Nucleic Acids Research, 2020 **1**
doi: 10.1093/nar/gkaa400

Deep sequencing of non-enzymatic RNA primer extension

Daniel Duzdevich [1,2,*,†], Christopher E. Carr [1,3,†] and Jack W. Szostak [1,2,4,5]

[1]Department of Molecular Biology, Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA 02114, USA, [2]Howard Hughes Medical Institute, Massachusetts General Hospital, Boston, MA 02114, USA, [3]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA, [4]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA and [5]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

Received February 20, 2020; Revised April 02, 2020; Editorial Decision April 30, 2020; Accepted May 05, 2020

ABSTRACT

Life emerging in an RNA world is expected to propagate RNA as hereditary information, requiring some form of primitive replication without enzymes. Non-enzymatic template-directed RNA primer extension is a model of the copying step in this posited form of replication. The sequence space accessed by primer extension dictates potential pathways to self-replication and, eventually, ribozymes. Which sequences can be accessed? What is the fidelity of the reaction? Does the recently illuminated mechanism of primer extension affect the distribution of sequences that can be copied? How do sequence features respond to experimental conditions and prebiotically relevant contexts? To help answer these and related questions, we here introduce a deep-sequencing methodology for studying RNA primer extension. We have designed and vetted special RNA constructs for this purpose, honed a protocol for sample preparation and developed custom software that analyzes sequencing data. We apply this new methodology to proof-of-concept controls, and demonstrate that it works as expected and reports on key features of the sequences accessed by primer extension.

INTRODUCTION

A central challenge of the RNA world hypothesis is replicating RNA without enzymes (1–7). Replication, in turn, requires a copying mechanism. Non-Enzymatic RNA Primer Extension (NERPE) is a model of template-directed RNA copying in which a primer is extended by the polymerization of nucleotides or the ligation of oligonucleotides (oligos) (8–10). Standard triphosphate nucleotides do not react with a primer-template complex because they are inert without evolved enzymes (11). Other activating moieties have been identified (12), including 2-aminoimidazole (2AI) (13,14). The mechanism of primer extension with imidazole-based nucleotide activation was only recently illuminated (15–19) (Supplementary Figure S1A–D). Briefly, 2AI-monoribonucleotides (2AIrN) react spontaneously to form a highly reactive 5′-5′phospho-imidazolium-phospho bridged dinucleotide intermediate in aqueous solution buffered to a pH $\cong$ pK_a of 2AI ($\cong$ 8.3). The dinucleotide intermediate binds the template through Watson–Crick base pairing, and the deprotonated oxygen of the primer 3′ hydroxyl attacks the proximal bridging phosphate displacing a 2AIrN as the leaving group. Although the reactive intermediate is a dinucleotide, primer extension proceeds one nucleotide at a time (Supplementary Figure S1C and D). A complementary pathway is non-enzymatic ligation of an activated oligo, which proceeds without a bridged intermediate and 2AI as the leaving group, yielding a primer extended by the length of the reacting oligo. Both pathways yield canonical 3′-5′ phosphodiester backbone linkages, but 2′-5′ linkages can also form, especially in the context of mismatched bases (20–23).

The prevailing technique for characterizing non-enzymatic RNA primer extension is denaturing polyacrylamide gel electrophoresis (PAGE) (Supplementary Figure S1E). The primer is fluorescently labeled, and the reaction products (primers extended to different lengths) are separated by size, typically yielding a characteristic banding pattern that can be mapped to +1, +2, +3, etc. extension events. PAGE analysis can be used to measure reaction kinetics (16), test novel reaction chemistries (14) and help determine mechanisms (19). A significant limitation of PAGE analysis is that information about the sequences accessed by primer extension is strictly determined by defined template and reactant identities. For example, interpreting a +1 product as an added rG requires rC in the template and activated 2AIrG as the

*To whom correspondence should be addressed. Tel: +1 617 643 6861; Fax: +1 617 724 2662; Email: duzdevich@molbio.mgh.harvard.edu
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

reactant. If the set of available template sequences harbors other bases in that position, and/or if additional activated nucleotides are used as reactants, then PAGE analysis cannot definitively establish product identity. Mixtures of activated nucleotides can yield products with mismatches to the templating sequence (errors), and mismatches are difficult or impossible to identify by PAGE analysis. Copying fidelity is of great interest because it will affect how heredity emerges (24–28). An ideal deep-sequencing methodology should be applicable to any experimentally defined templating sequence, access mismatch information (including relevant sequence context) and tolerate a range of reaction conditions.

Here we introduce NERPE-Seq, a deep-sequencing protocol and analysis pipeline for studying non-enzymatic RNA primer extension. We developed special RNA constructs, and a protocol to prepare multiple samples for simultaneous deep-sequencing (multiplexing). We identified biases arising from enzymatic reactions, RNA backbone heterogeneity and secondary structure, and then mitigated them to ensure that sequencing data reflect the properties of primer extension rather than any incidental selectivities of the protocol. We also created a custom analysis software that filters and sorts raw data, and generates standard measurements of template and product sequences. We show that NERPE-Seq can accurately characterize both non-enzymatic polymerization and ligation reactions, and identify mismatches.

MATERIALS AND METHODS

General

All reactions were performed with RNase-free water and salt solutions (not DEPC-treated, Ambion) in RNase-free 0.2 ml PCR tubes (VWR International), or 1.5 ml DNA LoBind tubes (Eppendorf). Bicine buffer was prepared from the Na^+ salt (Sigma-Aldrich) with RNase-free UltraPure Distilled Water (Invitrogen), adjusted to pH 8 with 2 M NaOH and syringe-filtered (Millex MP 0.22 μm, Millipore Sigma). Extreme caution was taken to prohibit contamination, including the use of barrier tips for all liquid handling (MultiMax, Sorenson BioScience, or TipOne, USA Scientific) except non-preparative gel loading. Enzymes and DNA/RNA ladders were purchased from New England BioLabs (NEB), and used as per the manufacturer's instructions unless otherwise indicated. All incubations with a specified temperature were performed in a Bio-Rad T100 thermal cycler.

2-aminoimidazole-activated monoribonucleotides and *2-aminoimidazole-activated 5′-CGG-3′ (2AI-CGG) trimer* were prepared essentially as previously described (14) with minor adjustments. Further details are in the Supplementary Data, Extended Materials and Methods.

Oligonucleotides (oligos) were ordered from Integrated DNA Technologies (IDT), NEB or synthesized in-house. All oligos were stored in TE (10 mM Tris-Cl, 1 mM ethylenediaminetetraacetic acid) buffer, pH 7 (Invitrogen) at −30°C. A table of oligos used in this study is included in the Supplementary Data (Table of Oligonucleotides). Oligos synthesized in-house, excepting the trimer described above, were prepared on an ABI Expedite instrument (standard phosphoramidites from ChemGenes; special phosphoramidites, synthesis reagents and supports from Glen Research) at the 1 μmole scale and purified by GlenPak (Glen Research) and PAGE. Further details are in the Supplementary Data, Extended Materials and Methods.

Standard primer extension reactions

Unless otherwise indicated, 1 μM primer and 1.2 μM template were mixed in water and 200 mM Na^+ bicine, pH 8, incubated at 85°C for 30 s, then cooled to 23°C at 0.2°C/s. Activated nucleotides or trimer were added to indicated concentrations (in cases with multiple nucleotide species, they were mixed first), and the reaction was initiated by the addition of 50 mM $MgCl_2$ (all concentrations indicate final concentrations, typically in a 20 μl volume). Timepoints were quenched by adding 1 μl of the reaction mixture to 20 μl Urea Load Buffer (8.3 M urea [Sigma-Aldrich], 1.3× TBE buffer [from a 10× autoclaved stock], 75 μM bromophenol blue [Sigma-Aldrich, from a 7.5 mM stock in DMSO], 880 μM orange G [Sigma-Aldrich, from an 88 mM stock in dimethyl sulfoxide (DMSO)], syringe-filtered). A total of 2.5 μl of the quenched material was then added to 1 μl of a 300 μM stock of RNA complementary to the template and incubated at 95°C for 3 min, then cooled to 25°C at 0.2°C/s. 16.5 μl additional Urea Load Buffer was added and the sample subjected to PAGE at 5 W for 10 min then 15 W for 1 h. In experiments with NPOM-caged bases in the template, the quench was into dye-free Urea Load Buffer, allowing for UV un-caging prior to the addition of RNA complementary to the template.

Gels and gel analysis

Polyacrylamide gels were prepared with the SequaGel–Urea Gel system (National Diagnostics). A 50 ml mix to yield a 20% gel was degassed under house vacuum with stirring for ∼10 mins before the addition of 10 μl tetramethylethylenediamine (TEMED, Sigma-Aldrich) and 100 μl of fresh 10% w/v ammonium persulfate (APS, Sigma-Aldrich) prepared with UltraPure Distilled Water. The gel was cast as 0.75 mm thick (1.5 mm for preparative) using 20 × 20 cm plates and allowed to set for 2 h. Gels were pre-run for at least 45 min at 20 W; all runs were at constant power, with heat-distributive aluminum or iron face-plates, in TBE prepared in Milli-Q water (Milli-Q Reference with Biopak Polisher, Millipore Sigma). Gels requiring staining (in particular, gels used to assay the sequencing constructs, which have no attached dyes) were removed from the glass plates and incubated in ∼140 ml of TBE + 14 μl SYBR Gold Nucleic Acid Gel Stain (Invitrogen) for several minutes. Gels were imaged with a Typhoon 9410 scanner (GE Healthcare) using the Auto PMT setting and at 50 μm resolution. For analysis and visualization, TIFF-formatted images were imported into Fiji (29). Bands were quantified using the Gel Analysis function, with relative band intensities reported as the ratio of the band intensity to the total lane intensity. Band edges were excluded. Contrast and color changes were applied uniformly to the entire gel image in all cases.

