An optimized kit-free method for making strand-specific deep sequencing libraries from RNA fragments

Erin E. Heyer1,2,3, Hakan Ozadam1,2,3, Emilio P. Ricci1,2,3, Can Cenik1,2,3,4 and Melissa J. Moore1,2,3,*

1Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605, USA, 2RNA Therapeutics Institute, University of Massachusetts Medical School, Worcester, MA 01605, USA, 3Howard Hughes Medical Institute, University of Massachusetts Medical School, Worcester, MA 01605, USA and 4Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

Received February 03, 2014; Revised October 09, 2014; Accepted November 10, 2014

ABSTRACT

Deep sequencing of strand-specific cDNA libraries is now a ubiquitous tool for identifying and quantifying RNAs in diverse sample types. The accuracy of conclusions drawn from these analyses depends on precise and quantitative conversion of the RNA sample into a DNA library suitable for sequencing. Here, we describe an optimized method of preparing strand-specific RNA deep sequencing libraries from small RNAs and variably sized RNA fragments obtained from ribonucleoprotein particle footprinting experiments or fragmentation of long RNAs. Our approach works across a wide range of input amounts (400 pg to 200 ng), is easy to follow and produces a library in 2–3 days at relatively low reagent cost, all while giving the user complete control over every step. Because all enzymatic reactions were optimized and driven to apparent completion, sequence diversity and species abundance in the input sample are well preserved.

INTRODUCTION

In cells, all RNA molecules interact with RNA binding proteins (RBPs) to form ribonucleoprotein particles (RNPs). An ever-increasing number of methodological approaches to map these protein-RNA interaction sites transcriptome-wide. Such techniques include ultraviolet-crosslinking methods (e.g., CLIP, PAR-CLIP (1, 2)) to map the ribonucleotides directly in contact with an individual RBP and RNP footprinting (e.g., Ribo-Seq, RIP-T-Seq; (3, 4)) to map the occupancy sites of larger complexes. Many projects in our laboratory are focused on transcriptome-wide RNP footprint analysis (5–7). Depending on the complex being examined and the RNA fragmentation method utilized (e.g., RNAase or sonication), bound RNA fragments can range from 10 to 200 nucleotides (nts). Therefore, we require a strand-specific library generation method that works for diverse RNA lengths, faithfully preserves their relative abundances in the original sample and excludes any contaminating DNA fragments.

Multiple commercial kits currently exist for strand-specific library preparation, but most are intended to capture either long RNAs (e.g., RNA-Seq) or short RNAs (e.g., miRNA-Seq), but not both. Further, commercial kits are regularly updated with new preparation methods. Because preparation method is the primary source of variability between deep sequencing libraries (8), quantitative comparisons are best done between identically generated libraries (i.e., with a single commercial kit version). However, the expense of commercial kits (and remaking libraries as new kits appear and older versions are phased out) is cost prohibitive for many academic laboratories. We therefore set out to develop an optimized, strand-specific RNA library preparation protocol that utilizes commercial only available reagents and works over a wide range of input amounts. We also wanted an approach that can be used to capture full-length RNP footprints as well as sites of reverse transcriptase stalling (e.g., sites of RNA-protein crosslinking from CLIP experiments or abasic/alkylated sites).

All current library preparation methods utilize enzymes to capture nucleic acid fragments by appending 5' and 3' adaptor sequences. Enzymes have inherent substrate preferences that are most significant at low substrate concentrations (K<sub>m</sub>/K<sub>c</sub> conditions) and at short reaction times (9). For ligation reactions, low tem rates can favor capture of sequences capable of base pairing with the adaptor (10). Low tem rates can also disfavor capture of sequences containing internal secondary structures. M any published library preparation protocols are suboptimal for
one or more of these factors, resulting in differential capture of an aRNA (e.g., RNA-Seq; [10–12]) and highly non-uniform (‘peaky’) coverage of long RNAs (e.g., RNA-Seq of RNA Pol II transcripts; [13]). For these reasons, we decided to re-examine 5- and 3-end capture conditions, with the goal of driving every reaction to completion.

Here, we present the detailed protocol for strand-specific RNA library preparation currently in use in our laboratory, as well as the titration and time course data we used to optimize each step. A key presented are deep sequencing data on (i) the effects of time and temperature on initial 3-end capture and (ii) capture uniformity for an equimolar pool of 29 miRNAs. Taken together, these data show that our method faithfully preserves fragment diversity and abundance in complex starting material and is minimally affected by fragment sequence or tailing potential.

MATERIALS AND METHODS

Gel analysis

All acrylamide gels were prepared using AccuGel reagents. Circularization reactions were performed and analyzed on native 8% PAGE gels. PCR products were similarly prepared and analyzed on the Double Wide Mini-Vertical system (C.B.S. Scientific) to limit the amount of heat denaturation. Gels were either exposed to a phosphor imager screen (Amersham Biosciences) or stained with SYBR Gold (Invitrogen) prior to visualization on a Typhoon Trio (Amersham Biosciences). Quantified cations were performed with ImageQuant (GE Healthcare).

