

METHOD

Genome-wide maps of polyadenylation reveal dynamic mRNA 3'-end formation in mammalian cell lineages

LI WANG,¹ ROBIN D. DOWELL,^{1,2,3} and RUI YI^{1,3}

¹Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, Colorado 80309, USA

²BioFrontiers Institute, University of Colorado, Boulder, Colorado 80309, USA

ABSTRACT

Post-transcriptional regulation, often mediated by miRNAs and RNA-binding proteins at the 3' untranslated regions (UTRs) of mRNAs, is implicated in important roles in the output of transcriptome. To decipher this layer of gene regulation, it is essential to measure global mRNA expression quantitatively in a 3'-UTR-specific manner. Here we establish an experimental and bioinformatics pipeline that simultaneously determines 3'-end formation by leveraging local nucleotide composition and quantitatively measures mRNA expression by sequencing polyadenylated transcripts. When applied to purified mouse embryonic skin stem cells and their daughter lineages, we identify 18,060 3' UTRs representing 12,739 distinct mRNAs that are abundantly expressed in the skin. We determine that ~78% of UTRs are formed by using canonical A[A/U]UAAA polyadenylation signals, whereas ~22% of UTRs use alternative signals. By comparing to relative and absolute mRNA abundance determined by qPCR, our RNA-seq approach can precisely measure mRNA fold-change and accurately determine the expression of mRNAs over four orders of magnitude. Surprisingly, only 829 out of 12,739 genes show differential 3'-end usage between embryonic skin stem cells and their immediate daughter cells, whereas the numbers increase to 933 genes when comparing embryonic skin stem cells with the more remotely related hair follicle cells. This suggests an evolving diversity instead of switch-like dynamics in 3'-end formation during development. Finally, core components of the miRNA pathway including *Dicer*, *Dgcr8*, *Xpo5*, and *Argonautes* show dynamic 3'-UTR formation patterns, indicating a self-regulatory mechanism. Together, our quantitative analysis reveals a dynamic picture of mRNA 3'-end formation in tissue stem cell lineages in vivo.

Keywords: 3'-end formation; polyadenylation signal; RNA-seq; cell lineages

INTRODUCTION

Patterns of gene expression within a complete transcriptome hold valuable information regarding the specific functions of a particular tissue or cell type. The key goals of transcriptome studies are to discover and catalog the complete set of transcripts, annotate the structure of these transcripts, and accurately quantify their absolute expression levels. It is widely recognized that the 3' UTRs of mRNAs harbor numerous regulatory elements that are implicated in important functions for mRNA metabolism (Di Giammartino et al. 2011). Alternative polyadenylation (APA) is of special interest because it generates 3' UTRs with different lengths, which provides a mechanism to control transcript stability, localization, and translational efficiency by modulating *cis*-elements on the 3' UTR (Di Giammartino et al. 2011). APA has been implicated in modulating proliferation and transformation in

cancer cells (Sandberg et al. 2008; Mayr and Bartel 2009; Lin et al. 2012). Recent studies further suggest that APA is widely present with cell-type-specific and tissue-specific patterns (Fu et al. 2011; Derti et al. 2012; Smibert et al. 2012; Ulitsky et al. 2012). However, most studies have focused on differences between in vitro cultured cells derived from different tissues, or whole organ-scale comparison. To date, there is little information about whether APA is involved in developmental transitions among closely related cell lineages in vivo, which is critical to understand how APA is initiated and controlled in biological processes. In this regard, mouse embryonic skin affords an ideal system to examine how APA is controlled in somatic cell lineages. First, embryonic skin stem cells and their daughter cells are abundantly available and functionally distinct (Blanpain and Fuchs 2009). Second, embryonic skin stem cells give rise to their differentiated daughter cells by an asymmetric cell division mechanism (Lechler and Fuchs 2005). Upon cell division, the differentiating daughter cells exit the cell cycle and embark on the epidermal differentiation program. Thus, the stem cell populations are separated from the differentiated cells by a single cell division (Supplemental Fig. S1). Taken together, profiling mRNA 3'-end

³Corresponding authors

E-mail robin.dowell@colorado.edu

E-mail yir@colorado.edu

Article published online ahead of print. Article and publication date are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.035360.112>.

formation in mouse embryonic skin provides an opportunity to examine APA during a developmental transition in vivo by global transcriptome analyses.

To determine the complexity of mRNA 3'-end formation and quantify their expression, many RNA-seq techniques have been recently developed (Ozsolak et al. 2009; Beck et al. 2010; Fu et al. 2011; Jan et al. 2011; Derti et al. 2012). In general, these techniques can be divided into three categories: (1) direct mRNA sequencing based on the Helicos platform (Ozsolak et al. 2009, 2010); (2) direct adaptor ligation to capture mRNA 3' ends followed by Illumina sequencing, also known as 3P-Seq (Jan et al. 2011); (3) oligo(dT) priming-based mRNA 3'-end capture followed by Illumina sequencing, usually known as 3Seq (Beck et al. 2010; Fu et al. 2011; Shepard et al. 2011; Derti et al. 2012). Among them, direct sequencing and 3P-Seq experimentally distinguish mRNA 3' ends, whereas oligo(dT) priming-based 3Seq can prime from any A-rich region in addition to authentic mRNA 3' ends that contain poly(A) tails. However, the quantitative performance in high throughput of direct sequencing is not clear, and the Helicos platform is yet to be widely accessible for extensive development. 3P-Seq requires multiple steps of enzymatic reaction including RNA adaptor ligation, which often yields poor quantification in deep sequencing because of significant bias introduced by RNA ligase (Hafner et al. 2011). In contrast, virtually all quantitative techniques for RNA molecules, including qPCR and microarray, are based on oligo(dT) or random priming followed by signal amplification. Therefore, oligo(dT) priming-based 3Seq has its unique advantage in mRNA quantification. We reason that if we can computationally distinguish authentic mRNA 3'-end signals from internal priming events, we can develop a tool to accurately and quantitatively measure mRNA 3' ends by 3Seq.

In this study, we leverage the distinct nucleotide composition patterns of mRNA 3'-end regions and establish a bioinformatics pipeline to identify mRNA 3' ends accurately. When benchmarked with the direct mRNA sequencing results, we determine that the overall successful rate of our analysis to detect authentic 3'-end signals is 88%. When applied to purified mouse skin stem cell lineages, we show that 3Seq allows accurate and cost-effective quantification of mRNA transcriptome, spanning at least four orders of magnitude. Our quantitative measurement also reveals a dynamic pattern of APA in closely related stem cell lineages in vivo.

RESULTS

Distinct nucleotide composition patterns at mRNA 3'-end regions

Among the three current 3'-end sequencing approaches, direct sequencing and 3P-Seq experimentally distinguish authentic 3'-end formation (Ozsolak et al. 2010; Jan et al. 2011). Interestingly, local nucleotide composition surround-

ing mRNA 3'-end regions determined by 3P-Seq in *Caenorhabditis elegans* showed a distinct pattern (Jan et al. 2011). Importantly, this nucleotide composition pattern was universally observed regardless of whether canonical or alternative polyadenylation signals (PAS) were used or whether proximal or distal 3' ends were analyzed (Jan et al. 2011). These observations suggest a requirement for multiple motifs with a specific positional distribution for 3'-end formation in *C. elegans*. To determine whether this pattern is conserved in mammals, we analyzed a previously published human data set generated by direct RNA sequencing (Ozsolak et al. 2010). As expected, the distinct pattern of nucleotide composition was observed (Fig. 1A). Specifically, we observed that (1) A is significantly enriched at +1 position; (2) an ~20-nt A-rich region with a peak at -18 position is identified at the -10 to -30 position, matching the position of the PAS; (3) a short U-rich region and a short A-rich region are observed at the -1 to -10 position; (4) an ~30-nt U-rich region is detected downstream from the cleavage site. In contrast, when we analyzed a previously published human data set generated by 3Seq (Jenal et al. 2012), the distinct pattern was not observed. Instead, the sequences downstream from the cleavage site were dominated by A-rich sequences, indicating widespread internal priming events in unfiltered 3Seq results (Fig. 1B). Together, these results suggested that authentic 3'-end formation requires a distinct nucleotide composition pattern, and we could develop a bioinformatics approach to leverage this pattern to computationally distinguish authentic mRNA 3'-end formation from the large number of internal priming events generated by 3Seq.

To test this possibility, we performed 3Seq and developed our bioinformatics pipeline with mouse embryonic skin lineages including basal stem cells and suprabasal differentiated cells (Supplemental Fig. S1). We also optimized experimental parameters for more effective library construction that improves cleavage site identification: (1) We fragmented RNA molecules into <150-nt pieces with the peak at 60–80 nt; (2) we specifically selected amplicons with <100-nt insertions. As a result, although the reads were sequenced from the 5' ends of the RNA fragments, the majority of mappable reads (89.1%) contained untemplated, tandem A's (Supplemental Fig. S2). We uniquely mapped 7.9 million and 7.4 million 100-mer reads for each library. Because of our optimization in the library construction, our reads were highly enriched around the cleavage site (nucleotide position 0) with a gradual slope from the 5' end and a sharp drop in the 3' end (Fig. 1C). To determine the accuracy of the cleavage site identification, we selected peaks whose cleavage site is followed by C, T, or G (non-A) in the genomic sequences, and plotted the reads density of these trimmed ends (Fig. 1D). Indeed, trimmed ends were significantly enriched at the defined 0 position. Thus, our optimized protocol has generated a very high percentage of reads with the untemplated A's that allows accurate determination of the cleavage site.

