PAT-seq: a method to study the integration of 3′-UTR dynamics with gene expression in the eukaryotic transcriptome

PAUL F. HARRISON,[1,2,6] DAVID R. POWELL,[1,2,6] JENNIFER L. CLANCY,[3] THOMAS PREISS,[3,4] PETER R. BOAG,[5] ANA TRAVEN,[5] TORSTEN SEEMANN,[1,2] and TRAUDE H. BEILHARZ[5]

[1]Victorian Bioinformatics Consortium, Monash University, Clayton 3800, Australia
[2]Life Sciences Computation Centre, Victorian Life Sciences Computation Initiative, Carlton 3053, Australia
[3]EMBL–Australia Collaborating Laboratory, Genome Biology Department, The John Curtin School of Medical Research (JCSMR), The Australian National University, Acton (Canberra) 2601, Australian Capital Territory, Australia
[4]Victor Chang Cardiac Research Institute, Darlinghurst (Sydney), New South Wales 2010, Australia
[5]Department of Biochemistry and Molecular Biology, Monash University, Clayton 3800, Australia
[6]Monash Bioinformatics Platform, Monash University, Clayton 3800, Australia

ABSTRACT

A major objective of systems biology is to quantitatively integrate multiple parameters from genome-wide measurements. To integrate gene expression with dynamics in poly(A) tail length and adenylation site, we developed a targeted next-generation sequencing approach, Poly(A)-Test RNA-sequencing. PAT-seq returns (i) digital gene expression, (ii) polyadenylation site/s, and (iii) the polyadenylation-state within and between eukaryotic transcriptomes. PAT-seq differs from previous 3′ focused RNA-seq methods in that it depends strictly on 3′ adenylation within total RNA samples and that the full-native poly(A) tail is included in the sequencing libraries. Here, total RNA samples from budding yeast cells were analyzed to identify the intersect between adenylation state and gene expression in response to loss of the major cytoplasmic deadenylase Ccr4. Furthermore, concordant changes to gene expression and adenylation-state were demonstrated in the classic Crabtree–Warburg metabolic shift. Because all polyadenylated RNA is interrogated by the approach, alternative adenylation sites, noncoding RNA and RNA-decay intermediates were also identified. Most important, the PAT-seq approach uses standard sequencing procedures, supports significant multiplexing, and thus replication and rigorous statistical analyses can for the first time be brought to the measure of 3′-UTR dynamics genome wide.

Keywords: RNA-seq; gene expression; polyadenylation; Ccr4; alternative polyadenylation; translational control; ePAT

INTRODUCTION

There are multiple points of regulation between mRNA transcription and translation by cytoplasmic ribosomes. Most of these have been selectively interrogated by high-throughput sequencing technologies to capture snapshots of system-wide control of gene expression. The convenient "hook" provided by the poly(A) tail on the vast majority of mRNA has also given rise to digital gene expression approaches that use 3′ focused sequencing based around SAGE (Velculescu et al. 1995) as a means to quantify the composition of the transcriptome (Ruzanov and Riddle 2010; Wu et al. 2010; Hong et al. 2011). Such approaches provide inexpensive and relatively simple tools to monitor eukaryotic gene expression. Moreover, the recent realization that condition-dependent alternative 3′-UTR cleavage and polyadenylation is common in eukaryotes, and can radically alter mRNA metabolism (Sandberg et al. 2008; Mayr and Bartel 2009; Di Giammartino et al. 2011), has led to further approaches to identify the frequency and position of alternative mRNA ends (Beck et al. 2010; Mangone et al. 2010; Ozsolak et al. 2010; Yoon and Brem 2010; Fu et al. 2011; Jan et al. 2011; Shepard et al. 2011; Ulitsky et al. 2012; Wilkening et al. 2013).

The poly(A) tail is more than just a convenient purification hook however; polyadenylation of protein-coding RNA is essential for eukaryotic life and normal protein translation. The length to which the poly(A) tail is extended after transcript cleavage is regulated; typically ~90 residues in yeast and ~300 residues in mammals. The exact length distribution at steady state, however, can reflect a number of metabolic activities that include normal transcript ageing, deadenylation associated with transcript silencing, and activation of

Corresponding author: traude.beilharz@monash.edu
Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.048355.114. Freely available online through the *RNA* Open Access option.

translation by cytoplasmic re-adenylation. Each of these processes is associated with specific disease states. For example, inappropriate cytoplasmic adenylation is found in cancer (Ortiz-Zapater et al. 2012), induced target deadenylation is associated with microRNA-mediated repression (Beilharz et al. 2009; Eulalio et al. 2009; Fabian et al. 2009), and mutations that result in hyperadenylation and nuclear retention of mRNA can cause intellectual disability (Pak et al. 2011). On the other hand, addition of a short poly(A) tail is also utilized by the RNA exosome during RNA decay, and thus many decay intermediates as well as non-coding transcripts are terminated by a short poly(A) tail (Wyers et al. 2005; Slomovic et al. 2010). Finally, widespread stutter activity of RNA Pol II surrounding transcriptional start and termination sites (Kapranov et al. 2010; Wei et al. 2011) forms a further source of adenylated RNA in the cell.