3' -adaptor ligation

Indicated amounts of either 5'-32P-labeled N24 RNA oligonucleotide (Dharmacon) or 28-mer oligonucleotide (5'-AUGUACGAGGAUGGACACCCGACAAGCCA-3'; IDT) were ligated to preadenylated adaptors in Cat-33 (5'-AppTGGAATTCCTGGTGCCAAGGdC-3' JDT) or EH-preadenylated adaptors in Cat-33 (5'-AppNNNNTGGAATTCCTGGTGCCAAGGdC-3' JDT) oligonucleotide (Dharmacon) or 28-mer oligonucleotide (5'-AUGUACGAGGAUGGACACCCGACAAGCCA-3'; IDT) using T4 RNA ligase (Invitrogen) with the conditions described in this paper. Due to the high viscosity of 50% PEG 8000, we found that low retention tips aided consistent pipetting while simultaneously preventing sample cross-contamination. Ligation efficiencies were calculated by dividing the quantified pixel signal of ligation products by the total amount of RNA signal (bands corresponding to both ligated and unligated RNA) in each lane, and multiplying by 100.

Reverse transcription

RT was performed with gel-purified RT primers (5'-GG -B-AGATCGGAGAAGGCTCGTGTAGGGAAAGAGTTGGT-3' SP18-CCTGGCAATCTCTGCTGAAACCTGCTTCCGATCT-CCTTGGCACCCGAGATTCCA-3', where B indicates a 5-nt barcode of sequence ATCAC, CGATG, TAGCT, GCTCC, ACAGT, CAGAT, TCCCG, GGCTA, AGTCA, CTCTT, TGAAT or GTAGA). RT products were detected by incorporating 32P-dCTP in the reaction. RT products intended for circularization were gel purified. For the data in Figures 4 and 5, we gel purified the cDNA from crushed gel pieces in 300 mM NaCl, 1 mM ethylenediaminetetraacetic acid (EDTA) during an overnight incubation at room temperature with constant rotation; eluted material was ethanol precipitated before circularization. We have since modified our approach to increase elution yield by eluting in TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0) and incubating at 37°C overnight with constant rotation. With this buffer, we can concentrate the eluate (either by butanol extraction or SpeedVac) before precipitating the sample in a single tube.

Circularization efficiency and PCR amplification

Circularization reactions were performed on gel-purified RT products as described in the text. The single-stranded DNA input was either body-labeled with 32P-dCTP in the RT reaction or end-labeled in an exchange reaction with 32P-ATP. Circularized RT products were separated from non-reactive, linear RT products on 10% denaturing PAGE gels, and the gels were exposed and quantified as described. The amount of circularization was determined by quantifying the pixel signal corresponding to the circularized product and dividing that value by the total pixel signal corresponding to the circularized product plus the linear input, and multiplying by 100.

PCR amplification from the circularized RT products was performed with KAPA HiFi Library Amplification Kit (Kapa Biosystems) according to manufacturer's instructions, except where otherwise noted. All PCR products were analyzed on native 8% PAGE gels and quantified as described above. Samples to be sequenced were excised and gel-extracted as described for RT products, precipitated, and quantified by gel analysis before sample submission.

N4 library construction and analysis

N4 libraries were constructed from 2 pmol of N24 RNA oligo using the optimized conditions shown in Supplementary Table S1, except for the described variations in 3' ligation conditions. In one case (22°C 6 hr library), a minute amount of 5'-32P-labeled N24 RNA oligo was added. All libraries were amplified with 7 PCR cycles and gel-purified prior to sequencing on a single Illumina HiSeq2000 lane (G new iz).

Deep sequencing data were analyzed with custom scripts unless otherwise noted. Data were parsed into individual
libraries by 5 barcode, allowing 1 m is match. The 3 adapt-
ors sequence was used from all libraries allowing 3 m is-
matches. Once individual sequence reads were identified, read lengths were calculated. All subsequent analysis utilized only 24 nt reads. For each library, we calculated the observed nt frequencies at each of the 24 positions. To de-
tem the expected values, we used the data across positions 5–20 from all libraries and used least squares lines to the frequency pattern for each nt. The equations for the line-
ts yielded the expected nt frequencies at all 24 positions. The chi-square statistic was calculated for each library by sum-
ing \[\frac{(\text{observed nt count} - \text{expected nt count})^2}{\text{expected nt count}}\] across all 4 nt counts at each N 24 position.