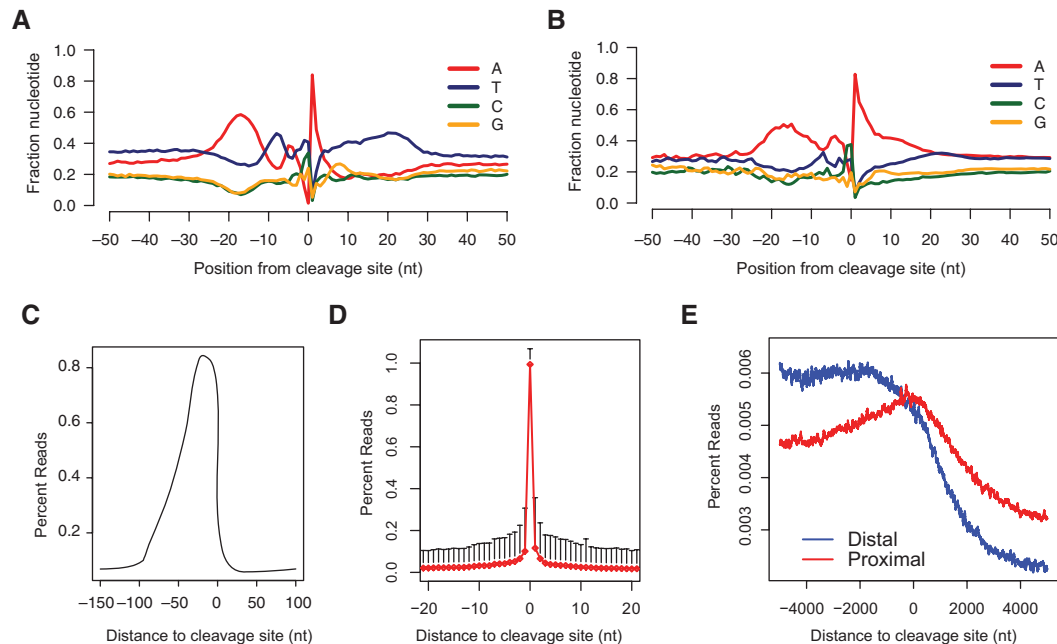


FIGURE 1. Distinct nucleotide composition patterns at mRNA 3'-end regions. Nucleotide composition plot of poly(A) site, -50 to $+50$ window for Direct RNA sequencing data from human liver RNA sample in *A* (Ozsolak et al. 2010) and 3Seq data from human U2OS cell line in *B* (Jenal et al. 2012). *(C)* 3Seq metagene analysis of -150 to $+100$ window centered at the cleavage site. The percent reads are calculated by dividing the read coverage on each position by all the number of reads mapped to this window. 3Seq reads are significantly enriched in the -100 to 0 position, with dramatic drop downstream from the 0 position. *(D)* Percent reads plot of 3' ends of trimmed reads centered at the cleavage site whose $+1$ position is T, G, or C. *(E)* H3K36me3 metagene plot centered at the cleavage site. Distal peaks (blue) are defined as the most 3' peak assigned to a given gene. Proximal peaks (red) are peaks mapped to RefSeq and Ensembl-annotated that are proximal in location to the distal peaks.

Because polyadenylation is a post-transcriptional process, the poly(A) site should be located within the actively transcribed region of RNA Pol II and close to the 3' end of the actively transcribed region of RNA Pol II (Lin et al. 2012). To provide insights into the link between 3'-end formation and RNA Pol II transcription, we performed ChIP-seq of histone H3K4me3 and H3K36me3 in the basal stem cells. The combination of these two histone methylation marks defines actively transcribed regions (Barski et al. 2007; Ernst and Kellis 2010). By matching the histone methylation marks and mRNA 3' ends as detected by 3Seq, we discovered that 9726 out of 9983 (97.4%) genes with H3K4me3 and H3K36me3 double-positive marks show 3Seq peaks in the basal stem cells. To define the relationship between 3'-end formation and Pol II transcription of mRNAs more accurately, we selected the most distal 3Seq peaks for all H3K4me3 and H3K36me3 double-positive genes and plotted the H3K36me3 density from the $+5$ kb to -5 kb region centered on the cleavage site. This metagene analysis revealed that the H3K36me3 signals are enriched upstream of the cleavage site and are weakened rapidly downstream from the cleavage site for all distal peaks (Fig. 1E). This indicated that transcription termination takes place shortly after the distal 3'-end cleavage site. In contrast, the drop-off of the H3K36me3 signals for proximal peaks was considerably milder, compared with that of the distal peaks (Fig. 1E). Collectively, these analyses strongly support that our optimized technique successfully

defines the 3' ends of mRNAs and provide new insights for the relationship between Pol II transcription, as detected by H3K36me3, and 3'-end formation.

Distinguish authentic 3'-end signals from internal priming signals

To systematically identify authentic 3'-end signals, we designed a bioinformatics pipeline to eliminate internal priming events by using a combination of motifs upstream of and downstream from the defined cleavage site. We divided all 3Seq peaks into four mutually exclusive categories based on the existence of the canonical PAS and the downstream A-rich sequences (Fig. 2A; Supplemental Fig. S3). Peaks in category 1 contained canonical PAS (A[A/U]UAAA) and showed no A-rich genomic sequences downstream from the cleavage site. Thus, they represented authentic 3'-end signals. Peaks in category 2 contained both canonical PAS and downstream A-rich sequences. Thus, some of them were derived from mRNA 3' ends, and the others were likely derived from internal priming. Peaks in category 3 were defined by the lack of both canonical PAS and downstream A-rich sequences. These peaks were likely derived from 3'-end formation using alternative PAS or sequencing artifacts as judged by the low reads number. Peaks in category 4 contained only downstream A-rich sequences and lacked canonical PAS upstream of the cleavage site. They should represent

