FFPEcap-seq: a method for sequencing capped RNAs in formalin-fixed paraffin-embedded samples

Jeffery M. Vahrenkamp,[1] Kathryn Szczotka,[2] Mark K. Dodson,[2] Elke A. Jarboe,[3] Andrew P. Soisson,[2] and Jason Gertz[1]

[1]*Department of Oncological Sciences,* [2]*Department of Obstetrics and Gynecology,* [3]*Department of Pathology, Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah 84112, USA*

The majority of clinical cancer specimens are preserved as formalin-fixed paraffin-embedded (FFPE) samples. For clinical molecular tests to have wide-reaching impact, they must be applicable to FFPE material. Accurate quantitative measurements of RNA derived from FFPE specimens is challenging because of low yields and high amounts of degradation. Here, we present FFPEcap-seq, a method specifically designed for sequencing capped 5′ ends of RNA derived from FFPE samples. FFPEcap-seq combines enzymatic enrichment of 5′ capped RNAs with template switching to create sequencing libraries. We find that FFPEcap-seq can faithfully capture mRNA expression levels in FFPE specimens while also detecting enhancer RNAs that arise from distal regulatory regions. FFPEcap-seq is a fast and straightforward method for making high-quality 5′ end RNA-seq libraries from FFPE-derived RNA.

[Supplemental material is available for this article.]

Cancer cells can easily be distinguished from normal tissue based on aberrant patterns of gene expression. Gene expression levels are used clinically to distinguish subtypes of cancer (Alizadeh et al. 2000; Perou et al. 2000), assess risk of recurrence (O'Connell et al. 2010), and determine the best treatment options (Paik et al. 2004). Although there is a lot of potential clinical benefit to gene expression profiling of tumor samples, technical hurdles remain and chief among them is the most common method of tumor sample storage, formalin-fixed paraffin-embedded (FFPE). It is estimated that more than 20 million FFPE specimens are collected each year in the United States alone (Waldron et al. 2012). Most tumor specimens are fixed in formalin and embedded in paraffin to preserve the morphology of the sample for histological analysis. FFPE samples are suitable for analyzing tumor histology and performing immunohistochemistry to assess protein expression; however, nucleic acids are harmed in the process with RNAs being the most labile of nucleic acids (Yakovleva et al. 2017). Methods that can overcome the low quality of RNA that is derived from FFPE samples and tap into this vast resource of archived tissues could have wide-reaching impact in both retrospective studies and clinical testing of samples to determine diagnosis, prognosis, and optimal treatments.

Standard RNA-seq approaches are not well designed for FFPE-derived RNA. Poly(A) selection can produce high 3′ bias that changes with the amount of degradation in the FFPE-derived RNA. Ribosomal RNA (rRNA) depletion-based methods are more appropriate, but suffer from higher rRNA contamination, lower numbers of alignable reads, and lower reproducibility (correlation) in gene expression (Adiconis et al. 2013; Zhao et al. 2014). There are methods that are better suited for analyzing RNA derived from FFPE samples. The nCounter technology from NanoString Technologies is capable of counting RNA molecules even when the RNA has been degraded (Veldman-Jones et al. 2015). Exome capture of RNA-seq from FFPE samples is an approach that is able to "rescue" the RNA-seq libraries by keeping most reads in exons (Cieslik et al. 2015). However, these methods do not allow for discovery, because both techniques require knowing the RNAs that you want to analyze when designing the experiment.

An alternative approach to quantifying transcription while also enabling discovery is to focus on the 5′ ends of transcripts, where a 5′ 7-methylguanylate cap is added to RNA polymerase II transcripts (Shuman 1997). Sequencing of 5′ transcript ends can significantly reduce the cost of RNA-seq while also providing precise information about transcription start sites. In addition, enhancer RNAs (eRNAs) that arise from distal regulatory elements (De Santa et al. 2010; Kim et al. 2010; Li et al. 2013; Mousavi et al. 2013; Hsieh et al. 2014) can also be identified by sequencing 5′ ends of capped RNAs (Andersson et al. 2014). There are several approaches that focus on sequencing the 5′ end of transcripts, and six of them were compared in a recent study (Adiconis et al. 2018). The protocols fall into three main categories: variations on nuclear run-on experiments that interrogate nascent RNA (e.g., GRO-seq and PRO-seq) (Core et al. 2008; Kwak et al. 2013); enzymatic or chemical modification of the 5′ cap, for example, CAGE (The FANTOM Consortium and RIKEN Genome Exploration Research Group and Genome Science Group 2005) and TSS-seq (Tsuchihara et al. 2009); and template switching, for example, STRT-seq (Islam et al. 2011) and nanoCAGE (Salimullah et al. 2011). Nuclear run-on techniques rely on viable cells that are actively performing transcription and are not suitable for fixed samples. Methods that use enzymatic and chemical modification of the 5′ cap have large input RNA requirements that may not be obtained from FFPE samples. Template switching involves the addition of nontemplated nucleotides on the 3′ end of cDNA by

Corresponding author: jay.gertz@hci.utah.edu

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.249656.119. Freely available online through the *Genome Research* Open Access option.

29:1826–1835 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/19; www.genome.org

reverse transcriptase; however, the addition of nucleotides is not dependent on a 5′ cap, and fragmented RNA can lead to sequencing of internal regions of RNA. Therefore, there is not an ideal method for sequencing 5′ ends of transcripts from FFPE-derived RNA.

Here, we describe the development of FFPEcap-seq, a method for sequencing 5′ ends of capped transcripts from FFPE-derived RNA. FFPEcap-seq combines the sensitivity and low input requirements of template switching (Salimullah et al. 2011; Ramsköld et al. 2012; Marinov et al. 2014), with a necessary enzymatic enrichment for 5′ capped RNAs, that is used in START-seq (Nechaev et al. 2010; Scruggs et al. 2015). We first aimed to make improvements to the published nanoCAGE protocol (Salimullah et al. 2011), including the addition of a unique molecular identifier. We then evaluated enzymatic enrichment of RNAs with 5′ caps, in freshly collected RNA, matched FFPE-derived RNA, and RNA from FFPE patient specimens. Correlation to RNA-seq data and the ability to detect eRNAs was assessed as well as the input requirements. Our results aim to establish FFPEcap-seq as a fast and inexpensive method for quantifying mRNAs and eRNAs in FFPE samples.

Results

Improvements to the nanoCAGE protocol

Figure 1 shows an overview of the FFPEcap-seq approach. The initial steps of FFPEcap-seq involve enzymatic cleanup of uncapped RNA fragments (Nechaev et al. 2010; Scruggs et al. 2015) and the library construction follows the published nanoCAGE protocol (Salimullah et al. 2011). Before combining the two approaches, we started by making nanoCAGE libraries with freshly extracted RNA from Ishikawa cells, an endometrial cancer cell line. The only modification that we initially made to the published protocol was the addition of a 6-bp sample-specific index to one of the final PCR primers. The final libraries had the expected yield (>600 ng) and size distribution (400–2000 bp); however, the sequencing results were poor. On a HiSeq 2500 lane, with an expected sequencing output of approximately 250 million reads, only 77 million reads passed the quality filter. Of the 77 million reads, fewer than 18 million contained the expected AGGG on their 5′ end, indicating that mostly unintended products were sequenced.

To sequence the initial libraries, a PhiX genome sequencing library was added to the lane to base balance the reads that should start with AGGG. Suspecting that the addition of the PhiX library reduced the amount of reads from our nanoCAGE libraries, we explored ways of base balancing the sequencing runs without the addition of a PhiX library. To base balance the 5′ end of the sequencing read and enable us to count molecules instead of assuming that sequencing depth is proportional to the initial RNA molecules, we added a unique molecular identifier (UMI) to the 5′ end of the library by adding nine equally mixed random bases adjacent to the AGGG on the 3′ end of the template switching (TS) oligo. This inline UMI scheme shortens the amount of sequencing coming from the RNA, but increases the quantitative resolution of an inherently low complexity library caused by the expected 5′ enrichment. Unfortunately, adding the UMI and eliminating PhiX library addition did not improve the sequencing results—18 million sequencing reads passed filter compared to an expected 250 million reads.