Non-preparative agarose gels were prepared as 1.4% w/v agarose (UltraPure Agarose, Invitrogen) in 50 ml TAE prepared in Milli-Q water + 5 μl SYBR Safe DNA Gel Stain (Invitrogen, added after microwave heating to form the agarose solution) and cast as 10 × 5 cm. Samples were loaded with Purple Loading Dye (NEB) and run at 80 V for 55 min at constant voltage. Gels were imaged on a Chemi-Doc imaging system (Bio-Rad). For visualization, TIFF-formatted images were imported into Fiji. Contrast and color changes were applied uniformly to the entire gel image in all cases.

RNA sample preparation for deep-sequencing

Water, 200 mM Na^+ bicine, pH 8, 1 μM hairpin construct and (where indicated) 1.2 μM 5′ Handle Block were incubated at 95°C for 3 min and cooled to 23°C at 0.2°C/s. The activated monoribonucleotides or activated trimer and 50 mM $MgCl_2$ were added and the mixture briefly vortexed. The final volume was 30 μl. The mixture was incubated at 23°C with a 42°C heated lid for 24 h. The reaction was quenched by the addition of 20 μl water followed by desalting in a MicroSpin G-25 spin column (GE Healthcare). The ∼53 μl eluate was transferred to a polymerase chain reaction (PCR) tube and placed under a 385 nm UV lamp (∼3 cm distance from the tops of the tubes, Spectroline ENF-240C, Spectronics) for 45 min with a manual mixing step halfway. An equal volume of Urea Load Buffer was added, the mixture was heated to 95°C for 3 min, cooled to 35°C and loaded on a preparative 20% polyacrylamide gel to completely remove residual unreacted nucleotides or oligos, which can interfere with downstream steps. The sample was subjected to PAGE at 10 W for 15 min and then 25 W for 1 h. The gel was stained in 120 ml TBE with 12 μl SYBR Gold Nucleic Acid Gel Stain for 10 min, visualized with a blue light transilluminator (Safe Imager 2.0, ThermoFisher Scientific) and the target band excised with a clean razor, purified with a ZR small-RNA PAGE Recovery Kit (Zymo Research) and eluted in 7 μl Tris-Cl, pH 7. The eluate was mixed with 0.36 μl RT Handle (from a 100 μM stock), heated to 95°C for 3 min and cooled to 25°C at 1°C/s. A total of 1.8 μl 10× RNA Reaction Buffer, 7.2 μl 50% PEG_{8000} and 3 μl T4 RNA Ligase 2, truncated KQ were added and the mixture incubated for 18 h at 25°C and 2 h at 4°C with a 42°C heated lid. A total of 0.7 μl 5 M NaCl, 0.5 μl sodium dodecyl sulphate (from a 10% w/v stock) and 0.25 μl Proteinase K were added and the mixture incubated for 20 min at room temperature. A total of 50 μl of 250 mM NaCl was added and the solution was extracted with 160 μl phenol–chloroform (UltraPure phenol:chloroform:isoamyl alcohol 25:24:1, Invitrogen) and twice with 200 μl chloroform (EMSURE, Millipore Sigma). The final extracted volume was ∼66 μl, which was run through an Oligo Clean & Concentrator spin column (Zymo Research) and eluted in 12 μl water.

Preparation of cDNA for deep-sequencing

A total of 0.7 μl RT Primer (from a 100 μM stock) was added to the purified sample and incubated at 75°C for 3 min, 37°C for 10 min and 25°C for 10 min. A total of 4 μl of ProtoScript II Buffer, 1 μl dNTP mix (NEB), 0.2 μl 1 M $MgCl_2$, 2 μl DTT (NEB) and 2 μl of ProtoScript II were added and the mixture incubated at 42°C for 12 h with a 105°C heated lid. (Alternatively, for standard conditions [see section on 2′-5′ linkages], no $MgCl_2$ was added and the incubation was at 50°C for 1 h.) Ten microliter of water was added and the mixture was run through an Oligo Clean & Concentrator spin column, including the RNA degradation step as per the manufacturer's instructions. The cDNA was eluted in 30 μl of TE, pH 7 and stored at 4°C. The cDNA stock concentration was measured with a NanoDrop 2000c spectrophotometer (ThermoFisher Scientific), and typically found to be in the fraction of a μg/μl range.

Preparation and isolation of barcoded PCR products for multiplexed deep-sequencing

A volume of isolated cDNA sample equivalent to 0.1 μg total cDNA was added to a 50 μl Q5 Hot Start High-Fidelity DNA Polymerase PCR reaction with 0.2 μM each of the NEBNext SR Primer for Illumina and NEBNext Index Primer for Illumina and run for six cycles with a 15 s 62°C extension step. Ten microliter Purple Loading Dye was added to the PCR product mixture and the entire volume was run on a preparative agarose gel. Gels were prepared as 1.4% w/v agarose (Certified Molecular Biology Grade, Bio-Rad) in 1× TAE (UltraPure, Invitrogen) and 4 μl SYBR Safe DNA Gel Stain, and cast as 10 × 5 cm with large wells sufficient to hold the 60 μl final sample volume. The gel was run at 70 V for 90 min at constant voltage. Target bands were visualized on a blue light transilluminator, excised and purified with a Quantum Prep Freeze 'N Squeeze spin column (Bio-Rad). The eluate was further purified with a magnetic bead clean-up (Agencourt AMPure XP) with a 1.7:1 ratio of beads to sample volume. Target material was eluted with 20 μl TE, pH 7 and submitted for sequencing. Samples were validated by TapeStation (Agilent) and qPCR before sequencing by Illumina MiSeq. A φX174 viral genome library was routinely spiked in to the pooled sequencing samples at 30% because the low diversity of the target library would otherwise make it difficult to differentiate among clusters in the sequencing flowcell during the first few rounds of sequencing.

Sequencing data analysis was performed using a custom code package written in MATLAB (MathWorks), as described in the 'Results' section and Supplementary Data. A comprehensive plain-language annotation of the code is provided in the Supplementary Data.

Average per-base error frequencies (f_{error}) based on the fraction of reads filtered ($R_{\text{filt.}}$) during a given step in data analysis were calculated by assuming a uniform average error rate across the length (in nucleotides, L) of the region being filtered such that $f_{\text{error}} = 1 - f_{\text{correct}}$; $(f_{\text{correct}})^L = (1 - R_{\text{filt.}})$.

RESULTS

A self-priming hairpin construct

A comprehensive picture of non-enzymatic RNA primer extension requires a sequencing approach that can supply

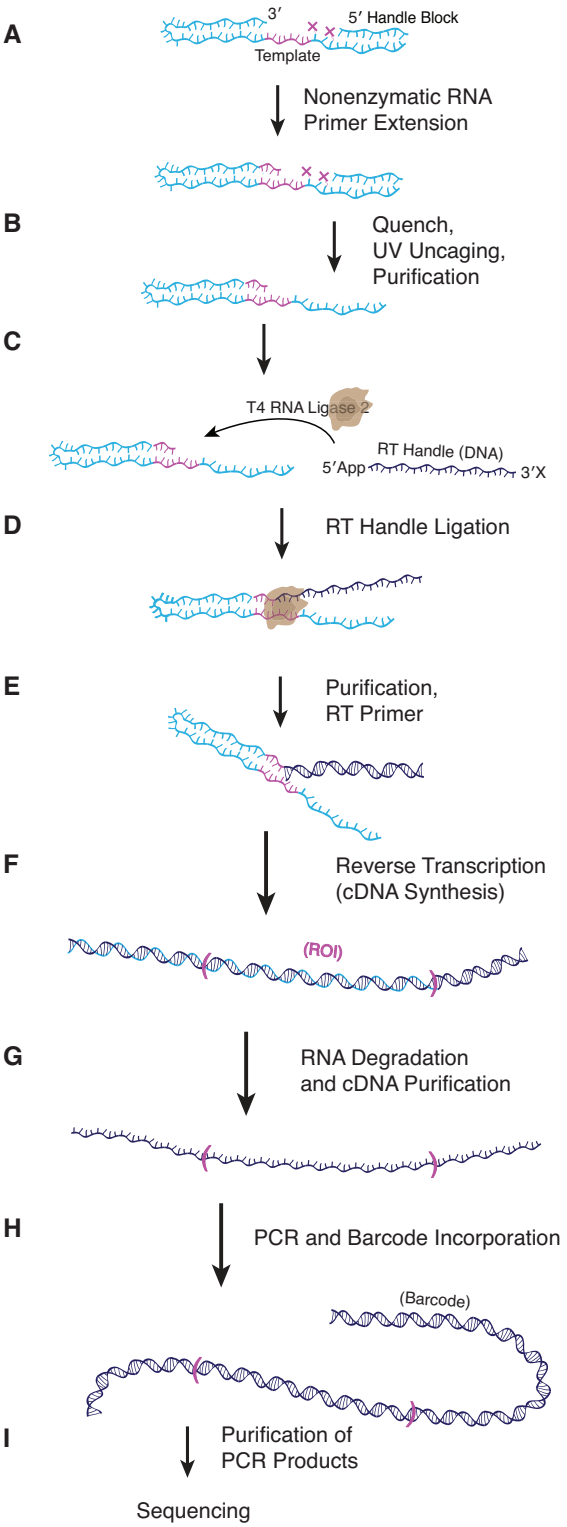

Figure 1. Protocol for preparing RNA hairpin constructs for sequencing. (**A**) NERPE-Seq RNA hairpin constructs contain a hairpin loop that connects the template to the primer so that the product of non-enzymatic primer extension and the corresponding template are on one continuous RNA strand. Two caged bases (magenta Xs) prevent primer extension from encroaching on the downstream 5′ Handle (Supplementary Figure S3). The 5′ Handle Block is complementary to the 5′ Handle and prevents it from interfering with primer extension (Supplementary Figure S8). (**B**) The primer extension reaction is quenched with a desalting size-exclusion spin column, the caged bases are uncaged and the target RNA is further gel purified. (**C** and **D**) The pre-adenylated DNA RT Handle (blocked on its 3′ end to prevent self-ligation) is ligated to the 3′ end of the RNA hairpin (the site of primer extension) (Supplementary Figure S4). (**E**) The ligase is removed by a Proteinase K digestion, the target RNA–DNA is phenol–chloroform extracted, and the RT primer is annealed to the RT Handle. (**F** and **G**) RT generates the cDNA (Supplementary Figure S5); the RNA is degraded, and the cDNA is isolated with a spin column. The ROI harbors the template, hairpin and any product sequences. (**H**) PCR is used to barcode the DNA and add flanking sequences (Supplementary Figure S6). Each barcode identifies DNA from a specific experiment and enables the sequencing of samples from multiple experiments at the same time. (**I**) The target PCR products are purified, and validated by automated electrophoresis and quantitative PCR prior to sequencing (Supplementary Figure S7).

