THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Genome-wide mapping of embedded ribonucleotides and other noncanonical nucleotides using emRiboSeq and EndoSeq

OPEN ACCESS

Genome-wide mapping of embedded ribonucleotides and other non-canonical nucleotides using emRiboSeq and EndoSeq

James Ding[1,2] - james.ding@postgrad.manchester.ac.uk

Martin S. Taylor[1] – martin.taylor@igmm.ed.ac.uk

Andrew P. Jackson[1] – andrew.jackson@igmm.ed.ac.uk

Martin A. M. Reijns[1]* – martin.reijns@igmm.ed.ac.uk

[1]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, The University of Edinburgh, Edinburgh, UK

[2]Current address: Arthritis Research UK Centre for Genetics and Genomics, Institute for Inflammation and Repair, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK

*Correspondence should be addressed to M. A. M. R.

tel +44 131 651 8633

fax +44 131 651 8800

Websites:

Andrew P. Jackson: http://www.hgu.mrc.ac.uk/people/a.jackson.html

Martin S. Taylor: http://taylor.igmm.ed.ac.uk/

KEYWORDS

ABSTRACT

Ribonucleotides are the most common non-canonical nucleotides incorporated into the genome of replicating cells. They are efficiently removed by ribonucleotide excision repair initiated by Ribonuclease (RNase) H2 cleavage. In the absence of RNase H2, such embedded ribonucleotides can be used to track DNA polymerase activity *in vivo*. To determine their precise location in *Saccharomyces cerevisiae* we developed embedded Ribonucleotide Sequencing (emRiboSeq), which uses recombinant RNase H2 to selectively create ligatable 3'-hydroxyl groups, in contrast to alternative methods that utilize alkaline hydrolysis. EmRiboSeq allows reproducible, strand-specific and potentially quantitative detection of embedded ribonucleotides at single-nucleotide resolution. This protocol can be adapted for the genome-wide mapping of other non-canonical bases by replacing RNase H2 with specific nicking endonucleases, a method we term Endonuclease Sequencing (EndoSeq). With the protocol taking <5 days to complete, these methods allow the *in vivo* study of DNA replication and repair, including the identification of replication origins and termination regions.

INTRODUCTION

Ribonucleotides are frequently incorporated into DNA by DNA polymerases[1-9] a subject of increasing interest discussed in a number of recent reviews[10-15]. Under normal conditions the cell can use their transient presence to its advantage: ribonucleotide incorporation by polymerase mu (Pol-μ) is thought to be beneficial for double-strand break repair by nonhomologous end joining[7,16], whereas ribonucleotide incorporation on the leading strand by Pol-ε promotes mismatch repair[17,18]. However, failure to remove embedded ribonucleotides, i.e. absence of RNase H2, results in genome instability[6,8,19-25]. This may be linked to heritable autoinflammatory disorders such as Aicardi–Goutières syndrome[21,25-27] and Systemic lupus erythematosus[21], and could have relevance for the neurodegenerative disorder Ataxia with Oculomotor Apraxia 1[28,29]. To map the genome-wide distribution of these embedded ribonucleotides we developed a next generation sequencing (NGS) method, emRiboSeq[30], which relies on cleavage of the phosphodiester bond directly 5' of a DNA-embedded ribonucleotide with recombinant RNase H2. In the cell, cleavage by RNase H2 initiates the removal of misincorporated ribonucleotides, a process termed ribonucleotide excision repair

(RER)[31]. Therefore, the use of RNase H2 deficient cells is essential to allow mapping of ribonucleotides. We successfully used this approach to establish that ribonucleotide incorporation patterns in the *Saccharomyces cerevisiae* genome are non-random, and to map the contribution of the replicative DNA polymerases[30] (Pol-ε, Pol-δ and Pol-α) using point mutations that alter their propensity to incorporate ribonucleotides[6,18,32]. In addition, this accurately identified origins of replication and termination regions at high resolution. In parallel three other groups developed different methods to map embedded ribonucleotides, yielding similar findings[33-36] (see also further comparison below).

Overview of emRiboSeq

Here we describe the details of the emRiboSeq protocol, which could potentially be applied to any organism, although we emphasise its application to *S. cerevisiae* with its relatively compact genome. In brief, the protocol (**Fig. 1**) includes isolation and sonication of genomic DNA, followed by end repair, deoxyadenosine-tailing and ligation of fragments to adapters with a deoxythymidine overhang on one end and a non-ligatable 2',3'-dideoxycytidine on the other. In order to minimise the contribution of nicks or double strand breaks to background, pre-existing 3'-hydroxyl (3'-OH) groups are blocked using terminal transferase and 2',3'-dideoxyadenosine-triphosphate (ddATP). The library of ligated fragments is then treated with RNase H2 (for emRiboSeq or another appropriate endonuclease for EndoSeq), generating 3'-OH and 5'-phosphate (5'-P) groups. Phosphates are removed to increase subsequent adapter ligation efficiency, followed by denaturation and annealing of single stranded fragments to double stranded adapters with a 3' random hexamer overhang. This allows ligation of the adapter to only those fragments with free 3'-OH ends. A conjugated biotin molecule in the random hexamer containing strand of the second adapter allows ligated fragments to be captured. Elution of the full length non-biotinylated strand is followed by second strand synthesis to produce a library that requires only size selection prior to sequencing.

Applications of the method

Genome-wide mapping of embedded ribonucleotides can offer insights into DNA replication and repair processes, as it allows *in vivo* tracking of DNA polymerase activities[30,33-35]. In addition to the potential for identifying replication origins and termination regions, emRiboSeq will facilitate the study of the causes and consequences of ribonucleotide misincorporation. Furthermore, to examine the distribution of other non-canonical nucleotides only a single step in the emRiboSeq protocol needs to be changed. In EndoSeq, appropriate nicking endonucleases that selectively cleave and generate 3'-OH ends can be used in place of RNase H2. This could be applied to lesions such as apurinic/apyrimidinic (AP) sites (e.g. using AP endonuclease 1, APE1), embedded deoxyinosines (e.g. using E. coli Endo V) or UV-induced damage such as cyclobutane pyrimidine dimers (CPDs, e.g. using UV damage endonuclease, UVDE). In other cases, multiple enzymes may be required to generate the required 3'-OH ends, such as the detection of embedded deoxyuracil by treatment with uracil-DNA glycosylase (UDG) to generate AP sites, which could then be cleaved by AP lyase enzymes (e.g.APE1 or Endo IV). Endonucleases that generate double strand breaks can also be used, although in this case strand-specific information is lost. We established proof of principle of the EndoSeq approach using sequence-specific endonucleases[30]. Together, emRiboSeq and EndoSeq add important tools to the molecular biology toolbox used to study DNA replication, damage and repair.

Comparison with other methods

Three other methods that allow embedded ribonucleotides to be mapped were developed independently[33-35], highlighting the current interest in studying ribonucleotide incorporation and *in vivo* polymerase contributions. A useful review and comparison of all four methods, emRiboSeq[30], HydEn-seq[33], ribose-seq[35,36] and Pu-seq[34], was made by Jinks-Robertson & Klein[37]. Whereas emRiboSeq uses enzymatic hydrolysis of embedded ribonucleotides (generating 5'-P and 3'-OH ends), the other three methods make use of alkaline hydrolysis of ribonucleotide-containing DNA (generating 5'-OH and 2',3'-cyclic phosphates). The advantage of chemical hydrolysis is the simplicity of the procedure; a disadvantage is the fact that abasic sites are also prone to hydrolysis[38], which may cause higher background. Post-hydrolysis, both emRiboSeq and ribose-seq capture the strand upstream of the nick, whereas HydEn-seq and Pu-seq capture the downstream fragment. The principle of specifically capturing these ends differs for each method. HydEn-seq phosphorylates the 5'-OH ends and uses this to ligate on a first adapter; a second random hexamer-containing adapter is used for second strand synthesis, followed by PCR amplification. Pu-seq uses random priming to perform second strand synthesis in the presence of deoxyuridine (dU) tri-phosphate, generating blunt ends that allow adapter ligation. The dU-containing strand is then degraded and the original strand, which now contains adapters, amplified. Ribose-seq initially ligates a two-sided adapter to the 5'-end of fragments, followed by circular ligation using *Arabidopsis thaliana* tRNA ligase (AtRNL), an enzyme that specifically allows single stranded ligation to 2',3'-cyclic phosphates. Subsequent degradation of linear DNA allows specific amplification of circular fragments. Each of the four methods is likely to introduce certain biases, which future in-depth side-by-side comparison will be useful to delineate. Notably, with the exception of emRiboSeq each of these library preparation methods uses PCR pre-amplification, itself a potential source of bias.