A previous template switching study discovered the formation of TS oligo concatemers on the 3′ end of the cDNA

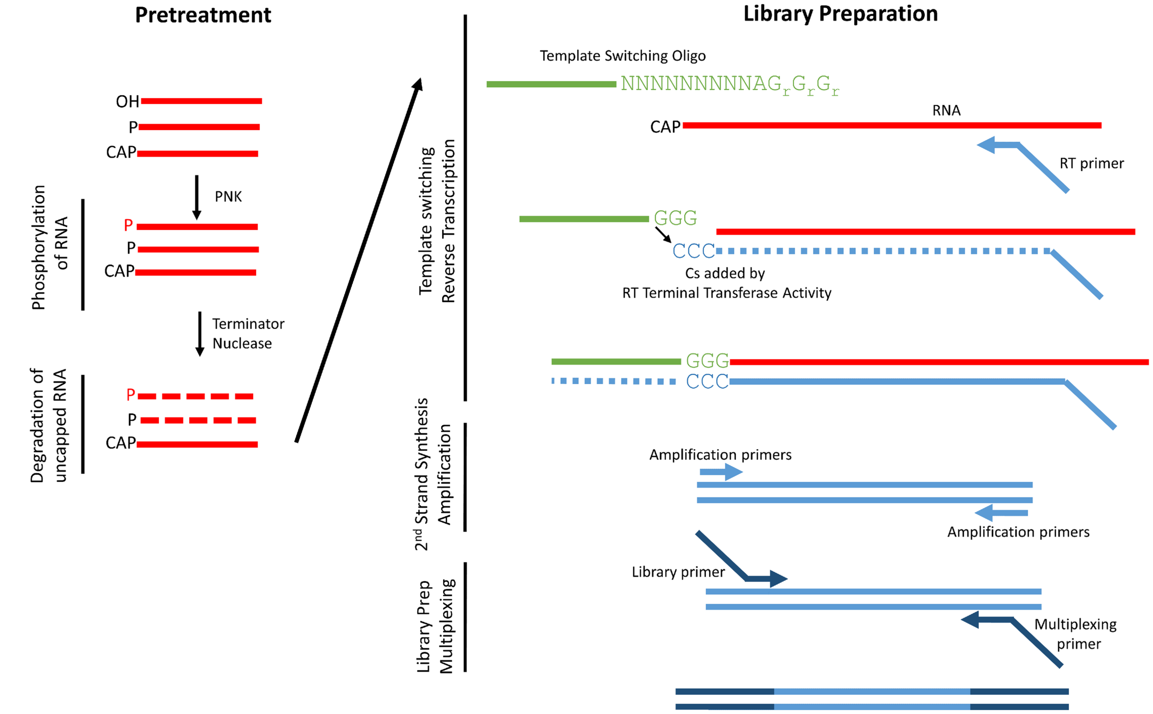


Figure 1. FFPEcap-seq overview. During enzymatic pretreatment, RNA is treated with T4 Polynucleotide Kinase (PNK) to phosphorylate 5′ hydroxyls and Terminator Nuclease to degrade uncapped RNAs. Sequence tags are added to the 5′ end of the cDNA during the reverse transcription reaction using template switching. cDNA is then amplified, and additional sequences are added during PCR amplification. (OH) Hydroxyl; (P) phosphate; (G_r) a guanine ribonucleotide; (N) an equal mix of nucleotides.

(Kapteyn et al. 2010). During the reverse transcription step, template switching can continue to occur on the end of the TS oligo, allowing for the formation of multiple concatemers of the reverse complement of the TS oligo being added to the 3′ end of the cDNA (Fig. 2A). This creates a single molecule with multiple sequencing primer binding sites followed by UMIs, which should lead to mixed signals and poor sequencing quality on an Illumina sequencer. To block the formation of concatemers, it is necessary to prevent reverse transcriptase from reaching the 5′ end of the TS oligo and causing terminal transferase activity. To test if concatemer formation was causing the sequencing issues, we replaced the TS oligo with a version that contained two unnatural bases at the 5′ end which are isomers of cytosine and guanine in which the hydrogen donors and receivers on the bases have been switched. Because these bases do not have cognate partners in the reverse transcription reaction, reverse transcriptase stalls, preventing terminal transferase activity. The use of this modified TS oligo greatly improved the sequencing results. The HiSeq lane produced 270 million reads, of which >80% aligned to the genome. The amount of detectable concatemers in the sequencing data decreased from 25% (SD = 14%) of the sequenced reads to <1% (SD = 0.25%) (Fig. 2B). It should be noted that concatemers likely made up a higher percentage of the previous libraries, but failed to pass the sequence quality filter.

The addition of isomer bases to the 5′ end of the template switching oligo was highly successful in blocking the formation of TS oligo concatemers; however, the modified bases made the oligo expensive and it suffered from low yields during synthesis. We analyzed the ability of other modifications to the 5′ end of the TS oligos to block the formation of TS concatemers and discovered that placing an 18-carbon spacer between the 5′ most base and the penultimate base of the TS oligo reduced TS oligo concatemers to 3% (SD = 0.28%). Because this modification was significantly more cost effective than the isomer containing oligo, we chose to use the internal 18-carbon spacer containing oligo in future experiments.

In addition to blocking the 5′ end of the TS oligo, we attempted to improve the template switching reaction by the addition of manganese (Mn), which has been reported to significantly increase the terminal transferase activity of reverse transcriptase (Schmidt and Mueller 1999). After the addition of 1 mM Mn to the template switching reaction, we discovered an increase in TS oligo concatemers to 11% (SD = 5.7%) when the carbon spacer was included, and an increase in the percentage of reads that map to ribosomal RNAs (rRNAs) from 21% (SD = 6.5%) to 48% (SD = 17.67%) (Fig. 2B,C). Because of these negative effects of Mn addition, we did not include it in any following experiments. Our results with different iterations of the nanoCAGE protocol resulted in significantly improved yield and library quality, while also incorporating sample and molecular barcodes.

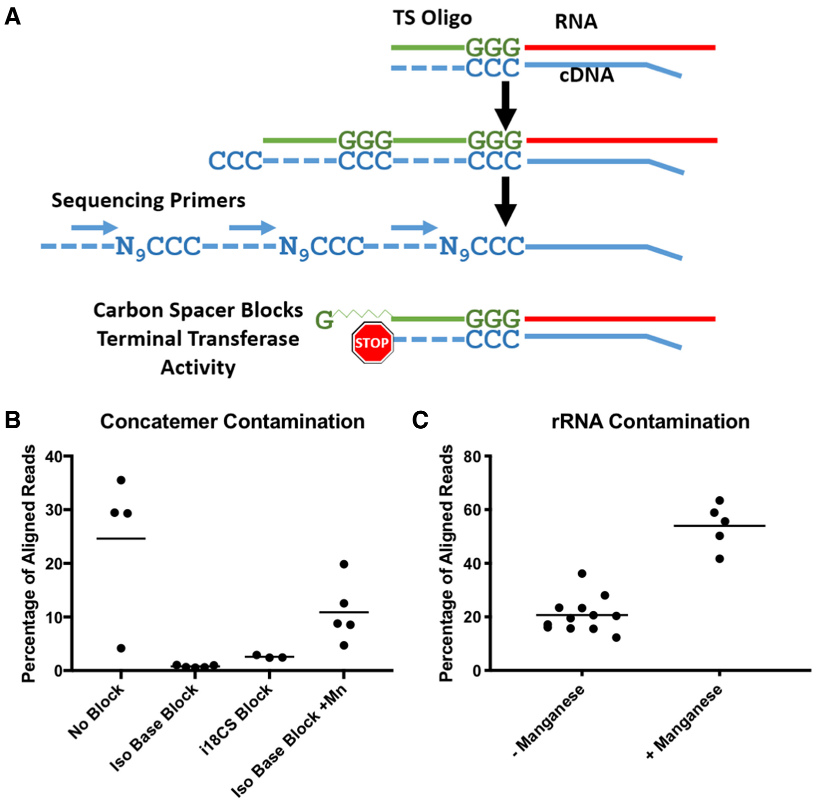


Figure 2. Improvements to the nanoCAGE protocol. (*A*) Schematic shows how template switching concatemers form during the template switching reaction and how they interfere with sequencing. Concatemers are blocked by the addition of an 18-carbon spacer between the two 5′-most bases on the template switching oligo. (*B*) Both 5′ isomer (Iso) bases and an internal 18-carbon spacer (I18CS) were effective at reducing the percentage of reads that included concatemers. The addition of manganese to the template switching reaction increased the percentage of reads that matched concatemers. (*C*) The addition of Mn to the template switching reaction significantly increased the number of contaminating rRNA reads in the final library. Lines on graph represent means.