both the primer extension product sequence and the corresponding template sequence. One strategy is to connect the primer to the template with a hairpin loop so that the construct primes itself (Figure 1A) (30–32). Each sequencing read will then contain the template, an intervening hairpin and any primer extension product sequence. We chose Illumina MiSeq as a deep-sequencing technology because of its low error rate, forward and reverse read functionality (see below), and capacity to process multiple experimental samples simultaneously (multiplexing) through barcoding (indexing).

RNA is prepared for deep-sequencing by modification with additional sequences (Figure 1A–E), reverse transcription (RT) (Figure 1E–G) and barcoding PCR (Figure 1G–I). The Illumina platform requires defined sequences (handles) on both 5′ and 3′ ends of the target. In this case the 3′ handle is also the binding site of the RT primer (RT Handle).

As a preliminary test of whether a hairpin construct could be effectively prepared for deep-sequencing (33–35), we adapted a standard RNA-Seq protocol in which the RT Handle is a 5′-adenylated (App) DNA oligo ligated by a derivative of T4 RNA Ligase 2, and the 5′ Handle is ligated by T4 RNA Ligase 1 (36,37). We designed three mock constructs (Mock 1–3) with the desired self-priming structure, an eight-base 'template' and varying extents of apparent primer extension (0, +3 and +7, respectively) (Supplementary Figure S2). RT Handle ligation worked on Mock 1 and Mock 2, but was less efficient on Mock 3 and also yielded off-target products (Supplementary Figure S2A). The subsequent 5′ Handle ligation was similarly flawed and most of the input material remained unligated (Supplementary Figure S2B). To compound these issues the barcoding PCR of the ligation products yielded only primer dimers for Mock 3 (Supplementary Figure S2C). We conclude that the standard RNA-Seq protocol is inadequate because it introduces a pronounced bias against longer non-enzymatic primer extension products.

For our application, the 5′ Handle sequence can be included as a component of the initial construct, thereby eliminating the need for a second ligation. However, the additional sequence downstream of the template introduces other challenges. First, if primer extension spans the template and then encroaches on the 5′ Handle sequence, the resultant product could interfere with sequencing and possi-

bly complicate data analysis. Second, the additional single-stranded RNA of the 5′ Handle could interfere with primer extension by physically interacting with the template. Addressing the first challenge requires a feature that stops primer extension at a defined position without otherwise affecting the reaction, but can be removed prior to RT. We chose 6-nitropiperonyloxymethyl (NPOM) amino caged deoxythymidine, a dT 'caged' on the N3 of the base with the NPOM moiety, which can be 'uncaged' by exposure to long-wavelength UV irradiation (38). We prepared an RNA test oligo with NPOM-caged dT bases (Supplementary Figure S3A) and exposed it to 385 nm UV light. PAGE analysis shows a gel shift after uncaging (Supplementary Figure S3B), and LC-MS identified UV exposure conditions under which most of the oligos are uncaged (Supplementary Figure S3C). We next tested whether the caged bases stop primer extension but allow RT when uncaged, and found that the caged oligo stops primer extension prior to the first caged dT, but allows full RT after uncaging (Supplementary Figure S3D–F).

Encouraged by the effectiveness of NPOM caging, we prepared a series of control hairpin constructs to facilitate protocol development (Supplementary Figure S3G–J). The Control Template Extended construct (CTEx) mimics full-length primer extension on a defined template. The CT and Control Template B (CTB) constructs harbor a defined 3′-GCC-5′ template sequence, which is known to work well for primer extension (14). The 6N construct harbors six randomized bases as a template. All constructs include a placeholding rA at the 5′ end of the template followed by two caged bases.

Optimizing RT handle ligation. Having defined a set of constructs we returned to the RT Handle ligation step, seeking to optimize it in the context of the encoded 5′ Handle sequence and NPOM caging. We tested several ligases and options for the adenylation and blocking of the RT Handle (Supplementary Figure S4A–C). 5′ pre-adenylation of a handle increases the efficiency and specificity of ligation (5′ adenylate is an intermediate of the T4 RNA Ligase 2 enzymatic pathway (39)), and a blocked or absent 3′-OH prevents handle self-ligation. Ultimately, our optimized protocol employs a 3′ dideoxy, 5′ pre-adenylated RT Handle and a derivative of T4 RNA Ligase 2 (T4 RNA Ligase 2, truncated KQ (40)). An ideal RT Handle ligation step would be equally efficient for all primer extension products while any hints of differential ligation, as previously observed for Mock 3 (Supplementary Figure S2A), would bias the proportions of specific products in sequencing data. To optimize the ligation conditions we leveraged a particular case in which the reaction was found to be inefficient: RT Handle ligation to the +3 products of non-enzymatic trimer ligation to CTB (Supplementary Figure S4D and E). In this positive control non-enzymatic ligation reaction, a 2AI-activated 5′-CGG-3′ (2AI-CGG) trimer is incubated with the CTB construct, which has a complementary 3′-GCC-5′ template. We used this case to screen RT Handle ligation conditions and found that a 25°C incubation temperature combined with either DMSO or excess PEG_{8000} yielded complete RT Handle ligation to all reaction products to within the resolution of PAGE analysis. Higher temperature and DMSO are expected to destabilize the hairpin whereas PEG functions as a crowding agent, presumably favoring enzyme-substrate interactions. We chose the excess PEG_{8000} condition for inclusion in the final protocol, but note the potential usefulness of DMSO for ligation reactions involving RNA with a base-paired 3′ end.

Preparing hairpin constructs for deep-sequencing

2′-5′ linkages and reverse transcription. With a set of test constructs and optimal RT Handle ligation conditions in place, we next sought to characterize the RT step (Figure 1E and F). We considered three possible RT issues that could bias the final sequencing data: the uncaged bases, secondary structure inhibition and 2′-5′ linkages in the products of primer extension. RT had already been shown to work normally once the NPOM caging is removed (Supplementary Figure S3F). RNA-Seq experiments routinely process RNAs with secondary structures and a short hairpin stem should not impede RT. Hairpin-mediated stalling would register during PAGE analysis of reverse-transcribed products, but we do not find this to be an issue (see below). Previous research has shown that a single 2′-5′ linkage can be read by reverse transcriptase (41), but given our interest in mismatches, which exhibit a higher proportion of 2′-5′ linkages (23,25), we tested whether multiple 2′-5′ linkages can also be read through by RT.

We began with a test oligo harboring a pair of 2′-5′ linkages separated by two bases (Supplementary Figure S5A): ProtoScript II reverse transcriptase is almost completely stalled by this substrate. However, a small amount of full-length DNA product (1.3%) suggested that although the enzyme is inefficient at reading through the 2′-5′ linkages, it is at least capable of doing so. This reaction was used to screen conditions with the goal of maximizing full-length reverse transcribed product (Supplementary Figure S5B). Increasing the $MgCl_2$ concentration from 3 to 10 mM combined with an overnight incubation at 42°C improved product yields, though stalling remained apparent (68% stalled) (Supplementary Figure S5C). Importantly, these conditions do not degrade the RNA template (Supplementary Figure S5D). We next prepared a series of RNA templates with increasingly separated pairs of 2′-5′ linkages (Supplementary Figure S5E). As previously reported, a single 2′-5′ linkage does not significantly stall RT (90% full-length product) (41). The most severe stall was measured for immediately adjacent 2′-5′ linkages (41% full-length product), whereas six intervening bases reduce the stalling effect (55% full-length product). A 2′-5′ linkage immediately adjacent to the primer consistently caused a stall regardless of the location of the subsequent 2′-5′ linkage. We conclude that primer extension products with multiple 2′-5′ linkages, especially at the first position after the RT primer or positioned tandemly, will be under-represented in the sequencing data.