Although these four methods achieve broadly the same outcome, the investigator may wish to consider the pros and cons of each before deciding which methodology to employ. The main strengths of emRiboSeq are that it is a robust, reproducible and validated technique, and the absence of pre-sequencing amplification means that no de-duplication is necessary during bioinformatics analysis, resulting in superior read depth. In addition, the lack of pre-amplification also gives it the potential to be used as a quantitative method, although the precision of quantitation will depend on the efficiency of all of the enzymatic steps in the library preparation (each should be performed to completion and under conditions that allow high specificity), as well as the quality of the starting material.

Potential disadvantages of our protocol are its relative length and its dependence on enzymatic cleavage, which could introduce enzyme-specific bias. In our published work we successfully employed recombinant human RNase H2 purified in our own laboratory. More recently, we have also generated emRiboSeq libraries using a commercially available preparation of RNase HII (New England Biolabs) with preliminary analyses showing that these produce similar data, indicating that type 2 RNase H enzymes from other sources can be used. However, careful experimental design and interpretation when using nucleases is important, and potential site specific variation in activity or risk of low level contaminating nuclease activity in both commercial and lab generated enzymes should be kept in mind. Optimal conditions for emRiboseq are those under which all ribonucleotides are cleaved whilst minimising non-specific cleavage. While the use of alkaline hydrolysis in other methods may result in detection of sites additional to ribonucleotides (e.g. abasic residues), this will

also circumvent potential problems such as sequence-specific biases in cleavage or contaminating nucleases.

Our method can be adapted to use endonucleases other than RNase H2 to study different non-canonical nucleotides. The EndoSeq method can be compared to Excision-seq, an alternative method that makes use of enzymatic cleavage at modified bases to determine their genome-wide distribution[39]. However, this method requires the modifications under investigation to be present at a high enough frequency to generate sufficiently small fragments. This is not a limitation for EndoSeq, as fragment size is determined by initial sonication, although the presence of modifications at too high a frequency could be problematic as discussed below.

Experimental Design

Sequencing platform. This protocol was developed using the Ion Torrent™ NGS platform (Life Technologies), but can be adapted to other platforms by changing the adapter sequences, which in our case were designed for compatibility with the capture and sequencing primers used for Ion Torrent™. The optimum length of fragments within a NGS library also differs according to the platform used and consequently the size selection range (Step 75) would also require adjustment. When applying our method to organisms with a substantially larger genome than that of *S. cerevisiae* read depth may become an issue, in which case the use of a sequencing methodology that generates larger numbers of reads will need to be considered.

EndoSeq: Use of alternative endonucleases. When substituting RNase H2 with alternative endonucleases to determine the genomic distribution of non-canonical nucleotides other than ribonucleotides there are several important considerations. The specificity of the endonuclease is crucial and where possible non-specific activity should be carefully avoided. It is therefore recommended that optimum conditions be established for Step 46. High frequency cleavage (i.e. >1 cut every 400 basepairs, bp) is not desirable as fragments that are cleaved multiple times are likely to be lost during final size selection. In this case, partial digestion may be possible, but this is likely to introduce bias, with sites that are preferentially cleaved overrepresented. Cleavage frequency of a particular nicking endonuclease can be assessed by alkaline gel electrophoresis of enzyme-treated DNA, followed by densitometry[8,30]. Existing estimates of ribonucleotide incorporation rates by *S. cerevisiae* polymerases[5,6,30] mean that multiple cleavage of sonicated fragments should be a rare event in the emRiboSeq protocol, although the occurrence of closely spaced ribonucleotides at specific positions cannot be ruled out. EmRiboSeq would not detect these with only the most 5' site captured.

Starting material. Isolation of genomic DNA (gDNA) is performed under conditions that minimise the generation of nicks. Where the endonuclease substrate is rapidly removed by endogenous processes it will be necessary to isolate gDNA from cells in which these processes are inactive, either through genetic (knock-out or knock-down) or chemical means (inhibitors). Although other, less efficient mechanisms exist for the removal of embedded ribonucleotides[40-42], the use of RNase H2 null *S. cerevisiae* in emRiboSeq was shown to be sufficient to allow mapping of ribonucleotides misincorporated by replicative DNA polymerases[30]. For similar experiments, the use of strains in which at least one of the genes encoding the RNase H2 subunits, *RNH201*, *RNH202* or *RNH203*, is

deleted is therefore recommended. We have used both mid-log phase ($A_{600\,nm}$ = 0.5) and stationary phase cultures ($A_{600\,nm}$ = 5 to 6) with similar results.

The emRiboSeq protocol was implemented in *S. cerevisiae*, which offers a number of advantages including ease of genetic engineering and a small, well defined genome. Additionally, RNase H2 deficient yeast proliferates at a normal rate, unlike RNase H2 null mammalian cells, although absence of RNase H2 activity in budding yeast does cause mutation rate increase and genome instability[6,19,24,42-45].

If using gDNA from a source other than *S. cerevisiae*, Steps 1-15 should be replaced with an appropriate protocol for isolating gDNA in which cleavage at embedded ribonucleotides is avoided prior to Step 46. In our protocol this was ensured by performing RNase treatment under high salt conditions (Step 15). Because our protocol does not use pre-amplification a large quantity of starting material is crucial: we recommend starting with 18 μg of DNA, so that 5 μg can be used as input for end repair (Step 32), although we have been successful with input amounts as low as 1 μg.

Controls. It is recommended that control experiments are performed when first carrying out this protocol, especially when using a new endonuclease. These include the preparation of an 'endonuclease negative' library where endonuclease treatment (Step 46) is omitted. This should result in the absence of 3'-OH groups available for ligation to the second adapter (Steps 52-54) and, therefore, a lack of amplification when using PCR to verify the quality of the library prior to sequencing (Steps 80-83). A 'positive' control experiment that allows the specificity of generated libraries to be estimated can be achieved through substitution of the endonuclease for one with well-defined sequence specificity. We originally performed such controls[30] using BciVI (New England Biolabs), which cleaves both strands, and the nicking endonuclease Nb.BtsI (New England Biolabs). The preparation of control libraries using sequence specific endonucleases provides the additional advantage of enabling the bioinformatic processing of sequence data to be optimised using a library with highly predictable results. It should, however, be noted that due to the sensitivity of the EndoSeq method even low levels of star activity can be observed.

MATERIALS

Reagents

For gDNA isolation from *S. cerevisiae*:
- RNase H2 deficient *S. cerevisiae*, such as SNM106 (a kind gift from T. Kunkel) or its derivatives[6,30] (5 $A_{600\,nm}$ unit pellets can be stored at -80°C).
- Glass beads, Ø 0.4-0.6 mm (Sartorius, cat. no. BBI-8541701)
- 10% (vol/vol) Triton X-100 (Sigma, cat. no. T8787) **CAUTION** Triton X-100 is corrosive. Wear personal protective equipment.
- 20% (wt/vol) Sodium dodecyl sulfate (SDS, Sigma, cat. no. 05030) **CAUTION** SDS is corrosive and toxic. Wear personal protective equipment and avoid inhalation.
- Sodium Chloride (NaCl, Sigma, cat. no. S3014)
- Tris(hydroxymethyl)methylamine (Tris base, Sigma, cat. no. T4661)