Effects of enzymatic treatment on fresh and FFPE RNA

The first step of FFPEcap-seq, which precedes library preparation, is enzymatic enrichment of 5′-capped RNAs. In FFPE-derived RNA, 5′ ends can be found in three major forms: methyl cap (RNA polymerase II transcripts) (Shuman 1997), hydroxyl (from internal hydrolysis during degradation), and mono-, di-, or triphosphates (RNA polymerase I and III transcripts) (Kwan et al. 2000). With a combination of T4 polynucleotide kinase to add phosphates to 5′ hydroxyls and RNA 5′ polyphosphatase to convert triphosphates to monophosphates, nearly every RNA should either harbor a 5′ cap or a 5′ monophosphate. Treating the sample with a 5′ phosphate-dependent exonuclease should then remove all uncapped RNAs, leaving only RNA that originally had 5′ caps. We compared two 5′ phosphate-dependent exonucleases: terminator (Epicenter/Lucigen) and XRN-1 (New England Biolabs). To prevent losses associated with multiple cleanup steps, we performed all enzymatic steps simultaneously because the enzyme buffers were compatible. Enzymatic treatments, as well as mock treatment controls, were applied to 400 ng total RNA before constructing libraries with the improved nanoCAGE protocol. We found that XRN-1-treated RNA was not substantially different from untreated RNA in terms of rRNA percentages (21%, SD = 4% vs. 16%, SD = 1.1%) and enrichment for the 10% most 5′ ends of transcripts (30%, SD = 1.5% vs. 35%, SD = 0.6%) (Supplemental Fig. S1). Terminator nuclease was effective at reducing rRNA to 5.26% (SD = 0.6%) and increased 5′ enrichment to 44.8% (SD = 0.3%). We also tested the necessity of RNA 5′ polyphosphatase treatment and did not find a significant effect from polyphosphatase treatment (Supplemental Fig. S1). Based on these results, we included only T4 polynucleotide kinase and terminator nuclease in our pretreatment protocol.

We next sought to test the enzymatic treatments on FFPE samples and created four matched sets of fresh derived RNA and FFPE-derived RNA from Ishikawa cells. We observed the expected degradation with RNA integrity numbers (Bioanalyzer) going from 9.6 when freshly derived to 3.4 when FFPE-derived. We created libraries with and without enzymatic treatment before applying the improved nanoCAGE protocol starting with 400 ng total RNA in all cases (Fig. 3A–C). Without enzymatic pretreatment, FFPE-derived RNA performed poorly when compared to freshly derived RNA, including significantly higher rRNA contamination (12.9%, SD = 3.1% fresh vs. 37%, SD = 7.1% FFPE) and lower 5′ end enrichment (55.4%, SD = 7% fresh vs. 22.8%, SD = 5.8% FFPE). Enzymatic treatment significantly improved the libraries from both freshly derived and FFPE-derived RNA samples (Fig. 3A–C). In the fresh RNA samples, enzymatic pretreatment decreased rRNA contamination (2.4%, SD = 1.1% treated vs. 12.2%, SD = 3.4% untreated) and increased 5′ enrichment (70.3%, SD = 11% treated vs. 55.4%, SD = 7% untreated). A similar improvement was observed with FFPE-derived RNA, with changes in rRNA levels (13%, SD = 7.1% treated vs. 37%, SD = 7.1% untreated) and 5′ enrichment (56.2%, SD = 7.7% treated vs. 22.8%, SD = 5.8% untreated) (Fig. 3D; Supplemental Fig. S2). Overall, the enzymatic treatment improves the FFPE-derived RNA libraries so that the quality metrics are similar to libraries created from freshly derived RNA without enzymatic treatment, whereas freshly derived RNA libraries are improved further by enzymatic treatment.

Enzymatic treatment did lead to a drop in the number of reads which aligned to the genome for both freshly derived and FFPE-derived RNA. We found that nearly all of these unaligned reads represented two different types of primer artifacts: TS oligo concatemers and primer artifact that was formed when the six random bases on 3′ end of the reverse transcription primer hybridized to the 9 base UMI on the 3′ end of the TS oligo (Supplemental Fig. S3). The increase in primer artifacts is likely the result of a decrease in RNA input into the nanoCAGE protocol because the enzymatic treatment should remove most of the total RNA. Despite making up between 15% and 35% of the total sequencing reads, when UMIs were considered these unaligned reads only made up 2%–5% (Supplemental Fig. S4) of total unique molecules sequenced, indicating that they are preferentially amplified during the PCR steps of the protocol or sequenced at a higher rate.

To benchmark our quality metrics to similar libraries, we analyzed CAGE data from The ENCODE Project Consortium (The ENCODE Project Consortium 2012) and nanoCAGE from The FANTOM Consortium (Andersson et al. 2014). We found that 5′