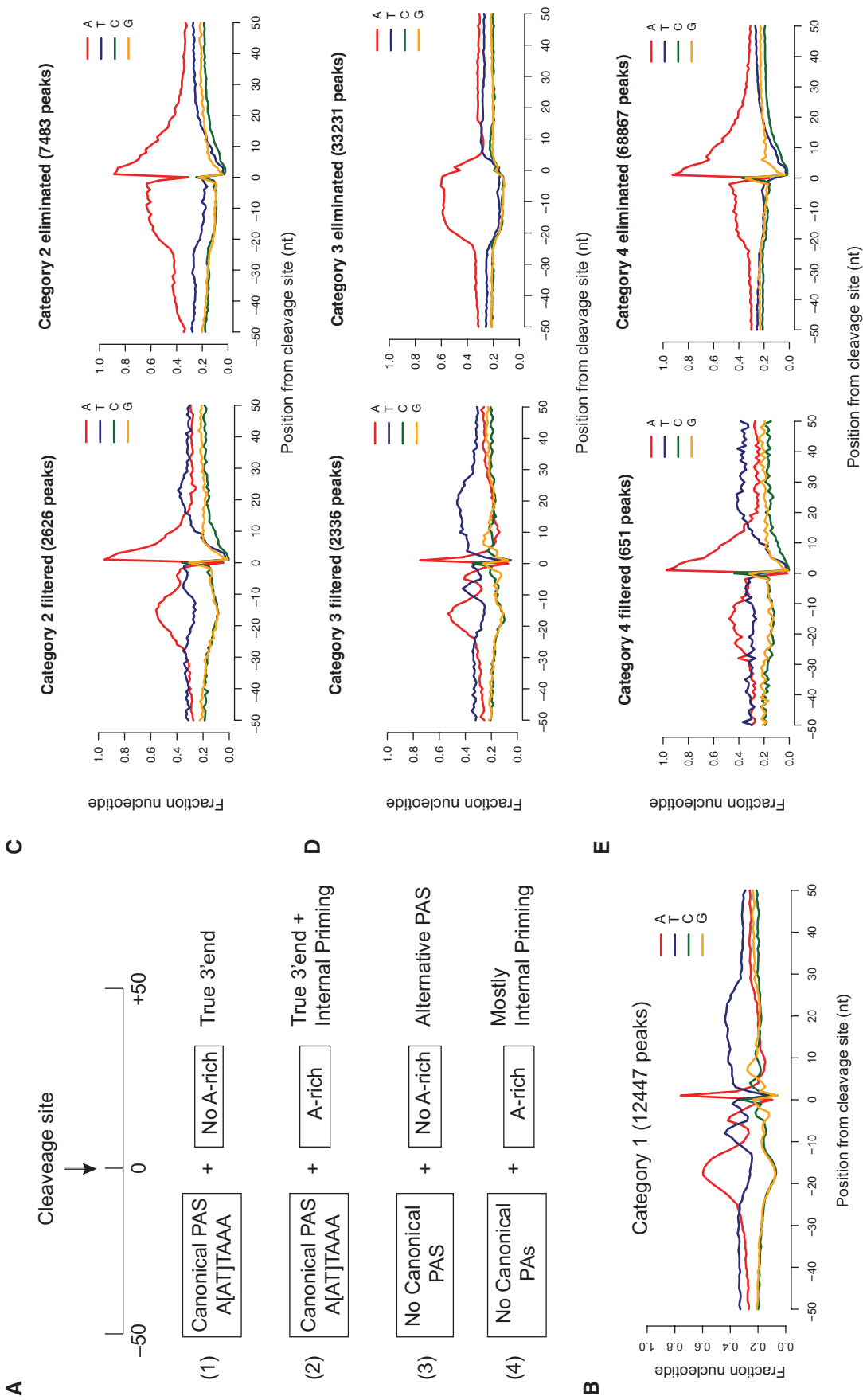


FIGURE 2. A bioinformatics pipeline distinguishes authentic 3'-end signals from internal priming events. (A) Classification of 3Seq data is based on the presence of polyadenylation signal in position -50 to 0 and downstream A-rich sequence in position 0 to +50. Interpretation of the four categories is noted on the right and discussed in detail in the text. (B-E) Nucleotide composition plot of category 1 in B, which is composed of true 3'-end signals and category 2 in C, category 3 in D, and category 4 in E. The fraction of nucleotide in a -50 to +50 window centered at the defined cleavage site is plotted. In C-E, each category is shown as a "filtered" panel, including all peaks that pass the filter, which is classified as the bona fide 3' end; and an "eliminated" panel, including all peaks that fail to pass the filter, which are classified as false signals.

mostly internal priming peaks and a few true 3'-end peaks that use alternative PAS.

To remove peaks that likely arise from sequencing noise, we first eliminated peaks from all four categories that have less than 10 reads covering the cleavage sites. Then, we analyzed nucleotide composition around the defined cleavage sites for peaks from all four categories. As expected, the nucleotide distribution among four categories was different from each other (Fig. 2). Notably, category 1 peaks showed the exact pattern for authentic 3' ends as observed in direct sequencing data (cf. Figs. 1A and 2B). We concluded that peaks in category 1 are derived from authentic 3'-end signals.

To analyze peaks in category 2 that contain both canonical PAS and downstream A-rich genomic sequences, we applied additional criteria to define high-confidence 3'-end signals. These filters were (1) positional distribution of PAS, e.g., at least one canonical PAS should be localized in position -10 to -30 upstream of the defined cleavage site; (2) no more than three canonical PAS should exist in the 50-nt window upstream of the cleavage site because multiple occurrence of PAS is likely due to interspersed non-A nucleotide in an A-rich sequence segment; (3) peaks should be mapped to either annotated 3'-UTR region (Refseq and Ensembl, with 10-kb extension in the 3' end) or intergenic regions. With these filters, we identified 2626 peaks in category 2 as genuine 3'-end signals and eliminated 7483 peaks. Strikingly, the nucleotide distribution of the filtered peaks closely resembled the pattern of true 3' ends as observed for category 1 peaks, whereas the eliminated peaks showed strong A-rich signals and lacked any specific PAS signals (Fig. 2C). This indicated that the eliminated peaks are characteristic of internal priming events.

We next analyzed peaks in category 3 with a focus on alternative PAS. We searched for enriched 10-nt motifs using MEME (Bailey et al. 2009) from position -1 to -50 sequences of category 3 peaks. Strikingly, our 10-nt motif search yielded results dominated by 6-nt motifs (Fig. 3A). We identified all enriched 6-nt motifs with $P < 1 \times 10^{-4}$ and selected 6-nt motifs that significantly enriched at the -10 to -30 position (Fig. 3A; Supplemental Fig. S4), consistent with the notion that the alternative PAS is also recognized and processed by the same polyadenylation machinery as the canonical PAS (Di Giammartino et al. 2011). Altogether, we identified 14 6-mer motifs as candidates for alternative PAS (Fig. 3B), the majority of which are consistent with previous studies (Beaudoing et al. 2000; Derti et al. 2012). We then used the presence of these 14 alternative PAS as a primary filter to screen for authentic 3'-end signals from category 3 ($P < 0.01$). We extracted 2336 true 3'-end peaks and eliminated 33,231 peaks from category 3. Although the eliminated peaks accounted for the majority of category 3 peaks, they showed significant enrichment for peaks with very low reads counts (less than 10 reads mapped to a peak), which is characteristic of sequencing/mapping noise (see below for detailed analysis). Importantly, when we plotted the nucleotide distribu-

tion of the filtered or eliminated peaks from category 3, only the pattern of the filtered peaks showed a strong resemblance to that of category 1, whereas the eliminated peaks showed a simple A-rich pattern without any other specific signals (Fig. 2D). These results further validated the robustness of our identification of the peaks that are derived from alternative PAS.

Since category 4 was composed of a large number of peaks with only A-rich sequence downstream from the cleavage site without canonical PAS upstream, we applied more stringent filters based on the presence of downstream U/GU-rich motifs for 3'-end identification ($P < 0.01$). In this manner, we identified 651 out of 68,841 peaks that were likely derived from authentic 3' ends. Importantly, when we plotted the nucleotide distribution surrounding the cleavage site for these 651 peaks, we again observed a pattern that strongly resembles that of category 1 peaks (Fig. 2E). As expected, the pattern of the eliminated peaks was consistent with internal priming events including prominent A-rich sequences downstream from the defined cleavage site and lacking any other specific signals (Fig. 2E).

Together, these results validated the effectiveness of our bioinformatics analysis to successfully distinguish authentic 3'-end signals from a large number of sequencing artifacts. Overall, we defined 18,060 3Seq peaks as mRNA 3' ends, corresponding to 12,739 genes (Supplemental Table S1). Although the eliminated peaks significantly outnumbered the retained peaks (86.4% of peaks are eliminated), the eliminated peaks only accounted for 57.5% of total mappable reads.

To further evaluate the performance of our bioinformatics pipeline, we applied our analysis to a previously published 3Seq data set generated from human U2OS cells (Jenal et al. 2012) and compared the results with the direct RNA sequencing (DRS) annotated 3' end generated from human liver (Ozsolak et al. 2010). Because these two data sets were not from the same RNA source, we only compared genes that were detected by both 3Seq (before filtering) and DRS (Supplemental Table S2). When a 3Seq peak intersected with a DRS-annotated 3'-end region, it was called a match. We defined successful classification as when the filtered peaks match the DRS annotation and the eliminated peaks do not match the DRS annotation, which were denoted as true-positive signals and true-negative signals, respectively. Among a total of 95,332 3Seq peaks, 10,891 peaks (11.4%) were true-positive signals, whereas 72,955 peaks (76.5%) were true-negative signals. This result indicated an 88% successful rate for our pipeline. 3Seq peaks that passed our filter but did not match the DRS annotation were classified as false positives, which constitute 3.6% of the total peaks. However, since these two data sets were derived from different cell types, false positives could also be derived from U2OS cell-type-specific gene expression. Finally, 8038 3Seq peaks (8.4%) were eliminated but had the DRS match and were therefore considered as false negatives. Despite a relative high proportion, however, 71.6% of the false-negative peaks had less than 10 reads covering the