PhiX reads were identified if they mapped to the Phi X 174 genome with a maximum of 6 errors within the 51 sequence reads. Mismatches were identified and counted if the PhiX reads were identical to each position by the total number of nt frequencies were calculated by dividing the observed nt count at each position by the total number of reads.

miRNA library construction and analysis

Sequencing data were parsed into individual libraries by 5 barcode using cutadapt version 1.3 (15), allowing 1 mismatch. The 3 adaptor ligations were selected and aligned by their 3 ends; nt frequencies were calculated by dividing the observed nt count at each position by the total number of reads.

mRNA sequencing analysis

Libraries were constructed from either 1 pmol or 50 fmol of an equimolar mix of 29 mRNAs (14) according to the optimal conditions shown in Supplementary Table S1. For each input amount, the ligation was performed with either the xed or N4 preadenylated 3 -adaptor. Libraries were pooled and sequenced on a single MiSeq lane. Deep sequencing data were parsed into individual libraries by 5 barcode using cutadapt version 1.3 (15), allowing 1 m is match. Reads were mapped to reference sequences using a custom script which (i) required that the 3 adaptor be present in the read and (ii) only counted reads mapping to reference mRNAs without 0 m is matches. Additionally, we counted the reads with 5 or fewer non-template 5' terminal additions and 5 or fewer 5' terminal deletions. Observed mRNA frequencies \(F_{\text{obs}}\) were calculated using the total numbers of reads for each mRNA sequence. Coefficients of variation (CV) were calculated by dividing standard deviation in mRNA counts by the mean mRNA counts. Terminal transese expression was assessed by dividing total number of reads in each 5 addition bin by the total full-length mRNA reads in each library. Error-free values from in silico folding were calculated using the Vienna RNA Package v.2.1.7 using the -T 30 parameter to obtain structure predictions at 30°C (16).

RESULTS

Protocol design

To generate strand-specific deep sequencing libraries, both ends of the captured RNA must be appended to 'read' sequences (adaptors) to enable primer hybridization for an -
**Figure 1. Method overview.**

**Step 1: Ligation.** RNA, shown in blue, is ligation to a preadenylated DNA adapter to form a RNA-DNA hybrid. The same tube, RT (Step 2), contains both the reverse and forward priming sequences for Illumina sequencing, as well as a barcode to uniquely identify the sample (Step 3). The RT product is gel purified, removing unligated adapters and unextended RT primers from the sample (Step 4). The gel purified RT product is circularized, forming a template for PCR (Step 5). The PCR product is then purified and used for deep sequencing (Step 6).

Protocol optimization

For optimization of each step, we used a pool of random-primed RNA 24mers (N24) to mimic the diversity of sequences in a biological sample. Ligation reactions were visualized using 5′-end 32P-labeled RNAs. RT products were visualized by including 32P-dCTP in the RT reaction. Circularization reactions were visualized using either body-labeled or 5′-end-labeled RT products.

**Step 1: preadenylated 3′-adaptor ligation.** When we initiated this project, the manufacturer’s (NEB) suggested conditions for RNLI2Tr.K227Q ligation reactions were 500 nM single-stranded RNA, 1 M 3′ adaptor, 10 U/μl enzyme, and 15% w/v PEG 8000 in 1× reaction buffer at 16°C overnight. As our goal was to create a robust protocol that could be successfully employed over a wide range of RNA input concentrations, we set out to explore the limits of these parameters (Figure 2). For all experiments below, we pre-mixed the RNA and 3′-adaptor in water and incubated them at 65°C for 10 min prior to enzyme addition.

Ligation efficiency depends on successful collision of multiple components. Such collisions can be increased by molecular crowding agents (e.g. PEG) and/or dehydrating co-solutes (e.g. dimethyl sulfoxide (DMSO)), and published 3′-adaptor ligation protocols vary with regard to PEG 8000 and DMSO inclusion (34–38). Consistent with a recent report that 25% PEG 8000 enhances ligation efficiency (see Figure 4B in (38)), we found that 25% PEG 8000 resulted in near complete N24 ligation at 16°C (Figure 2A). However, increasing DMSO had no effect, regardless of PEG 8000 absence or presence (Figure 2B). Thus, all subsequent ligation reactions included 25% PEG 8000 but no DMSO.

We next titrated preadenylated 3′-adaptor, N24 and enzyme concentrations. Using two different N24 concentrations, near complete ligation was observed at all adaptor concentrations above 130 nM (Figure 2C). A t470 nM adaptor, ligation was highly efficient with N24 concentrations above 50 nM (Figure 2D) and enzyme concentrations above 6 U/μl (Figure 2E). A greater dependence of ligation efficiency on enzyme concentration at 10 nM N24 does suggest, however, that additional enzyme will increase yields for very dilute RNA samples (39).

