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

Erin E. Heyer  
*University of Massachusetts Medical School*

Hakan Ozadam  
*University of Massachusetts Medical School*

Emiliano P. Ricci  
*University of Massachusetts Medical School*

See next page for additional authors

Follow this and additional works at: http://escholarship.umassmed.edu/gsbs_sp

Part of the Biochemistry Commons, Bioinformatics Commons, Computational Biology Commons, and the Nucleic Acids, Nucleotides, and Nucleosides Commons

Repository Citation

http://escholarship.umassmed.edu/gsbs_sp/1890

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in GSBS Student Publications by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
An optimized kit-free method for making strand-specific deep sequencing libraries from RNA fragments

Authors

Comments
© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

Creative Commons License
This work is licensed under a Creative Commons Attribution-Noncommercial 4.0 License

Rights and Permissions
Citation: Nucleic Acids Res. 2015 Jan;43(1):e2. doi: 10.1093/nar/gku1235. Epub 2014 Dec 12. Link to article on publisher's site

This article is available at eScholarship@UMMS: http://escholarship.umassmed.edu/gsbs_sp/1890
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 an 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 methodologies employ deep sequencing 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 complexity being examined and the RNA fragment entation method utilized (e.g. RNAase or sonication), bound RNA fragments can range from 10 to 200 nucleotides (nt). Therefore, we require a strand-specif c 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. mRNA-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 common only available reagents and works over a wide range of input amounts. We also wanted an approach that could 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 otherwise 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 (kcat/Km conditions) and at short reaction times (9). For ligation reactions, low temperatures can favor capture of sequences capable of base pairing with the adaptor (10). Low temperatures can also disfavor capture of sequences containing internal secondary structures. Many published library preparation protocols are suboptimal for
one or both of these factors, resulting in differential capture of an aRNA, e.g. 32P RNA-Seq (10–12) and highly non-uniform (‘peaky’) coverage of long RNAs e.g. RNA-Seq of aRNA PolII 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 rationale and rationale data we used to optimize each step. As presented are deep sequencing data on (i) the effects of the 5' and 3' primer on initial 3'-end capture and (ii) capture uniformity by analysis for an equimolar pool of 29 mRNAs. Taken together, these data show that our method faithfully preserves fragment diversity and abundance in complex starting mixtures and simultaneously affected by fragment sequence or folding potential.

MATERIALS AND METHODS

Gel analysis

All acrylamide gels were prepared using A cugel reagents (National Diagnostics). Ligation samples were prepared in an equal volume of 2× denaturing load buffer (12% Ficoll Type 400-DL, 7 M Urea, 1× TBE, 0.02% Bromophenol Blue, 0.02% Xylene Cyanol), denatured for 5 min at 95°C and cooled on ice prior to loading on denaturing 15% polyacrylamide (19:1)–8 M Urea–1× TBE gels. Reverse transcription reactions were performed as described in the text. The single-stranded RT product was diluted in one-third volume of 3× denaturing load buffer (8% Ficoll Type 400-DL, 10.5 M Urea, 1.5× TBE, 0.02% Bromophenol Blue, 0.02% Xylene Cyanol), denatured for 5 min at 95°C, and analyzed on 10% denaturing polyacrylamide (19:1)–8 M Urea–1× TBE gels. Reverse transcription RT samples were diluted in one-third volume of 3× denaturing load buffer (8% Ficoll Type 400-DL, 10.5 M Urea, 1.5× TBE, 0.02% Bromophenol Blue, 0.02% Xylene Cyanol), denatured for 5 min at 95°C, and analyzed on 10% denaturing polyacrylamide gel electrophoresis (PAGE) gels. Circularization reactions were performed similarly to ligation reactions and analyzed on 10% denaturing PAGE gels. Polymerease chain reaction (PCR) products for gel analysis were mixed with 5× non-denaturing load buffer (12% Ficoll Type 400-DL, 1× TBE, 0.02% Bromophenol Blue, 0.02% Xylene Cyanol) before separation on native 8% PAGE gels. PCR products to be sequenced were similarly prepared and analyzed on the DNA using the optimized conditions shown in Supplementary Table S1, except for the described variations in ligation conditions. In one case (22°C 6hr library), am mix containing 28-mers was added, and the product was analyzed on 8% PAGE gels as described above. Samples were sequenced using custom 384-well plates and analyzed with ImageQuant software.