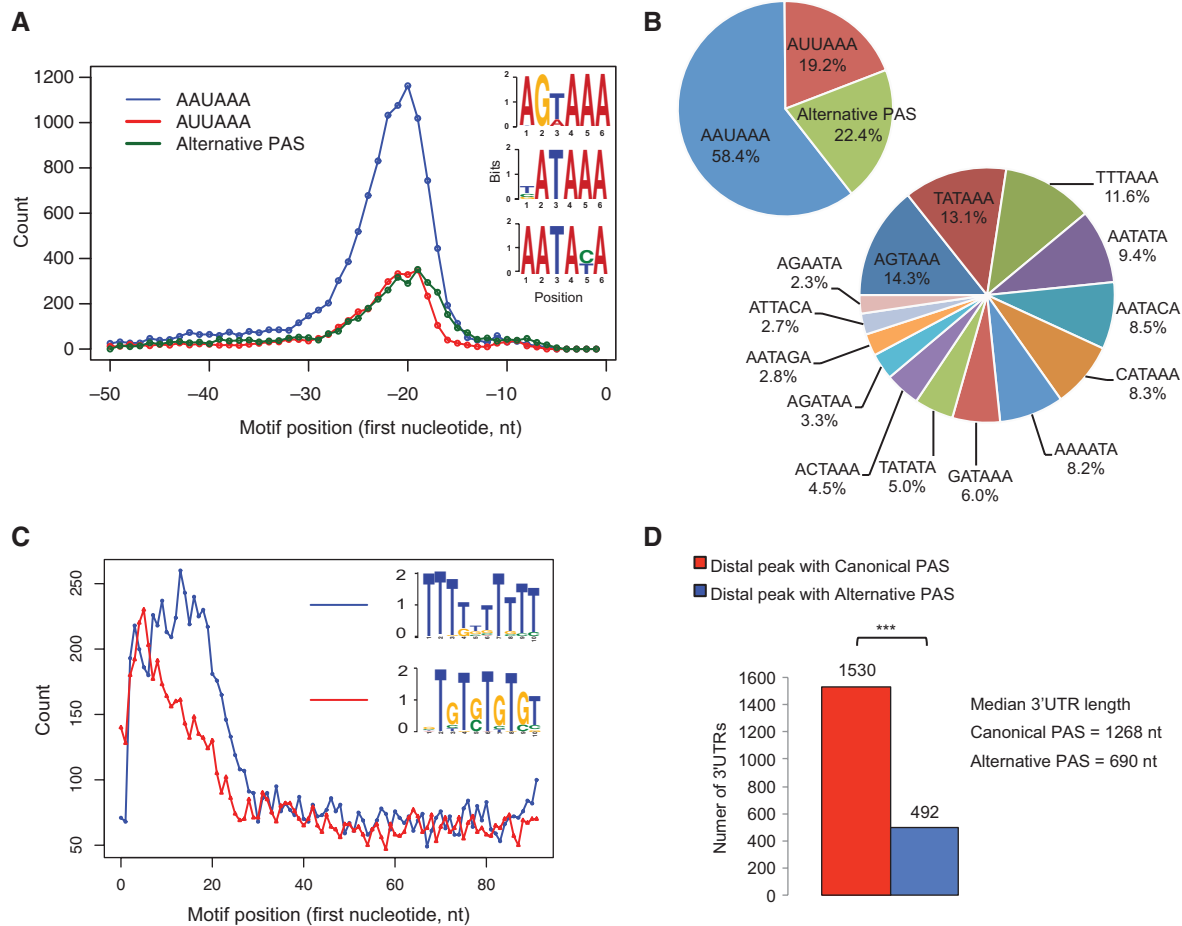


FIGURE 3. Characterization of sequence elements surrounding the polyadenylation site. (A) Positions of PAS are plotted in the -50 to 0 window upstream of the cleavage site. The position of the first nucleotide in the 6-mer motif is plotted on the x -axis. (Right) Three enriched 6-mer motifs identified in the -50 to 0 window upstream of the cleavage site. T is shown instead of U. (B) Genome-wide usage of PAS. All alternative PAS are displayed in detail in a subset of the pie chart. (C) Positions of the U-rich (blue line) and GU-rich motifs (red line) are identified in the 0 to $+100$ window downstream from the cleavage site; T is shown instead of U, as in A. (D) Preference for canonical and alternative PAS by distal and proximal 3' UTRs. The number of distal 3' ends that use canonical PAS (red) is significantly larger than that of the distal 3' ends that use alternative PAS (blue). (***) $P < 0.001$, two-sample Z-test.

cleavage site, which resulted in the exclusion of these peaks by our analysis (Supplemental Table S2). We expected that these peaks should be correctly classified with a deeper sequencing depth. Among the remaining false negatives (28.4% of the false negatives e.g., 2.4% of the all peaks), the leading causes for the false classification were (1) failure to detect alternative PAS (15.8% of the total false negatives) and (2) failure to define the cleavage site (12.6% of the total false negatives). Thus, with the sequencing depth and stringency of filtering, only 2.4% of 3Seq peaks were misclassified by our analysis.

Genome-wide analysis of polyadenylation signals and downstream sequence motifs in epidermal lineages

After successfully classifying 3Seq peaks and distinguishing authentic 3' ends of mRNAs from sequencing artifacts, we performed genome-wide analyses of the utilization of PAS

and other prominent motifs for 3'-end formation in the skin lineages. Canonical PAS, A[A/U]UAAA, were dominantly used, e.g., AAUAAA peaks count for 58.4% and AUUAAA peaks count for 19.2% of the total peaks, and only 22.4% of mRNAs use the alternative PAS for 3'-end formation (Fig. 3B). Taken together, these results indicate that the utilization of PAS is strongly biased toward canonical PAS in mouse.

Next, we analyzed the sequences that are downstream from the cleavage sites, where U/GU-rich sequences have been reported as Cstf binding sites (MacDonald et al. 1994; Beyer et al. 1997; Takagaki and Manley 1997). These U/GU-rich motifs have been proposed to be present within 50 nt downstream from the cleavage site, and their positional enrichment is less well defined compared with the PAS. We searched for 10-nt motifs that are enriched in position 0 to $+100$ using true 3'-end peaks in categories 1 and 3. We

identified a U-rich motif and a GU-rich motif as shown in Figure 3C. Furthermore, the U/GU-rich motifs also showed significant positional enrichment at 0 to +30 nt immediately downstream from the cleavage site (Fig. 3C). Overall, we observed a strong enrichment of these motifs in 3'-end formation when examining the presence of these motifs in category 1–3 peaks. For example, 90.8% of category 1 peaks, 80.4% of filtered category 2 peaks, and 91.4% of filtered category 3 peaks contained at least one U/GU-rich motif at the 0 to +100 position downstream from the cleavage site ($P < 0.01$). These results suggest that for most genes, both upstream PAS and downstream U/GU-rich motifs are used to direct 3'-end formation.

Because for many genes more than one 3' UTR is formed, we next asked whether there is a preference for the canonical or alternative PAS at proximal and distal ends. We identified 2022 genes using both canonical and alternative PAS. Interestingly, significantly more canonical PAS were used to generate longer 3' UTRs than alternative PAS. For example, 1530 distal ends (75.7%) were formed by using canonical PAS, whereas 492 distal ends (24.3%) used alternative PAS (Fig. 3D). The median length for 3' UTRs formed by the canonical PAS was 1268 nt, whereas it was 690 nt for the 3' UTRs formed by alternative PAS. Taken together, these results show a strong bias toward canonical PAS by the distal peaks.

Genome-wide quantification of the transcriptome

We next benchmarked the quantitative performance of 3Seq. We first chose a group of genes with known differential expression between E14 basal stem cells and suprabasal differentiating cells including basal cell markers, e.g., transcription factors (*Lef1* and *Trp63*), extracellular basement membrane gene ($\alpha 6$ *integrin*), and basal structural genes (*Krt14* and *Krt5*), as well as a Wnt receptor (*Fzd10*) and suprabasal markers, e.g., structural genes (*Loricrin*, *Krt1*, and *Krt10*) and a cell cycle inhibitor (*Cdkn1a*, also known as *p21*). We quantified their expression by counting reads mapped to each gene and normalizing to the total mappable reads from each library. Remarkably, the fold-change data determined by 3Seq measurement showed very robust correlation with the fold-change results obtained by qPCR for all of these genes (Pearson correlation, $r = 0.996$) (Fig. 4A). This result indicates that the 3Seq quantification is suitable for differential gene expression studies, and for individual genes its performance is comparable to RT-qPCR quantification.

In the genome-wide analysis, we noticed that the highest reads number for a single gene in each library was greater than 30,000. This observation suggested that with the overall sequencing depth at 8 million uniquely mappable reads, 3Seq could provide a quantitative dynamic range of up to 10^4 . To test this hypothesis, we first examined normalized reads count per million mappable reads (RPM) for 12,739 genes that showed detectable expression in the skin. The result showed that 3Seq can detect gene RPM over four orders of

magnitude in both basal and suprabasal cells (Fig. 4B). The distribution of mRNA expression levels is also consistent with recent studies that suggest that the dynamic range for mRNA copy number in a single mammalian cell is $\sim 10^4$ with the median expression of 17 copies per cell (Schwanhauser et al. 2011; Djebali et al. 2012). Because our analysis can simultaneously determine 3'-end formation and quantitative mRNA expression, we examined the correlation between 3'-end formation by either canonical PAS or alternative PAS and mRNA expression level. Intriguingly, mRNAs using A[A/U]UAAA showed considerably higher expression than mRNAs using alternative PAS in both basal stem cells and suprabasal differentiating cells in the global analysis (Fig. 4C). Furthermore, mRNAs using AUUAAA were generally expressed at a lower level than mRNAs using AAUAAA. These observations suggest: (1) canonical PAS may be processed more robustly than alternative PAS; or (2) highly expressed mRNAs (e.g., structural genes) may prefer canonical PAS. To support the second suggestion, we analyzed PAS composition for mouse housekeeping genes (Hsiao et al. 2001). As expected, 363 genes out of 419 3' UTRs of these housekeeping genes (86.6%) used A[A/U]UAAA as PAS, which was a significantly higher percentage than the utilization ratio of 77.6% for all 3' UTRs genome-wide ($P < 1 \times 10^{-4}$, hypergeometric test) (Supplemental data set S1).

To evaluate the accuracy of absolute copy number quantification by 3Seq, we used two different sets of in vitro cultured keratinocytes, which provided us an ample amount of total RNA, to construct 3Seq libraries and cDNAs for extensive quantification. With these two different keratinocytes samples, we uniquely mapped 1.6 million and 1.2 million 50-mer 3Seq reads, respectively. Despite the low sequencing depth of these experiments, we still observed the dynamic range of $\sim 10^4$. We selected 16 genes whose expression spanned the entire dynamic range for quantitative analysis. To ensure accurate copy number quantification by qPCR, we characterized amplification efficiency and determined the dynamic range of linear amplification for each qPCR primer set before calculating the copy number for each gene. Strikingly, when we compared the normalized 3Seq quantification for each gene with their absolute copy number, it showed a remarkable consistency between these two measurements. In the two independent experiments, Pearson's correlation coefficient is 0.8104 and 0.8227, respectively (Fig. 4D). In general, we also observed an average of 1.48-fold overestimation by 3Seq quantification, with a standard deviation of 0.79. This was probably due to the sequencing depth, e.g., only 1.6 million mappable reads and the short reads length (50-mer) for these experiments. We expect that with a depth of ~ 20 – 30 million total reads per library and longer reads, e.g., 100-mer instead of 50-mer, we could observe more robust absolute quantification by 3Seq. Nonetheless, these results demonstrated that 3Seq enables highly accurate quantification for mRNA throughout the entire dynamic range of mRNA expression.