Protocol trials and PCR optimization. To evaluate the protocol by PAGE, we used the CTEx construct, which mimics efficient primer extension on a defined template (Figure 2A). CTEx was prepared under experimental conditions, incubated for 24 h at 23°C and desalted. The CTEx RNA runs as a well-defined and undegraded band at the ex-

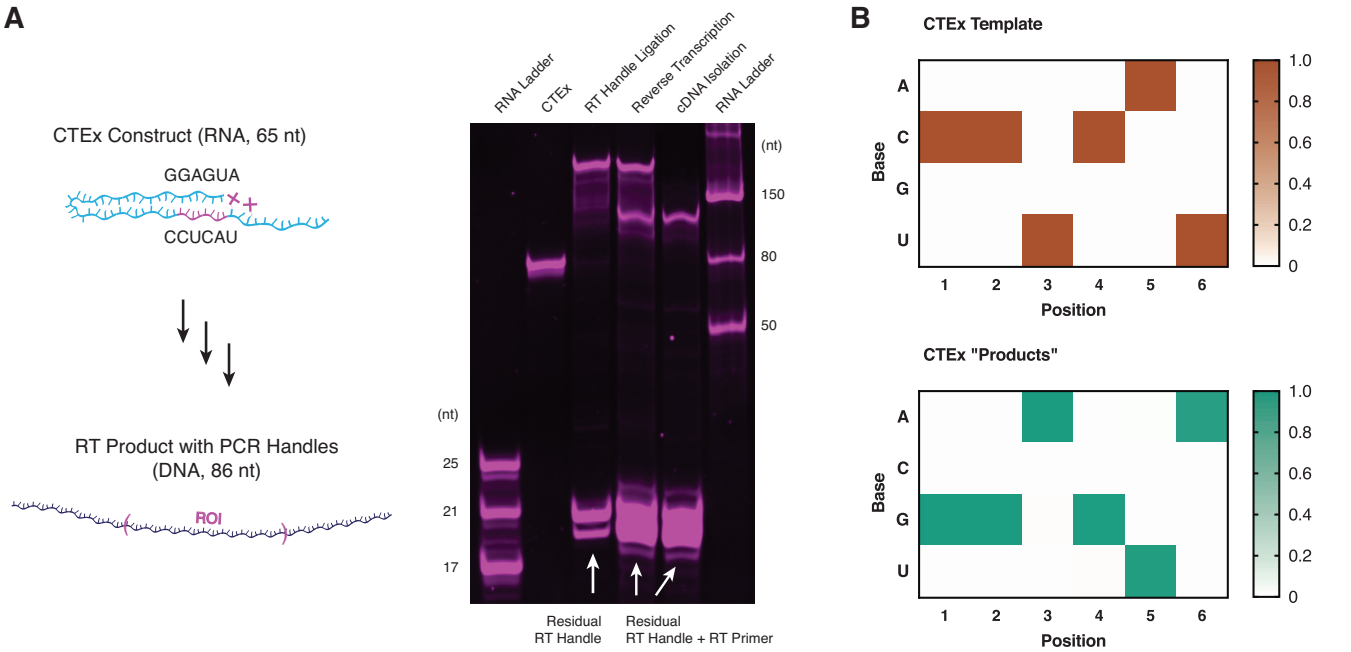

Figure 2. Sequencing a Construct that Mimics Full-length Non-enzymatic Primer Extension. (**A**) The hairpin construct CTEx was designed to mimic an efficient non-enzymatic primer extension reaction. The RNA was converted into cDNA using the optimized protocol. The ROI includes the template and 'product' sequences. After exposure to mock primer extension conditions and desalting, CTEx ran as a well-defined band at the expected position in PAGE. After uncaging and additional purification, the RT Handle was ligated to the 3′ end with very high efficiency. The product was purified and reverse transcribed, yielding a distinct DNA band, which was then isolated. (**B**) The NERPE-Seq protocol accurately measures the CTEx template and 'product' sequences without the template or product identities being provided during analysis. The heat maps show the frequency of each base at the indicated position.

pected size after this treatment. It was then uncaged, PAGE-purified, the RT Handle ligated and the product isolated by Proteinase K treatment, phenol–chloroform extraction and spin column purification. The RT Handle ligation was extremely efficient. The ligated product was then reverse transcribed, yielding a distinct cDNA product band. Finally, the RNA was degraded and the cDNA isolated by spin column purification, yielding a well-defined band of pure material. Although CTEx mimics very efficient non-enzymatic primer extension, it exhibits none of the biases identified with the Mock 3 construct and the non-optimized protocol (compare Figure 2A with Supplementary Figure S2A and B).

The final steps prior to sequencing submission are barcoding and purification of PCR products. We optimized the PCR conditions (Supplementary Figure S6), and then prepared a series of trials using the CTB construct (Supplementary Figure S7). As a negative control the CTB construct was exposed to primer extension conditions but without activated nucleotides so that no extension products were expected in the sequencing data (CTB Control, CTBC). As a positive control for canonical primer extension CTB was incubated with 20 mM each of 2AIrG and 2AIrC (CTB 40 mM-activated nucleotides, CTB40). As a positive control for non-enzymatic ligation CTB was incubated with 500 μM 2AI-CGG trimer (CTB Ligation, CTBL). PAGE analysis was used to evaluate key steps of the protocol for each experiment, as for CTEx above, and all three show the expected banding patterns (Supplementary Figure S7A). They also yielded well-defined PCR products as measured by agarose gel electrophoresis (Supplementary Figure S7B).

Samples submitted for sequencing are routinely quantified by automated electrophoresis, which is more sensitive than traditional agarose gel analysis. This enabled us to test a variety of PCR purification strategies. We compared purification by spin column, magnetic beads and agarose gel (Supplementary Figure S7C, 1–3). The spin column was largely ineffective at eliminating off-target bands, magnetic beads eliminated lower molecular weight off-target bands and agarose gel purification eliminated higher molecular weight off-target bands. We therefore combined the magnetic bead and agarose gel purifications (Supplementary Figure S7C, 4). Satisfied with the optimization of this final step, we were prepared to submit samples for deep-sequencing and began formulating a toolkit to analye the raw data.

Data processing and analysis

Sequencing data from non-enzymatic RNA primer extension is unique and pre-existing software cannot answer questions we are interested in. We therefore developed a custom code package implemented in MATLAB (MathWorks) to transform raw sequencing reads into a useful format (Pre-processing) and then make a series of key measurements (Characterization) (Figure 3).

Pre-processing. Sequencing by MiSeq (Illumina, MGH Department of Molecular Biology Next-generation Sequencing Core) with a MiSeq Reagent Kit v3 (150 cycle) typically produced ∼20 million barcoded reads of which ∼95% passed the instrument's quality filter. Pre-processing

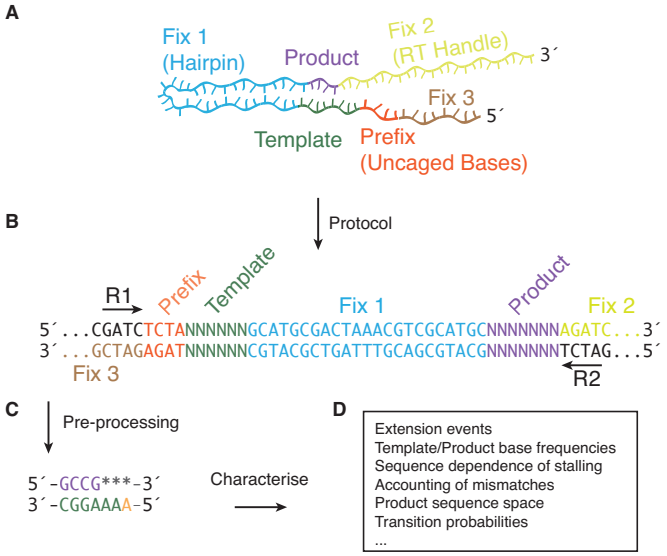

Figure 3. Data Analysis. (**A**) Cartoon of a hairpin construct after RT Handle ligation. The labels and color coding indicate the various sequence regions used during data processing. The hairpin and handles are defined sequences ('fixed'), the Prefix is a four-base motif with the two caged bases, the Template is of defined length but the analysis does not specify a defined sequence, and the Product is of indeterminate length and sequence. (**B**) The double-stranded DNA generated by barcoding PCR (the barcode is to the 3′ of Fix 2 and not shown). The final location of each region from the construct is labeled and color-coded. Paired-end sequencing provides both the forward (R1) and reverse (R2) sequences, which are compared against each other for quality control. A series of checks identifies the fixed sequences, filters out low-quality reads, and extracts the Template and Product. (**C**) The end result of Pre-processing is a set of Template-Product pairs. Unextended bases are indicated by an asterisk, and a placeholding A is included as an internal marker. (**D**) Template-Product pairs are queried in the Characterize stage to assay the sequence properties of templates and complementary products, and indicate the positions and contexts of mismatches.

then converts the raw sequencing reads into high quality template-product pairs (Figure 3A). We use paired-end sequencing so that each hairpin construct is sequenced once in the 'forward' direction (Read 1, R1) and once in the 'reverse' direction (Read 2, R2) (Figure 3B). R1 and R2 are compared to each other throughout Pre-processing to increase the quality of the output data. Pre-processing is initiated with debarcoded raw sequencing files. These are broken up into 10 000 reads/block to enable random access and parallel processing. Then, each read pair is subjected to a series of rigorous quality checks, and if it passes the template and product sequences are extracted.

Sequence accuracy is checked using per-base quality scores (42,43), q, provided by the sequencing run. These are converted to per-base error probabilities, p, ($p = 10^{-q/10}$) averaged across the read ($\bar{p} = \text{mean}(p)$), and transformed back to an overall quality score for the entire length of the read, $q_{\bar{p}} = -10\log(\bar{p})$. A default threshold of $q_{\bar{p}} = 26$ was selected for the index (44), and $q_{\bar{p}} = 30$ for R1 and R2, which corresponds to a per-base error probability of 0.1%.

An additional check on accuracy assumes that a read with an error in one position is likely to have other errors. We therefore require that sections of a read from defined (fixed) sequences in the hairpin construct are correct. The fixed sequences are also used to locate the Template and Product, which can be of variable sequence and variable length. The trimmed forward and reverse reads are then required to agree, and the final Template-Product pairs are generated (Figure 3C). Each pair is oriented as in the source hairpin construct to intuitively reflect primer extension, with the primer extension product shown 5′-to-3′ on a 3′-to-5′ template. If any step fails the read pair is discarded. For a typical multiplexed MiSeq run with eight barcoded samples, pre-processing yields approximately half a million read pairs per sample.