- Boric acid (Sigma, cat. no. B7901)
- Ethylenediaminetetraacetic acid (EDTA, Sigma, cat. no. E6758)
- Phenol pH 7.9 (Sigma, cat. no. P4557) **CAUTION** Phenol is corrosive and toxic. Wear personal protective equipment and handle it in a fume hood.
- Phenol:chloroform:isoamylalcohol (25:24:1) (Sigma, cat. no. P3803) **CAUTION** Phenol:chloroform:isoamylalcohol is corrosive and toxic. Wear personal protective equipment and handle it in a fume hood.
- Chloroform (Sigma, cat. no. C2432) **CAUTION** Chloroform is toxic. Wear personal protective equipment and handle it in a fume hood.
- 100% Ethanol (EtOH, Sigma, cat. no. E7023) **CAUTION** Ethanol is highly flammable. Handle away from potential sources of ignition.
- Nuclease free distilled water (H_2O, Life-Technologies, cat. no. 10977)
- RNase, DNAse-free (Roche, cat. no. 11 119 915 001)
- Poly(ethylene glycol) 8000 ($PEG_{8,000}$, Sigma, cat. no. 89510)
- Agencourt Ampure XP beads (Beckman Coulter, cat. No. A63880)

For library preparation from gDNA
- Agarose (electrophoresis grade, Sigma, cat. no. A9539)
- SYBR® Gold nucleic acid stain, 10,000x in DMSO (Life Technologies, cat. No. S-11494)
- 100 bp DNA ladder (e.g. Promega, cat. no. G210A, supplied with Blue/Orange 6X Loading Dye)
- NEBNext® end repair module (New England Biolabs, cat. no. E6050)
- NEBNext® dA-tailing module (New England Biolabs, cat. no. E6053)
- NEBNext® Quick Ligation module (New England Biolabs, cat. no. E6056)
- Custom oligonucleotides; ours were synthesised by Eurogentec and purified using manufacturer recommended methods (Table 1).
- Terminal Transferase, supplied with 10x reaction buffer and 2.5 mM $CoCl_2$ (New England Biolabs, cat. no. M0315)
- 10 mM ddATP, Sequencing Grade Na-salt pH 8.3 (Roche, cat. no. 03 732 738 001)
- RNase HII (New England Biolabs, cat. no. M0288) or recombinant human RNase H2 (purified as previously described[46]) for emRiboSeq; other recombinant (nicking) endonuclease of interest for EndoSeq. CRITICAL: alternative sources of recombinant type 2 RNase H enzymes may be used
- 10% (wt/vol) Bovine Serum Albumin Fraction V (BSA, Roche, cat. no. 10 735 086 001)
- Magnesium chloride ($MgCl_2$, Sigma, cat. no. M2670)
- Nb.BtsI, Supplied with 10x CutSmart® buffer (New England Biolabs, cat. no. R0707)
- BciVI, Supplied with 10x CutSmart® buffer (New England Biolabs, cat. no. R0596)
- Shrimp Alkaline Phosphatase (SAP), supplied with 10x reaction buffer (Affymetrix USB®, cat. no. 70092Z)
- Dynabeads® M-280 Streptavidin (Life Technologies, cat. no. 11205D)
- Tri-sodium citrate (Sigma, cat. no. C8532)
- Glycogen (Roche, cat. no. 10 901 393 001)
- Sodium acetate (NaOAc, Sigma, cat. no. S2889)

- Sodium hydroxide (NaOH, Sigma, cat. no. 38215) **CAUTION** Sodium hydroxide is corrosive. Wear personal protective equipment.
- Phusion Flash High-Fidelity PCR Master Mix (Thermo Scientific, cat. no. F-548)
- E-Gel® SizeSelect™ Agarose Gels, 2% (Life Technologies, cat. no. G6610-02)
- Agilent High Sensitivity DNA Kit (Agilent Technologies, cat. no. 5067-4626)
- Potassium Chloride (KCl, Sigma, cat. no. P9541)
- Ion 318™ Chip Kit v2 (Life Technologies, cat. no. 4484534) or Ion PI™ Chip Kit v2 (Life Technologies, cat. no. 4482413)
- Ion PGM™ Template OT2 200 Kit (Life Technologies, cat. no. 4480974) or Ion PI™ Template OT2 200 Kit v2 (Life Technologies, cat. no. 4485146)
- Ion PGM™ Sequencing 200 Kit v2 (Life Technologies, cat. no. 4482006) or Ion PI™ sequencing kit (Life Technologies, cat. no. 4485149). CRITICAL: These Ion Torrent sequencing products are under constant development. We therefore suggest using the most appropriate equivalent recommended products.

Equipment

- Nuclease-free 1.5 ml tubes (Starlab, cat. no. E1415-1500)
- Vortex MS 3 basic (IKA, cat. no. 0003617000)
- Table-top centrifuge (Eppendorf, cat. no. 5424 R)
- Agarose gel electrophoresis equipment
- DynaMag™-2 Magnet, magnetic stand (Life Technologies, cat. no. 12321D)
- Nanodrop spectrophotometer (Thermo Scientific)
- Bioruptor® Plus with 1.5 ml tube holder (Diagenode, cat. no. B01020001)
- Rotator (IKA, cat. no. 0004015000)
- Thermocycler (Biorad, cat. no. C1000)
- E-Gel® iBase™ Power System (Life Technologies, cat. no. G6400)
- 2100 Electrophoresis Bioanalyzer Instrument (Agilent Technologies, cat. no. G2939AA)
- Ion OneTouch™ 2 System (Life Technologies, cat. no. 4474779)
- Ion PGM™ Sequencer (Life Technologies, cat. no. 4462921) or Ion Proton™ Sequencer (Life Technologies, cat. no. 4476610)

Reagent Setup

Solutions
CRITICAL: Prepare all of the following solutions using nuclease-free water.

- Lysis buffer (2% Triton X-100, 1% SDS, 0.5 M NaCl, 10 mM Tris, 1 mM EDTA, pH8.0; store at room temperature (RT) for up to six months)
- TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0; store at RT for up to six months)
- 0.5x TBE (44.5 mM Tris, 44.5 mM Boric acid, 1 mM EDTA, pH 8.0; store at RT for up to six months)
- 0.5 M NaCl (store at RT for up to six months)
- 75% (vol/vol) EtOH (store at RT for up to six months)
- 3 M NaOAc pH 5.2 (store at RT for up to six months)

- 40% (wt/vol) $PEG_{8,000}$, 1.25 M NaCl (filter sterilise using a 0.22 µm filter and store at 4°C for up to six months; warm to RT before use)
- 5x Oligonucleotide annealing buffer (300 mM KCl, 250 mM Tris, pH 8.0; store at RT for up to six months)
- 2x RNase H2 reaction buffer for use with recombinant human RNase H2[46] (120 mM KCl, 100 mM Tris-HCl pH 8.0, 20 mM $MgCl_2$, 0.02% BSA, 0.02% Triton X-100; store at 4°C for up to six months)
- 0.15 M NaOH (make up fresh)
- Saline sodium citrate (SSC) (0.15 M NaCl, 0.015 M sodium citrate, pH 7.0; store at RT for up to six months)
- 1x Bind and Wash buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 2 M NaCl; store at RT for up to six months)
- 2x Bind and Wash buffer (20 mM Tris-HCl pH 7.5, 2 mM EDTA, 4 M NaCl; store at RT for up to six months)

Double stranded adapters

Prepare the first adapter (dstrP1) by combining oligonucleotides trP1-top and trP1-bottom, and the second adapter (dsA) by combining A-top and A-bottom (see Table 1 for oligonucleotide sequences).

In each case, prepare 100 µl 40 µM double stranded adapter as follows: Combine 40 µl of each single stranded oligonucleotide (100 µM) with 20 µl 5x oligonucleotide buffer, incubate at 95°C for 5 min in a thermocycler, before removing and allowing to cool gradually to RT (store at -80°C for up to one year).

Table 1: Oligonucleotides used in library preparation and quality control. All sequences are given 5' to 3'.

Oligonucleotide	Sequence
trP1-top	CCTCTCTATGGGCAGTCGGTGAT-phosphorothioate-T
trP1-bottom	phosphate-ATCACCGACTGCCCATAGAGAGG-dideoxyC
A-top	phosphate-CTGAGTCGGAGACACGCAGGGATGAGATG-dideoxyG
A-bottom	biotin-CCATCTCATCCCTGCGTGTCTCCGACTCAGNNNNNNN-C3-phosphoramidite
primer A	CCATCTCATCCCTGCGTGTCTCCGAC
primer trP1	CCTCTCTATGGGCAGTCGGTGATT

PROCEDURE

gDNA isolation from yeast TIMING 2-3 h

1   Add 200 µl of lysis buffer and ~0.2 ml of glass beads to the pellet from a yeast culture (5 $A_{600\ nm}$ units), in a 1.5 ml tube. To isolate sufficient gDNA for a particular sample, we usually carry out Steps 1-14 with 8 tubes per sample (i.e. a total of 40 $A_{600\ nm}$ units). These can be pooled in Step 14, and subsequent steps carried out for multiple samples at once, making appropriate adjustments for the larger volume. Amounts given in Steps 1-23 are for a single tube containing 5 $A_{600\ nm}$ units.