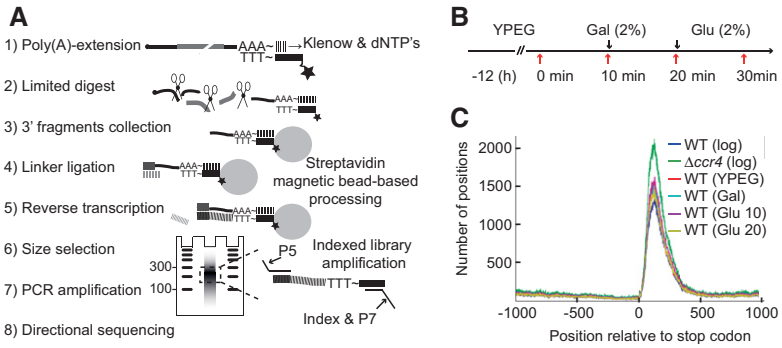

FIGURE 1. Poly(A)-Test sequencing. (*A*) Schematic representation of the PAT-seq approach. (*B*) Schematic of the experimental approach for the Crabtree Warburg metabolic shift of yeast cells transitioning from respiratory to fermentative growth. Red arrows indicate times of cell harvest; YPEG, Gal, and Glu refer to ethanol/glycerol, galactose, and glucose as carbon source. (*C*) The position of each adenylation site relative to the annotated transcript stop codon (0). Note the peak position for adenylation sites is ~100 bases after the stop. The increased number of positions in the Δ*ccr4* sample derives from loci that are silent in the wild-type strain.

Here we harness the efficiency of Klenow-mediated 3′ tagging (Janicke et al. 2012) to measure the dynamics of the adenylated transcriptome. The PAT-seq [for Poly(A)-Test RNA-sequencing] approach depends on the initially counter-intuitive notion of including the poly(A) tail in 3′ focused RNA-seq libraries. The potential loss of fidelity within homopolymers is avoided by directional sequencing from the 5′ end of fragments. We show here that this approach can provide an efficient method for the measure of 3′-UTR dynamics. Using just 1 µg of total RNA from each of 13 biological samples for library preparation, and multiplexed over a single lane of Illumina Hiseq sequencing, the PAT-seq approach accurately detected statistically significant changes in poly(A) tail-length distribution, reported digital gene expression, and clearly identified polyadenylation-site usage within and between transcriptomes.

RESULTS AND DISCUSSION

The PAT-seq methodology

To build a quantitative method for measure of 3′-UTR dynamics in eukaryotic transcriptomes, we adapted the ePAT approach (Janicke et al. 2012) to NGS. A schematic representation of the approach is shown in Figure 1A. Briefly, adenylated RNA is extended by dNTPs using the Klenow fragment of DNA polymerase I with an annealed anchor-oligo as a template. Importantly, any undesirable priming to internal poly(A) tracts in RNA is avoided by a requirement for this 3′ extension in subsequent steps. Here, we applied an anchor sequence compatible with the Illumina index primers and included a 5′ biotin moiety to facilitate handling. In a second step, the 3′ tagged RNA is subjected to limited fragmentation by RNase T_1 which cleaves after G-residues and thus ensures that cleavage is only possible within the body of the RNA, leaving the poly(A)-tract and the DNA-based 3′-tag protected. The extended 3′ fragments were collected on streptavidin magnetic beads and 5′ phosphorylated to allow ligation of an Illumina compatible splinted 5′-linker. Reverse transcription was primed from the bead-bound anchor sequence. The PAT-seq cDNA libraries were eluted from beads, size-selected by 6% urea-PAGE and amplified with primers that introduce the features for strand-specific Illumina sequencing and indexing. For the *Saccharomyces cerevisiae* mRNA analyzed here, the window of selection was 120–300 bases accommodating inserts of ~60–240 bases in length. This range was selected to ensure sufficient 3′-UTR sequence to unambiguously align reads to the yeast genome and to extend well into poly(A) sequence, allowing the generation of a surrogate score of adenylation. Because all reads run 5′ to 3′, from unique sequence into a variable length of poly(A) homopolymers, color balance is preserved and any loss of sequencing register caused by PCR slip is limited to the end of the read.

PAT-seq as a tool to study 3′-UTR dynamics

To demonstrate the versatility of the PAT-seq approach, we took advantage of the rapid and widespread transcriptional change in yeast cultures responding to carbon source shifts (Fig. 1B). The sequential addition of first galactose, and then glucose to cells growing with glycerol/ethanol as a carbon source induces a massive shift in transcription as cells rewire their metabolism from respiratory to fermentative growth, in what is termed the Warburg and Crabtree effect (Diaz-Ruiz et al. 2011). As an additional control for the fidelity of the poly(A) tail measurement, we also profiled wild-type cells and cells lacking the major deadenylase, Ccr4 (Tucker et al. 2001). Biological replicates of each strain were profiled, utilizing 1 µg of total RNA as input into

PAT-seq library preparation (see Materials and Methods and Supplemental Material). The libraries were amplified using 16 cycles of Illumina-indexing PCR, pooled and sequenced on a single lane of an Illumina Hiseq 1500 in rapid-run mode using 100-bp single-end chemistry. This returned an average of 8 M reads per biological sample for aligning to the *S. cerevisiae* genome. We developed an open-source software-pipeline called *tail-tools* pipeline for analysis of PAT-seq data (http://rnasystems.erc.monash.edu/). To avoid poly(A) driven mismapping, 3′ homopolymer stretches were masked prior to alignment to the reference genome sequence, and alignments were subsequently extended if part of the homopolymer stretch was genome encoded. The position of the first nontemplated adenosine, within a run of more than three, was taken as the site of adenylation.