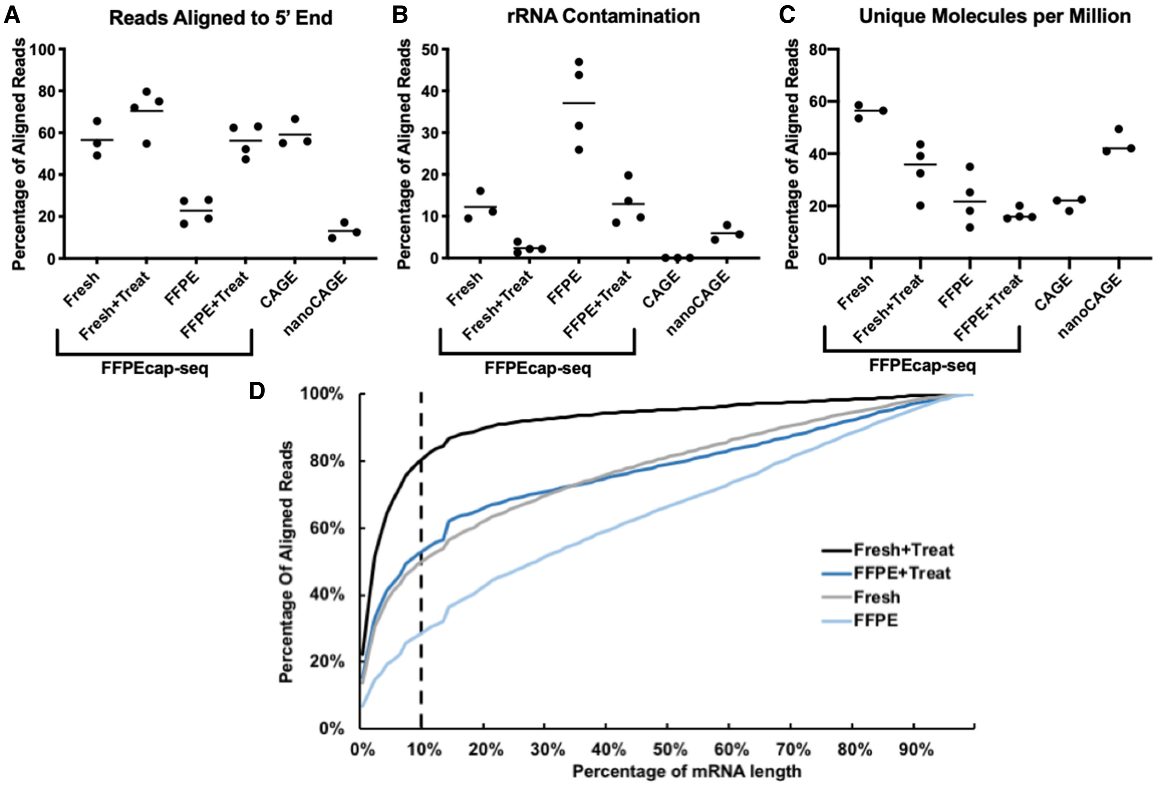


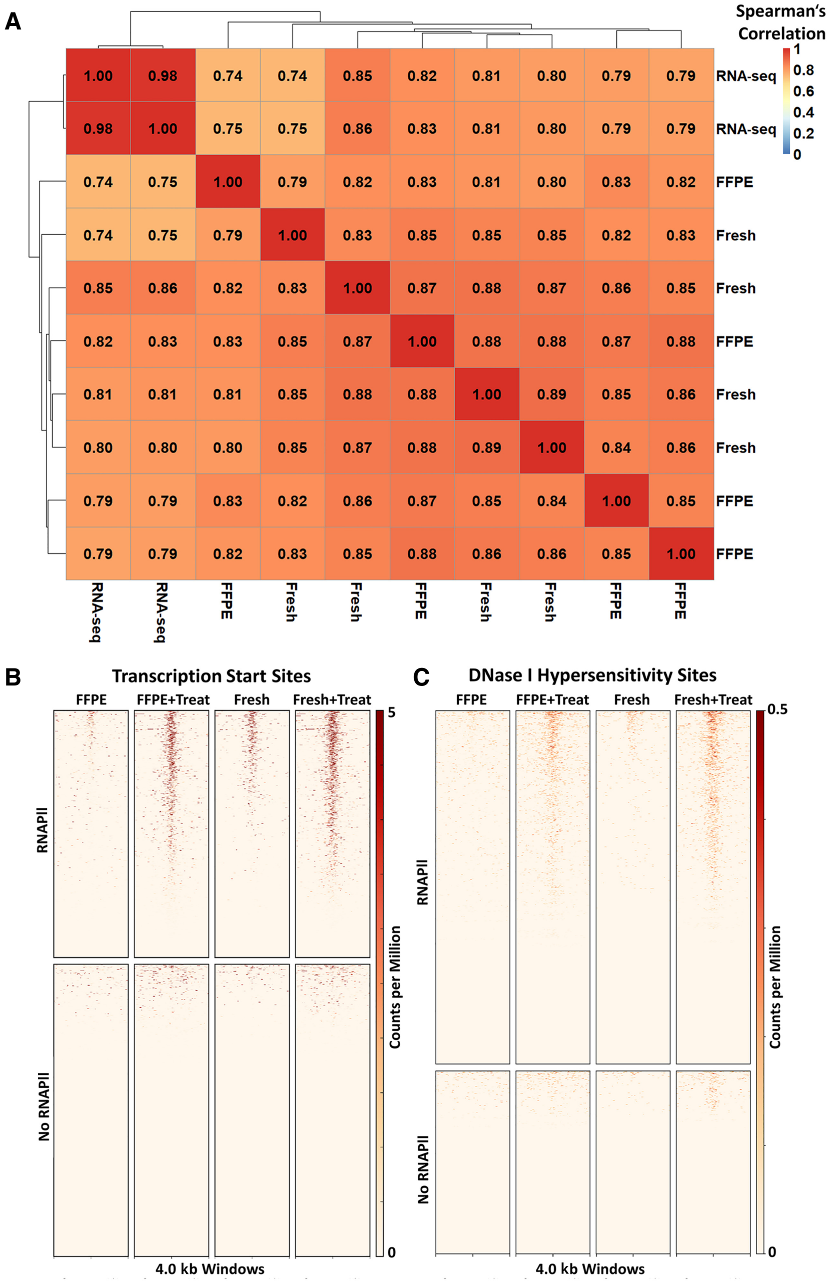
Figure 3. Enzymatic pretreatment improves library quality. Quality metrics, including percentage of reads mapping to the 10% most 5′ end of transcripts (*A*), percentage of reads mapping to rRNAs (*B*), and the percentage of unique reads from a random sampling of 1 million reads (*C*), are comparable to CAGE and nanoCAGE libraries and show improvement following enzymatic treatments for libraries made with either freshly derived or FFPE-derived RNA. Lines in graph represent means. (*D*) The cumulative percentage of reads within a certain percentage of mRNA length, starting at the 5′ end, is shown.

enrichment, rRNA levels, and unique number of molecules, were similar between FFPEcap-seq, CAGE, and nanoCAGE. Enrichment of reads mapping to the 5′ end of transcripts was similar between CAGE libraries (59.1%) and FFPEcap-seq libraries (70% in enzymatically treated libraries from fresh RNA), but nanoCAGE libraries showed less 5′ enrichment (13.2%) (Fig. 3A). CAGE libraries harbored very few reads mapping to rRNAs (0.05%) followed by nanoCAGE (6%), which was between FFPEcap-seq enzymatically treated freshly derived RNA (2.4%) and FFPE-derived RNA (12.9%) (Fig. 3B). The percentage of unique molecules in a random sampling of 1 million reads was higher for nanoCAGE (44.1%) and FFPEcap-seq applied to freshly derived RNA (33.8–56.2%) compared to CAGE (20.9%) (Fig. 3C). When the total number of unique molecules was quantified, we did not observe a significant difference between FFPEcap-seq (1.43 million, SD = 148,027), CAGE (1.59 million, SD = 408,340), and nanoCAGE (1.15 million, SD = 372,425). Because the CAGE and nanoCAGE libraries did not include UMIs, unique molecules were estimated based on mapped genomic locations. Overall, these results indicate that FFPEcap-seq produces libraries of comparable quality to CAGE and nanoCAGE.

The initial success of the FFPEcap-seq protocol on RNA derived from FFPE blocks that were <1 yr old, led us to explore the applicability to patient samples stored as FFPE blocks for longer periods of time. We performed FFPEcap-seq using RNA derived from two grade II endometrial adenocarcinoma FFPE blocks that were 10 yr old (collected in 2009) and 14 yr old (collected in 2005). The quality metrics were comparable to RNA from the 1-yr-old cell line FFPE samples (Supplemental Fig. S5A). The 5′ enrichment was 45% and 64.1%, rRNA levels were 1.1% and 0.8%, and the percentage of unique molecules from a random sampling of 1 million reads was 27.6% and 64.2%, for the 2005 and 2009 samples, respectively. We also observed an enrichment of reads mapping to transcription start sites and distal regulatory elements that showed RNA polymerase II (RNAPII) binding in Ishikawa cells based on ChIP-seq (Supplemental Fig. S5B,C; Carleton et al. 2017). These results indicate that FFPEcap-seq can be applied to FFPE specimens that have been stored for several years.

FFPEcap-seq data correlate with RNA-seq and uncover enhancer RNAs

The FFPEcap-seq libraries showed favorable quality metrics, and we next sought to determine if FFPEcap-seq produces reliable gene expression measurements. Reproducibility was examined by comparing replicate FFPEcap-seq libraries, and high concordance was observed with an average Spearman's rank correlation of 0.84 (Fig. 4A). We next compared FFPEcap-seq gene expression measurements to standard RNA-seq data from Ishikawa cells grown under the same conditions (Vahrenkamp et al. 2018) and found an average Spearman's rank correlation of 0.78 (Fig. 4A). RNA-seq from the patient samples described above was performed using the RNA Exome approach developed by Illumina.