Published reports using T4 RNA ligases for library preparation employ a wide range of reaction times (1 h to overnight) and temperatures (5°C–37°C) (1,23,34,37–43). However, higher temperatures should stabilize both intra- and inter-molecular secondary structures, potentially biasing ligations against internally structured RNAs and toward RNA sequences that partially base pair with the 3′-adaptor (10–11,27). Higher temperatures should alleviate these issues, but could decrease enzyme stability and increase RNA degradation. Using our N24 pool, we assessed ligation efficiencies across a range of incubation times and temperatures (Figure 2F). Both 4°C and 37°C yielded poor ligation (Step 1) and RT (Step 2) were carried out in a single tube without any cleanup or buffer exchange in between, and the sample is only subjected to a single gel purification (Step 3) after RT.
Figure 2. 3’ adaptor ligation optimization. (A) Ligation efficiency versus % PEG 8000 (w/v) (n=2; black line, mean). (B) Comparison of DMSO and PEG as ligation enhancers. Absence or presence of indicated species are indicated by − and +; ligation efficiencies are indicated below each lane. N24 RNA was 5’-end labeled with [32P]-ATP. (C) Ligation efficiency versus 3’-adaptor concentration (n=1). (D) Ligation efficiency versus N24 concentration (n=1). (E) Ligation efficiency versus RNL2 concentration at four different N24 RNA concentrations (n=1). (F) Ligation efficiency versus time and temperature (n=3; error bars, standard deviation). Circles indicate ligation conditions for N24 libraries. In all panels, data were generated by quantification of denaturing polyacrylamide gels similar to that shown in panel B; ligation efficiency = (ligated RNA : DNA product)/(unligated RNA + ligated RNA : DNA product) in each lane.

Ligation efficiencies at all incubation times. Using radioactively labeled RNA, we determined that the lower yields at 37°C were not due to increased RNA degradation (data not shown); rather, the plateau reached after 2 h suggests that enzyme is unstable at 37°C. All reactions incubated between 16°C and 30°C ultimately resulted in near complete ligation. However, the 16°C and 22°C reactions took longer to reach completion (10–14 h) than did the 25°C and 30°C reactions (4–6 h).

Based on all of the above data, we adopted the following as our standard ligation reaction conditions: 470 nM adaptor, 50-330 nM RNA, 6 U/1 RN L2 K227Q, 1× RNL2 reaction buffer (from NEB; 50 mM Tris HCl, pH 7.5 @ 25°C, 10 mM MgCl2, 1 mM DTT) plus an additional 1 mM DTT to ensure a reducing environment, incubated for 6 h at 30°C and then 20 min at 65°C (to heat inactivate the enzyme). These conditions yield efficient ligation over the wide range of RNA fragment lengths we generally obtain when footprinting endogenous RNP complexes (4–6).

Step 2: reverse transcription. A number of high fidelity reverse transcriptases are commercially available. For our purposes, we wanted an enzyme that produced a high yield of full-length product with minimal side products when added directly to the heat-inactivated/diluted 3’-adaptor ligation reaction from Step 1. We tested Accuscript (Agilent), AMV RT (Finnzymes), Superscript III (Invitrogen) and Transcriptor (Roche) (Figure 3A). In all cases, ligation reactions were diluted and supplemented with either (i) the appropriate amount of manufacturer-supplied 5× or 10× RT buffer or (ii) the same buffer minus MgCl2, as the Step 1 reaction already contains MgCl2, and concentrations of
MgCl₂ above 3 mM can inhibit RT (44). For all four enzymes (tested at the manufacturer’s recommended concentration), we observed more full-length RT product when no Mg²⁺ was added beyond that supplied by the diluted ligation reaction. As SuperScript III gave the highest RT product yield, we chose it for subsequent optimization. By varying the amount of the heat-inactivated Step 1 reaction in the Step 2 reaction, we determined that an all RT product yield was obtained when the ligation reaction constituted one-third of the total volume of the RT reaction (data not shown). This resulted in a nanomolar MgCl₂ concentration of 3.3 mM. At this 3-fold dilution, we found no inhibitory effect on RT by the PEG 8000 present in the Step 1 reaction; rather, Step 1 reactions containing 25% PEG 8000 gave the highest Step 2 yields (Figure 3B).

We next varied RT primer, enzyme and RNA input amounts. To maximize RT product yield, it is important that the RT primer concentration be greater than the 3’-adapter concentration but not excessively so, as this would favor empty circle formation in the subsequent circularization reaction (Step 4). We observed no advantage for RT yield when the RT primer:3’-adapter ratio was significantly higher than 1:3 (Figure 3C). Further, all SuperScript III concentrations above 3 U/μl gave comparable product yields (Figure 3D). Varying the temperature (50°C, 55°C and 60°C) and time (30 min and 1 h) of the RT reactions revealed 55°C for 30 min to be optimal (data not shown). When the input RNA was varied between 3.3 and 133 nM, the yield of RT product increased linearly across this range (Figure 3E and F). Thus, like the ligation reaction, the RT reaction proved highly robust and amenable to library construction over a wide range of input amoun
ts.