Aligning the number of adenylated positions relative to the stop codon of all annotated yeast genes, shows that the vast majority of the PAT-seq reads map to 3′ UTRs, and confirms previous estimates that the average length of a yeast 3′ UTR is ∼100 bases (Fig. 1C; see also Supplemental Fig. S3e; Nagalakshmi et al. 2008). Simple exploratory analysis within the integrated genome browser (IGV) (Thorvaldsdóttir et al. 2012) highlights that most PAT-seq reads map to "peaks" adjacent to sites of polyadenylation (Supplemental Fig. S1) and because the PAT-seq reads are directional, they are readily mapped to their genomic locus of origin. Many loci showed additional evidence for noncoding 3′ and 5′ sense and antisense transcription as has been previously noted (Supplemental Fig. S1b; Nagalakshmi et al. 2008; Ozsolak et al. 2010; Yoon and Brem 2010). Furthermore, since RNA can become adenylated during exosome-mediated decay (Slomovic et al. 2010), noncoding and structural RNA was also detected (Supplemental Fig. S1c). When reads were assigned to annotated protein-coding genes, 6111 out of the 6486 (94%) annotated genes were detected in our combined data set. However, when reads containing a poly(A) stretch were clustered into adenylation sites across the genome, 23,636 adenylation sites (or peaks) were identified in the *S. cerevisiae* transcriptome. This increase in number of adenylation sites relative to annotated genes reflects the complex interplay between adenylation of the coding and noncoding transcriptome. Raw and normalized data are available (GEO accession GSE53461).

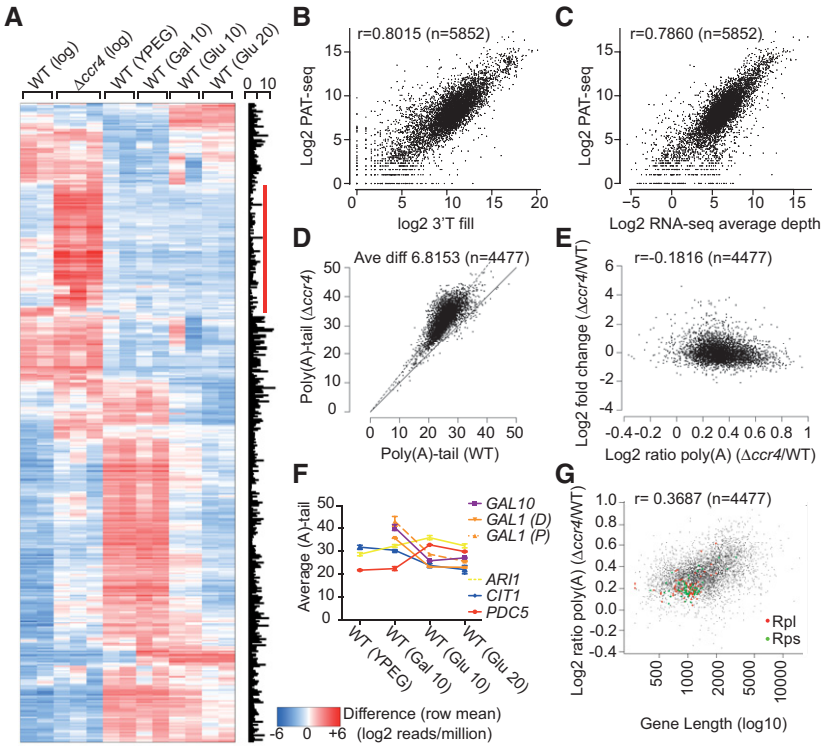

FIGURE 2. PAT-seq for DGE and polyadenylation state. (*A*) Differential expression of peaks (adenylated sites) with greater than sixfold change in expression line average and ≥10 reads. The red bar indicates normally silent genes deregulated in the Δ*ccr4* mutant. (*B*) The Pearson's correlation between PAT-seq read count and 3′ T-fill. Each black spot represents one of *n* genes. (*C*) The correlation between PAT-seq read count and the per gene average depth of coverage by RNA-seq. (*D*) The correlation between the average (per gene) andenylation-state of the Wt and the Δ*ccr4* transcriptomes. The solid line indicates the line of tail-length parity; the dashed line indicates the average change in adenylation-state ratio between the wild-type and the Δ*ccr4* transcriptome. (*E*) The correlation between tail-length change and expression-level change between the wild-type and the Δ*ccr4* transcriptome. (*F*) The adenylation-state change in average tail sequenced for candidate mRNA during the metabolic shift. The change is homo-directional to gene expression change. (*G*) The correlation between transcript length and adenylation-state ratio (Δ*ccr4* versus wild type). Large ribosomal subunit genes are marked red, and small ribosomal subunit genes are green in the figure on the *right*. Note: All data presented have an associated *P* value <0.0001.