2 Add 200 μl of phenol and vortex for 2 min (continuous mixing at 3,000 rpm).

3 Add 200 μl of TE and vortex for 1 min (continuous mixing at 3,000 rpm).

4 Centrifuge at 5,000 g for 5 min, 4°C.

5 Transfer the aqueous phase to a new 1.5 ml tube, add 400 μl of phenol:chloroform:isoamylalcohol (25:24:1) and vortex for 30 s.

6 Centrifuge for 5 min at 5,000 g, 4°C.

7 Add 400 μl of chloroform and vortex for 30 s.

8 Centrifuge at 5,000 g for 2 min, 4°C.

9 Transfer the aqueous phase to a new 1.5 ml tube, add 1 ml of ice-cold 100% EtOH and mix thoroughly by repeated inversion.

10 Centrifuge at ≥ 13,000 g for 10 min, 4°C and remove the supernatant.

11 Wash the pellet with 0.5 ml 75% EtOH.

12 Centrifuge at ≥ 13,000 g for 2 min, 4°C and remove the supernatant.

13 Air-dry the pellet at RT.

14 Resuspend the pellet containing total nucleic acids in 50 μl of 0.5 M NaCl. CRITICAL STEP: Replicates can be pooled at this stage (making appropriate adjustments for the larger volume in subsequent steps).

15 Add 1 μl of DNase-free RNase (0.5 μg/μl) and incubate for 1 h at RT.

DNA purification

16 Add 50 μl of Ampure XP after allowing it to equilibrate to RT for at least 5 min (1x volume).

17 Mix thoroughly and incubate for >1 min at RT. If necessary centrifuge briefly (1,500 g for 1 s at RT) to collect droplets.

18 Place on a magnetic stand for >1 min. The beads will collect near the magnet and the solution will become clear. Remove and discard the supernatant.

19 Without removing the tube from the stand, wash twice with 500 μl 75% EtOH. Mixing can be achieved by rotating the tube on the stand. Remove and discard the supernatant.

20 Ensure the beads are dry before resuspending in the required volume and diluent as detailed in the next step. This can take a few min at RT.

21 Elute DNA in 50 μl H_2O, by vortexing to resuspend and incubating at RT for 1 min.

22 Using the magnetic stand to retain the beads, transfer the eluate to a new 1.5 ml tube and discard the beads.

23 Measure the concentration of eluted gDNA using 1.5 μl on a nanodrop spectrophotometer. Usually ~3 μg of DNA can be expected per 5 $A_{600 nm}$ units of culture. CRITICAL STEP: If visualised on an agarose gel an additional band, which migrates at around 4.5 kb, is present. This appears to be RNA resistant to RNase cleavage at 0.5 M NaCl, but sensitive to subsequent degradation by RNase A at low salt concentration.

PAUSE POINT

gDNA can be stored at -80°C for at least 7 d (storage at -80°C is recommended to prevent any unwanted RNase activity on embedded ribonucleotides).

Sonication TIMING 2-3 h

24 Prepare 300 μl aliquots of 10 ng/μl DNA in H_2O in 1.5 ml tubes. Ideally, 6 aliquots (18 μg) of DNA should be sonicated per library, however using as few as 3 aliquots can give enough material to proceed.

25 Sonicate each aliquot to achieve an average length of ~400 bp using a Bioruptor® Plus, following the manufacturer's recommendations.

26 Confirm the distribution of fragments by gel electrophoresis: Load 10 µl per aliquot alongside 100 bp DNA ladder on a 0.5x TBE, 1% agarose gel. Electrophorese and visualise using a nucleic acid stain (e.g. SYBR® Gold). **CRITICAL STEP** The distribution of fragment sizes will have a strong influence on the concentration of the library following size selection. In our experience, an average length of ~400 bp can be achieved with 18 cycles of 30 s sonication and 60 s cooling on the 'high' setting, however this is likely to require optimisation in each lab. Once the desired distribution of fragments is achieved, ethanol precipitate the DNA. Assuming 290 µl per tube, add 1 µl glycogen, 29 µl 3M NaOAc and 800 µl 100% EtOH.

27 Precipitate for 1 h at -20°C and centrifuge at ≥ 13,000 g for 30 min, 4°C and remove the supernatant.

28 Wash the pellet with 0.5 ml 75% EtOH.

29 Centrifuge at ≥ 13,000 g for 2 min, 4°C and remove all supernatant.

30 Air dry pellet, dissolve and pool pellets from all aliquots of each library in a total volume of 100 µl H_2O.

31 Purify DNA by repeating Steps 17-23, using 120 µl Ampure XP (1.2x volume) and eluting in 86.5 µl H_2O.

   CRITICAL STEP The large amount of small fragments generated during sonication can lead to a loss of up to 70% of the initial DNA. Whilst we have prepared libraries using as little as 1 µg, we recommend taking through 5 µg of sonicated DNA to the next steps. Subsequent steps, including reaction and purification conditions, as well as elution volumes, assume 5 µg of sonicated DNA is used and should be scaled accordingly if this is not the case.

PAUSE POINT
Sonicated DNA can be stored at -80°C for at least 7 d.

End repair, dA-tailing and adapter ligation TIMING ~3 h + overnight incubation

32 To 5 µg of sonicated DNA in 85 µl H_2O add 10 µl 10x buffer and 5 µl enzyme mix from the NEBNext® end repair module.

33 Incubate for 1 h at 20°C.

34 Purify DNA by repeating Steps 17-21, using 120 µl Ampure XP (1.2x volume) and eluting in 42 µl H_2O.
   CRITICAL STEP: Do not separate from the beads as they can be reused and do not interfere with the dA-tailing reaction.

35 Add 5 µl 10x buffer and 3 µl Klenow Fragment (3´→ 5´ exo⁻) from the NEBNext® dA-tailing module.

36 Incubate for 1 h at 37°C.

37 Add 17 µl 5 M NaCl and 22.5 µl 40% $PEG_{8,000}$, 1.25 M NaCl.

38 Purify DNA by repeating Steps 17-21, eluting in 20 µl H_2O.
   CRITICAL STEP: Do not separate from the beads as they can be reused and do not interfere with adapter ligation.

39 Add 10 µl 5x buffer, 15 µl 40 µM dstrP1 adapter and 5 µl T4 Quick Ligase from the NEBNext®
    quick ligation module.

40 Incubate overnight or for at least 12 h at 16°C. Ligation for shorter times may be possible, but we
    have not tested this.

41 Add 17 µl 5 M NaCl and 22.5 µl 40% $PEG_{8,000}$, 1.25 M NaCl. CRITICAL STEP: 1x NEBNext Ligation
    buffer contains 6% $PEG_{6,000}$ and may therefore alter the size selection effect, however this
    nucleation procedure has proved sufficient to significantly deplete the adaptors.

42 Purify DNA by repeating Steps 17-22, eluting in 195 µl H_2O.

3' blocking and endonucleolytic treatment TIMING ~4.5 h

43 Add 25 µl 10x Terminal Transferase reaction buffer, 25 µl 2.5 mM $CoCl_2$, 2.5 µl 10 mM ddATP, 2.5
    µl 20 U/µl Terminal Transferase.

44 Incubate at 37°C for 2 h.

45 Purify DNA by repeating Steps 17-23, using 300 µl Ampure XP (1.2x volume) and eluting the 3'
    blocked DNA in 51.5 µl H_2O.

    Use the obtained concentration to scale subsequent reactions. Normally 2-3 µg of DNA remains
    at this point. **CRITICAL STEP** Subsequent steps, including reaction and purification conditions, as
    well as elution volumes, assume 2 µg of DNA is used and should be scaled accordingly if this
    deviates significantly.