Characterization. This stage extracts the sequence properties of the template-product pairs generated by pre-processing. First, products of different lengths are quantified. The resultant histogram is equivalent to the data accessed by PAGE analysis (see below). Next, the position-dependent base distribution of the template is measured. In a randomized template each of the four bases would be equally represented at each position, but it is impossible to achieve a perfect composition by solid-state oligonucleotide synthesis (see below) so template normalization factors are calculated. To eliminate bias associated with a particular experimental condition (for example, an excess of 2′-5′ linkages), a comparison is included to template frequencies from a negative control experiment in which no primer extension is expected.

Primer extension products are grouped into three non-overlapping classes to facilitate analysis: those with no extension events (Unextended Set), those with extensions that are perfectly complementary to the template (Complementary Set) and those with extensions that contain one or more mismatches (Mismatch Set). The position-dependent base composition of the Complementary Set products are tabulated on an absolute, relative and normalized basis.

Identifying mismatches is a major goal of using deep-sequencing to study primer extension. The fidelity of primer extension (or, conversely, the error rate) is important for evaluating conditions under which emergent genetic information could be accurately propagated. In the Mismatch Set, position-dependent counts of mismatches and the distribution of terminal mismatches are tabulated. These measures will reveal condition-dependent trends in overall primer extension fidelity and the impact of mismatches on extension termination. Finally, the occurrence of each possible type of mismatch at each position is measured and normalized.

To visualize the sequence space accessed by primer extension all position-dependent combinations of three-base stretches in the products are counted (see Supplemental Data). Any differences between the results for all products and just the Complementary Set will show the extent to which mismatches contribute to the sequence space. Finally, sequential base transition frequencies for the product and template are computed. Here, we define sequential base transition frequencies as the probabilities of each subsequent base identity given the current base. Collectively, these metrics amount to a broad and deep accounting of non-enzymatic RNA primer extension sequence space.

Deep sequencing results

Because the ultimate experimental goal of NERPE-Seq is to routinely analyze data from primer extension reactions on random template sequences, the software does not assume a specific template sequence or any particular primer extension product. We can therefore evaluate NERPE-Seq by submitting control constructs and reactions that *do* have specific templates and products, and ask if the results are as expected. We first tested the CTEx construct, which mimics primer extension on a defined template, and found that NERPE-Seq accurately measures both the defined template and product sequences (Figure 2B). Encouraged by this result, we next considered a control experiment that included a primer extension reaction.

The encoded 5′ handle can interfere with NERPE. The defined-template construct Control Template (CT, Supplementary Figure S3H) was used to compare the results of primer extension as measured by PAGE analysis with the distribution of products as measured by NERPE-Seq. Disappointingly, the two datasets do not agree, with +2 products over-represented and +3 products under-represented in the NERPE-Seq analysis (Supplementary Figure S8A and B). We designed a slightly different construct with a longer template region and two caged bases instead of one (CTB, Supplementary Figure S3I), but the PAGE and NERPE-Seq results are in even greater disagreement (Supplementary Figure S8C and D). Next we considered the possibility that the hairpin primer–template complex does not anneal correctly, and screened annealing conditions coupled to primer extension. We anticipate that if the primer–template complex is sensitive to annealing conditions, then subsequent primer extension reactions will show different distributions of products. However, all the primer extension reactions yielded the same characteristic pattern by PAGE analysis regardless of the annealing condition (Supplementary Figure S8E), suggesting that the primer-template adopts the same conformation across tested conditions. Finally, we considered that for the specific template sequence used in CT and CTB (3′-GCC-5′), the 5′ Handle may interfere with the template region. The RNAstructure web server (45) was used to query the secondary structures of CT and CTB, and in both case the 5′ Handle sequence was predicted to interact with the template and block +3 primer extension products, as observed in the experimental results. We therefore included an oligo complementary to the 5′ Handle sequence (5′ Handle Block Test) to a primer extension reaction with CTB, and the expected product pattern was restored (Supplementary Figure S8F). We conclude that portions of the 5′ Handle can transiently base pair with the template and inhibit primer extension, but if the ssRNA in the handle is occluded by a complementary strand to form duplex RNA, then primer extension proceeds without interference (Supplementary Figure S8G).

NERPE-Seq reproduces known results and can report key measurements. With the insight that blocking the 5′ Handle prevents it from interfering with primer extension, we again compared the extent of primer extension as measured by PAGE analysis (Figure 4A) with the extent of primer extension as measured by NERPE-Seq (Figure 4B). The reactions were incubated for 24 h with 20 mM 2AIrG and 20 mM 2AIrC. The PAGE analysis shows predominantly +3 products, with some +2 and +4 (over-extended) products, as expected (Figure 4C). The NERPE-Seq analysis shows the same distribution (Figure 4D), in excellent agreement with the PAGE results.

We next tested whether non-enzymatic ligation can also be measured correctly by NERPE-Seq. Primer-templates were incubated for 24 h with 500 μM 2AI-CGG. Again, the product distributions as measured by PAGE and NERPE-Seq are in excellent agreement (Figure 4E and F). These results demonstrate that NERPE-Seq can accurately make the same measurements of primer extension as PAGE analysis. NERPE-Seq can also plot product base distributions for each position (Figure 4G and H), dominated by 5′-CGG-3′, as expected for a 3′-GCC-5′ template.

Beyond the power to supply reliable data, we sought an assay that can be applied across a variety of conditions. To test whether we can monitor the consequences of distinct primer extension reactions, we performed a narrow titration of activated nucleotide concentrations from 10 mM each of 2AIrG and 2AIrC (CTB20), through 20 mM each (CTB40, described above), to 30 mM each (CTB60). We expect a lower proportion of +3 products for CTB20 than CTB40, which is what we observe: 70% for CTB20 versus 75% for CTB40. The proportion of +3 products for CTB60 does not follow this trend, with 67% +3, but also slightly more +2 and +4 than observed in the other two experiments. This suggests that increasing the activated nucleotide concentration is not necessarily efficient at driving primer extension to the ideal products. Future work will explore potential explanations for such phenomena. We also expect over-extension to the +4 position, which harbors a templating rA, to increase with concentration. We observe that each increase in activated nucleotide concentration also increases over-extension, with +4 extensions accounting for 7% of CTB20 products, 13% of CTB40 products and 16% of CTB60 products. Taken together, these results indicate that NERPE-Seq analysis is sensitive to changes in experimental conditions.

An essential application of NERPE-Seq is to characterize mismatches. Mismatches were measured in the cases described above, but those experiments were not designed to specifically yield them. We therefore used a control experiment expected to specifically generate G-U and C-U mismatches: a 3′-GCC-5′ template was incubated with 20 mM 2AIrU for 24 h and analyzed by PAGE or NERPE-Seq. The two product distributions agree (Figure 5A and B), suggesting that NERPE-Seq is not biased by mismatches. In the sequencing data, all products should be U bases incorrectly paired across the 3′-GCC-5′ template, and this is what we observe (Figure 5C). As expected for a necessarily incorrect primer extension, the overall efficiency was much lower than in the experiments above, with only 6.1% of read pairs registering extension events ($n_{\text{total read pairs}} = 6.6 \times 10^5$). We can also visualize all position-dependent mismatches in the same experiment (Figure 5D). The G-U mismatch dominates position 1 and the C-U mismatch dominates position 2, both as expected. The additional apparent mismatches in downstream positions result from experimental errors. Notably, the vast majority of the data in this case (>98% of all