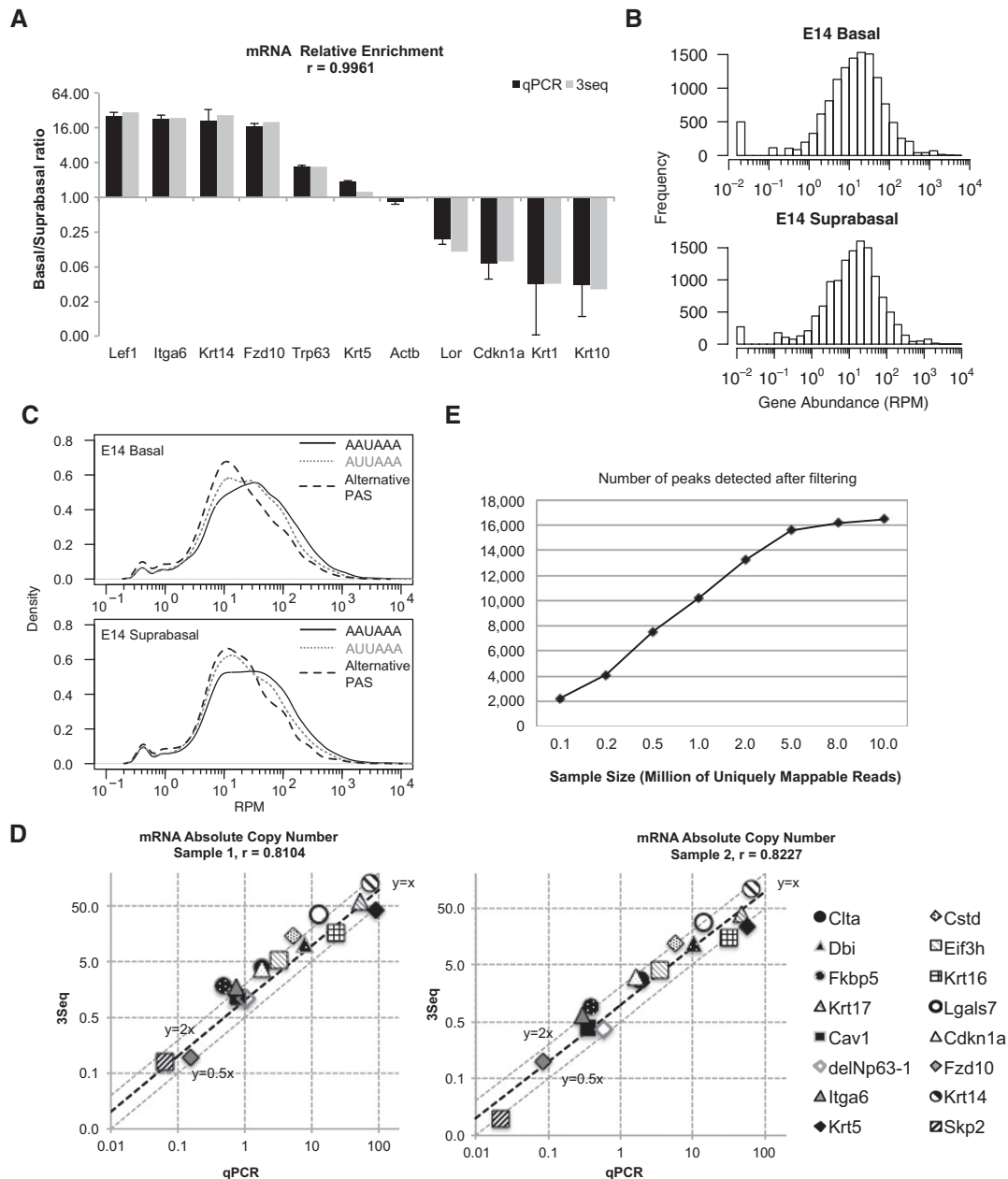


FIGURE 4. 3Seq provides accurate measurement for both relative and absolute mRNA quantification. (A) Measurement of mRNA relative abundance between E14 basal and suprabasal samples. The \log_2 enrichment ratio of basal/suprabasal for selected genes is plotted. Basal markers (*Lef1*, *Itga6*, *Krt14*, *Krt5*, *Fzd10*, and *Trp63*) are significantly enriched in basal cells, and suprabasal markers (*Lor*, *Cdkn1a*, *Krt1*, and *Krt10*) are enriched in suprabasal cells. Measurement by qPCR and 3Seq is highly consistent, with Pearson correlation coefficient, $r = 0.9961$. (B) \log_{10} histogram of 3Seq RPM measurement (read count per million mappable reads, RPM) for individual genes in embryonic day 14 (E14) basal and suprabasal samples. Note that *Hprt* has an RPM of 137.8 in the E14 basal sample and 124.2 in the E14 suprabasal sample. (C) Smoothed \log_{10} density plot of 3Seq RPM measurement for individual peaks that use canonical AAUAAA (black), AUUAAA (gray dashed), or alternative PAS (gray dotted). (D) \log_{10} ratio of absolute copy number for 16 genes as measured by qPCR is compared with the measurement by 3Seq (normalized to *Hprt*). Two individual experimental results are plotted. 3Seq and qPCR experiments are performed using in vitro cultured mouse primary keratinocytes. Measurement by qPCR and 3Seq are highly consistent, with Pearson correlation coefficient, $r = 0.8104$ and 0.8227 . (E) Impact of sequencing depth on 3'-end detection and quantification. Subsampling of the different size of mappable reads is used to perform 3Seq peak filtering (solid line).

We then evaluated the impact of sequencing depth on the quantification power of 3Seq. To this end, we performed subsampling from all of our E14 basal 3Seq data (11.2 million uniquely mappable reads). By subsampling from 0.1 million

to 10 million reads following the exact filtering analysis, we determined that with 5 million uniquely mappable reads, it was sufficient to detect 90% of 3'-end events and the detection of authentic 3' ends reached a plateau with 8–10 million

reads (Fig. 4E). Taken together, we estimated that with 8–10 million mapped reads, our experimental and bioinformatics pipeline represents a robust tool for genome-wide mRNA quantification. With the output of Illumina HiSeq routinely approaching 250 million reads per lane, we could profile a minimum of 8–10 samples with a single lane for the cost of ~\$1,000.

Differential 3'-UTR usage in the epidermal lineages

Transcriptome quantification by 3Seq enables the detection of differential 3'-UTR usage in a quantitative manner. For example, if gene A forms two 3'-UTR isoforms with a proximal and distal PAS, respectively, it would be important to determine if gene A prefers the proximal isoform to the distal isoform in one cell lineage versus the other. These insights may uncover a molecular basis for dynamic regulation of mRNA, e.g., stability, localization, and translation through differential 3'-UTR expression in different cell lineages. Indeed, it has been reported that differential 3'-UTR usage is associated with cell type specificity (Flavell et al. 2008; Shepard et al. 2011) and cell proliferation (Sandberg et al. 2008). However, differential 3'-UTR usage between closely related cell lineages in vivo has not been well characterized by sequencing. As a result, the dynamics of 3'-end formation during developmental transitions remains unclear. Since 3Seq was able to quantify abundance of each 3'-end isoform independently, we directly examined the data to identify 3'-UTR switching events. We analyzed 3Seq data for differential 3'-end formation between the embryonic day 14 (E14) basal and suprabasal cells.