Circularization efficiency and PCR amplification

Circularization reactions were performed 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-dATP. Circularized RT products were separated from non-circular, 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 linear input, and multiplying by 100.

PCR amplification from the circularized RT product was performed with KAPA HiFi Library Amplification Kit (Kapa Biosystems) according to manufacturer’s instructions, except where otherwise noted. A 12 PCR reactions were analyzed on native 8% PAGE gels and quantified as described above. Samples were excised and extracted as described for RT products, precipitated and quantified by gel analysis before sample submission.

N24 library construction and analysis

N24 libraries were constructed from 0.1 pmol of N24 RNA oligo using the optimized conditions shown in Supplementary Table S1, except for the described variations in ligation conditions. In one case (22°C 6hr library), a mix containing 28-mers was added. A 12 PCR reactions were performed with 7 PCR cycles and gel purified prior to sequencing on a single lane of an Illumina HiSeq2000 lane (E/new 1).

Deep sequencing data were analyzed with custom scripts unless otherwise noted. Data were parsed into individual
libraries by 5 barcode, allowing 1 mismatch. The 3 adapter sequence was appended from all libraries allowing 3 mismatches. Once individual sequence reads were identified, read lengths were calculated. A subsequent analysis utilized only 24 nt reads. For each library, we calculated the observed nt frequencies at each of the 24 positions. To determine expected values, we used the data across positions 5-20 from all libraries and fitted least squares lines to the expected patterns at each nt. The equations for the lines yielded the observed nt frequencies at all 24 positions. The chi-square statistic was calculated for each library by summing [(observed nt count - expected nt count)^2]/(expected nt count) across all 4 nts at each N 24 position.

PhiX reads were identified if they mapped to the PhiX 174 genome with a maximum of 6 errors within the 51 nt sequence length. Mismatches were identified and counted if the reads were identical to the genome with a maximum of 6 errors within the 51 nt sequence length. Mismatches were counted at each position by the total number of reads. For analysis of nt distribution across ribosome binding sites, nt frequencies were calculated by dividing the observed nt count at each position by the total number of reads for each miRNA.

mRNA library construction and analysis
Libraries were constructed from either 1 pmol or 50 fmol of an equimolar mix of 29 miRNAs (14) according to the optimized conditions shown in Supplemental 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, allowing 1 mismatch. 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 with 0 mismatches. A library was considered to be present if it met these criteria.

RESULTS
Protocol design
To generate strand-specific deep sequencing libraries, both ends of the captured RNA must be appended to 3′ sequences (adaptors) to enable primer hybridization for amplification and sequencing. These adaptors generally correspond to the 3′ and reverse primer sequences used for cloning and sequencing. Therefore, we decided to adopt a 3′-end adaptor ligation approach widely used in the small RNA field (23) - direct ligation of a preadenylated DNA adaptor to the 3′-end of RNA fragments using RNA ligase (Figure 1, Step 1). We chose to use a truncated and mutant form of T4 RNA Ligase 2 (RNL2 Tr Tr. K227Q) because published reports indicated it has less substrate bias and produces fewer side products than the full-length wild-type enzyme (12,26), and RNL2 is known to be less affected by nt identity at the ligation site than T4 RNA Ligase 1 (27). Following 3′ adaptor ligation, a highly efficient method for amplifying the 5′ adapter is to reverse transcribe the RNA from the 3′ adapter with an RT primer containing the 5′ adapter sequence at the other end and then circularize the resulting single-stranded cDNA using CircLigase (3) (Figure 1, Step 2 and 4). An excellent linker (Spacer 18, an 18-atom hexa-ethyleneglycol spacer) placed between the 3′ and 5′ adapter sequences to minimize structural constraints for circularization and preclude the possibility of rolling circle PCR (28).
Figure 1. Method overview. Step 1: Ligation. RNA, shown in blue, is ligated to a preadenylated DNA adaptor to form a RNA-DNA hybrid. In the same tube, RT is performed (Step 2). The RT primer 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 adaptors 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).

For optimization of each step, we used a pool of randomized RNA 24mers (N 24) 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 T7 K227Q ligation reactions were 500 nM single-stranded RNA, 1 M 3'-adaptor, 10 U/lenzyme 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 this mixture 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. dimethylsulfoxide (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 2B in (38)), we found that 25% PEG 8000 resulted in near complete N 24 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, N 24 and enzyme concentrations. Using two different N 24 concentrations, near complete ligation was observed at all adaptor concentrations above 130 nM (Figure 2C). A t470 nM adaptor ligation was highly efficient with N 24 concentrations above 50 nM (Figure 2D) and enzyme concentrations above 6 U/1 µl (Figure 2E). A greater dependence of ligation efficiency on enzyme concentration at 10 nM N 24 does suggest, however, that additional enzyme will increase yields for every dilute RNA sample (39).