	RNA-seq	RNA-seq	FFPE	Fresh	Fresh	FFPE	Fresh	Fresh	FFPE	FFPE
RNA-seq	1.00	0.98	0.74	0.74	0.85	0.82	0.81	0.80	0.79	0.79
RNA-seq	0.98	1.00	0.75	0.75	0.86	0.83	0.81	0.80	0.79	0.79
FFPE	0.74	0.75	1.00	0.79	0.82	0.83	0.81	0.80	0.83	0.82
Fresh	0.74	0.75	0.79	1.00	0.83	0.85	0.85	0.85	0.82	0.83
Fresh	0.85	0.86	0.82	0.83	1.00	0.87	0.88	0.87	0.86	0.85
FFPE	0.82	0.83	0.83	0.85	0.87	1.00	0.88	0.88	0.87	0.88
Fresh	0.81	0.81	0.81	0.85	0.88	0.88	1.00	0.89	0.85	0.86
Fresh	0.80	0.80	0.80	0.85	0.87	0.88	0.89	1.00	0.84	0.86
FFPE	0.79	0.79	0.83	0.82	0.86	0.87	0.85	0.84	1.00	0.85
FFPE	0.79	0.79	0.82	0.83	0.85	0.88	0.86	0.86	0.85	1.00

Figure 4. FFPEcap-seq results are correlated with RNA-seq results and uncover eRNAs. (*A*) Clustered Spearman's correlation matrix shows the correlation of read counts across genes between multiple FFPEcap-seq, all undergoing enzymatic treatment, and RNA-seq libraries. (*B*) Heatmap shows read depth at transcription start sites. Transcription start sites are split based on their overlap with RNA Polymerase II (RNAPII) ChIP-seq sites. (*C*) Heatmap of FFPEcap-seq read depth at DNase I hypersensitivity sites. Sites were split based on RNAPII overlap as measured by ChIP-seq.

For the patient samples, a Spearman's rank correlation of 0.57 (2005 sample) and 0.68 (2009 sample) was observed between the FFPEcap-seq measurements and RNA Exome results.

Considering the different type of information provided by FFPEcap-seq and RNA-seq, we consider this correlation to be high, but wanted to analyze the FFPEcap-seq in an orthogonal manner by analyzing RNAPII ChIP-seq from the same cells grown in the same conditions (Carleton et al. 2017). When we separated promoter regions based on the presence of RNAPII, we find that a majority of RNAPII-bound regions show FFPEcap-seq signal, but those promoters unoccupied by RNAPII show very little FFPEcap-seq signal (Odds ratios [Fresh untreated, FFPE untreated, Fresh treated, FFPE treated] = 81.58, 105,08, 132.05, 170.73; all *P*-values $< 2.2 \times 10^{-16}$, Fisher's exact test). This pattern was especially clear in the enzymatically treated libraries compared to untreated libraries (Fig. 4B). Taken together, these results show that FFPEcap-seq expression measurements are reproducible and correlate well with standard RNA-seq and RNAPII genome binding.

FFPEcap-seq also has the potential to identify eRNAs, because RNAs that derive from enhancers are RNAPII-produced transcripts that harbor 5′ caps. eRNA production correlates with enhancer activity; however, their levels are much lower than mRNAs (Andersson et al. 2014). To focus on potential regulatory regions, we examined DNase I hypersensitivity sites in Ishikawa cells (Gertz et al. 2013) that were distal from promoter regions (>2 kbp from annotated transcription start sites). In samples that did not undergo enzymatic treatment to enrich for capped 5′ ends, there was minimal signal at these candidate regulatory regions; however, we saw a marked increase in signal at these sites when enzymatic treatment was performed in either freshly derived or FFPE-derived RNA (Fig. 4C). When we split the DNase I hypersensitivity sites based on overlap with RNAPII binding, we find significant enrichment of signal at RNAPII-bound regions for both FFPE-derived RNA (*P*-value $< 2.2 \times 10^{-16}$, odds ratio = 6.47, Fisher's exact test) and freshly derived RNA (*P*-value $< 2.2 \times 10^{-16}$, odds ratio = 5.76, Fisher's exact test). These data suggest that we are able to detect eRNAs when FFPEcap-seq is applied to either freshly derived or FFPE-derived RNA.

To look at eRNA detection in more depth, we used an estrogen response, which has been shown to induce eRNA production at thousands of enhancers (Hah et al. 2013; Li et al. 2013). We used FFPEcap-seq to analyze RNA from Ishikawa cells, an estrogen responsive cell line (Carleton et al. 2017; Vahrenkamp et al. 2018; Rodriguez et al. 2019), that were treated with 17β-estradiol (E2) or vehicle (DMSO) for 1 h (Supplemental Fig. S6). By focusing on previously published ESR1-bound sites discovered with ChIP-seq (Gertz et al. 2013), we found that an average of 1386 loci (16% of ESR1-bound sites) showed FFPEcap-seq signal in the DMSO-treated samples and an average of 2876 loci (30.3%) had FFPEcap-seq signal in the E2-treated samples. After separating ESR1-bound sites based on RNAPII binding, we found a 2.4-fold enrichment in the percentage of loci with FFPEcap-seq signal at RNAPII-bound sites (51.3%) compared to sites not bound by RNAPII (21%). When signal at ESR1-bound sites was partitioned by mapping strand, we observed a shift (Supplemental Fig. S6) characteristic of bidirectional transcription at enhancers (Hah et al. 2013). These results reinforce the conclusion that eRNAs can be discovered by FFPEcap-seq.

Low input performance

After establishing that FFPEcap-seq can consistently measure gene expression and identify eRNAs, we wanted to determine the limitations of the approach in terms of RNA input requirements. We created FFPEcap-seq libraries with decreasing amounts of either freshly derived or FFPE-derived RNA, going from 400 ng to 6.25 ng of total RNA with twofold dilutions in between. Based on library yield, we sequenced all the fresh RNA libraries and down to the 12.5 ng input library for the FFPE RNA. The number of unique molecules decreased in a mostly linear fashion from 400 ng down to 100 ng for FFPE RNA and from 400 ng down to 25 ng for fresh RNA (Fig. 5A). Approximately fourfold more FFPE-derived total RNA is required to recover the same number of molecules as freshly derived RNA. We also observed an increase in primer artifacts with decreasing input. The number of unaligned reads, mostly primer artifacts, represents the majority of reads in low input libraries from 25 ng or less, but as we observed previously, the primer artifacts make up much smaller fractions of the total molecules (Fig. 5B). The increase in unaligned reads with decreasing input amounts affects the fresh samples differently than the FFPE samples. In the FFPE samples, transcription start site molecules remain a steady proportion of the total molecules, whereas reads aligning outside of exons and to rRNA decreased with decreasing input amounts. For fresh RNA, the group that decreased the most with lower input RNA was transcription start site mapped reads, whereas the unaligned reads increased. Based on the results from these RNA input experiments, we recommend using at least 25 ng of freshly derived RNA or 100 ng of FFPE-derived RNA, which should yield roughly 250,000 unique molecules.

Discussion

There are numerous opportunities for the clinical use of gene expression; however, accurately quantifying RNA from FFPE samples is a unique challenge owing to the high levels of degradation. In this study, we developed FFPEcap-seq, a method designed for sequencing capped 5′ RNA ends from FFPE-derived RNA. To develop the method, we needed to make several improvements to the previously published nanoCAGE protocol (Salimullah et al. 2011). The original protocol was relatively easy to implement and did enrich for the 5′ ends of known genes, but the sequencing results were very poor as a result of concatemers of the TS oligo introducing multiple sequencing primer binding sites. Our issues sequencing the libraries are consistent with Adiconis et al. (2018) reporting low sequencing yields with nanoCAGE. By modifying the TS oligo, we were able to block these concatemers, which resulted in a large increase in sequence yield and quality. We also added a UMI in the TS oligo to enable counting of molecules as opposed to sequence reads. The UMI allowed us to uncover that primer artifacts are preferentially amplified and sequenced, likely owing to their shorter overall length.