46 Endonuclease treatment: This should be optimised for each new endonuclease. The following
    reaction conditions were successfully employed for *Escherichia coli* RNase HII (option A), human
    RNase H2 (option B; note that alternative sources of type 2 RNase H enzymes may be used) and
    Nb.BtsI (option C).

    A RNase HII treatment
        i   To 2 µg of 3' blocked DNA in 50 µl add 10 µl 10x reaction buffer (provided), 50 U
            RNase HII and H_2O up to 100 µl.
        ii Incubate for 2 h at 37°C and then inactivate by incubating for 20 min at 80°C.

    B RNase H2 treatment
        i   To 2 µg of 3' blocked DNA in 50 µl add 50 µl 2x RNase H2 reaction buffer and 20
            pmol RNase H2. CRITICAL STEP: This large excess of enzyme was used to ensure
            complete digestion, but lower amounts are likely sufficient.
        ii Incubate for 2 h at 37°C and then inactivate by incubating for 20 min at 80°C.

    C Nb.BtsI treatment
        i   To 2 µg of 3' blocked DNA in 50 µl add 10 µl 10x CutSmart® buffer, 0.5 µl 10 U/µl
            Nb.BtsI and 39.5 µl H_2O.
        ii Incubate for 1 h at 37°C and then inactivate by incubating for 20 min at 80°C.
            CRITICAL STEP: This enzyme exhibits star activity, particularly when high
            concentrations of enzyme or long incubation times are used.

PAUSE POINT
Endonuclease treated samples can be stored at -20°C for at least 7 d.

Dephosphorylation and second adapter ligation TIMING 2 h + overnight incubation

47 Purify DNA by repeating Steps 17-21, using 120 µl Ampure XP (1.2x volume) and eluting in 85 µl H_2O.
   CRITICAL STEP: Do not separate from the beads as they can be reused and do not interfere with the de-phosphorylation reaction.

48 Add 10 µl 10x SAP reaction buffer and 5 µl 1 U/µl SAP.

49 Incubate at 37°C for 1 h and then inactivate by incubating for 15 min at 65°C.

50 Add 34 µl 5 M NaCl and 45 µl 40% $PEG_{8,000}$, 1.25 M NaCl.

51 Purify DNA by repeating Steps 17-22, eluting in 29 µl H_2O.

52 Incubate the library at 95°C for 5 min in a thermocycler and then snap cool in an icy water bath.
   CRITICAL STEP Libraries must be denatured immediately prior to ligation.

53 Once cooled, immediately add 10 µl 5x buffer, 6 µl 40 µM dsA adapter and 5 µl T4 Quick Ligase from the NEBNext® quick ligation module.

54 Incubate overnight or for at least 12 h at 16°C. CRITICAL STEP: Ligation for shorter times may be possible, but we have not tested this.

55 Purify DNA by repeating Steps 17-22, using 90 µl Ampure XP (1.8x volume) and eluting in 40 µl H_2O.

PAUSE POINT
Eluted DNA fragments can be stored at -20°C for at least 7 d.

Single stranded library preparation, second strand synthesis and size selection TIMING 4-5 h

56 Resuspend Dynabeads® M-280 Streptavidin and transfer 20 µl to a new 1.5 ml tube.

57 Using a magnetic stand to retain the beads, discard the supernatant and replace with 1 ml 1x Bind and Wash buffer.

58 Using a magnetic stand to retain the beads, discard the supernatant and replace with 40 µl 2x Bind and Wash buffer.

59 Add the whole 40 µl of adapter-ligated library from Step 55 and incubate at RT for 15 min on a rotator.

60 Using a magnetic stand to retain the beads, discard the supernatant and replace with 50 µl SSC.

61 Incubate at RT for 5 min on a rotator.

62 Repeat Steps 60-61 to ensure no unbound fragments are carried through.

63 Using a magnetic stand to retain the beads, discard the supernatant and replace with 40 µl 0.15 M NaOH.

64 Incubate at RT for 10 min on a rotator.

65 Using a magnetic stand to retain the beads, transfer the eluate to a new 1.5 ml tube.

66 Repeat the elution: add 40 µl 0.15 M NaOH to the beads from Step 65, incubate at RT for 5 min on a rotator, use a magnetic stand to retain the beads and pool the eluate from this step with that collected in Step 65.

67 Add 120 µl H_2O to the pooled eluate and ethanol precipitate by adding 1 µl glycogen, 20 µl 3 M NaOAc and 550 µl 100% EtOH.

68 Precipitate and wash the DNA by repeating Steps 27-30.

69 Dissolve the pellet containing the single stranded library in 9.5 μl H_2O.

70 Add 0.5 μl 10 μM primer A and 10 μl 2x Phusion Flash Master mix.

71 In a thermocycler, incubate the 20 μl reaction at 98°C for 1 min, followed by 58°C for 30 seconds, and 72°C for 1 min.

72 Purify DNA by repeating Steps 17-22, using 36 μl of Ampure XP (1.8x volume) and eluting in 20 μl H_2O.

73 Set up a 2% E-Gel® SizeSelect™ agarose gel on the iBase™ Power System according to manufacturer's instructions.

74 Load the library directly into the desired well; loading dye is unnecessary. Use 100 bp DNA ladder as a marker.

75 Once the 200 bp fragment has entered the collection well start collecting the library. Pause the electrophoresis every five seconds, mix and remove the liquid in the collection well, transferring it to a 1.5 ml collection tube, then replenish the well (20 μl H_2O). Stop once the 300 bp fragment has completely entered the collection well. A total of ~500 μl is usually collected per sample. Alternatively, the sample can be separated by conventional agarose gel electrophoresis and fragments between 200 and 300 bp purified by gel extraction, e.g. we have successfully applied the QIAquick Gel Extraction Kit (QIAGEN, cat.no. 28704) following the manufacturer's instructions (if this alternative is used, proceed to Step 78).

76 Assuming 500 μl has been collected, divide size-selected DNA into 2 x 250 μl in 1.5 ml tubes and ethanol precipitate by adding 25 μl of 3 M NaOAc and 700 μl 100% EtOH.

77 Precipitate and wash by repeating Steps 27-30.

78 Dissolve each pellet in 20 μl H_2O and purify by repeating Steps 17-22, using 36 μl Ampure XP (1.8x volume) and eluting in 15 μl H_2O.
    CRITICAL STEP Contaminants, such as carried over Ampure beads, can disrupt library quantification using the Bioanalyzer.

PAUSE POINT
Prepared libraries can be stored at -20 °C overnight, or if necessary, for up to 7 d.

Library quality control and sequencing

79 Transfer 0.5 μl to a new tube for further library verification by PCR. Dilute this aliquot 1 in 20, by adding 9.5 μl H_2O.

80 Set up three PCR reactions per library, each using 0.5 μl of the 1 in 20 library dilution, 0.5 μl 10 μM primer trP1, 0.5 μl 10 μM primer A, 8.5 μl H_2O and 10 μl 2x Phusion Flash master mix.

81 Using a thermocycler incubate the reactions at 98°C for 10 sec, followed by 15 cycles for one reaction, 16 for the second and 17 for the third, with each cycle 1 s at 98°C, 5 s at 58 °C and 5 s at 72°C; finally, incubate at 72°C for 1 min. ?Troubleshooting

82 Load 10 μl of each reaction alongside 100 bp DNA ladder on a 0.5x TBE, 1% agarose gel. Electrophorese and visualise using a nucleic acid stain, allowing a qualitative as well as semi-quantitative assessment: Successfully generated libraries yield a smear between ~200 and 300 bp that is usually visible after 15 cycles of amplification.

83 Following the manufacturer's protocol, confirm the size distribution of the library and estimate its concentration using a Bioanalyzer and High Sensitivity DNA Kit (Agilent). ?Troubleshooting

84 Assuming the Bioanalyzer trace shows a quantifiable distribution of material between ~200 and 300 bp, which is confirmed by PCR, the library can then be prepared for Ion PGM™ or Ion

Proton™ sequencing following the manufacturer's protocol. CRITICAL STEP: The Bioanalyzer trace and PCR amplification from a typical library are shown in **Fig. 2b and c**. The required input for Ion OneTouch™ emulsion PCR is 100 µl of 12 pM for the Ion Proton™ and 26 µl of 25 pM for the Ion PGM™. We recommend using the Ion PI™ Chip to achieve the maximum number of reads. Alternatively the Ion 318™ Chip can be used.