Published reports using T4 RNA ligases for library preparation employ a wide range of reaction times (1 h to overnight) and tem peratures (5°C–37°C) (1,23,34,37,40–43). However, colder tem peratures should stabilize both intra- and intermolecular secondary structures, potentially biasing ligation against internally structured RNAs and toward RNA sequences that partially base pair with the 3'-adaptor (10–11,27). Higher tem peratures should alleviate these issues, but could decrease enzyme stability and increase RNA degradation. Using our N 24 pool, we assessed ligation efficiencies across a range of incubation times and tem peratures (Figure 2F). Both 4°C and 37°C yielded poor
**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 radiactively 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 a manufacturer-supplied 5x or 10x RT buffer or (ii) the same buffer minus MgCl₂ as the Step 1 reaction already contains MgCl₂, and concentrations of...
Figure 3. RT optimization. (A) Comparison of high-fidelity reverse transcriptases for the amount of RT product generated ± M gCl₂ in the RT buffer. Absence or presence of indicated species are indicated by − and +. (B) RT product signal versus % PEG 8000 (w/v) in the ligation reaction (n = 3; black line, mean). (C) RT product signal versus RT primer concentration (n = 1). (D) RT product signal versus SSIII concentration (n = 1). (E) RT product signal varies with RNA input concentration, ranging from 3.3 nM (lane 2) to 133 nM (lane 6). (F) RT product signal versus RNA input concentration (n = 1). Replicate of panel E, incorporating ³²P in the RT for quantitation. In panels B, C, D and F, data were generated by quantitation of denaturing polyacrylamide gels similar to panels A and E.

M gCl₂ above 3 mM can inhibit RT (44). For all four enzymes (tested at the manufacturer's recommended concentration), we observed more full-length RT products when no M g²⁺ 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 a 3-fold dilution of the RT product yield was obtained when the ligation reaction constituted one-third of the final volume of the RT reaction (data not shown). This resulted in a 3-nM gCl₂ 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'-adaptor 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'-adaptor 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 amounts.

Based on the above data, we adopted the following as our standard Step 2 reaction conditions: 3-fold dilution of the
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 followed by heat inactivation at 75°C for 15 min.

Step 3: gel purification. See Materials and Methods.

Step 4: circularization. There are currently two commercially available 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, was recommended by Epicentre for use with CircLigase I. However, as no amount of betaine improved CircLigase II 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 had no significant effect on circularization efficiency (data not shown), so we continued to use the manufacturer’s suggested conditions. A time course revealed complete circularization with 5 U/1 enzyme and 50 nM input ssDNA required at least 2 h at 60°C (Figure 4B). Titration of the 24 N4 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 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: 1X CircLigase buffer (Epicentre), 1 M betaine, 50 M adenosine triphosphate, 2.5 mM MgCl₂ and 5 U/l CircLigase I in 20 μl containing all of the ssDNA isolated in Step 3. This mixture is incubated...
bated at 60°C for 3 h followed by heat inactivation at 80°C for 10 min.

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 (final volume) 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: Pfu Ultra 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 Pfu Ultra II (Figure 5A and B). Pfu Ultra 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 its robustness with regard to GC content (45), 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 was 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- to 4-fold increase in input from 1 to 2 µl produced only a 1.5- to 2-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’-adaptors at the 5’-end of the 3’ adaptor (N4 adaptor) 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 (<7%) in the fraction of full-length species compared to the 20 min and 1 h incubation times (Figure 6A, inset 1). 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 temperatures (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 (>10^14) is vastly outnum bered by the reads obtained per library (10^7), 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 fragments (Figure 6B). A cross all libraries, there was 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/miscal lling at the sequencing level, we determined the m/m atch frequency in the PhiX fragments sequenced alongside our N24 libraries (Supplementary Figure S2B). A cross all positions corresponding to our N24 inserts, the PhiX m/m atch frequency was no greater than 0.000049 for any of the 4 nt, with G being the least frequently miscal lled base (0.0002). Additionally, when analyzing the nt frequency per position in ribosome footprinting libraries obtained 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 phosphoramidite overincorporation during oligonucleotide synthesis (47).