Even with these protocol modifications, nanoCAGE is not suitable for FFPE-derived RNA because template switching does not discriminate between capped 5′ ends and hydroxyl 5′ ends that are the result of degradation. This limitation of template switching was corroborated by our results in which 5′ enrichment was significantly reduced in FFPE-derived RNA that lacked enzymatic pretreatment. To overcome this limitation and create high-quality libraries from FFPE-derived RNA, we used enzymatic enrichment of 5′ transcript ends as described by Karen Adelman's laboratory (Nechaev et al. 2010; Scruggs et al. 2015). Enzymatic treatment of RNA before nanoCAGE greatly improved the quality of the libraries and resulted in 5′ enrichment and rRNA levels similar to those observed in libraries from untreated freshly derived RNA. Unexpectedly, libraries from freshly derived RNA were also improved by the enzymatic treatments, which could be

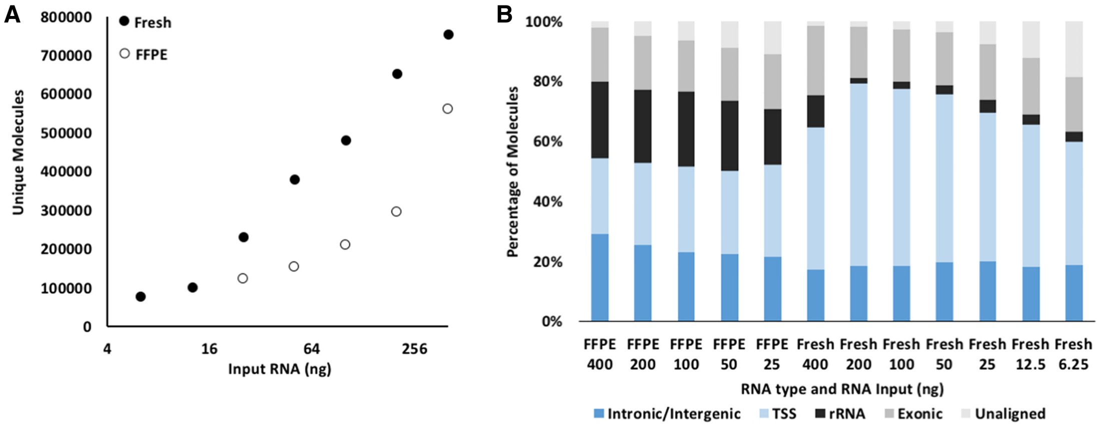


Figure 5. Evaluation of FFPEcap-seq input requirements. (*A*) The number of unique molecules sequenced is shown for different input amounts of freshly derived or FFPE-derived RNA. (*B*) The distribution of molecules across mapping categories changes with differing input amounts. (TSS) Transcription start sites.

attributable to a cleanup of low-level degradation and/or a reduction in rRNA amounts added to the nanoCAGE library construction. The one disadvantage to the enzymatic pretreatment is an increase in primer artifacts that we believe stems from less total RNA going into the nanoCAGE reaction, because the enzymatic treatment removes upward of 90% of total RNA owing to rRNA removal.

The mRNA gene expression measurements derived from FFPEcap-seq were reproducible and were well correlated with standard RNA-seq results. In addition, we were able to detect RNAs that were generated from promoter distal regulatory regions, likely representing eRNAs. The production of both mRNAs and eRNAs as measured by FFPEcap-seq was supported by RNAPII binding, providing evidence that FFPEcap-seq is discovering the 5′ ends of RNAPII transcribed RNAs. In testing the input limits of FFPEcap-seq we found that amounts as low as 100 ng for FFPE-derived RNA and 25 ng for freshly derived RNA are adequate for FFPEcap-seq library production and the identification of a quarter million molecules. Additional improvements, in terms of input requirements, might be achievable by combining FFPEcap-seq with recent adaptations of 5′ end detection methods. For example, SLIC-CAGE uses degradable carrier RNAs to enable library construction from <5 ng total RNA (Cvetesic et al. 2018), and NanoPARE uses a unique transposition approach to construct high-quality libraries from 5 ng of total RNA (Schon et al. 2018). Overall, our results establish FFPEcap-seq as a new method for accurately quantifying transcripts from FFPE-derived RNA. Because FFPEcap-seq can be performed in less than a day with standard laboratory equipment and costs less than $15 to construct, we believe it is an attractive method for interrogating gene expression in FFPE samples.

Methods

Cell culture

Ishikawa cells (Sigma-Aldrich) were maintained at 37°C with 5% CO_2 in RPMI-1640 media containing 10% fetal bovine serum, 50 units/mL penicillin, and 50 mg/mL streptomycin. Cells were transferred to hormone-depleted media 5 d before RNA harvest to match previously performed RNA-seq experiments. Hormone-depleted media consisted of phenol red-free RPMI-1640, 10% charcoal-dextran treated fetal bovine serum (Sigma-Aldrich), 50 units/mL penicillin, and 50 mg/mL streptomycin. For E2 inductions, cells were treated with 10 nM E2 or 0.02% DMSO as a vehicle control for 1 h before harvest. Cells were processed into FFPE samples or RNA was immediately extracted. For immediate extraction, cells were lysed in RLT Plus buffer (QIAGEN) with 1% beta-mercaptoethanol. RNA was extracted with the Quick-RNA Miniprep kit (Zymo Research).

Patient samples

Deidentified patient specimen FFPE blocks were obtained from the Huntsman Cancer Institute's Biorepository and Molecular Pathology Shared Resource. Slides stained with hematoxylin and eosin were used to identify regions of tumor. Two cores that were 2 mm in diameter were extracted from each FFPE block in the tumor-identified regions. RNA was extracted from the cores using the RNeasy FFPE kit (QIAGEN).

FFPE preparation

Cells were scraped from a 15 cm dish in phosphate buffered saline and centrifuged at 2000*g* for 6 min at 4°C. Cells for the FFPE treatment were placed in 10% neutral buffered formalin (NBF) and centrifuged in a 1.5 mL tube at 200*g* for 6 min at 4°C. The 1.5-mL tubes were submerged in 15-mL tubes flooded with 10% NBF for 24 h at room temperature. After 24 h, the NBF was rinsed with 70% ethanol, and cell pellets were dislodged from the 1.5-mL tube and placed in a tissue cassette with histology paper and submerged in 70% ethanol. The cassettes were embedded in paraffin by the Biorepository and Molecular Pathology Shared Resource at the Huntsman Cancer Institute. Four 20 µm scrolls were collected from these samples, and RNA was extracted using the RNeasy FFPE kit (QIAGEN).

FFPEcap-seq library preparation

A detailed FFPEcap-seq protocol can be found in the Supplemental Protocol. RNA concentration was measured using a Qubit 2.0 with the RNA high sensitivity or RNA broad range assay (Thermo Fisher Scientific), depending on the concentration. Enzyme pretreatment was performed using 1 µL (1 unit) Terminator Nuclease (Lucigen), 2 µL (20 units) T4 Polynucleotide Kinase (New England Biolabs), 2 µL 10× Terminator Buffer A, 2 µL 10 mM ATP, 1 µL SUPERase In (Thermo Fisher Scientific), input RNA, and water in a 20 µL reaction. Samples were incubated for 45 min at 30°C before being cleaned up with the RNA Clean and Concentration kit (Zymo Research). RNA was eluted using 6 µL water, with 5 µL being collected after centrifugation. One microliter of eluted RNA was used for quantification with the Qubit 2.0 RNA high sensitivity kit (Thermo Fisher Scientific). For enzymatic treatment controls, the preceding procedure was followed except Terminator and T4 Polynucleotide Kinase were replaced with water. When polyphosphatase was tested, 1 µL (20 units) RNA 5′ polyphosphatase (Lucigen) was included in the reaction.