Overall, 39.1% of genes contain more than one 3' end (Fig. 5A). We first asked whether differential PAS utilization is associated with 3'-UTR switching. We studied the PAS distribution in splicing-independent, differential 3'-end formation and observed a strong preference for the alternative PAS in these 3'-UTR switching events, compared with the genome-wide pattern ($P < 0.0001$) (Fig. 5B). Interestingly, increased preference for the alternative PAS was mainly balanced by the decreased usage of AAUAAA, but not AUUAAA (Fig. 5B). On the other hand, the usage of alterna-

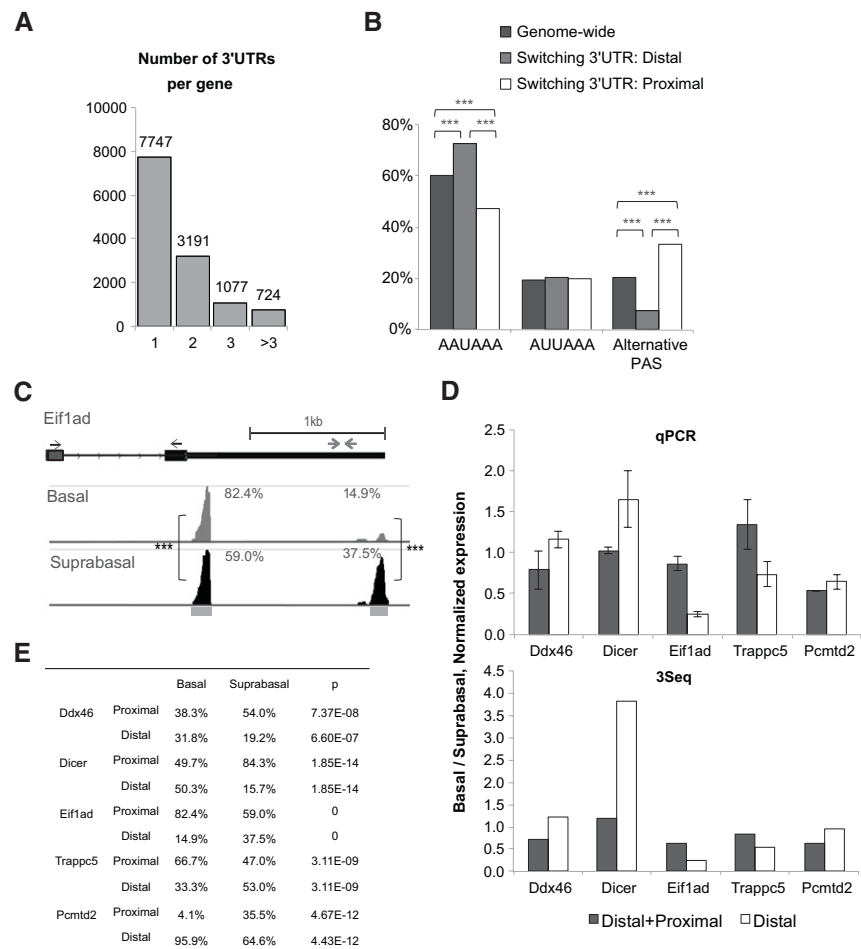


FIGURE 5. Differential 3'-UTR formation is identified in embryonic skin stem cell lineages. (A) Histogram of the number of 3' UTRs per gene. (B) Positional preference of PAS in splicing-independent 3'-UTR switching. The proportions of 3' UTRs containing AAUAAA, AUUAAA, or alternative PAS are plotted in three categories: genome-wide distribution (black), distal switching 3' UTRs (gray), and proximal switching 3' UTRs (white). (C) Examples of splicing-independent 3'-UTR switching of the *Eif1ad* gene. (Left) 3Seq coverage with RefSeq annotation is plotted. The positions of qPCR primers for detecting both proximal and distal 3'-end expression (gray) and distal 3'-end only (black) are shown (arrows). Gray bars below the coverage track indicate defined the 3Seq peak region. The proportions of each 3' end are labeled. (Right) qPCR validation and comparison with 3Seq measurements; (***) $P < 0.001$, (**) $P < 0.01$. (D) qPCR validation of switching 3' UTRs. A universal primer (black arrows in C) is used to detect both proximal and distal isoforms, while another pair of primers only measures the distal isoforms (gray arrows in C). The error bar is standard deviation from biological replicates. (E) Quantification of switching 3'-UTR isoforms. The proportion of each 3' UTR (proximal or distal) is calculated by dividing 3Seq quantification of each 3' end by the whole gene quantity. The P-value is calculated using a two-sample z-test, with multiple comparison correction (Benjamini and Hochberg).

tive PAS at the distal 3' ends for these 3'-UTR switching events is drastically reduced (7.1%) in contrast to the genome-wide usage rate (20.3%). To systematically identify 3'-UTR switching events among these genes, we focused on the proportion of each 3'-end isoform for a single transcript. Overall, we identified a total of 1228 3'-end isoforms from splicing-independent, authentic alternative polyadenylation that are differentially enriched between the basal and suprabasal lineages ($P < 0.05$, Benjamini and Hochberg correction). These 3' ends belong to 829 genes (Supplemental data set S2).

To validate these 3'-UTR switching events experimentally, we selected six genes (*Ddx46*, *Dicer*, *Eif1ad*, *Trappc5*, *Pcmdt2*, and *Opa1*) that showed differential usage of 3' ends. We performed qPCR to validate the detection by 3Seq. After validating qPCR primers and expression variation between cell sorting experiments (Fig. 5C), six out of seven genes showed the same differential usage of the distal 3' ends in the qPCR analysis, consistent with the results obtained by the 3Seq analysis (Fig. 5D,E). This result further validated the ability of 3Seq in quantitative analysis for individual 3' ends.

To extend our study to more remotely related cell types, we compared the E14 embryonic skin lineages with the postnatal day 4 (P4) hair follicle lineages, which originated from the E14 basal stem cells (Blanpain and Fuchs 2009), for 3'-UTR switching events. We identified 1394 3' ends that switched independently of splicing between these two lineages, corresponding to 933 genes (Supplemental data set S2). Therefore, 3'-UTR switching was slightly more widespread between the E14 and P4 lineages than that between the E14 basal and suprabasal lineages. We applied GO term analysis by using Gene Set Enrichment Analysis (Subramanian et al. 2005) for the genes with switched 3' UTRs. However, we did not observe any strong enrichment for any particular functional cluster with a stringent *P*-value cutoff (1×10^{-4}). This result suggests that differential 3'-end formation takes place in many genes with diverse functions without any particular functional enrichment.

3'-End formation patterns for genes in the miRNA pathway

The presence of a very short 3' UTR (154 nt) and a long 3' UTR (3850 nt) in *Dicer* suggested a possibility of regulating *Dicer* mRNA by the length of the 3' UTR. Indeed, a previous study suggested that shortening of the *Dicer* 3' UTR in tumor cells provides a mechanism to stabilize *Dicer* mRNA and promote tumor development (Mayr and Bartel 2009). Interestingly, *Dicer* mRNA also showed a dramatic 3'-UTR shortening during epidermal differentiation (Fig. 6A). To test whether the differential 3'-UTR usage is a general pattern for genes involved in the miRNA biogenesis pathway, we examined mRNAs for core components of the miRNA pathway (Krol et al. 2010). Interestingly, we found extreme cases in which a very long 3' UTR is used (*Ago2*, 11 kb) or a very short 3' UTR is used (*Drosha*, 123 nt) (Fig. 6B). Seven genes in addition to *Dicer* (*Dgcr8*, *Xpo5*, *Ago1*, *Ago2*, *Ago4*, *Tnrc6a*, *Tnrc6b*) showed the

"short-long" 3'-UTR pattern, whereas *Ago3* and *Tnrc6c* each showed a single 3' end (Fig. 6C). The proportion of genes in the miRNA biogenesis pathway containing multiple 3' UTRs (eight out of 11, 72.7%) is significantly higher than the genome-wide percentage (39.1%, $P < 0.01$, two-sample *Z*-test). Interestingly, the remarkable heterogeneity of 3'-UTR formation for genes involved in the miRNA biogenesis pathway was also observed in human cells when we examined the direct sequencing results (Supplemental Table S3). These 3'-UTR dynamics suggested a possibility that conserved and extensive feedback regulation exists to modulate the expression of these genes. For example, mRNAs with very short 3' UTRs, e.g., *Drosha* and the short isoforms of *Xpo5* and *Dicer*, could be largely immune to miRNA-mediated regulation, whereas mRNAs with very long 3' UTRs, e.g., *Ago2* and the long isoforms of *Xpo5*, *Dicer*, and *Dgcr8*, could be sensitive to extensive miRNA-mediated regulation.

DISCUSSION

We have optimized the 3Seq technique with a bioinformatics pipeline to identify accurately and quantitatively the 3' end of mRNA. By leveraging the local sequence composition characteristic of mRNA 3'-end formation, we can accurately distinguish "false" peaks derived from internal priming from

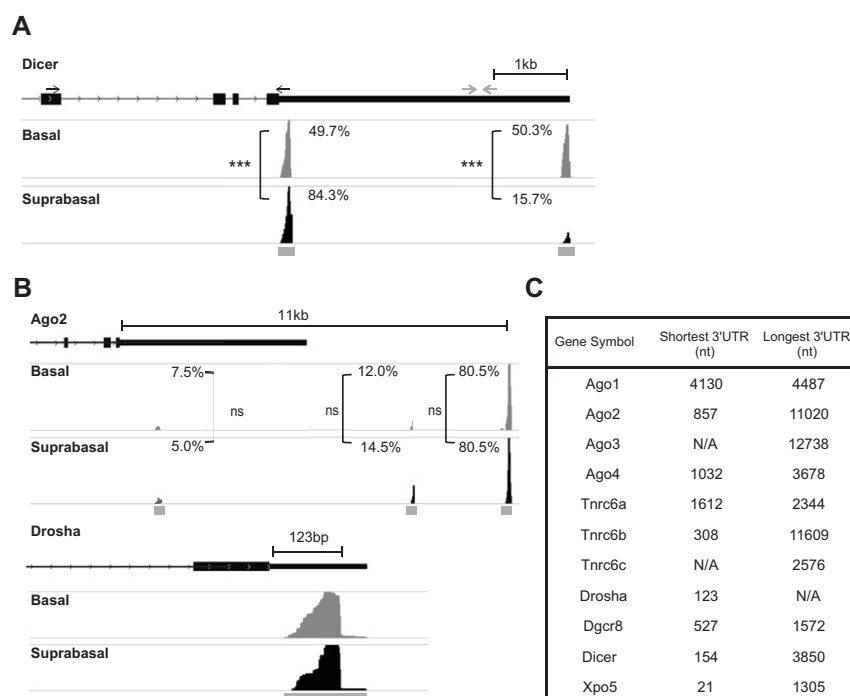


FIGURE 6. Dynamics of 3'-end formation for core components in miRNA biogenesis pathways. (A) 3'-UTR switching between short (154 nt) and long (3850 nt) isoforms of *Dicer* is detected between the stem cells and differentiated cells. 3Seq coverage with RefSeq annotation is plotted. (B) Long 3' UTRs of *Ago2* (8314 nt and 11,020 nt) and a very short 3' UTR of *Drosha* (123 nt) are detected in the skin. Gray bars below the coverage track indicate the defined 3Seq peak region, and proportions of each 3' end are labeled in both A and B. (C) Table of shortest and longest 3' UTRs for genes in the miRNA biogenesis pathway. (***) $P < 0.001$, (ns) not significant.

authentic poly(A) sites. The accurate annotation of 3'-end formation for specific cell populations will provide precise 3'-UTR information for studies focusing on the regulatory function of the 3' UTR, e.g., cell-type-specific identification of miRNA targets and characterization of cell-type-specific binding motifs for RNA-binding proteins.