Examination of Figure 6B reveals that the majority of inter-library variance occurred at the 3’ terminus of captured RNAs (positions 21–24). To assess expected nt frequencies (Fexp) at these terminal positions, we used the observed frequency (Fobs) data from all libraries to generate four best-line 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-line lines to calculate expected nt counts at every nt position for each library. Calculating the chi-square statistic allowed us to quantify the 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 > 4°C–18 h > 22°C–1 h > 30°C–1 h > 30°C–6 h > 30°C–6 h N4 > 22°C–6 h. That is, the libraries exhibiting the greatest deviation from expected were those wherein 3’-adapter 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’-adaptors at the 5’-end of the 3’-adaptor (N4) had no additional benefit in reducing position 21–24 deviation compared to the 3’-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 Fobs/Fexp 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 G 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 f mol or 1 pmol of an equimolar 29 m RNA pool previously used to benchmark small RNA library preparation (SR 899527 and SR 899530; 14). Barcoded libraries were generated using either the fixed or N4 preadenylated 3’-adaptor; then pooled and sequenced on a single MiSeq lane (Table 1). Plotting Fobs vs. Fexp (where Fexp = 1/29 = 0.0345) revealed no recurring over- or underrepresentation pattern for any individual RNA across four libraries (Figure 7A). In contrast, all four of our libraries exhibited lower variability than both the previous benchmark (14) (Figure 7B) and a new library preparation protocol for capturing scarce mRNAs (39). In our libraries, the lowest CV in Fobs were obtained with the fixed 3’-adaptor at 1 pmol input and the N4 3’-adaptor at 50 fmol and 1 pmol input. At 50 fmol input, however, the fixed 3’-adaptor did result in some higher variability. Therefore, the N4 3’-adaptor may be preferable when using our protocol to construct libraries from very low input RNA.

It has previously been noted that both secondary structures internal to individual mRNAs and the ability of
Figure 6. N24 length and bias analysis. (A) Distribution of read lengths, shown as a percent of the total sequences. Dashed line indicates ideal 25% incorporation and capture of all four nts. (B) Nt frequency versus N24 sequence position. (C) Total bias at each N24 sequence position.

Table 1. miRNA libraries

<table>
<thead>
<tr>
<th>Input</th>
<th>Adaptor</th>
<th>Sequencing platform</th>
<th>Mapped reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 pmol</td>
<td>Fixed</td>
<td>M Seq</td>
<td>1 044 234</td>
</tr>
<tr>
<td></td>
<td>N 4</td>
<td></td>
<td>1 393 238</td>
</tr>
<tr>
<td>50 fmol</td>
<td>Fixed</td>
<td>M Seq</td>
<td>1 389 911</td>
</tr>
<tr>
<td></td>
<td>N 4</td>
<td></td>
<td>676 609</td>
</tr>
<tr>
<td>SRR 899527</td>
<td></td>
<td>HiSeq 2000</td>
<td>715 728</td>
</tr>
<tr>
<td>SRR 899530</td>
<td></td>
<td></td>
<td>1 424 004</td>
</tr>
</tbody>
</table>

Individual miRNAs to hybridize to the 3'-adapter can affect capture efficiency (10,27). To address this possibility, we made scatter plots of read frequency versus individual miRNA features and calculated both slope and -value for the line best fitting the data (Supplementary Figure S4). We note that a slope other than 0 is potentially indicative of bias, with the magnitude of the slope indicating the strength of the bias dependent on the particular feature being plotted. The -value indicates only how well the line fits the data. These plots revealed no correlation with a | -value | > 0.5 between Fobs, G+C content, or between Fobs and the calculated folding energies (G) for each miRNA alone or each miRNA co-ligated with the adapter in any of our four libraries. We could also detect no apparent folding energy effects in the previous benchmark libraries. With the latter samples, however, there were readily observable trends with regard to nt composition, the most significant being a negative correlation (m = 0.058; m = −0.72) between Fobs and the number of U’s in the last 10 nts of each miRNA (Supplementary Figure S5). This is consistent with our N24 data showing an increased bias against U’s in the last few nts when ligation reactions conditions are suboptimal (Supplementary Figure S3). The absence of the same trend in our miRNA libraries highlights the more even coverage provided by our optimized ligation conditions.