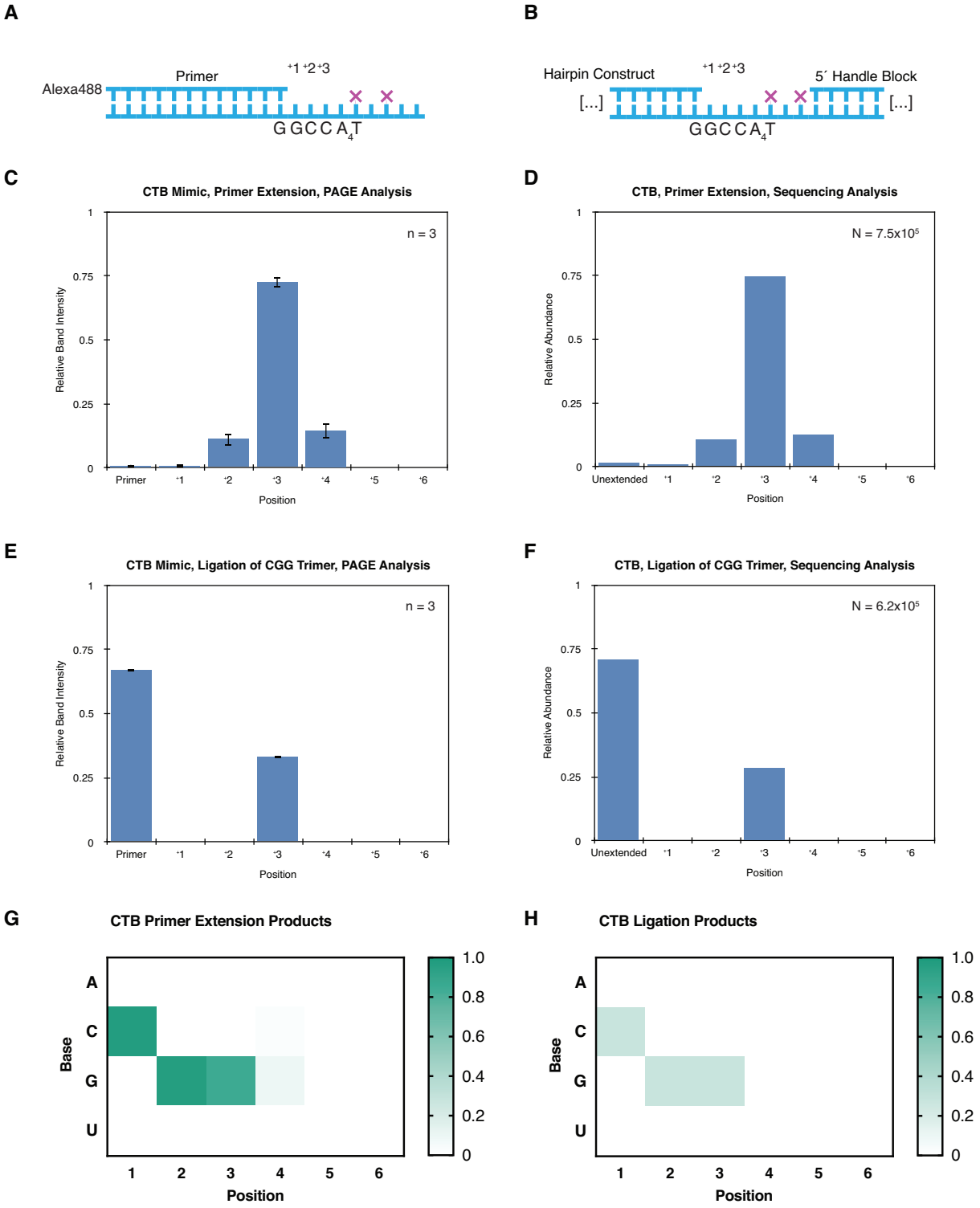

Figure 4. NERPE-Seq Can Accurately Measure Primer Extension on a Defined Template. (**A**) CTB Mimic, used in PAGE analysis, is a construct that mimics the CTB hairpin (**B**). (**C**) The extent of primer extension with 20 mM each of 2AIrG and 2AIrC after 24 h using CTB Mimic, as measured by PAGE analysis. (**D**) The same reaction as in (C), but measured by NERPE-Seq on CTB. (Number of read pairs prior to filtering = 1.8×10^6.) (**E**) The extent of primer extension with 500 μM 2AI-CGG after 24 h using CTB Mimic, as measured by PAGE analysis. (**F**) The same reaction as in (E), but measured by NERPE-Seq on CTB. (Number of read pairs prior to filtering = 1.6×10^6.) (**G**) NERPE-Seq reveals the expected pattern of products from the polymerization reaction (same experiment as in D) and (**H**) from the non-enzymatic ligation reaction (same experiment as in F). The heat maps show the frequency of each base at each position, including nulls. Note in (G) the mismatched over-extension of G and C products across the templating A in position 4. In (H) the relative intensities at each position are equivalent because all products result from trimer ligation (i.e.: all ligation events contribute equivalently to the +1, +2 and +3 positions). Finally, the overall intensities in (G) are higher than in (H) because primer extension by polymerization is in this case more efficient than ligation.

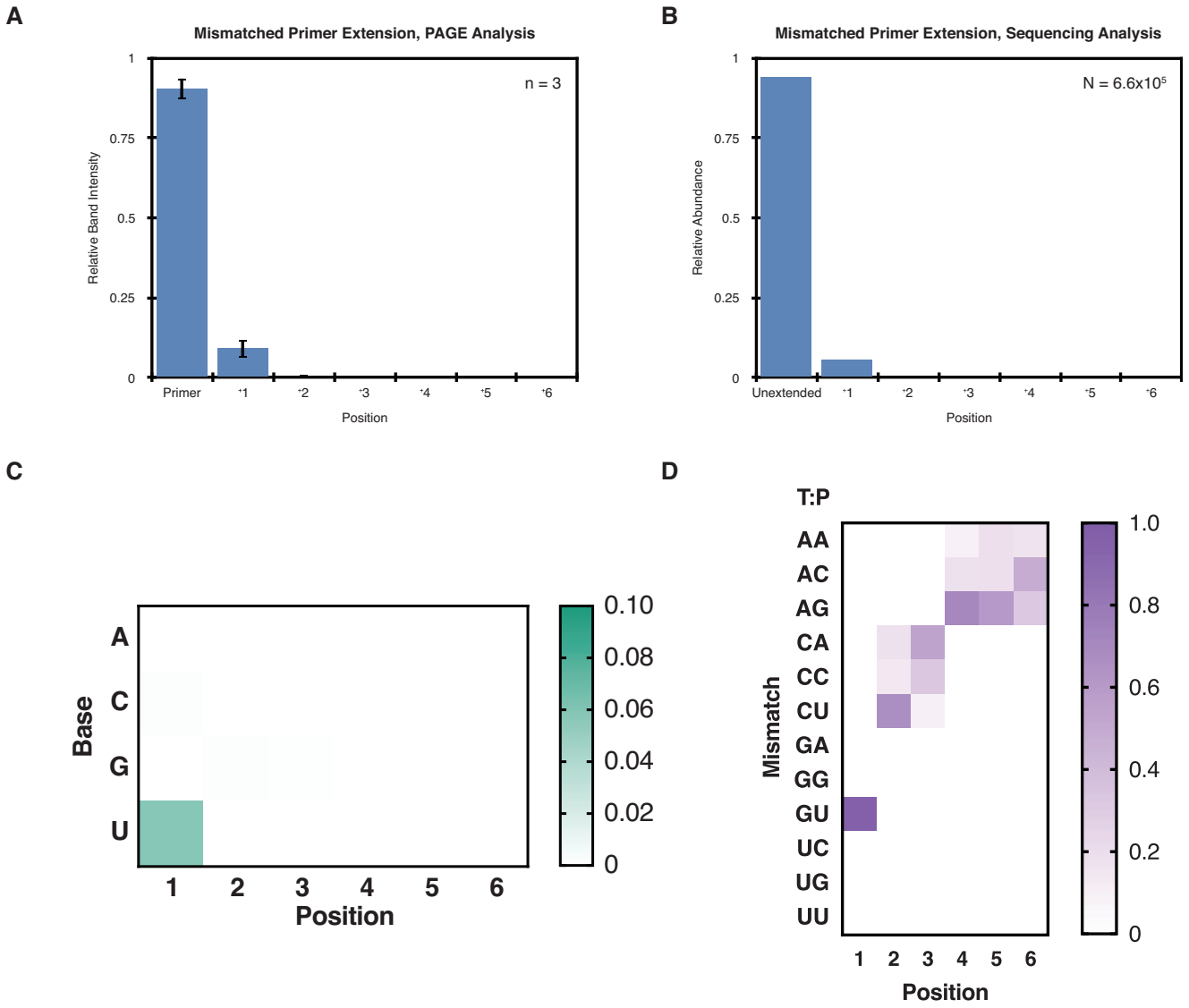

Figure 5. NERPE-Seq can measure mismatch frequencies. (**A**) The extent of primer extension on a 3′-GCC-5′ template incubated with 20 mM 2AIrU for 24 h as measured by PAGE analysis. (**B**) The same experiment as in (A), but measured by NERPE-Seq. (Number of read pairs prior to filtering = 1.6×10^6.) (**C**) NERPE-Seq reveals the expected pattern of products. The heat map shows the proportion of each base at each position, including nulls. Note the intensity scale bar maximum value is set to 0.1 because there are relatively few products. (**D**) The position-dependent distribution of all possible mismatches. The heat map shows the proportion of each mismatch at each position relative to all mismatches in that position. As expected, the Template:Product (T:P) mismatch of G:U in the first position dominates the data, and C:U is also evident in the second position. The data become noisy even by position 2 because there are very few extension events beyond +1 (>98% of the data is in position 1).

mismatches; $n_{\text{mismatches}} = 3.9 \times 10^4$) falls in the position 1 bin, with just under 1% in the position 2 bin. This indicates that mismatch data begins to encounter error noise when a given position harbors <1% of all mismatches, a useful metric for future analyses.

The 6N template and oligonucleotide synthesizer errors. Our data analysis assumes that a randomized synthetic template will not be truly random, and calculates normalization factors to adjust product proportions. The phosphoramidites used in solid-state oligo synthesis are known to exhibit distinct coupling efficiencies (46), but even if these are accounted for by mixing appropriately scaled molar quantities of each base, we cannot assume that the product will be exactly random. NERPE-Seq automatically measures the template base distributions, so we performed a negative control experiment with the 6N hairpin construct, which has a 6-nt random template prepared with an equimolar mixture of all four phosphoramidites. 6N was not exposed to activated nucleotides or oligonucleotides but was otherwise prepared and analyzed exactly as the other samples. The template is measured to be not fully random (Supplementary Figure S9A), with G over-represented and C under-represented. The base distribution also changes as a function of position. To better understand this observation, we measured the position-dependent sequential

base transition frequencies (the probability of each subsequent base given the current base, Supplementary Figure S9B). Sequential transitions from C-to-G (that is, 3′-CG-sequences in the template) are favored, and there are other more subtle effects. This suggests that truly random distributions by solid-state oligo synthesis are inherently impossible because pairwise base identities influence coupling efficiencies. These results highlight the necessity of the normalization approach we implemented in our data analysis.