Based on the above data, we adopted the following as our standard Step 2 reaction conditions: 3-fold dilution of the
Step 3: gel purification. See Materials and Methods.

Step 4: circularization. There are currently two commercial enzymes for ssDNA circularization: CircLigase I and II (Epicentre). We tested both at 50 nM input ssDNA and found that CircLigase I gave much higher circularization efficiency (98–99%) than CircLigase II (45–61%) (Figure 4A). Betaine, a compound commonly used in PCR reactions to eliminate the energy difference between A–T and G–C base pairs, is recommended by Epicentre for use with CircLigase I. However, as no amount of betaine in the circularization reaction improved the efficiency to that obtained with CircLigase I, we decided to proceed with CircLigase I.

To explore the limits of CircLigase I performance, we tested a range of conditions. Changing the enzyme concentration and doubling or reducing by half the reaction volume did not significantly affect on circularization efficiency (data not shown), so we continued to use the manufacturer’s suggested conditions. A time course revealed that complete circularization with 5 U/l enzyme and 50 nM input ssDNA required at least 2 h at 60°C (Figure 4B). Titration of the 24 RT product indicated that ligation efficiency dropped off precipitously below 25 nM ssDNA (Figure 4C). This dropoff was unaffected by either increasing or decreasing the enzyme concentration (data not shown), but was substantially rescued by the inclusion of 1 M betaine in the circularization reaction (Figure 4D). In this case, circularization of <5 nM ssDNA by 24 RT product could not be detected by direct observation of the 32P-labeled substrate and product on a gel, relative PCR product yields served as a proxy for circularization yields, with cycle number adjusted for RNA input amount. In order to exclude the possibility of betaine stimulating the yield of the PCR reaction instead of the circularization reaction, we added betaine subsequent to heat inactivation of CircLigase I; under these conditions, no betaine-dependent increase in PCR signal was observed (data not shown).

Based on the above data, we adopted the following as our standard Step 4 reaction conditions: 1× CircLigase buffer (Epicentre), 1 M betaine, 50 M adenosine triphosphate, 25 mM MgCl2, and 5 U/l CircLigase I in 20 µl containing all of the ssDNA isolated in Step 3. This mixture is incubated

heat-denatured ligation reaction from Step 1, supplemented with 333 nM RT primer, 5.33 U/l SuperScript III (to ensure consistent results and allow for some variability in nucleic acid concentration determination and enzyme activity), 50 mM Tris-HCl (pH 8.3 at room temperature), 75 mM KCl, and 5 mM DTT. This mixture is incubated at 55°C for 30 min by heat inactivation at 75°C for 15 min.

![Figure 4. Circularization optimization.](image-url)

- **A** Circularization efficiency versus betaine concentration (n=1) for CircLigase I and II (n=1).
- **B** Circularization efficiency versus time and betaine concentration (n=1).
- **C** Circularization efficiency versus N24 RT product concentration (n=2).
- **D** N24 PCR signal versus N24 RT product concentration prior to circularization (n=2; line, mean) at 0 M and 1 M betaine. In all panels, data were generated by quantitation of polyacrylamide gels (denaturing, panels A–C; non-denaturing, panel D). Circularization efficiency = (circularized RT product)/(linear RT product + circularized RT product) in each lane. N24 PCR signal = intensity of N24 PCR product band.
Step 5: PCR. To eliminate another gel purification step, we decided to use a portion of the completed and inactivated circularization reaction as direct input to PCR amplification. Adding 1.5 µl of a heat-inactivated circularization reaction containing 88 nM input RT product directly to a 25 µl PCR reaction, we tested the following high fidelity polymerases, each using their respective manufacturer’s supplied buffer and recommended cycling conditions (i.e. time and temperatures) for 8 cycles: PfuUltra II (Stratagene), Herculase II (Stratagene), Phusion (Finnzymes), KAPA HiFi (Kapa Biosystems), Advantage HD (Clontech), PrimeSTAR Max (Clontech) and AccuPrime Pfx (Invitrogen). Addition of DMSO, a PCR enhancing agent, did not significantly increase PCR amplification with any enzyme, perhaps with the exception of PfuUltra II (Figure 5A and B). PfuUltra II, Herculase II, Phusion, PrimeSTAR Max and KAPA HiFi all gave comparable product yields, but KAPA HiFi generated the least amount of slower migrating side products (indicated by *) just above the desired product (Figure 5A and B). Because of this and an independent report demonstrating robustness with regard to GC content, we decided to proceed with KAPA HiFi.