PAT-seq returns digital gene expression data

To visualize expression change within our data, the Tail-Tools pipeline generates heatmaps of expression, built from either read-counts associated with annotated genes, or from individual peaks mapped to the genome (as in Fig. 2A). In general, RNA-seq is considered highly quantitative (Nookaew et al. 2012). Several 3′ focused RNA-seq methods have been developed for cleavage and adenylation site mapping and RNA quantitation. Of these, the 3′ T-fill approach has been suggested by Wilkening et al. (2013) to be the most robust. To confirm that our PAT-seq approach accurately estimates mRNA abundance, we performed a comparison to the wild-type yeast transcriptome analyzed by the 3′ T-fill approach or regular RNA-seq under equivalent experimental conditions (Wilkening et al. 2013). Comparing the read-counts between PAT-seq and 3′ T-fill for the measure of digital gene expression the correlation is strong ($r = 0.8015$)

(Fig. 2B), as is the correlation between PAT-seq and regular RNA-seq ($r = 0.7860$) (Fig. 2C). Indeed the latter correlation is slightly higher than the internal correlation between 3′ T-fill and regular RNA-seq performed within the Steinmetz laboratory ($r = 0.7185$; Supplemental Fig. S2a). The correlation of these data collected by laboratories on opposite sides of the globe strongly supports the power of PAT-seq for digital gene expression. Internal reproducibility between biological replicates was also very strong ($r = 0.9670$ and $r = 0.9294$ for gene expression and polyadenylation state, respectively) (Supplemental Fig. S2b–d). Genome-wide, adenylation site mapping was essentially identical between PAT-seq and 3′ T-fill (Supplemental Fig. S2e).

The integration of adenylation-state with gene expression

The changing distribution of poly(A) tails associated with mRNA can be diagnostic of certain aspects of RNA metabolism. Newly synthesized mRNA is usually long tailed, but any particular mRNA may display a spectrum of poly(A) tail lengths at steady state. In yeast, the global distribution ranges between ~10 and 80 A residues in wild-type cells (Minvielle-Sebastia et al. 1998; Lee et al. 2014), but can be restricted for specific transcripts, for example, *RPL46*, *PGK1*, and *MFA2* have been analyzed in high resolution to show a maximal poly(A) length of ~55, ~60, and ~70, respectively (Brown and Sachs 1998). The detection limit for data analyzed here was ~80 A (Supplemental Fig. S3). To determine if meaningful poly(A) tail-length distributions can be extracted from PAT-seq data, we calculated the average number of nontemplated adenylate residues terminating each mapped read, and then compared the adenylation-state of the transcriptome of wild-type cells with cells lacking the major cytoplasmic deadenylase Ccr4. The average length of the poly(A) tail sequenced in wild-type cells was 25.6 adenosines; in Δ*ccr4* cells the average was 32.4. Most transcripts have an extended poly(A) tail in Δ*ccr4* cells (data points above the diagonal line in Fig. 2D). Moreover, the longer the average sequenced tail-length in wild-type cells, the greater the increase in the mutant (dashed line in Fig. 2D).

Given that mRNA decay is initiated by poly(A)-trimming, a natural expectation was that poly(A)-stabilization in Δ*ccr4* cells would correspond to an increase in mRNA abundance. This was not the case however, if anything, a negative correlation was observed between expression change and poly(A) tail length-change (Fig. 2E). These data support recent evidence for transcript buffering (at the level of transcription) in the absence of normal mRNA turnover (Sun et al. 2013) and point further toward translational regulation as a source for the phenotypic differences that have been observed in Δ*ccr4* cells. Note: Random fragmentation by RNase T1 and a tight window of size selection, combined with 100 base Illumina reads, meant that not all reads were sequenced to the end of the poly(A) tract in our libraries for this experiment. In effect this means the poly(A) distribution of the transcriptome was subsampled. It is important to note that this still allowed detection of dynamic changes in adenylation-state between transcriptomes. For example, changes between the adenylation state of specific transcripts in Δ*ccr4* and wild-type cells were as easily detected when the data were further subsampled for only reads that include an in-phase 3′ anchor and thus represent a complete native tail (Supplemental Fig. S3).

Condition-dependent changes in poly(A) tail length were also clearly recorded. The metabolic shift applied to wild-type cells (Fig. 1B) is accompanied by well-characterized changes in the adenylation-state of specific genes (Decker and Parker 1993; Beilharz and Preiss 2007; Janicke et al. 2012). In general, the adenylation-state changes observed are homo-directional to changes in transcription. Thus, transcripts required for galactose catabolism (e.g., *GAL1*) are induced with a long poly(A) tail that is shortened after transcriptional inhibition by glucose as the transcript population ages (Janicke et al. 2012), and transcriptional repression of mRNAs encoding respiratory proteins is accompanied by age-related poly(A) shortening (e.g., *CIT1*). mRNA encoding fermentative mRNA, on the other hand, (e.g., *PDC5*) increase in poly(A) length as new transcripts replace aged ones (Fig. 2F).

We and others have previously reported that longer transcripts tend to have shorter poly(A) tails at steady state (Beilharz and Preiss 2007; Lackner et al. 2007; Subtelny et al. 2014). Here we extend this observation showing that longer transcripts exhibit a bigger proportional difference in tail length between wild-type cells and Δ*ccr4* cells (Fig. 2G). This could mean that the Ccr4–Not complex is recruited to such transcripts earlier in their metabolism. Or, that the poly(A) tails are less protected from the Ccr4–Not complex in longer transcripts. Notably, the generally short ribosomal protein genes tend to exhibit only moderate poly(A) extension in the absence of Ccr4 and hints at yet another example of the specialized control of this tightly regulated group of transcripts. The coregulated ribosomal biogenesis cluster on the other hand, does not show this trend (data not shown).