To begin the modified nanoCAGE library preparation, trehalose-sorbitol solution was made by adding 5 mL of a saturated trehalose solution (Sigma-Aldrich) to 10 mL 1.28 g/mL sorbitol solution (Sigma-Aldrich). One gram celex 100 (Sigma-Aldrich) was added to the solution and vortexed vigorously and allowed to sit for 3 h at room temperature. The tube was then centrifuged for 10 min at 600*g* and the supernatant was aliquoted and stored at −30°C. Oligo master mix was made by combining 8 µL Trehalose-Sorbitol solution with 1 µL 1 mM Template switching oligo, and 1 µL 100 µM reverse transcriptase oligo (for all oligo sequences, see Supplemental Table S1). All oligonucleotides were ordered from Integrated DNA Technologies. One microliter of the oligo master mix was combined with the remaining 4 µL RNA and incubated at 65°C for 10 min, then immediately moved to ice for 2 min. The template switching reaction contained 1 µL Prime Script Reverse Transcriptase (Takara), 2 µL 5× Prime Script buffer, 1.5 µL 5M Betaine, 0.5 µL 200 mM DTT, and 0.625 µL 10 mM dNTP mixture. The template switching reaction mixture was combined with the RNA oligo mixture and heated for 10 min to 22°C, for 30 min at 40°C, and then for 15 min at 75°C. The sample was then cleaned up using Ampure XP beads (Beckman Coulter) with a ratio of 1.5:1 of bead volume to sample volume. The cDNA was eluted in 40 µL water. For second strand synthesis, 7.5 µL purified cDNA was added to 25 µL Phusion HF master mix (New England Biolabs), with 2.5 µL 10 µM second strand forward primer, 2.5 µL 10 µM second strand reverse primer, and 12.5 µL of water. The reaction was heated for 1 min at 95°C, followed by 24 cycles for 15 sec at 95°C, for 10 sec at 65°C, and for 2 min 68°C. This reaction was purified using Ampure XP beads (Beckman Coulter) with a 0.8:1 ratio of bead volume to sample volume and eluted in 30 µL water. The final library was created by taking 40 ng purified double-stranded cDNA and mixing it with 25 µL 2× Phusion HF master mix (New England Biolabs), 1.25 µL 10 µM library forward primer, 1.25 µL 10 µM library reverse primer (containing a sample index), and water to a total volume of 50 µL. Samples were heated for 1 min at 95°C, followed by two cycles of 15 sec at 95°C, 10 sec at 55°C, and 2 min at 68°C, then followed by eight cycles of 15 sec at 95°C, 10 sec at 65°C, and 2 min at 68°C. This reaction was purified using Ampure XP beads (Beckman Coulter) with a 0.8:1 ratio of bead volume to sample volume and eluted in 30 µL water. Samples were then measured by Qubit 2.0 DNA high sensitivity (Thermo Fisher Scientific) before sample pooling of equal weights. Libraries were sequenced on an Illumina HiSeq 2500 as single end 50-bp reads using a custom read one sequencing primer (Supplemental Table S1) and standard index primer.

FFPEcap-seq analysis

The sequencing reads were first processed by removing the first 13 bases (N_9AGGG) and storing the 9 base UMI. We also removed up to two additional Gs from the 5′ end of the trimmed sequence to account for additional Cs added during the template switching reaction. The trimmed reads were then scanned for sequences that match the template switching oligo or the reverse transcription oligo. If a perfect match of at least eight bases of either oligo was found in the read, the read was considered to be a primer artifact and removed from downstream analysis. The remaining reads were aligned to the hg19 build of the human genome using Bowtie (Langmead et al. 2009) with the following parameters: Bowtie -S --chunkmbs 512 -m 1 -t --best -q -l 32 -e 80 -n 2. The hg19 build of the human genome was used for all genomic analyses. We do not believe that realigning reads to the current genome build (GRCh38) would substantially change results, because we are restricting our analyses to uniquely alignable regions of the genome. The reads were also aligned to the RefSeq reference for hg19 using Bowtie with the following parameters: Bowtie -S -n 2 -a -m 10 -S --chunkmbs 512. Reads that had the same UMI and aligned to the same place were considered to originate from a single molecule of RNA and were counted as a single molecule in downstream analysis.

CAGE libraries were obtained from Gene Expression Omnibus accession number GSM849364. NanoCAGE libraries were obtained from The FANTOM5 Consortium (http://fantom.gsc.riken.jp/5/datafiles/latest/basic/). MCF-7 CAGE samples were used. For nanoCAGE data sets (referred to as CAGEscan by The FANTOM Consortium), we used two cervical cancer replicates and a renal carcinoma. Only the forward read was used for comparisons, and the reads were analyzed as described for FFPEcap-seq libraries with the exception that no read trimming was performed owing to a lack of UMIs. Postalignment analysis was performed using AWK and R (R Core Team 2017). Ribosomal RNAs were identified by their alignment to the rRNA genes in the RefSeq alignment. 5′ enrichment was calculated as the percentage of reads that align to the 5′-most 10% of RefSeq mRNAs in the RefSeq alignments. We calculated the percentage of unique molecules per million reads by taking a random sampling of one million reads and determining the number of unique reads, based on the UMI sequence and genomic mapping location. To perform gene expression comparisons, reads mapping to hg19 UCSC knownGene gene models (Kent et al. 2002) were counted using featureCounts from the SubRead package (Liao et al. 2014). Correlation between FFPEcap-seq libraries and RNA-seq libraries (GSE109892) were calculated using normalized counts generated using DESeq2 (Love et al. 2014). All statistics were calculated using R. Heatmaps of FFPEcap-seq signal at RNAPII- and ESR1-bound sites were centered on ChIP-seq peak summits from RNAPII (GSE99905) and ESR1 (GSE32465) ChIP-seq data.

RNA sequencing from patient samples

To perform RNA-seq from patient samples, we used Illumina's TruSeq RNA Exome kit, which combines RNA-seq library construction with exome capture. Libraries were constructed from 100 ng total RNA and sequenced on the NovaSeq as paired end 50-bp reads. For comparison with FFPEcap-seq samples, only the forward reads were aligned and analyzed. Reads were aligned and analyzed using the same programs as the FFPEcap-seq samples, with the exception of the removal of UMIs and their associated analysis. Gene expression analysis was performed in the same manner as the aforementioned RNA-seq analysis.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE126346. FFPEcap-seq analysis software is available at GitHub (https://github.com/jeffpkamp/FFPEcap-seq) and as Supplemental Code.

Acknowledgments

This work was supported by the National Institutes of Health/National Cancer Institute (NIH/NCI) R21 CA196455 to J.G. Research reported in this publication used the High-Throughput Genomics Shared Resource and the Biorepository and Molecular Pathology Shared Resource at the University of Utah and was supported by NIH/NCI award P30 CA042014. We thank K-T Varley, Eric Bogenschutz, Schuyler O'Brien, Scott Kuersten, Jesse Gray, Sunil Sharma, and Phil Bernard, as well as members of the Gertz and Varley laboratories, for their suggestions and comments during the development of FFPEcap-seq.

Author contributions: Conceptualization and methodology for the project were done by J.M.V. and J.G.; resources were acquired by K.S., M.K.D., and A.P.S.; investigation was done by J.M.V. and E.A.J.; the original draft was written by J.M.V.; review editing was done by J.M.V. and J.G.; and J.G. acquired funding.