When preparing deep sequencing libraries, higher amounts of input DNA and low cycle numbers are desirable to amplify the greatest number of unique species. However, as with the RT reaction (Step 2), we were concerned that the diluted circularization buffer might affect PCR efficiency. Therefore, we titrated the volume of CircLigase reaction included in each PCR reaction. When this volume varied from 0.5 to 3.5 µl in a 15 µl PCR reaction, the PCR band intensity increased with increasing input, but not to scale (i.e. a 2-fold increase in input from 1 to 2 produced only a 1.5-fold increase in output; Figure 5C), likely indicating some inhibitory effect of the CircLigase reaction on PCR efficiency. We therefore limit the amount of added CircLigase reaction to one-fifth of the total PCR reaction volume.
Consequences of incomplete 3′-adaptor ligation

Having optimized each step in the protocol (Supplementary Table S1), we next wanted to assess the quality of libraries it generates. Because many published protocols use lower 3′-adaptor ligation temperatures and/or shorter incubation times than our optimized conditions (Figure 2F), we also wanted to test the effects of these variables. Therefore, we prepared seven different libraries using our synthetic N24 pool. All libraries were prepared identically except for the 3′-adaptor ligation step, for which the conditions are shown in Figure 2F and Supplementary Figure S2A. In one library, we also included four random 3′-adapters at the 5′-end of the 3′-adaptor (N4 adapter) to assess whether this would reduce 3′-end capture bias, as has been previously suggested (10,14,33).

To eliminate possible sequencing variability, all libraries were barcoded, mixed together and sequenced to similar depth within a single Illumina HiSeq 2000 lane (Supplementary Figure S2A). A library included in this lane was a library of random 500 nt fragments generated from the PhiX174 genome (15% of total sequences); PhiX inclusion increases the nt diversity at every position, thereby increasing the base calling accuracy (46).

To address the concern that long incubation times at higher temperatures could lead to significant RNA hydrolysis, we first examined the lengths of the captured sequences (Figure 6A). In all libraries, the majority of captured sequences were 24 nts. As expected, however, incubation at 22°C or 30°C for 6 h did result in a small decrease (1%) in the fraction of full-length species compared to the 20 min and 1 h incubation times (Figures 6A, inset I). A library as expected, this effect was somewhat less apparent at 4°C. Nonetheless, the impact of this material loss must be weighed against the higher capture variability introduced by shorter ligation times and lower nt concentrations (see below).

For further analysis we focused solely on full-length (24 nt) reads. Because the number of possible sequences in a 24-nt random oligo (10^24) is vastly outnumbered by the reads obtained per library (10^6), unique species constituted >99.5% of each library and >99.6% of the entire pooled data set (Supplementary Figure S2A). Because each library captured a unique sequence set, it was not possible to calculate the capture frequency for individual species. Therefore, to assess capture bias driven by nt identity, we measured nt frequency at each position in our captured fragment ends (Figure 6B). We already had a notable enrichment in G that decreased linearly in the 5′→3′ direction. To determine the extent to which this might be due to base misincorporation/misincorporating at the sequencing level, we also measured the miscalled frequency in the PhiX fragment ends sequenced alongside our N24 libraries (Supplementary Figure S2B). Across all positions corresponding to our N24 inserts, the PhiX misincorporation frequency was no greater than 0.000499 for any of the 4 nts, with G being the least frequently miscalled base (<0.00021). Additionally, when analyzing the nt frequency per position in ribosomal footprinting libraries ad lib with our optimized ligation conditions, we see no 3′→5′ trend toward G enrichment (Supplementary Figure S2C). Thus, the most likely explanation for the overabundance of G in the N24 libraries was guanosine phosphoribosyltransferase overincorporation during oligonucleotide synthesis (47).

Examination of Figure 6B reveals that the majority of inter-library variance occurred at the 3′ termini of captured RNAs (positions 21–24). To estimate expected nt frequencies (F_exp) at these terminal positions, we used the observed frequency (F_obs) data from all libraries to generate four best-fit lines (one for each nt) through positions 5–20 (Figure 6B), as these internal positions should be least affected by enzyme preference during 3′-adaptor ligation and circularization. We then used these best-fit lines to calculate expected nt counts at every nt position for each library. Calculating the chi-square statistic allowed us to quantitatively compare deviation in observed nt count from expected nt count (Figure 6C). This analysis revealed that the chi-square statistic at positions 21–24 decreased in the following order: 30°C–20 min > 30°C–20 min > 30°C–1 h > 30°C–1 h > 30°C–6 h > 30°C–6 h. That is, the libraries exhibiting the greatest deviation from expected were those wherein 3′-adaptor ligation was only 30–85% complete (Figure 2F), either because of insufficient incubation time or a suboptimal ligation temperature. For reactions that did proceed to apparent completion (the three 6-h libraries), inclusion of four random 3′-adapters at the 5′-end of the 3′-adaptor (N4) had no additional benefit in reducing position 21–24 deviation compared to the xed-sequence 3′-adaptor (although see miRNA data below).