Little correlation was observed between the average sequenced poly(A) tail and mRNA abundance or protein expression (Supplemental Fig. S3B,C). This is in contrast to our pervious observations using poly(U)-chromatography and microarrays (Beilharz and Preiss 2007). However, multiple factors likely explain this difference. Including for example, the plasticity of adenylation state, and the genetic background of yeast strains utilized, previously W303a, versus BY4741 in the current study. An important further difference is that previous approaches depended on relative changes in the proportion of long versus short-tailed mRNA as measured by competitive-hybridization on microarrays. This measurement was weighted toward the longest tails within a population and the array technology favored

abundant transcripts (Beilharz and Preiss 2007; Lackner et al. 2007). Here we report the average length of tail-length distribution at high resolution and with digital precision. Recently, an accurate but technically complex alternative approach to the measure of poly(A) tail length, PAL-seq, was described (Subtelny et al. 2014). The average tail length of the yeast-transcriptome differed by only 1 nt between the PAL-seq and PAT-seq methods, and the average-length distributions were similar, with a moderate gene-to-gene correlation ($r = 0.3339$; Supplemental Fig. S3c). Such modest correlation is not uncommon for between-laboratory comparisons as exemplified by Grigull et al. (2004) in a study comparing mRNA stability.

Bringing statistical rigor to tests of 3′-UTR dynamics

The cost-effective nature of PAT-seq means biological replication is feasible and the data are thus readily analyzed for statistical significance using a combination of standard and custom tools. To identify statistically significant polyadenylation changes, we modified the limma software package to account for a depth-dependent variation in average poly(A) tail-length measurements (see Materials and Methods). For example, between Δ*ccr4* cells and BY4741 wild-type cells, 135/5607 genes, 277 /147750 adenylation sites, and the poly(A) tails of 4108/5229 genes, were statistically significantly differentially expressed (FDR ≤ 0.05).

Within the four samples representing the metabolic shift from respiration to fermentation, 1947/5696 annotated genes, 2721/17113 adenylation sites and the poly(A)-tails of 499/5273 genes were statistically significantly differentially expressed (by ANOVA, FDR ≤ 0.05). To confirm that this approach correctly identified the expected transcripts, we looked within the 499 statistically differentially adenylated transcripts. As expected, the galactose regulon (*GAL1*, *GAL2*, *GAL3*, *GAL7*, *GAL10*, and *GAL80*) was identified as significantly regulated; moreover, the metabolic shift to fermentation signals a major change in adenylation-state. The gene ontology terms associated with increased or decreased poly(A) length after 10 min of glucose addition were cytoplasmic translation (GO: 0002181 $P = 6.791^{46}$) and oxidation–reduction process (GO: 0055114 $P = 1.836^{24}$), respectively (Funassociate: Berriz et al. 2009). Interactive visualization of our complete data set is available here (http://rnasystems.erc.monash.edu/).

Cells lacking Ccr4 fail to silence repressed loci

The genes up-regulated in Δ*ccr4* cells, were those that are typically silent in rich media (Fig. 2A). The mating pathway was the major deregulated gene-set associated with loss of Ccr4. However, in addition to the aberrant expression of mating specific genes such as *PRM3*, Δ*ccr4* cells also overexpress *GPG1*, a morphogenic regulator of pseudohypal growth, and *GSC2* the catalytic subunit of 1,3 β-glucan synthase, normally involved in the spore wall formation (Fig. 3A, and Additional file 3). Moreover, a number of differentially expressed adenylated noncoding transcripts were identified as emanating from the long terminal repeats (LTRs) of yeast retro-transposons (TY) and ribosomal RNA. One such transcript extended from the *Ty3 LTR* (YORWsigma3) in Δ*ccr4* cells overlapping the 3′ UTR, and reducing the abundance of the major chromatin remodeler *SNF2* (Fig. 3A,B). Similarly, a transcript extending from the *YLRWsigma2* LTR overlapped the secretory regulator *AVL9* (Supplemental Fig. S1c). Failure to appropriately silence these loci in rich media may explain the diverse phenotypes that have been assembled for this mutant (Panepinto et al. 2013).

Alternative adenylation of coding and noncoding RNA

Widespread alternative polyadenylation (APA) provides a mechanism whereby single transcripts can switch between different 3′-UTR-encoded signals regulating translation (Ji and Tian 2009; Beck et al. 2010; Mangone et al. 2010; Ozsolak et al. 2010; Yoon and Brem. 2010; Fu et al. 2011; Haenni et al. 2012; Jan et al. 2011; Shepard et al. 2011; Ulitsky et al. 2012). PAT-seq is exquisitely sensitive to the presence of alternative cleavage and adenylation sites within the 3′ UTRs of mRNA (Fig. 3C; Supplemental Fig. S1). Because PAT-seq depends on extension of the 3′ end of adenylated RNA, priming to internal poly(A)-tracts is rare and adenylation sites identified by PAT-seq represent bona fide adenylated 3′ ends of RNA (Fig. 3A; double bands