Under some conditions, reverse transcriptases can exhibit terminal transferase (TdT) activity, resulting in non-template nt addition to cDNA 3' ends (48). Examination of our miRNA libraries revealed that, while some nterplates 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).
During preparation, these samples were immediately gel purified after RT (Supplementary Table S1). With one set of libraries, we observed more extensive TdT activity when the RT reaction was maintained at 4°C overnight following the heat inactivation step (data not shown). This suggests that SuperScript III is not completely inactivated by the manufacturer’s suggested heat inactivation regimen and will continue to add untemplated nts during long, low temperature incubations.

**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 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 any 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

<table>
<thead>
<tr>
<th>Sample</th>
<th>Observed Frequency (F_{obs})</th>
<th>Expected Frequency (F_{exp})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1pmol-fixed</td>
<td>0.08</td>
<td>0.10</td>
</tr>
<tr>
<td>1pmol-N4</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>50pmol-fixed</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>50pmol-N4</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>SRR899527</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>SRR899530</td>
<td>0.02</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Figure 7. miRNA pool libraries. (A) Proportion of each miRNA in each library. Line represents perfectly even capture with each miRNA representing 1/29th of the reads. (B) Box plot showing the distribution of proportions. CV = standard deviation (miRNA counts)/mean (miRNA counts). (C) Terminal transferase activity. Bar chart showing percent of 5′ additions and subtractions as a percentage of full-length reads.

During preparation, these samples were immediately gel purified after RT (Supplementary Table S1). With one set of libraries, we observed more extensive TdT activity when the RT reaction was maintained at 4°C overnight following the heat inactivation step (data not shown). This suggests that SuperScript III is not completely inactivated by the manufacturer’s suggested heat inactivation regimen and will continue to add untemplated nts during long, low temperature incubations.
contrasted with the expected nt identity bias at the -3 position relative to the 3'-ligation reaction (49). Although we observed no advantage of the N4 adaptor over a randomized adaptor (10,14,33). Inclusion of degenerate nt substrates allows for identification of species that are preferentially amplified during the PCR reaction (25). Nonetheless, our N24 data highlight the importance of driving the ligation reaction as close to completion as possible.

Contrary to expectation (10,27), we could detect no effects on N24 or N12 RNA capture efficiency that could be attributed to either internal secondary structure forming propensity or the ability of captured sequences to hybridize with the adaptor (Supplementary Figure S4 and data not shown). In our N24 data, however, we did detect an unexpected nt identity bias at the -3 position relative to the 3'-ligation site (F19B, Figure S3). This is consistent with a previous report demonstrating -3 substrate bias by both RNL1 and RNL2 (27). Currently, there is no clear explanation for this effect, as a crystal structure of RNL2 bound to substrate suggests that RNL2 substrate specificity is dictated solely by the nt at positions -3 and -2 (50). Nonetheless, our N24 data highlight the importance of driving the 3'-ligation reaction as close to completion as possible.

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(s), libraries can be prepared for any deep sequencing platform by simply modifying the 3' and 3'-adapter sequences. Our method employs a variety of RT primer sets that differ only by their 5' barcode, allowing multiple samples to be sequenced on the same flow cell lane. Barcoding the same sample during the RT step minimizes opportunities for accidental mixing or cross-contamination of samples. We currently use a set of twelve 5' barcodes (see Materials and Methods) that were chosen such that the nt position is balanced (to increase in silico base calling accuracy by Illumina platform(s) and there is no possibility for barcode misidentication, even with two sequencing errors. A more 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 requirements; too many cycles will result in PCR jackpots that can overwhelm the library and introduce signal contamination. A recently published qPCR approach for identifying the correct number of cycles can easily be applied to our method (17).

Supplementary 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'-adapter ligation reaction; no additional MgCl2 in the RT reaction; a single gel purification step; inclusion of 1M betaine in the C-terminal ligation I; and optimized conditions 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 any of the conditions (Supplementary Table S1) and placed a detailed protocol at http://www.umassmed.edu/coxelab/resources/protocols/.

ACKNOWLEDGMENTS

We thank Catherine Sterling, Victor Ambros and Phillip Zamore for helpful discussions related to method development. Rui Yikindly provided the 29 miRNA mix. Fred H. Yde and Hank Daum at Epicentre aided in troubleshooting the circularization reaction.

FUNDING

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

CONFLICT OF INTEREST STATEMENT

None Declared.

REFERENCES