Bioinformatics

85 Align sequence reads to the appropriate genome (we use sacCer3) with bowtie2 (we used version 2.0.0). Command:

```
bowtie2 -x sacCer3 -U runName.fastq runName.sam
```

86 Perform alignment quality filtering and format conversion using Samtools (we used version 0.1.18) and BEDTools (we used version 2.16.2). A Bam file (denoted runName.master.bam below) containing all mapped and unmapped reads is a convenient way to store the complete raw sequence and quality information, as well as the alignment to the reference genome. Samtools view with the –q 30 option only retains reads with a mapping quality score >30 (misalignment probability < 0.001) for analysis. The ribonucleotide incorporation site is one nucleotide upstream and on the opposite strand to the mapped read 5' end, a transformation that can be achieved with the short Perl code illustrated below. All command syntaxes shown assume a Bash shell and standard Unix tools in addition to the specific software mentioned above. Commands:

```
samtools view -b -S runName.sam > runName.master.bam
samtools view -b -q 30 -S runName.sam > runName.bam
bamToBed -i runName.bam > runName.bed
   perl -alne \\
       'if($F[5]=~s/\+/-/){$F[2]=$F[1];$F[1]--;} \\
       else{$F[5]=~s/-/\+/;$F[1]=$F[2];$F[2]++} \\
       print join "\t", @F' \\
       < runName.bed | sort -k1,1 -k2,2n -k 6 > runName.ribo.bed
```

87 To facilitate comparison between libraries of differing read depth, normalise read counts to sequence tags per million (TPM =(fpe/total_fpe)*1×10^6 ,where fpe is 5' end reads at site and total_fpe is total number genome mapped 5' ends). Conversion to ribonucleotide counts per site (and strand) and the TPM normalisation can be achieved with the following commands:

```
bedtools groupby -i runName.ribo.bed -g 1,3,6 -c 1 -full \\
       -o count > runName.counts.bed
export totalfpe=`wc -l runName.ribo.bed | awk '{print $1}'`
cat runName.counts.bed | perl -alne \\
       '$F[6]=($F[6]/$ENV{totalfpe})*1e6; print join "\t", @F;' \\
       > runName.tmp.bed
```

88 Subsequent analysis can be performed in the R statistical environment (or Excel) by importing the tab delimited Bed file runName.tmp.bed. Such files can also be loaded into common genome browsers for visualisation and provided as "processed data files" to the Gene Expression Omnibus (GEO) and similar repositories along with the raw data (fastq file). CRITICAL STEP: Automation of emRiboSeq sequence processing (Steps 85-87) can be achieved using the script found at https://github.com/taylorLab/LaggingStrand/blob/master/emRiboSeqProcessor.sh with further details available in ref. 30 and examples of code performing analysis at https://github.com/taylorLab/LaggingStrand/

TIMING

Day 1
- Steps 1-23, gDNA isolation from yeast: 2-3 h.

Day 2
- Steps 24-31, sonication: 2-3 h.
- Steps 32-40, end repair, dA-tailing and adapter ligation: ~3 h + overnight incubation.

Day 3
- Steps 41-46, 3' blocking and endonucleolytic treatment: ~4.5 h.
- Steps 47-54, dephosphorylation and second adapter ligation: ~2 h + overnight incubation.

Day 4
- Steps 55-78, single stranded library preparation, second strand synthesis and size selection: 4-5 h.
- Steps 79-84, library quality control (~2h).

Day 5
- Steps 85-88, sequencing followed by bioinformatics processing of data (several days).

Several potential pause points have been highlighted within the protocol. If necessary, it is also possible to pause at many other points. However, it should be noted that Ampure XP beads should be removed prior to freezing and reintroduced before the next purification step.

TROUBLESHOOTING

Step	Problem	Possible reason	Possible solution
81	Amplification of 'endonuclease negative' control library	Incomplete removal of 3'-OH groups	Ensure that the 3' blocking reaction conditions and timing are optimal for the terminal transferase used (Steps 43-44).
83	Poor bioanalyzer trace	Carry over of Ampure beads or other contaminant	Additional purification (a further round of Ampure Bead purification or spin column based purification) may help, but is not guaranteed to improve the trace. N.B. we have successfully generated high quality sequencing data from samples that failed to produce good Bioanalyzer traces. In these cases we estimated the concentration based on the semi-quantitative PCR performed in Step 81, and comparing it with results from a previous library.

ANTICIPATED RESULTS

EmRiboSeq libraries generated using this protocol typically had a concentration of 400-2,000 pM. The total number of reads for the Ion 318™ Chip was typically 4-6 million, and an order of magnitude

higher for the Ion PI™ Chip at 50-90 million reads. Of these reads 70-80% mapped unambiguously (phred scaled mapping quality >30) to the yeast genome, resulting in an average read coverage of >35-fold for the Ion 318™ Chip and >400-fold for the PI™ Chip, with 150 base reads. However, for embedded ribonucleotide mapping it is the read 5' end coverage over the genome rather than sequenced read coverage that matters. For this measure, 0.2x to 0.4x genome average coverage for the Ion 318™ Chip and 3x to 6x coverage with the PI™ Chip was typical. Results for libraries prepared using other endonucleases are likely to be similar, as was the case for control libraries prepared using sequence specific endonucleases. Independent libraries yielded highly reproducible data, and demonstrated >99.9% site and strand specificity and >99% single nucleotide resolution (**Fig. 2c and d**).

It was previously unknown if embedded ribonucleotides were randomly distributed throughout the genome. The advent of a number of high-throughput sequencing methods, including emRiboSeq, has started to shed light on this. It is now clear that this distribution is non-random, with certain patterns emerging, including those related to polymerase-dependent nucleotide preferences, as well as sequence context[30,33-35]. The way is now open to investigate genome-wide ribonucleotide incorporation, uncover which rules underlie the observed distribution and determine the (patho)physiological consequences.

AUTHOR CONTRIBUTIONS

M.A.M.R, M.S.T. and A.P.J. conceived and designed the original protocol. M.A.M.R. and J.D. modified and updated the protocol to its current state. M.S.T. performed all computational analyses. J.D. and M.A.M.R. wrote the manuscript with assistance from M.S.T and A.P.J.

ACKNOWLEDGEMENTS

COMPETING FINANCIAL INTERESTS

The authors declare that they have no competing financial interests.

REFERENCES

1. Clausen, A.R., Zhang, S., Burgers, P.M., Lee, M.Y. & Kunkel, T.A. Ribonucleotide incorporation, proofreading and bypass by human DNA polymerase delta. *DNA Repair (Amst)* **12**, 121-7 (2013).
2. Goksenin, A.Y. et al. Human DNA polymerase epsilon is able to efficiently extend from multiple consecutive ribonucleotides. *J Biol Chem* **287**, 42675-84 (2012).
3. Gosavi, R.A., Moon, A.F., Kunkel, T.A., Pedersen, L.C. & Bebenek, K. The catalytic cycle for ribonucleotide incorporation by human DNA Pol lambda. *Nucleic Acids Res* **40**, 7518-27 (2012).

4.   Kasiviswanathan, R. & Copeland, W.C. Ribonucleotide discrimination and reverse transcription by the human mitochondrial DNA polymerase. *J Biol Chem* **286**, 31490-500 (2011).

5.   Nick McElhinny, S.A. et al. Abundant ribonucleotide incorporation into DNA by yeast replicative polymerases. *Proc Natl Acad Sci U S A* **107**, 4949-54 (2010).

6.   Nick McElhinny, S.A. et al. Genome instability due to ribonucleotide incorporation into DNA. *Nat Chem Biol* **6**, 774-81 (2010).

7.   Nick McElhinny, S.A. & Ramsden, D.A. Polymerase mu is a DNA-directed DNA/RNA polymerase. *Mol Cell Biol* **23**, 2309-15 (2003).

8.   Reijns, M.A. et al. Enzymatic Removal of Ribonucleotides from DNA Is Essential for Mammalian Genome Integrity and Development. *Cell* **149**, 1008-1022 (2012).

9.   Yao, N.Y., Schroeder, J.W., Yurieva, O., Simmons, L.A. & O'Donnell, M.E. Cost of rNTP/dNTP pool imbalance at the replication fork. *Proc Natl Acad Sci U S A* **110**, 12942-7 (2013).