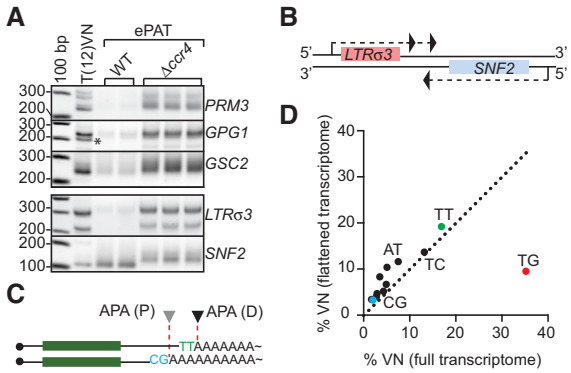

FIGURE 3. Alternative cleavage and adenylation. (*A*) To validate PAT-seq data, gene-by-gene T_{12}VN-PAT and ePAT assays were performed. The T_{12}VN-PAT assays indicate the size of the PCR amplicons with a limiting (A12)-poly(A) tail whereas the ePAT assay includes the full-native poly(A) tail in amplicons. Note the up-shift in amplicons sizes in the Δ*ccr4* mutant samples. (*B*) Schematic of the antiparallel orientation and the Ty3 LTR *YORWsigma3 (LTRσ3)* transcript and of *SNF2*. (*C*) APA shifts the transcript cleavage and adenylation between Proximal (P) and Distal (D) recognition sites. (*D*) The dinucleotide preceding the adenylation site is nonrandom. The flattened transcriptome indicates the percentage of dinucleotide usage at unique adenylation sites, comparing abundant and rare sites equally. The full transcriptome indicates all reads encompassing the adenylation site, incorporating transcript abundance.

PRM3 and *LTRσ3)*. To validate examples of APA in our data, we used a classic anchored $T_{(12)}$-VN oligonucleotide, commonly used in 3′ RACE (Janicke et al. 2012) and purpose-built 3′ focused RNA-seq approaches (Beck et al. 2010; Yoon and Brem 2010; Shepard et al. 2011; Derti et al. 2012) including the 3′T-fill method used for comparisons (Fig. 2B,C). While generally validating all the APA sites we detect by PAT-seq, a shortcoming of this anchored approach is internal priming (see Fig. 3A, *GPG1**). Moreover, the $T_{(12)}$-VN approach sometimes failed to support the stoichiometry of APA forms suggested by PAT-seq and gene-by-gene ePAT measurements. For example, the amplicons for *SNF2* generated by $T_{(12)}$-VN priming, appear less abundant than the WT and Δ*ccr4* samples prepared by ePAT, while this is not the case for the other genes shown (Fig. 3A).

We reasoned that nonrandom dinucleotide usage immediately prior to the adenylation sites in abundant transcripts might deplete specific combinations of the variable nucleotides (VN; N = any, V = any but T) in the $T_{(12)}$-VN approach. To address this issue, we extracted dinucleotide frequency immediately preceding the polyadenylation site in either all adenylated reads, or all distinct 30 base sequences immediately preceding adenylation in reads (full or flattened transcriptome, Fig. 3D). This uncovered a strong bias in sequences immediately prior to the site of adenylation. In the full wild-type yeast transcriptome (all adenylated reads), the frequency distribution was as follows: TG (35.2%) > TT (16.9%) > TC (13.2%) > AT (7.5%) > AC (5.1%) > CT (4.9%) > CC (4.3%) > AG (3.5%) > GT (2.9%) > GG (2.8%) > CG (2%) > GC (1.7%). We next compared these proportions with the unique adenylation sites (flattened transcriptome) encoded in the genome. For 11/12 variable dinucleotides, a high correlation was observed between the full or flattened transcriptome ($r = 0.94$). A single outlier (TG) was ~3.7 times over-represented in the full transcriptome (Fig. 3B) and likely represents highly abundant transcripts. Together, TT and TG precede >50% of the adenylation sites in the transcriptome. Indeed the poly(A) tail of *SNF2* is preceded by the TT nucleotide pair, likely explaining its under-representation by the $T_{(12)}$-VN approach in the validation data. By the Klenow-mediated extension approach, in contrast, there is no sequence selection beyond a requirement for adenylation, and thus it provides an unbiased tagging strategy for quantitation of adenylated RNA molecules.

Rapid advances in sequencing technology mean longer and more accurate reads through the poly(A) tails are possible. At the time of revision of this manuscript we utilize 150 base reads, and broader size selection, to detect statistically significant 3′-UTR dynamics in the human, murine, and nematode worm transcriptomes. The sum attributes described here lead us to propose PAT-seq as a powerful new addition to the family of RNA-seq methodologies and particularly for the measurement of 3′-UTR dynamics within eukaryotic transcriptomes.

MATERIALS AND METHODS

Yeast culture and RNA extraction

The yeast strains BY4741 and Δccr4::KanMX were grown in rich media (2% peptone, 1% yeast extract, and 2% glucose) to an OD_{600} of 0.8 at 30°C with shaking. To induce a carbon source shift, BY4741 cells were grown overnight in 100 mL rich glycerol/ethanol media (2% peptone, 1% yeast extract, 3% glycerol, and 2% ethanol) to an OD_{600} of 0.8 at 30°C with shaking. At the start of the experiment 10 mL of culture was harvested, washed in 1 mL of ice-cold dH_2O and snap frozen. To induce Galactose catabolic gene expression, 40% Galactose was added to the culture to a final concentration of 2% (w/v). After 10 min, 10 mL of culture was harvested and 40% glucose was added to a final concentration of 2% (w/v). Additional samples were harvested after 10 and 20 min of growth in the presence of glucose. Total RNA was extracted from snap frozen cell pellets by hot phenol extraction as previously described (Beilharz and Preiss 2009).