Our results indicate that the quantification of mRNA levels by global 3Seq analysis has comparable accuracy as qPCR for differential expression of individual genes. In addition, 3Seq quantification shows robust correlation with the absolute copy number of individual genes over the dynamic range of $\sim 10^4$. This accuracy was obtained by less than 3 million uniquely mappable reads (~ 10 million raw reads). It should be noted that the proportion of mappable reads could be further increased by sequencing longer fragments or using a paired-end strategy. Considering that the current yield for the Illumina HiSeq platform is approaching 250 million reads per lane, our 3Seq protocol would enable multiplexing of more than 10 samples per lane while retaining quantification accuracy. Practically, this significantly lowers the cost of transcriptome profiling. In particular, 3Seq is suitable for transcriptome profiling in gene functional studies in which the main goal is usually global measurement of transcripts' abundance, and detection of full-length transcripts is not required.

The simultaneous detection of 3' ends and quantification of mRNA expression gives us a unique opportunity to examine the dynamics of 3'-end formation in a quantitative manner and reveals several hitherto unappreciated insights. For example, we show that mRNAs using canonical PAS, especially AAUAAA, tend to express at a considerably higher level than mRNAs using alternative PAS (Fig. 4C). Furthermore, in 3'-UTR switching events, the distal 3' ends are strongly biased toward AAUAAA, whereas the proximal 3' ends are biased toward alternative PAS (Fig. 5B). These observations suggest an intrinsic relationship between polyadenylation and gene expression. Because of the simplicity and cost-effectiveness of 3Seq, we anticipate that 3Seq, when combined with robust bioinformatics analysis, can be widely applied in numerous studies for gene expression.

The quantitative profiling of polyadenylated RNAs in skin stem cell lineages yields important insights into the dynamics of mRNA 3'-end formation in closely related somatic cells in vivo. Among 12,739 genes that are detected in our analysis, 4992 of them (39.1%) contain more than one 3' end. Interestingly, this percentage is comparable to the proportion of mRNAs with alternative 3' ends that has been detected in *C. elegans* (30%) by 3P-Seq (Jan et al. 2011). Of note, our study examined the alternative 3' UTRs in a single mammalian lineage at a specific time point, whereas the *C. elegans* study used the whole organism at different developmental stages. This result indicates that a large number of genes use different 3' UTRs to control their mRNAs, e.g., stability and location as well as translation. We were particularly interested in detecting mRNA transcripts that differentially use

distinct 3'-UTR isoforms between the basal stem cells and the suprabasal differentiating cells. It could provide critical insights for the role of differential 3'-UTR usage during developmental transitions. Surprisingly, we only detected 829 genes that show differential 3'-end usage between these two lineages. It indicates that the impact of mRNA 3'-end switch could be limited during developmental transitions. However, we could not rule out the possibility that a few master regulators, which use differential 3'-end formation, could have a significant impact on the process. For example, the 3' UTR of *Dicer* shows one of the most dramatic shortenings once the stem cells embark on the differentiation program (Fig. 6A). In contrast to our findings, several previous studies suggest that differential 3'-end formation might be a widespread mechanism for cell fate specification (Tian et al. 2005; Sandberg et al. 2008; Mayr and Bartel 2009). However, these studies used very different samples for such comparison, e.g., (1) a cohort of studies from a large number of cell lines (Tian et al. 2005); (2) in vitro cultured tumor cells (Mayr and Bartel 2009; Fu et al. 2011; Shepard et al. 2011); or (3) a mixed cell population isolated from a large tissue or organ (Ozsolak et al. 2010). As a result, those samples were either heterogeneous or highly transformed, which may not be able to reflect transcriptome diversity accurately in a single lineage in vivo. Unlike these cells, the basal stem cells and suprabasal differentiating cells are spatiotemporally well defined and closely linked with a much shorter distance in their developmental timeline. To examine if an increased developmental gap could lead to a more dramatic change in 3'-end usage, we also compared embryonic skin stem cells with neonatal hair follicle lineages, which originate from the stem cells. We found that 933 genes show 3'-end switching among these two cell types (Supplemental data set S2). Thus, our results indicate that differential usage of alternative polyadenylation could be either a specific event for a certain type of developmental transition and cell transformation or a process that gradually takes place through in vivo differentiation.

Finally, our results reveal a complex 3'-end formation pattern for core components of the miRNA pathway (Fig. 6). Eight out of 11 core components of the pathway contain a short and long 3'-UTR isoforms. In particular, the short forms of *Dicer* and *Xpo5* and the only 3' UTR of *Drosha* are among the shortest 3' UTRs, e.g., 154 nt, 21 nt, and 123 nt, respectively, and are likely refractory to any miRNA-mediated regulation. In contrast, the long forms of *Dicer* and *Xpo5* are 3850 nt and 1305 nt, respectively. It is tempting to speculate that the mRNAs with a short 3' UTR provide a "constant" output for these proteins, whereas the mRNAs with a long 3' UTR provide a "regulated" output that is negatively correlated with the strength of the miRNA pathway. In our recent study, we have demonstrated that the function of the core components of the miRNA pathway e.g., Ago proteins, shows strong correlation to their expression level (Wang et al. 2012). Therefore, the control over the expression of these

genes with the short-long 3' UTRs through miRNA-mediated regulation can provide a self-regulatory mechanism for the overall function of the miRNA pathway.

MATERIALS AND METHODS

Flow cytometry

Mouse embryonic epidermis is dissected from E14 embryo, cut into small pieces (2 mm × 2 mm), and trypsinized for 10 min at 37°C. Digested sample is resuspended with PBS and 3% chalexed serum and filtered to get a single-cell suspension. To isolated basal and suprabasal populations from embryonic epidermis, transgenic mouse line K14H2BGFP is used, and the single-cell suspension is further stained with rat anti-human CD49f PE antibody (BD Pharmingen) and then undergoes flow cytometry (Supplemental Fig. S1B).

Cell culture, RNA extraction, and cDNA construction and quantitative PCR

Primary keratinocytes are isolated and cultured as previously described (Yi et al. 2008). RNA from in vitro cultured cells or in vivo sorted cells are extracted by TRIzol and precipitated with isopropanol. cDNA construction is performed using the Invitrogen SuperScript III cDNA Construction Kit. Quantitative PCR is performed using the Bio-Rad SYBR Green system. For absolute copy number qPCR, PCR amplicons of candidate genes are gel-purified, and concentration is measured. A standard curve of each amplicon is determined by performing a series dilution of the amplicon and calculating the linear relationship of the Ct values and copy number input per reaction. Copy number values are normalized to internal control *Hprt*. The qPCR primers are listed in Supplemental Table S4.

Chromatin immunoprecipitation

ChIP experiments were performed using the ChIP-IT Express kit from Active Motif (catalog #53008), following the manufacturer's suggested procedure. H3K4me3 antibody was purchased from Active Motif (catalog #39159), and H3K36me3 antibody was purchased from Abcam (catalog #ab9050).

Sequencing library preparation

For the 3Seq experiment, 0.5–2 µg of total RNA are poly(A) selected twice using the Invitrogen Dynabeads mRNA DIRECT Kit. The poly(A) RNA is fragmented for 5 min using the Ambion RNA Fragmentation Kit. Reverse transcription is performed using fragmented RNA and an anchored oligo, P7T20V (Supplemental Methods), and second strands are further synthesized using RNase H and DNA polymerase I. The double-strand DNA library was constructed following the Illumina standard dsDNA library protocol. Briefly, the dsDNA ends were repaired using PNK and Klenow enzyme, followed by treatment with Klenow 3'-to-5' exo⁻ to generate an A-overhang used for adaptor ligation. Ligated products are gel-purified with size selection of 150–400 bp and PCR-amplified by 18–26 cycles, depending on the resulting library quantity. Amplified libraries were sequenced at the Illumina HiSeq2000 platform. For ChIP-

Seq experiments, ChIP-ed DNA was purified and followed by the Illumina standard dsDNA library protocol described above.