10.   Caldecott, K.W. Molecular biology. Ribose--an internal threat to DNA. *Science* **343**, 260-1 (2014).

11.   Dalgaard, J.Z. Causes and consequences of ribonucleotide incorporation into nuclear DNA. *Trends Genet* **28**, 592-7 (2012).

12.   Potenski, C.J. & Klein, H.L. How the misincorporation of ribonucleotides into genomic DNA can be both harmful and helpful to cells. *Nucleic Acids Res* **42**, 10226-34 (2014).

13.   Vaisman, A. & Woodgate, R. Redundancy in ribonucleotide excision repair: Competition, compensation, and cooperation. *DNA Repair (Amst)* **29**, 74-82 (2015).

14.   Wallace, B.D. & Williams, R.S. Ribonucleotide triggered DNA damage and RNA-DNA damage responses. *RNA Biol* **11**, 1340-6 (2014).

15.   Williams, J.S. & Kunkel, T.A. Ribonucleotides in DNA: origins, repair and consequences. *DNA Repair (Amst)* **19**, 27-37 (2014).

16.   Martin, M.J., Garcia-Ortiz, M.V., Esteban, V. & Blanco, L. Ribonucleotides and manganese ions improve non-homologous end joining by human Polmu. *Nucleic Acids Res* **41**, 2428-36 (2013).

17.   Ghodgaonkar, M.M. et al. Ribonucleotides misincorporated into DNA act as strand-discrimination signals in eukaryotic mismatch repair. *Mol Cell* **50**, 323-32 (2013).

18.   Lujan, S.A., Williams, J.S., Clausen, A.R., Clark, A.B. & Kunkel, T.A. Ribonucleotides are signals for mismatch repair of leading-strand replication errors. *Mol Cell* **50**, 437-43 (2013).

19.   Allen-Soltero, S., Martinez, S.L., Putnam, C.D. & Kolodner, R.D. A saccharomyces cerevisiae RNase H2 interaction network functions to suppress genome instability. *Mol Cell Biol* **34**, 1521-34 (2014).

20.   Cho, J.E., Kim, N., Li, Y.C. & Jinks-Robertson, S. Two distinct mechanisms of Topoisomerase 1-dependent mutagenesis in yeast. *DNA Repair (Amst)* **12**, 205-11 (2013).

21.   Gunther, C. et al. Defective removal of ribonucleotides from DNA promotes systemic autoimmunity. *J Clin Invest* **125**, 413-24 (2015).

22.   Hiller, B. et al. Mammalian RNase H2 removes ribonucleotides from DNA to maintain genome integrity. *J Exp Med* **209**, 1419-26 (2012).

23.   Kalhorzadeh, P. et al. Arabidopsis thaliana RNase H2 deficiency counteracts the needs for the WEE1 checkpoint kinase but triggers genome instability. *Plant Cell* **26**, 3680-92 (2014).

24.   Kim, N. et al. Mutagenic processing of ribonucleotides in DNA by yeast topoisomerase I. *Science* **332**, 1561-4 (2011).

25.   Pizzi, S. et al. Reduction of hRNase H2 activity in Aicardi-Goutieres syndrome cells leads to replication stress and genome instability. *Hum Mol Genet* **24**, 649-58 (2015).

26.   Crow, Y. et al. Mutations in genes encoding ribonuclease H2 subunits cause Aicardi-Goutières syndrome and mimic congenital viral brain infection. *Nat Genet* **38**, 910-6 (2006).

27.   Reijns, M.A. & Jackson, A.P. Ribonuclease H2 in health and disease. *Biochem Soc Trans* **42**, 717-25 (2014).

28. Schellenberg, M.J., Tumbale, P.P. & Williams, R.S. Molecular underpinnings of Aprataxin RNA/DNA deadenylase function and dysfunction in neurological disease. *Prog Biophys Mol Biol* **117**, 157-165 (2015).

29. Tumbale, P., Williams, J.S., Schellenberg, M.J., Kunkel, T.A. & Williams, R.S. Aprataxin resolves adenylated RNA-DNA junctions to maintain genome integrity. *Nature* **506**, 111-5 (2014).

30. Reijns, M.A. et al. Lagging-strand replication shapes the mutational landscape of the genome. *Nature* **518**, 502-6 (2015).

31. Sparks, J.L. et al. RNase H2-initiated ribonucleotide excision repair. *Mol Cell* **47**, 980-6 (2012).

32. Williams, J.S. et al. Evidence that processing of ribonucleotides in DNA by topoisomerase 1 is leading-strand specific. *Nat Struct Mol Biol* **22**, 291-7 (2015).

33. Clausen, A.R. et al. Tracking replication enzymology in vivo by genome-wide mapping of ribonucleotide incorporation. *Nat Struct Mol Biol* **22**, 185-91 (2015).

34. Daigaku, Y. et al. A global profile of replicative polymerase usage. *Nat Struct Mol Biol* **22**, 192-8 (2015).

35. Koh, K.D., Balachander, S., Hesselberth, J.R. & Storici, F. Ribose-seq: global mapping of ribonucleotides embedded in genomic DNA. *Nat Methods* **12**, 251-7, 3 p following 257 (2015).

36. Koh, K.D., Hesselberth, J. & Storici, F. Ribose-seq: ribonucleotides in DNA to Illumina library. (2015).

37. Jinks-Robertson, S. & Klein, H.L. Ribonucleotides in DNA: hidden in plain sight. *Nat Struct Mol Biol* **22**, 176-8 (2015).

38. Bailly, V., Derydt, M. & Verly, W.G. Delta-elimination in the repair of AP (apurinic/apyrimidinic) sites in DNA. *Biochem J* **261**, 707-13 (1989).

39. Bryan, D.S., Ransom, M., Adane, B., York, K. & Hesselberth, J.R. High resolution mapping of modified DNA nucleobases using excision repair enzymes. *Genome Res* **24**, 1534-42 (2014).

40. Potenski, C.J., Niu, H., Sung, P. & Klein, H.L. Avoidance of ribonucleotide-induced mutations by RNase H2 and Srs2-Exo1 mechanisms. *Nature* **511**, 251-4 (2014).

41. Vaisman, A. et al. Removal of misincorporated ribonucleotides from prokaryotic genomes: an unexpected role for nucleotide excision repair. *PLoS Genet* **9**, e1003878 (2013).

42. Williams, J.S. et al. Topoisomerase 1-mediated removal of ribonucleotides from nascent leading-strand DNA. *Mol Cell* **49**, 1010-5 (2013).

43. Clark, A.B., Lujan, S.A., Kissling, G.E. & Kunkel, T.A. Mismatch repair-independent tandem repeat sequence instability resulting from ribonucleotide incorporation by DNA polymerase epsilon. *DNA Repair (Amst)* **10**, 476-82 (2011).

44. Huang, M.E., Rio, A.G., Nicolas, A. & Kolodner, R.D. A genomewide screen in Saccharomyces cerevisiae for genes that suppress the accumulation of mutations. *Proc Natl Acad Sci U S A* **100**, 11529-34 (2003).

45. Lazzaro, F. et al. RNase H and Postreplication Repair Protect Cells from Ribonucleotides Incorporated in DNA. *Mol Cell* **45**, 99-110 (2012).