References

Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM, Sivachenko A, Thompson DA, Wysoker A, Fennell T, et al. 2013. Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat Methods* **10:** 623–629. doi:10.1038/nmeth.2483

Adiconis X, Haber AL, Simmons SK, Levy Moonshine A, Ji Z, Busby MA, Shi X, Jacques J, Lancaster MA, Pan JQ, et al. 2018. Comprehensive comparative analysis of 5′-end RNA-sequencing methods. *Nat Methods* **15:** 505–511. doi:10.1038/s41592-018-0014-2

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, et al. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403:** 503–511. doi:10.1038/35000501

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507:** 455–461. doi:10.1038/nature12787

Carleton JB, Berrett KC, Gertz J. 2017. Multiplex enhancer interference reveals collaborative control of gene regulation by estrogen receptor α-bound enhancers. *Cell Syst* **5:** 333–344.e5. doi:10.1016/j.cels.2017.08.011

Cieslik M, Chugh R, Wu YM, Wu M, Brennan C, Lonigro R, Su F, Wang R, Siddiqui J, Mehra R, et al. 2015. The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. *Genome Res* **25:** 1372–1381. doi:10.1101/gr.189621.115

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322:** 1845–1848. doi:10.1126/science.1162228

Cvetesic N, Leitch HG, Borkowska M, Müller F, Carninci P, Hajkova P, Lenhard B. 2018. SLIC-CAGE: high-resolution transcription start site mapping using nanogram-levels of total RNA. *Genome Res* **28:** 1943–1956. doi:10.1101/gr.235937.118

De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei CL, Natoli G. 2010. A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. *PLoS Biol* **8:** e1000384. doi:10.1371/journal.pbio.1000384

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74. doi:10.1038/nature11247

The FANTOM Consortium and RIKEN Genome Exploration Research Group and Genome Science Group. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559–1563. doi:10.1126/science.1112014

Gertz J, Savic D, Varley KE, Partridge EC, Safi A, Jain P, Cooper GM, Reddy TE, Crawford GE, Myers RM. 2013. Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol Cell* **52:** 25–36. doi:10.1016/j.molcel.2013.08.037

Hah N, Murakami S, Nagari A, Danko CG, Kraus WL. 2013. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res* **23:** 1210–1223. doi:10.1101/gr.152306.112

Hsieh CL, Fei T, Chen Y, Li T, Gao Y, Wang X, Sun T, Sweeney CJ, Lee GS, Chen S, et al. 2014. Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. *Proc Natl Acad Sci* **111:** 7319–7324. doi:10.1073/pnas.1324151111

Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P, Linnarsson S. 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* **21:** 1160–1167. doi:10.1101/gr.110882.110

Kapteyn J, He R, McDowell ET, Gang DR. 2010. Incorporation of non-natural nucleotides into template-switching oligonucleotides reduces background and improves cDNA synthesis from very small RNA samples. *BMC Genomics* **11:** 413. doi:10.1186/1471-2164-11-413

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12:** 996–1006. doi:10.1101/gr.229102

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465:** 182–187. doi:10.1038/nature09033

Kwak H, Fuda NJ, Core LJ, Lis JT. 2013. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339:** 950–953. doi:10.1126/science.1229386

Kwan S, Gerlach VL, Brow DA. 2000. Disruption of the 5′ stem-loop of yeast U6 RNA induces trimethylguanosine capping of this RNA polymerase III transcript in vivo. *RNA* **6:** 1859–1869. doi:10.1017/S1355838200991325

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi:10.1186/gb-2009-10-3-r25

Li W, Notani D, Ma Q, Tanasa B, Nunez E, Chen AY, Merkurjev D, Zhang J, Ohgi K, Song X, et al. 2013. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498:** 516–520. doi:10.1038/nature12210

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30:** 923–930. doi:10.1093/bioinformatics/btt656

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15:** 550. doi:10.1186/s13059-014-0550-8

Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ. 2014. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res* **24:** 496–510. doi:10.1101/gr.161034.113

Mousavi K, Zare H, Dell'orso S, Grontved L, Gutierrez-Cruz G, Derfoul A, Hager GL, Sartorelli V. 2013. eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Mol Cell* **51:** 606–617. doi:10.1016/j.molcel.2013.07.022

Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, Adelman K. 2010. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science* **327:** 335–338. doi:10.1126/science.1181421

O'Connell MJ, Lavery I, Yothers G, Paik S, Clark-Langone KM, Lopatin M, Watson D, Baehner FL, Shak S, Baker J, et al. 2010. Relationship between tumor gene expression and recurrence in four independent studies of patients with stage II/III colon cancer treated with surgery alone or surgery plus adjuvant fluorouracil plus leucovorin. *J Clin Oncol* **28:** 3937–3944. doi:10.1200/JCO.2010.28.9538

Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, et al. 2004. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* **351:** 2817–2826. doi:10.1056/NEJMoa041588

Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al. 2000. Molecular portraits of human breast tumours. *Nature* **406:** 747–752. doi:10.1038/35021093

R Core Team. 2017. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Ramsköld D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, et al. 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30:** 777–782. doi:10.1038/nbt.2282

Rodriguez AC, Blanchard Z, Maurer KA, Gertz J. 2019. Estrogen signaling in endometrial cancer: a key oncogenic pathway with several open questions. *Horm Cancer* **10:** 51–63. doi:10.1007/s12672-019-0358-9

Salimullah M, Sakai M, Plessy C, Carninci P. 2011. NanoCAGE: a high-resolution technique to discover and interrogate cell transcriptomes. *Cold Spring Harb Protoc* **2011:** pdb.prot5559. doi:10.1101/pdb.prot5559

Schmidt WM, Mueller MW. 1999. CapSelect: a highly sensitive method for 5′ CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. *Nucleic Acids Res* **27:** e31. doi:10.1093/nar/27.21.e31

Schon MA, Kellner MJ, Plotnikova A, Hofmann F, Nodine MD. 2018. NanoPARE: parallel analysis of RNA 5′ ends from low-input RNA. *Genome Res* **28:** 1931–1942. doi:10.1101/gr.239202.118

Scruggs BS, Gilchrist DA, Nechaev S, Muse GW, Burkholder A, Fargo DC, Adelman K. 2015. Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Mol Cell* **58:** 1101–1112. doi:10.1016/j.molcel.2015.04.006

Shuman S. 1997. Origins of mRNA identity: capping enzymes bind to the phosphorylated C-terminal domain of RNA polymerase II. *Proc Natl Acad Sci* **94:** 12758–12760. doi:10.1073/pnas.94.24.12758

Tsuchihara K, Suzuki Y, Wakaguri H, Irie T, Tanimoto K, Hashimoto S, Matsushima K, Mizushima-Sugano J, Yamashita R, Nakai K, et al. 2009. Massive transcriptional start site analysis of human genes in hypoxia cells. *Nucleic Acids Res* **37:** 2249–2263. doi:10.1093/nar/gkp066

Vahrenkamp JM, Yang CH, Rodriguez AC, Almomen A, Berrett KC, Trujillo AN, Guillen KP, Welm BE, Jarboe EA, Janat-Amsbury MM, et al. 2018. Clinical and genomic crosstalk between glucocorticoid receptor and estrogen receptor α in endometrial cancer. *Cell Rep* **22:** 2995–3005. doi:10.1016/j.celrep.2018.02.076

Veldman-Jones MH, Brant R, Rooney C, Geh C, Emery H, Harbron CG, Wappett M, Sharpe A, Dymond M, Barrett JC, et al. 2015. Evaluating robustness and sensitivity of the NanoString technologies nCounter platform to enable multiplexed gene expression analysis of clinical samples. *Cancer Res* **75:** 2587–2593. doi:10.1158/0008-5472.CAN-15-0262

Waldron L, Simpson P, Parmigiani G, Huttenhower C. 2012. Report on emerging technologies for translational bioinformatics: a symposium on gene expression profiling for archival tissues. *BMC Cancer* **12:** 124. doi:10.1186/1471-2407-12-124

Yakovleva A, Plieskatt JL, Jensen S, Humeida R, Lang J, Li G, Bracci P, Silver S, Bethony JM. 2017. Fit for genomic and proteomic purposes: sampling the fitness of nucleic acid and protein derivatives from formalin fixed paraffin embedded tissue. *PLoS One* **12:** e0181756. doi:10.1371/journal.pone.0181756

Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM. 2014. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* **15:** 419. doi:10.1186/1471-2164-15-419

Received February 20, 2019; accepted in revised form October 3, 2019.