The data analysis filters out sequencing reads with an average quality score below 30, a threshold that corresponds to a per-base error rate of 0.1%. The analysis also checks reads for perfect matches to the fixed sequences from the hairpin construct. These checks typically flag about half of the data for disposal, which is significantly more than would be expected if only sequencing errors were involved. For example, 23.4% of read pairs that had already passed with the requisite average quality score were filtered out of the 6N data because they did not harbor a correct Fix 1 sequence in R1. This corresponds to a 1.15% per-base error, much higher than expected after initial quality filtering. We suggest that this additional error is primarily due to imperfect oligo synthesis. This interpretation implies that the errors are genuine incorrect bases present in the sequence rather than mistakes made by the sequencer, and should therefore occur in both R1 and R2. (If they were sequencing errors, then the error frequency in R1 and R2 would be uncorrelated or only weakly correlated.) The data analysis filters sequencing reads in pairs: if an R1 read is filtered out, then the corresponding R2 read is also discarded. Therefore, if an error occurs in both R1 and R2, then the R1 filter pass should eliminate the corresponding R2 reads. When R2 is subsequently filtered for correct Fix 1 sequences, only an additional 0.021% of read pairs are discarded, *corresponding to an average per-base error of only 0.00089%* (one estimate of sequencing error). This strongly suggests that Fix 1 is sequenced correctly but contains errors from the synthesizer: read pairs with deviations from the known Fix 1 sequence are discarded during the R1 pass, and the R2 pass only results in a small additional loss due to sequencing errors. A comprehensive accounting of error sources can be found in the Supplementary Data.

To see whether synthesizer errors show any interpretable patterns we plotted the base distributions of Fix 1 errors from the filtered R1 reads (Supplementary Figure S9C). Each read in this data contains at least one base that is incorrect compared with the defined Fix 1 sequence. The errors appear somewhat random overall, but in more than half of positions the highest frequency incorrect base is the same as the *preceding* correct base in the 3′-to-5′ direction, which is the direction in which the oligonucleotide was synthesized. The simplest explanation is contamination during a given synthesis cycle from the previously added phosphoramidite, which would result in a substitution mutation. Alternatively, one phosphoramidite base could be added twice in a single synthesis cycle, resulting in an insertion mutation. Both types of errors are filtered out by pre-processing. We conclude that the large proportion of filtered read pairs results primarily from errors during oligo synthesis and that a discernible fraction of those errors can be traced to line contamination or double addition during synthesis cycles.

DISCUSSION

Experimentally capturing the properties of an emergent life-like replicating system in an RNA world scenario will require accommodating the complexity of the sequence space accessed by non-enzymatic copying. It is this sequence space that will determine how, or even whether, primitive RNA phenotypes can be explored. Although the chemistry of non-enzymatic polymerization and ligation is now well-understood, much less is known about how it will influence copied sequence properties generally, especially in the context of random templates and oligomers and all four canonical ribonucleotides. The fidelity of primer extension will dictate whether favorable sequences can be maintained across cycles of copying. Current methods cannot routinely measure the sequence features of non-enzymatic RNA primer extension products across arbitrary templates.

Primer extension experiments are traditionally assayed by HPLC or electrophoresis, especially denaturing PAGE. Both techniques separate products based on size, and interpreting the results requires knowledge of the template and potential product sequences. Mass spectroscopy can precisely identify multiple products if the number of possible outcomes is not so high that the spectrogram becomes too densely populated to interpret (25,47–48). Here too the template must be defined for the results to be unambiguous. HPLC, PAGE and mass spectroscopy work on non-canonical RNA (49–52), unlike sequencing, which relies on biological enzymes. Deep-sequencing has previously been used to characterize the products of non-enzymatic primer extension on a small set of defined templates (26). That study isolated the product strands and inferred mismatch identities based on the known template sequences. To date, there has been no generalizable method to directly measure individual template-product pairs across sequence space.

The NERPE-Seq analysis generates comprehensive data about the templates, products and mismatches in a primer extension experiment, and multiplexing enables reports from several experiments at once. We also gain the benefit of high throughput because deep-sequencing generates millions of sequencing reads. Quality filtering jettisons about half of the data, primarily because of errors from oligo synthesis. In the case of an RNA hairpin construct with six randomized template bases there are $4^6 = 4096$ possible template sequences, so the approximately half million read pairs typically retained per sample after filtering affords over 100× coverage. NERPE-Seq is applicable to any experiment compatible with RNA. The constructs are made by solid-state oligo synthesis and can therefore harbor any template sequence chosen by the experimenter, within the limits of synthesis capabilities. This versatility allows for the study of any standard RNA primer extension experiment on any template, and the use of substrates in any combination so long as they are compatible with RT (53). We anticipate that NERPE-Seq could easily be modified for DNA-based non-enzymatic reactions (54), as well as biological or putative ribozyme-catalyzed primer extension (55).

As we begin to probe higher-order primer extension experiments, the set of reactants will only multiply. Experiments with all four nucleotides are rare, partly because the reaction is inefficient and partly because the results are

challenging to interpret without sequencing data (56,57). NERPE-Seq may help us address the relationship between inefficient primer extension with all four bases and the role of the bridged intermediate. A unique feature of the bridged intermediate pathway is that sequential pairs of templating bases are relevant to each additional added base. This means that the +2 templating position, and whatever substrates occupy it, or attempt to occupy it, can affect what happens at the +1 product position. NERPE-Seq will directly report on these features. Of similar interest is the role, if any, of the ratios of activated and unactivated nucleotide species, and reactant heterogeneity (47,56–59). NERPE-Seq will enable the routine analysis of such experiments in outstanding detail.

DATA AVAILABILITY

The NERPE-Seq analysis code is available in the GitHub repository: https://github.com/CarrCE/NERPE-Seq.

Sequencing reads and NERPE-Seq analysis output data is available at OSF.io:
https://osf.io/uk9t4/?view_only=5096935d030d48a4b7437cd27d02b3f0.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the staff of the MGH Department of Molecular Biology Next-generation Sequencing Core for sample validation, MiSeq runs and communicating technical knowledge. We are grateful to Saurja DasGupta, Li Li, Stephanie Zhang, Lijun Zhou and Aleksandar Radaković for comments on the manuscript, and members of the Szostak laboratory for feedback and sharing experimental expertise.

FUNDING

Simons Foundation [290363 to J.W.S.,]; National Science Foundation [CHE-1607034 to J.W.S.]; National Aeronautics and Space Administration [NNX15AF85G, 80NSSC19K1028, 80NSSC18K1301 to C.E.C.]. J.W.S. is an Investigator of the Howard Hughes Medical Institute. Funding for open access charge: Howard Hughes Medical Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Gilbert,W. (1986) Origin of life—the RNA world. *Nature*, **319**, 618–618.
2. Orgel,L.E. (1989) Was RNA the first genetic polymer? *Evolutionary Tinkering in Gene Expression*. Vol. **169**, pp. 215–224.
3. Joyce,G.F. (2002) The antiquity of RNA-based evolution. *Nature*, **418**, 214–221.
4. Robertson,M.P. and Joyce,G.F. (2012) The origins of the RNA world. *Cold Spring Harb.Perspect. Biol.*,**4**, a003608.
5. Krishnamurthy,R. (2015) On the emergence of RNA. *Isr. J. Chem.*, **55**, 837–850.
6. Szostak,J.W. (2017) The narrow road to the deep past: in search of the chemistry of the origin of life. *Angew. Chem. Int. Ed.*, **56**, 11037–11043.
7. Joyce,G.F. and Szostak,J.W. (2018) Protocells and RNA self-replication. *Cold Spring Harb.Perspect. Biol.*, **10**, a034801.
8. Weimann,B.J., Lohrmann,R., Orgel,L.E., Schneide,H. and Sulston,J.E. (1968) Template-directed synthesis with Adenosine-5'-Phosphorimidazolide. *Science*, **161**, 387.
9. Sulston,J., Lohrmann,R., Orgel,L.E. and Miles,H.T. (1968) Nonenzymatic synthesis of oligoadenylates on a polyuridylic acid template. *Proc. Natl. Acad. Sci. U.S.A.*, **59**, 726–733.
10. Lohrmann,R., Bridson,P.K., Bridson,P.K. and Orgel,L.E. (1980) Efficient metal ion catalyzed template-directed oligonucleotide synthesis. *Science*, **208**, 1464–1465.
11. Rohatgi,R., Bartel,D.P. and Szostak,J.W. (1996) Kinetic and mechanistic analysis of nonenzymatic, template-directed oligoribonucleotide ligation. *J. Am. Chem. Soc.*, **118**, 3332–3339.
12. Kervio,E., Sosson,M. and Richert,C. (2016) The effect of leaving groups on binding and reactivity in enzyme-free copying of DNA and RNA. *Nucleic Acids Res.*, **44**, 5504–5514.
13. Fahrenbach,A.C., Giurgiu,C., Tam,C.P., Li,L., Hongo,Y., Aono,M. and Szostak,J.W. (2017) Common and potentially prebiotic origin for precursors of nucleotide synthesis and activation. *J. Am. Chem. Soc.*, **139**, 8780–8783.
14. Li,L., Prywes,N., Tam,C.P., O'Flaherty,D.K., Lelyveld,V.S., Izgu,E.C., Pal,A. and Szostak,J.W. (2017) Enhanced nonenzymatic RNA copying with 2-Aminoimidazole activated nucleotides. *J. Am. Chem. Soc.*, **139**, 1810–1813.
15. Walton,T. and Szostak,J.W. (2016) A highly reactive imidazolium-bridged dinucleotide intermediate in nonenzymatic RNA primer extension. *J. Am. Chem. Soc.*, **138**, 11996–12002.
16. Walton,T. and Szostak,J.W. (2017) A kinetic model of nonenzymatic RNA polymerization by Cytidine-5′-phosphoro-2-aminoimidazolide. *Biochemistry*, **56**, 5739–5747.
17. Zhang,W., Tam,C.P., Walton,T., Fahrenbach,A.C., Birrane,G. and Szostak,J.W. (2017) Insight into the mechanism of nonenzymatic RNA primer extension from the structure of an RNA-GpppG complex. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 7659–7664.
18. Zhang,W., Walton,T., Li,L. and Szostak,J.W. (2018) Crystallographic observation of nonenzymatic RNA primer extension. *eLife*, **7**, e36422.
19. Walton,T., Zhang,W., Li,L., Tam,C.P. and Szostak,J.W. (2019) The mechanism of nonenzymatic template copying with imidazole-activated nucleotides. *Angew. Chem. Int. Ed.*, **58**, 10812–10819.
20. Lohrmann,R. and Orgel,L.E. (1978) Preferential formation of (2'-5')-linked internucleotide bonds in non-enzymatic synthesis. *Tetrahedron*, **34**, 853–855.
21. Rohatgi,R., Bartel,D.P. and Szostak,J.W. (1996) Nonenzymatic, template-directed ligation of oligoribonucleotides is highly regioselective for the formation of 3'-5' phosphodiester bonds. *J. Am. Chem. Soc.*, **118**, 3340–3344.
22. Sheng,J., Li,L., Engelhart,A.E., Gan,J.H., Wang,J.W. and Szostak,J.W. (2014) Structural insights into the effects of 2 '-5 ' linkages on the RNA duplex. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 3050–3055.
23. Giurgiu,C., Li,L., O'Flaherty,D.K., Tam,C.P. and Szostak,J.W. (2017) A mechanistic explanation for the regioselectivity of nonenzymatic RNA primer extension. *J. Am. Chem. Soc.*, **139**, 16741–16747.
24. Rajamani,S., Ichida,J.K., Antal,T., Treco,D.A., Leu,K., Nowak,M.A., Szostak,J.W. and Chen,I.A. (2010) Effect of stalling after mismatches on the error catastrophe in nonenzymatic nucleic acid replication. *J. Am. Chem. Soc.*, **132**, 5880–5885.
25. Leu,K., Kervio,E., Obermayer,B., Turk-MacLeod,R.M., Yuan,C., Luevano,J.M., Chen,E., Gerland,U., Richert,C. and Chen,I.A. (2013) Cascade of reduced speed and accuracy after errors in enzyme-free copying of nucleic acid sequences. *J. Am. Chem. Soc.*, **135**, 354–366.
26. Heuberger,B.D., Pal,A., Del Frate,F., Topkar,V.V. and Szostak,J.W. (2015) Replacing uridine with 2-Thiouridine enhances the rate and fidelity of nonenzymatic RNA primer extension. *J. Am. Chem. Soc.*, **137**, 2769–2775.
27. Prywes,N., Michaels,Y.S., Pal,A., Oh,S.S. and Szostak,J.W. (2016) Thiolated uridine substrates and templates improve the rate and fidelity of ribozyme-catalyzed RNA copying. *Chem. Commun.*, **52**, 6529–6532.
28. Hanle,E. and Richert,C. (2018) Enzyme-free replication with two or four bases. *Angew. Chem. Int. Ed.*, **57**, 8911–8915.