Unexpectedly, position 22 exhibited equal or greater deviation than position 24 in all seven libraries. When comparing F_obs−F_exp for each nt, another feature readily observable in the 30°C–20 min library, and to a lesser extent in the 30°C–1 h library, is a tendency toward higher GC content at positions 11–15 (Supplementary Figure S3). Currently, we have no clear explanations for either of these effects (see Discussion), but both strengthen the point that uneven capture is accentuated by short ligation times.

Method validation

To assess how our optimized protocol performs on a known RNA sample, we made libraries from 50 fmol of Poly(A) mRNA pool previously used to benchmark small RNA library preparation (SR 899527 and SR 899530; 14). Barcoded libraries were generated using either the xed or N4 preadenylated 3′-adaptor; pooled and sequenced on a single MiSeq lane (Table 1). Plotting F_obs versus F_exp (where F_exp = 1/29 = 0.0345) revealed no recurring over- or underrepresentation pattern for any individual miRNA across four libraries (Figure 7A). In contrast, analysis of Figure 6B revealed a striking over-representation pattern for any individual miRNA across four libraries (Figure 7B). Interestingly, both the previous benchmark (14) (Figure 7B) and a new library preparation protocol for capturing scarce miRNAs (39) in our libraries, the lowest CV in F_obs were obtained with the xed adaptor at 1 pmol input and the N4 adaptor at 50 fmol and 1 pmol input. At 50 fmol input, however, the xed adaptor did result in somewhat higher variability. Therefore, the N4 adaptor may be preferable when using our protocol to construct libraries from very low input RNA.

It has previously been noted that both secondary structure internal to individual miRNAs and the ability of in-
A

Figure 6. N 24 length and bias analysis. (A) Distribution of read lengths, shown as a percent of the total sequences. (B) Nt frequency versus N 24 sequence position. Dashed line indicates ideal 25% incorporation and capture of all four nts. (C) Total bias at each N 24 sequence position.

Table 1. miRNA libraries

| Input | Adaptor | Sequencing platform | Mapped reads |
|-------|---------|--------------------|--------------|
| 1 pmol | Fixed | M Seq | 1 044 234 |
| | N 4 | | 1 393 238 |
| 50 fmol | Fixed | N 4 | 1 389 911 |
| | | | 676 609 |
| SRR 899527 | Hiseq 2000 | | 715 728 |
| SRR 899530 | | | 1 424 004 |

Under some conditions, reverse transcriptases can exhibit terminal transferase (TdT) activity, resulting in non-templated nt addition to cDNA 3' ends (48). Examination of our miRNA libraries revealed that, while some non-templated addition did occur, extensions were generally limited to a single nt and these extended species were 20- to 50-fold less abundant than full-length species (Figure 7C).
Figure 7. mRNA pool libraries. (A) Proportion of each mRNA in each library. Line represents perfectly even capture with each mRNA representing 1/29th of the reads. (B) Boxplot showing the distribution of proportions. CV = standard deviation (mRNA counts)/mean (mRNA counts). (C) Terminal transferase activity. Box chart showing percent of 5' additions and subtractions as a percentage of full-length reads.

Discussion

In this study, we set out to develop a method that yields robust strand-specific deep sequencing libraries from diverse RNA inputs. Our method involves 3 ligation of a preadenylated adaptor followed by RT, circularization and PCR. This approach combines features of several previously published protocols (3,23,43), with modifications to enhance capture efficiency and minimize sample loss. Our method works across a range of input amounts, is easy to follow, and produces a library in 2–3 days at a relatively low reagent cost (<$25 per sample), all while giving the user complete control over every step. Because the input to our method is generic single-stranded RNA with a 3' hydroxyl, it can be used to capture many different sized RNA footprints. Our approach can also be used to map sites of RNA-protein crosslinking (e.g. from CLIP experiments) and other base modifications that cause reverse transcriptase to either stall (e.g. abasic or alkylated sites) or incorporate the wrong base (e.g. PAR-CLIP). To date, various members of our laboratory have used this method to generate multiple footprinting libraries for Ribo-Seq and other RNA-protein
complexes, as well as RNA-Seq libraries (6 and unpublished results). Input fragment sizes have ranged from 20 to 200 nts, input amount has ranged from 400 pg to 200 ng RNA and all resulted in high complexity libraries. Our method is highly reproducible, with both read counts and RPKM for Ribo-Seq and RNA-Seq biological replicates having correlation coefficient of 0.93–0.99 (6 and unpublished results).