46. Reijns, M.A. et al. The structure of the human RNase H2 complex defines key interaction interfaces relevant to enzyme function and human disease. *J Biol Chem* **286**, 10530-9 (2011).
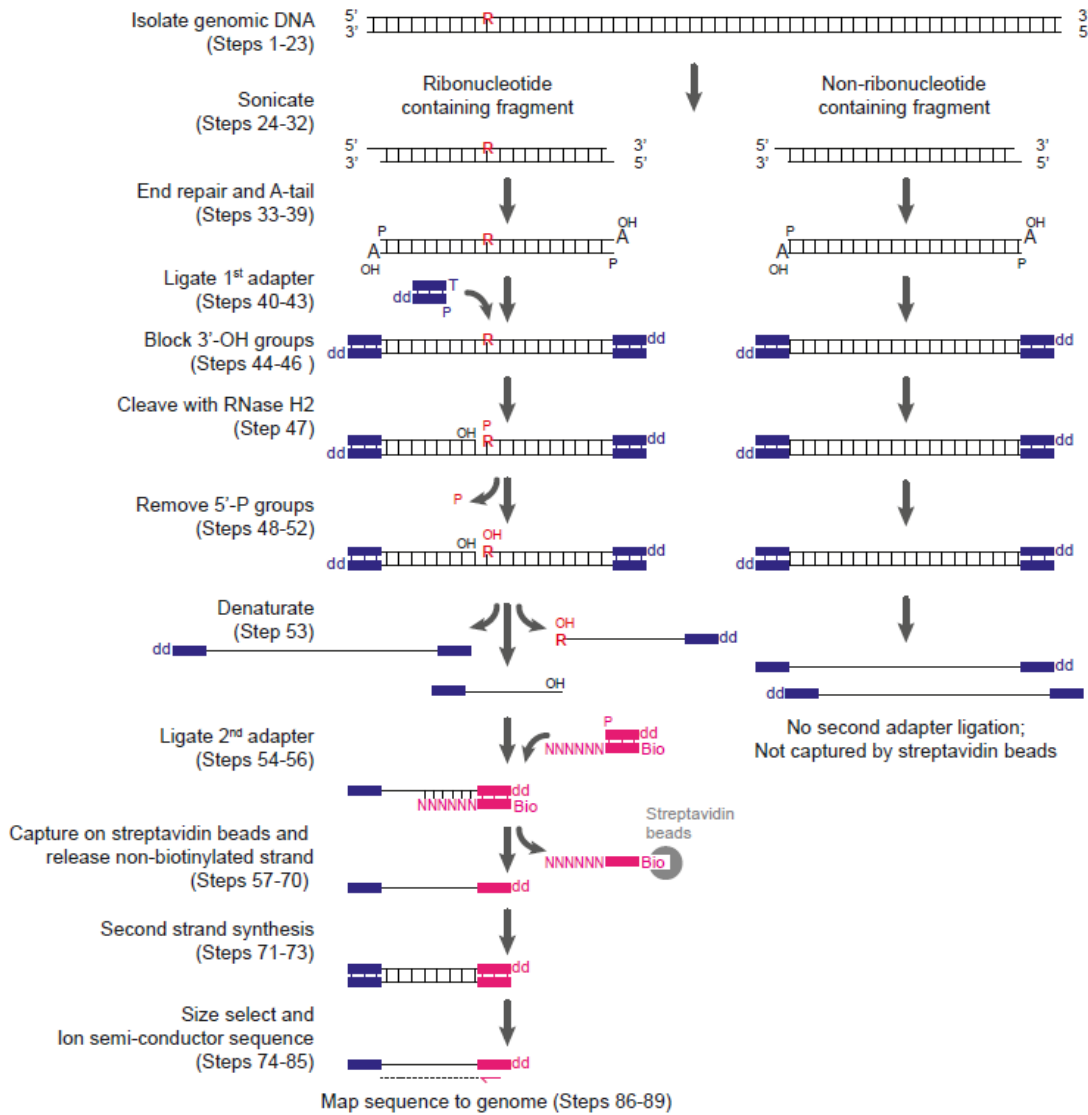
Figure 1 | The emRiboSeq protocol. Schematic depicting processing of ribonucleotide (red 'R') and non-ribonucleotide containing DNA fragments. The first adapter (blue) attaches to both fragments, but only those that contain ribonucleotides will be captured for sequencing after ligation with the second adapter (pink). For EndoSeq, a different endonuclease is used in place of RNase H2 (Step 46). OH, hydroxyl; P, phosphate; A, deoxyadenosine; T, deoxythymidine; dd, dideoxynucleoside; Bio, biotinylated nucleoside; NNNNNN, random deoxynucleotide hexamer. Adapted from ref. 30.
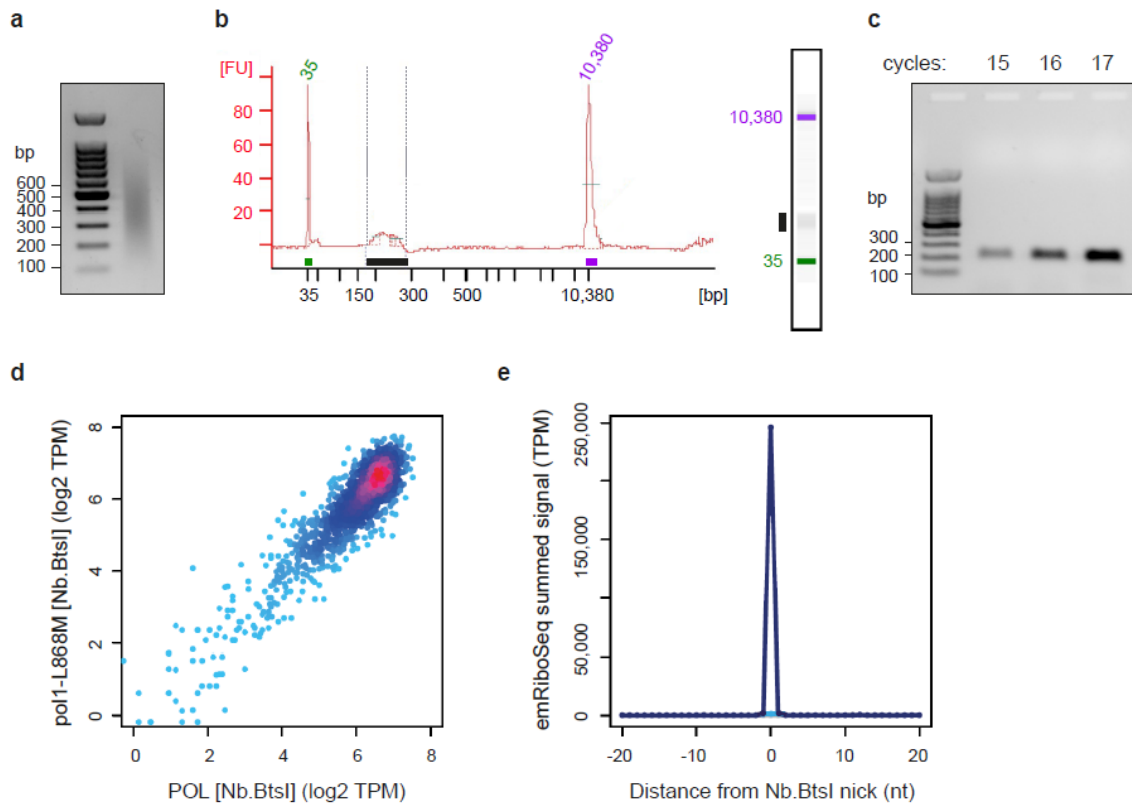
Figure 2 | Library quality control and anticipated results. (**a**) Sonicated DNA separated by agarose gel electrophoresis (Step 26) shows an average fragment size of approximately 400 bp. (**b**) Bioanalyzer result (Step 83) for an emRiboSeq library shows a typical trace (left) and gel-like image (right) with a peak for fragments between ~180 and ~300 bp in size (black bar). Standards (green and purple bars) of defined size and amount allow quantification. FU, arbitrary fluorescence units. (**c**) Agarose gel electrophoresis of PCR products after 15, 16 and 17 cycles of amplification (Steps 80-82) of the same library shows product between 200 and 300 bp in size. (**d**) Sequencing results for libraries generated using Nb.BtsI are highly reproducibility between different strains (POL, wildtype polymerase; pol1-L868M, increased Pol-α ribonucleotide incorporation) after normalizing read counts to sequence tags per million (TPM). The majority of bona fide Nb.BtsI sites were present at maximal frequency, although some sites were present at lower frequencies. This is the result of partial loss during size selection because of their close proximity to other cleavage sites, a highly reproducible finding between independent libraries (Spearman's rho=0.82, p <2.2-e16). Due to star activity Nb.BtsI-like sites were also detected. Other libraries prepared using BciVI restriction enzyme digestion did not show such star activity (data not shown), and allowed calculation of the site specificity for the method (>99.9%). (**e**) Summed signal at Nb.BtsI sites (dark blue, correct strand; light blue, opposite strand) highlights the strand specificity (>99.9%) and single nucleotide resolution (>99%) of this method. Panels d and e adapted from ref. 30.