29. Schindelin,J., Arganda-Carreras,I., Frise,E., Kaynig,V., Longair,M., Pietzsch,T., Preibisch,S., Rueden,C., Saalfeld,S., Schmid,B. *et al.* (2012) Fiji: an open-source platform for biological-image analysis. *Nat. Methods*, **9**, 676–682.
30. Wu,T.F. and Orgel,L.E. (1992) Nonenzymatic template-directed synthesis on oligodeoxycytidylate sequences in hairpin oligonucleotides. *J. Am. Chem. Soc.*, **114**, 317–322.
31. Wu,T.F. and Orgel,L.E. (1992) Nonenzymatic template-directed synthesis on Oligodeoxycytidylate sequences in hairpin Oligonucleotides 2. Templates containing cytidine and guanosine residues. *J. Am. Chem. Soc.*, **114**, 5496–5501.
32. Wu,T. and Orgel,L.E. (1992) Nonenzymatic template-directed synthesis on Oligodeoxycytidylate sequences in hairpin Oligonucleotides 3. Incorporation of adenosine and uridine residues. *J. Am. Chem. Soc.*, **114**, 7963–7969.
33. Hafner,M., Renwick,N., Brown,M., Mihailovic,A., Holoch,D., Lin,C., Pena,J.T.G., Nusbaum,J.D., Morozov,P., Ludwig,J. *et al.* (2011) RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA*, **17**, 1697–1712.
34. Zhuang,F.L., Fuchs,R.T., Sun,Z.Y., Zheng,Y. and Robb,G.B. (2012) Structural bias in T4 RNA ligase-mediated 3 '-adapter ligation. *Nucleic Acids Res.*, **40**, e54.
35. Fuchs,R.T., Sun,Z.Y., Zhuang,F.L. and Robb,G.B. (2015) Bias in ligation-based small RNA sequencing library construction is determined by Adaptor and RNA structure. *PLoS One*, **10**, e0126049.
36. Ozsolak,F. and Milos,P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98.
37. Zhang,Z.J., Lee,J.E., Riemondy,K., Anderson,E.M. and Yi,R. (2013) High-efficiency RNA cloning enables accurate quantification of miRNA expression by deep sequencing. *Genome Biol.*, **14**, R109.
38. Lusic,H. and Deiters,A. (2006) A new photocaging group for aromatic N-heterocycles. *Synthesis*, **8**, 2147–2150.
39. Nandakumar,J., Shuman,S. and Lima,C.D. (2006) RNA ligase structures reveal the basis for RNA specificity and conformational changes that drive ligation forward. *Cell*, **127**, 71–84.
40. Viollet,S., Fuchs,R.T., Munafo,D.B., Zhuang,F. and Robb,G.B. (2011) T4 RNA Ligase 2 truncated active site mutants: improved tools for RNA analysis. *BMC Biotech.*, **11**, 72.
41. Lorsch,J.R., Bartel,D.P. and Szostak,J.W. (1995) Reverse-transcriptase reads through a 2'-5'-linkage and a 2'-thiophosphate in a template. *Nucleic Acids Res.*, **23**, 2811–2814.
42. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
43. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
44. Wright,E.S. and Vetsigian,K.H. (2016) Quality filtering of Illumina index reads mitigates sample cross-talk. *BMC Genomics*, **17**, 876.
45. Bellaousov,S., Reuter,J.S., Seetin,M.G. and Mathews,D.H. (2013) RNAstructure: Web servers for RNA secondary structure prediction and analysis. *Nucleic. Acids. Res.*, **41**, W471–W474.
46. Ellington,A., Jack,D. and Pollard,J. (2000) Introduction to the Synthesis and Purification of Oligonucleotides. *Curr. Protoc. Nucleic Acid Chem.*, doi:10.1002/0471142700.nca03cs00.
47. Deck,C., Jauker,M. and Richert,C. (2011) Efficient enzyme-free copying of all four nucleobases templated by immobilized RNA. *Nat. Chem.*, **3**, 603–608.
48. Sosson,M. and Richert,C. (2018) Enzyme-free genetic copying of DNA and RNA sequences. *Beilstein J. Org. Chem.*, **14**, 603–617.
49. Blain,J.C., Ricardo,A. and Szostak,J.W. (2014) Synthesis and nonenzymatic template-directed polymerization of 2′-Amino-2′-deoxythreose Nucleotides. *J. Am. Chem. Soc.*, **136**, 2033–2039.
50. O'Flaherty,D.K., Zhou,L.J. and Szostak,J.W. (2019) Nonenzymatic template-directed synthesis of mixed-sequence 3′-NP-DNA up to 25 nucleotides long inside model protocells. *J. Am. Chem. Soc.*, **141**, 10481–10488.
51. Kim,S.C., O'Flaherty,D.K., Zhou,L.J., Lelyveld,V.S. and Szostak,J.W. (2018) Inosine, but none of the 8-oxo-purines, is a plausible component of a primordial version of RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 13318–13323.
52. Wright,T.H., Giurgiu,C., Zhang,W., Radakovic,A., O'Flaherty,D.K., Zhou,L.J. and Szostak,J.W. (2019) Prebiotically plausible "Patching" of RNA backbone cleavage through a 3′-5′ pyrophosphate linkage. *J. Am. Chem. Soc.*, **141**, 18104–18112.
53. Lelyveld,V.S., O'Flaherty,D.K., Zhou,L.J., Izgu,E.C. and Szostak,J.W. (2019) DNA polymerase activity on synthetic N3′ → N5′ phosphoramidate DNA templates. *Nucleic Acids Res.*, **47**, 8941–8949.
54. Bhowmik,S. and Krishnamurthy,R. (2019) The role of sugar-backbone heterogeneity and chimeras in the simultaneous emergence of RNA and DNA. *Nat. Chem.*, **11**, 1009–1018.
55. Le Vay,K., Weise,L.I., Libicher,K., Mascarenhas,J. and Mutschler,H. (2019) Templated self-replication in biomimetic systems. *Adv. Biosyst.*, **3**, 1800313.
56. Prywes,N., Blain,J.C., Del Frate,F. and Szostak,J.W. (2016) Nonenzymatic copying of RNA templates containing all four letters is catalyzed by activated oligonucleotides. *eLife*, **5**:e17756.
57. Zhou,L.J., Kim,S.C., Ho,K.H., O'Flaherty,D.K., Giurgiu,C., Wright,T.H. and Szostak,J.W. (2019) Non-enzymatic primer extension with strand displacement. *eLife*, **8**, e51888.
58. Mariani,A., Russell,D.A., Javelle,T. and Sutherland,J.D. (2018) A light-releasable potentially prebiotic nucleotide activating agent. *J. Am. Chem. Soc.*, **140**, 8657–8661.
59. Sosson,M., Pfeffer,D. and Richert,C. (2019) Enzyme-free ligation of dimers and trimers to RNA primers. *Nucleic Acids Res.*, **47**, 3836–3845.