PAT-seq library preparation and Illumina sequencing

Briefly, the 3′ tag addition was based on our previous work (Janicke et al. 2012) except that a template oligonucleotide compatible with Illumina adaptor sequences (PAT-seq end-extend: [Bio]CAGACGT GTGCTCTTCCGATCTTTTTTTTTTTTTTTT) was used. The 5′ biotin facilitates enrichment of 3′-end-extended RNA fragments. For limited RNase T1 digestion, the extended RNA was mixed with a dilute (1/1000) solution of RNase T1 (100,000 units/mL; Roche) for 1 min on ice followed by immediate phenol/chloroform extraction in phase-lock tubes to stop the reaction. The extended 3′ RNA fragments were collected on streptavidin beads and 5′ phosphorylated with T4 PNK. A splinted 5′ linker was prepared by stoichiometrically pre-annealing PAT-seq Splint A (5′-CCCTACACGACGCTCTTC CG(rA)(rT)(rC)(rT)-3′) and PAT-seq Splint B (3′-GGGATGT GCTGCGAGAAGGCTAGANNNN-5′). This was ligated to the 5′ end of the 3′ fragments with T4 RNA ligase 2 (New England Biolabs) overnight at 16°C. Excess 5′ splint was removed by washing the magnetic beads prior to reverse transcription from the PAT-seq end-extend primer on the magnetic matrix using Super Script III (Life Technologies). The cDNA was size selected by elution from the beads in 2× formamide gel loading buffer and electrophoresis (6% urea-PAGE) alongside a 25-bp DNA ladder. The gel was stained with gel-star nucleic acid stain (Lonza) and imaged (Fugi LAS3000 and printed 1:1 to facilitate gel excision. Library cDNAs were eluted by the "crush and soak" method and then ethanol precipitated with the aid Glycoblue co-precipitant (Life Technologies). One-third of the purified cDNA was used as input for 16 cycles of amplification with PAT-seq Universal forward sequencing primer (5′-AATGATA CGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCT CTTCCG-3′) and ScriptSeq Index PCR reverse primers (Epicentre) and AmpliTaq Gold 360 Master Mix (Life Technologies). A detailed laboratory-ready protocol for library preparation is supplied (Supplemental file 3). PAT-seq libraries were sequenced on a single lane of the Illumina Hiseq1500 platform with 100 base rapid chemistry according to the manufacturer's instructions at the Gandel Charitable Trust Sequencing Centre (Monash University). T_{12}VN-PAT and ePAT assays were performed as previously described (Lee et al. 2014). Figures were prepared using Adobe Photoshop, Illustrator, and GraphPad Prism.

Analysis of PAT-seq data

We have automated analysis of PAT-seq data with an open-source pipeline, tail-tools (http://rnasystems.erc.monash.edu/). This software automates alignment to reference, identification of adenylation sites, read counting, and poly(A) tail-length estimation, production of visualizations to assess data quality, and statistical analysis. Reads were first clipped of poly(A) and adaptor sequence: The read was searched for a run of "A"s extending to the end of the read, or a run of "A"s extending into the adaptor sequence. An error rate of one base in five was allowed, and read bases with quality below 10 were ignored. Clipped reads were then aligned to the reference genome using Bowtie 2 (Langmead and Salzberg 2012). Where a read had several equal best alignments, one was chosen at random. Alignments which were followed by "A"s in the reference genome were extended to cover these "A"s if they were also seen in the original read. We refer to the number of nontemplated "A"s in a read as its tail length. Reads with tail length of at least four are referred to as poly(A) reads below. Reads were assigned to genes if their alignment overlapped the region from the 5′ end of the gene to 200 bases 3′ of the 3′ end of the gene. If this would assign a read to multiple genes, the gene minimizing the distance between the 3′ end of the alignment and the 3′ end of the gene was chosen. From this a count of reads per gene was obtained. Where a gene had at least 10 poly(A) reads, the average tail length of poly(A) reads is also calculated for that gene. It is expected to be an underestimate as the whole poly(A) tail is not always read, except in the case of poly(A) tails shorter than 12 bases, in which case it may be an overestimate.

Adenylation sites were called where the 3′ end of the alignments of at least 50 poly(A) reads occurred within 10 bases of each other. Where multiple candidate sites exist within 50 bases of each other, only the site with the greatest number of poly(A) reads is called. Reads were assigned to adenylation sites if their alignment overlapped a region from 100 bases 5′ of the site to the site itself. Again, if a read could be assigned to multiple adenylation sites the site minimizing the distance to the 3′ end of the alignment was chosen. As with genes, read counts and average tail lengths are calculated for each called adenylation site.

Statistical analysis and differential expression testing

Since each adenylated RNA molecule generates only a single read, raw counts were simply converted to $\log_2$ reads per million (RPM) without further normalization to transcript length. Normalization of read counts between samples was performed using TMM normalization (Robinson and Oshlack 2010) as implemented in Bioconductor package edgeR, to obtain reads per million (RPM) values. To visualize expression data in heatmaps, RPM values are transformed using the variance-stabilized log transformation (Durbin et al. 2002) to suppress excess variation in genes or adenylation sites with low read count. Significant differential expression was detected using a moderated *t*-test on log transformed count data, using the Bioconductor package limma (Smyth 2004), and using voom to log-transform and weight read counts. Before testing for differential expression, we filter out all features where there is no sample with at least 10 reads. For comparisons to existing data from Wilkening et al. (2013), files were extracted from GSE40110 using data pile-ups of start positions for 3′T fill, and depth of coverage for RNA-Seq. For comparisons to PAL-seq, data were extracted from GSE52809 and relied on at least 10 reads in each data set.