One of our major goals in developing this protocol was to minimize capture bias. We did so by identifying conditions wherein both the RN L2 and CN ligase reactions were driven to apparent completion, thereby minimizing ligase sequence preferences and any intra- and inter-molecular secondary structure effects. Our analysis of the effects of time and temperature on 3’-adaptor ligation clearly indicates that incorrect ligation exacerbates capture bias (Figures 2, 6B and 6C and Supplementary Figure S3). Nonetheless, even under conditions where the ligation reaction appeared to proceed to completion, apparent 3’-end biases were not fully eliminated (Figure 6C). Three recent papers reported that 3’-end capture bias can be reduced by including a short (2–4 nt) randomized region at the 5’-end of the 3’-adaptor (10, 14, 33). Inclusion of degenerate nts in the adaptor also allows for identification of species that are preferentially amplified during the PCR reaction (49). Although we observed no advantage of the N4 adaptor over the N24 or miRNA pool input (Figures 6C, 7A and B and Supplementary Figure S3), the 4’-adaptor was clearly superior when the RNA pool input was lowered to 50 fmol (Figure 7A and B). Therefore, using a 5’-randomized adaptor is recommended (to increase initial base calling accuracy by Illumina sequencing all in one single-end sequencing run).

Following ligation, RT of the captured RNA attaches a sequence tag to the 3’-end of the RNA, allowing for PCR amplification and deep sequencing. Although the adaptor sequences used here are for sequencing on Illumina platform, libraries can be prepared for any deep sequencing platform by simply modifying the 5’ and 3’-adaptor sequences. Our method employs a variety of RT primers that differ only by their 5’ barcode, allowing multiple samples to be sequenced on the same flow cell lane. Barcoding the same samples during the RT step minimizes opportunities for accidental mixing or cross-contamination of sample pools. We currently use a set of 12 5’-barcodes (see Materials and Methods) that were chosen such that the n/t position is balanced (to increase initial base calling accuracy by Illumina platform s) and there is no possibility for barcode misidentication, even with two sequencing errors. After circularization, the barcode is positioned 5’ to the captured cDNA sequence, allowing for barcode identification and fragment sequencing all in one single-end sequencing run.

Following circularization, one must determine the optimal number of PCR cycles for each sample. Cycle number is highly dependent on the original RNA input amount. Our current approach is to empirically determine the correct number of PCR cycles by gel analysis; too few cycles will result in product yield below the sequencing input requirement; too many cycles will result in PCR jackpots that can overwhelm the library and introduce significant bias. A recently published qPCR approach for identifying the correct number of cycles can easily be applied to our method (17).

Tw o similar protocols for making strand-specific libraries were recently published (51, 52), speaking to the overall strength of this strategy. Nonetheless, the modifications we describe here (i.e. inclusion of 25% PEG in the 3’-adaptor ligation reaction; no additional MgCl2 in the RT reaction; a single gel purification step; inclusion of 1M betaine in the CN ligase I reaction; and optimized time and temperature to ensure completion of all reactions) offer significant improvements over similar methods. To assist the reader in implementing our protocol, we have included a short summary of the conditions (Supplementary Table S1) and placed a detailed protocol at http://www.umassmed.edu/moorelab/resources/protocols/.

ACCESSION NUMBERS

High-throughput sequencing data have been deposited in the GEO database under accession number GSE63606.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

We thank Catherine Sterling, Victor Ambros and Phillip Zamore for helpful discussions related to method development. Rui Yang kindly provided the 29S RNA m. x, Fred Hudey and Hank Dau at Epicentre aided in troubleshooting the circularization reaction.

FUNDING

M.J.M. is a Howard Hughes Medical Institute Investigator. Funding for open access charge: Howard Hughes Medical Institute.

Con ict of interest statement: None declared.

REFERENCES

1. Ule, J., Jensen, K.B., Ruggiu, M., and Darnell, R.B. (2003) CLIP identifies OVA-regulated RNA networks in the brain. Science, 302, 1212–1215.

2. Hafner, M., Langthaler, M., Burger, L., Khoriati, M., Hauser, J., Bemigger, P., Rothkaessle, A., Rosch, M., and Jönsson, P.-O. (2010) Transcriptional eQTL identification of RNA-binding protein and mRNA RNA target sites by PAR-CLIP. Cell, 141, 129–141.
46. Illumina (2013) Technical note: using a PhiX control for HiSeq sequencing runs. http://see.illumina.com/documents/products/technote/technote.phicontrolv3.pdf. (October 2013, date last accessed).

47. Bartel, D. P. and Szostak, J. W. (1993) Isolation of new ribozymes from a large pool of random sequences. Science, 261, 1411–1418.

48. Chen, D. and Patton, J. T. (2001) Reverse transcriptase adds nontemplated nucleotides to cDNA during 5’RACE and primer extension. BioTechniques, 30, 574–582.

49. König, J., Zarnack, K., Zupan, B., Turner, D. J., Luscombe, N. M. and Ule, J. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol., 17, 909–915.

50. 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.

51. Ingolia, N. T., Brar, G. A., Rouskin, S., McGachy, A. M. and Weissman, J. S. (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat. Protoc., 7, 1534–1550.

52. Epicentre Technologies Corporation (2012) ARTseq Ribosome Profiling Kit. http://www.epibio.com/applications/rna-sequencing/ribosome-profiling/artseq-ribosome-profiling-kits?protocols. (June 2013, date last accessed).