotides. *Nat Biotechnol*, **19**, 773-776.
63. Ralser, M., Kuhl, H., Ralser, M., Werber, M., Lehrach, H., Breitenbach, M. and Timmermann, B. (2012) The Saccharomyces cerevisiae W303-K6001 cross-platform

genome sequence: insights into ancestry and physiology of a laboratory mutt. *Open Biol*, **2**, 120093.

64. Garfinkel, D.J., Nyswaner, K.M., Stefanisko, K.M., Chang, C. and Moore, S.P. (2005) Ty1 copy number dynamics in Saccharomyces. *Genetics*, **169**, 1845-1857.

65. Vengrova, S. and Dalgaard, J.Z. (2004) RNase-sensitive DNA modification(s) initiates S. pombe mating-type switching. *Genes Dev*, **18**, 794-804.