Bioinformatics analysis

Read mapping is performed using the short reads aligner Bowtie (version 0.12.7) (Langmead et al. 2009). We trim off reads containing adapter sequences and all continuous adenosines from the 3' end, and then align against the mouse reference genome. Aligned reads cluster to form peaks. To define the reads-enriched region, we use the peak calling algorithm MACS (version 1.4.0) (Zhang et al. 2008). Peaks are further processed using PeakSplitter (version: 0.1, embedded in MACS), which separates very close peaks, and thus called a single peak by MACS. The cleavage site of a given peak is determined using all trimmed reads within the peak. The positions of 3'-terminal nucleotides of all trimmed reads are intersected with 3Seq peak regions. The nucleotide position that has the maximum count of the 3' terminal of trimmed reads is defined as the cleavage site. See the detailed bioinformatics pipeline in the Supplemental Methods.

mRNA quantification and switching 3'-UTR detection

Peaks classified as authentic 3' ends are used for transcript quantification. We use Refseq and Ensembl annotation to assign peaks to gene symbols (database downloaded on 4 April 2012). Peaks within 10 kb downstream from an annotated 3' end and not overlapping with any annotated gene body are assigned as an extended 3' UTR. The expression level for a given gene is calculated by summing the reads count of all authentic 3'-end peaks mapped to the gene normalized to the total number of mappable reads (millions). For absolute quantification comparison with qPCR, 3Seq quantification is further normalized to internal control *Hprt*. Peak lengths are not normalized. Lengths of 3' UTRs are reannotated using the distance between the stop codon and the cleavage site of 3' seq peaks, with intron size subtraction. Genes with more than two major 3' UTRs (quantity is larger than 10% of total gene quantity) are candidates for switching 3'-UTR detection. The *P*-value is calculated using a two-sample Z-test with multiple comparison correction (Benjamini and Hochberg). Switching 3' UTRs are defined as at least 1.5-fold difference of proportion with *P* < 0.05.

DATA DEPOSITION

3Seq raw read sequences and processed data of E14 basal, E14 suprabasal, P4 hair follicle, and in vitro cultured keratinocyte samples are available at NCBI/GEO (study GSE37641).

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank the members of the Yi and Dowell laboratory for discussion; L. Greiner for maintaining the mice; Y. Han for flow cytometry and cell sorting; J. Huntley, J. Castoe, and J. Dover for Illumina Sequencing; E. Fuchs (HHMI, Rockefeller University) for providing

K14-H2BGFP mice; and S. Jackson for preparing RNA samples from in vitro cultured keratinocytes. This publication was made possible by Grant Numbers R00AR054704 and R01AR059697, a seed grant from the Butcher program (to R.Y.), start-up funds provided by the University of Colorado (to R.Y. and R.D.D.), and by funds from the Boettcher Foundation's Webb-Waring Biomedical Research Program (to R.D.D.) and by NIH Predoctoral Training Grant T32 GM08759 (to L.W.).

Received July 3, 2012; accepted November 27, 2012.

REFERENCES

- Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202–W208.
- Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**: 823–837.
- Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. 2000. Patterns of variant polyadenylation signal usage in human genes. *Genome Res* **10**: 1001–1010.
- Beck AH, Weng Z, Witten DM, Zhu S, Foley JW, Lacroute P, Smith CL, Tibshirani R, van de Rijn M, Sidow A, et al. 2010. 3'-End sequencing for expression quantification (3SEQ) from archival tumor samples. *PLoS One* **5**: e8768.
- Beyer K, Dandekar T, Keller W. 1997. RNA ligands selected by cleavage stimulation factor contain distinct sequence motifs that function as downstream elements in 3'-end processing of pre-mRNA. *J Biol Chem* **272**: 26769–26779.
- Blanpain C, Fuchs E. 2009. Epidermal homeostasis: A balancing act of stem cells in the skin. *Nat Rev Mol Cell Biol* **10**: 207–217.
- Derti A, Garrett-Engel P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**: 1173–1183.
- Di Giammartino DC, Nishida K, Manley JL. 2011. Mechanisms and consequences of alternative polyadenylation. *Mol Cell* **43**: 853–866.
- Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* **489**: 101–108.
- Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**: 817–825.
- Flavell SW, Kim T-K, Gray JM, Harmin DA, Hemberg M, Hong EJ, Markenscoff-Papadimitriou E, Bear DM, Greenberg ME. 2008. Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection. *Neuron* **60**: 1022–1038.
- Fu Y, Sun Y, Li Y, Li J, Rao X, Chen C, Xu A. 2011. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res* **21**: 741–747.
- Hafner M, Renwick N, Brown M, Mihailovic A, Holoch D, Lin C, Pena JT, Nusbaum JD, Morozov P, Ludwig J, et al. 2011. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* **17**: 1697–1712.
- Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, Lee KF, Clark KE, Haverly P, et al. 2001. A compendium of gene expression in normal human tissues. *Physiol Genomics* **7**: 97–104.
- Jan CH, Friedman RC, Ruby JG, Bartel DP. 2011. Formation, regulation and evolution of *Caenorhabditis elegans* 3'UTRs. *Nature* **469**: 97–101.
- Jenal M, Elkon R, Loayza-Puch F, van Haaften G, Kuhn U, Menzies FM, Vrielink JA, Bos AJ, Drost J, Rooijers K, et al. 2012. The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. *Cell* **149**: 538–553.
- Krol J, Loedige I, Filipowicz W. 2010. The widespread regulation of microRNA biogenesis, function and decay. *Nat Rev Genet* **11**: 597–610.
- Langmead B, Trapnell C, Pop M, Salzberg S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.
- Lechler T, Fuchs E. 2005. Asymmetric cell divisions promote stratification and differentiation of mammalian skin. *Nature* **437**: 275–280.
- Lin Y, Li Z, Ozsolak F, Kim SW, Arango-Argoty G, Liu TT, Tenenbaum SA, Bailey T, Monaghan AP, Milos PM, et al. 2012. An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Res* **40**: 8460–8471.
- MacDonald CC, Wilusz J, Shenk T. 1994. The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. *Mol Cell Biol* **14**: 6647–6654.
- Mayr C, Bartel DP. 2009. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**: 673–684.
- Ozsolak F, Platt AR, Jones DR, Reifemberger JG, Sass LE, McInerney P, Thompson JF, Bowers J, Jarosz M, Milos PM. 2009. Direct RNA sequencing. *Nature* **461**: 814–818.
- Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM. 2010. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143**: 1018–1029.
- Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. 2008. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* **320**: 1643–1647.
- Schwanhauser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. 2011. Global quantification of mammalian gene expression control. *Nature* **473**: 337–342.
- Shepard PJ, Choi E-A, Lu J, Flanagan LA, Hertel KJ, Shi Y. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**: 761–772.
- Smibert P, Miura P, Westholm JO, Shenker S, May G, Duff MO, Zhang D, Eads BD, Carlson J, Brown JB, et al. 2012. Global patterns of tissue-specific alternative polyadenylation in *Drosophila*. *Cell Rep* **1**: 1–13.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* **102**: 15545–15550.
- Takagaki Y, Manley JL. 1997. RNA recognition by the human polyadenylation factor CstF. *Mol Cell Biol* **17**: 3907–3914.
- Tian B, Hu J, Zhang H, Lutz CS. 2005. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* **33**: 201–212.
- Ulitsky I, Shkumatava A, Jan CH, Subtelny AO, Koppstein D, Bell GW, Sive H, Bartel DP. 2012. Extensive alternative polyadenylation during zebrafish development. *Genome Res* **22**: 2054–2066.
- Wang D, Zhang Z, O'Loughlin E, Lee T, Houel S, O'Carroll D, Tarakhovsky A, Ahn NG, Yi R. 2012. Quantitative functions of Argonaute proteins in mammalian development. *Genes Dev* **26**: 693–704.
- Yi R, Poy M, Stoffel M, Fuchs E. 2008. A skin microRNA promotes differentiation by repressing 'stemness'. *Nature* **452**: 225–229.
- Zhang Y, Liu T, Meyer C, Eickhout J, Johnson D, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.



RNA

A PUBLICATION OF THE RNA SOCIETY

Genome-wide maps of polyadenylation reveal dynamic mRNA 3'-end formation in mammalian cell lineages

Li Wang, Robin D. Dowell and Rui Yi

RNA 2013 19: 413-425 originally published online January 16, 2013

Access the most recent version at doi:[10.1261/rna.035360.112](https://doi.org/10.1261/rna.035360.112)

Supplemental Material

<http://rnajournal.cshlp.org/content/suppl/2012/12/27/rna.035360.112.DC1>

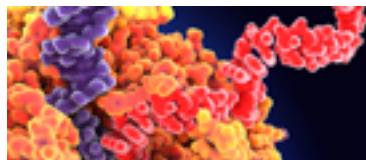
References

This article cites 35 articles, 12 of which can be accessed free at:
<http://rnajournal.cshlp.org/content/19/3/413.full.html#ref-list-1>


License

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



Use CRISPRmod for targeted modulation of endogenous gene expression to validate siRNA data



To subscribe to *RNA* go to:
<http://rnajournal.cshlp.org/subscriptions>