Differential tail-length testing

We tested for differential tail lengths, using a custom modification of limma. The accuracy to which each estimated tail length is known is quite variable, depending on the number of poly(A) reads available and the native distribution, and this needs to be taken into account before limma can be used. We therefore modified the limma package as follows: Let X be the design matrix for a linear model we wish to fit to the data. We first find a basis N for the null space of the design matrix X. That is, a matrix N such that N^TX is zero, and for which concatenating the columns of X and N produces a square matrix of full rank. Multiplying a data vector $\vec{y}$ by N^T eliminates any contribution the linear model may have made to $\vec{y}$. Assuming the tail lengths Y_{ij} for a feature i in the different samples j are independent and normally distributed with variances σ^2_{ij}, we model the vector $\vec{y_i}$ as being drawn from a multivariate normal distribution with mean given by the linear model and covariance by the diagonal matrix $\Sigma_i = \mathrm{diag}(\sigma^2_{i1}, \sigma^2_{i2}, \ldots, \sigma^2_{in})$. For each feature i, $N^T\vec{y_i}$ is drawn from a multivariate normal distribution with mean 0 and covariance $N^T\Sigma_i N$. We seek an assignment of σ^2_{ij} by maximum likelihood estimation, by maximizing the total of the log probability densities of each $N^T\vec{y_i}$ being drawn from the multivariate normal distribution $\mathcal{N}(0, N^T\Sigma_i N)$. Let r_{ij} be the number of reads used to calculate the average tail length y_{ij}. We expect each tail length observed in a read to have some technical variance σ^2_T. We further expect there to be per-sample biological variance σ^2_B, which was found to scale as the square of the tail length. Hence our model of variance in the tail lengths is

$$\sigma^2_{ij} = \frac{\sigma^2_{\mathrm{T}}}{r_{ij}} + \sigma^2_{\mathrm{B}} y^2_{ij}$$

with σ^2_T and σ^2_B chosen by MLE as described above.

The variances σ^2_{ij} are used to calculate weights $1/\sigma^2_{ij}$ for use with the limma software package. By using limma we ensure that features with biological variance larger than the σ^2_{B} fitted globally are not falsely counted as significant. Before testing for differential tail lengths, we filter out all features where there are insufficient samples with at least 10 poly(A) reads to allow the fitting of the linear model.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

Dr. Stuart Archer and the Beilharz laboratory are acknowledged for fruitful discussions and their insightful critique of the manuscript. T.H.B. was supported by an Australia Research Fellowship from the Australian Research Council (DP0878224), and grants from the Australian National Health and Medical Research Council (APP1042851, APP1042848). T.P. acknowledges funding through an ARC Discovery Grant (DP1300101928) and a NHMRC Senior Research Fellowship (SRF514904). P.F.H., D.R.P., and T.S. were supported by the VLSCI's Life Sciences Computation Centre, a collaboration between Melbourne, Monash, and La Trobe Universities and an initiative of the Victorian Government, Australia.

Received October 23, 2014; accepted April 20, 2015.

REFERENCES

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.

Beilharz TH, Preiss T. 2007. Widespread use of poly(A) tail length control to accentuate expression of the yeast transcriptome. *RNA* **13:** 982–997.

Beilharz TH, Preiss T. 2009. Transcriptome-wide measurement of mRNA polyadenylation state. *Methods* **48:** 294–300.

Beilharz TH, Humphreys DT, Clancy JL, Thermann R, Martin DI, Hentze MW, Preiss T. 2009. microRNA-mediated messenger RNA deadenylation contributes to translational repression in mammalian cells. *PLoS One* **4:** e6783.

Berriz GF, Beaver JE, Cenik C, Tasan M, Roth FP. 2009. Next generation software for functional trend analysis. *Bioinformatics* **25:** 3043–3044.

Brown CE, Sachs AB. 1998. Poly(A) tail length control in *Saccharomyces cerevisiae* occurs by message-specific deadenylation. *Mol Cell Biol* **18:** 6548–6559.

Decker CJ, Parker R. 1993. A turnover pathway for both stable and unstable mRNAs in yeast: evidence for a requirement for deadenylation. *Genes Dev* **7:** 1632–1643.

Derti A, Garrett-Engele 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.

Diaz-Ruiz R, Rigoulet M, Devin A. 2011. The Warburg and Crabtree effects: on the origin of cancer cell energy metabolism and of yeast glucose repression. *Biochim Biophys Acta* **1807:** 568–576.

Durbin BP, Hardin JS, Hawkins DM, Rocke DM. 2002. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* **18 Suppl 1:** S105–S110.

Eulalio A, Huntzinger E, Nishihara T, Rehwinkel J, Fauser M, Izaurralde E. 2009. Deadenylation is a widespread effect of miRNA regulation. *RNA* **15:** 21–32.

Fabian MR, Mathonnet G, Sundermeier T, Mathys H, Zipprich JT, Svitkin YV, Rivas F, Jinek M, Wohlschlegel J, Doudna JA, et al. 2009. Mammalian miRNA RISC recruits CAF1 and PABP to affect PABP-dependent deadenylation. *Mol Cell* **35:** 868–880.

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.

Grigull J, Mnaimneh S, Pootoolal J, Robinson MD, Hughes TR. 2004. Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Mol Cell Biol* **24:** 5534–5547.

Haenni S, Ji Z, Hoque M, Rust N, Sharpe H, Eberhard R, Browne C, Hengartner MO, Mellor J, Tian B, et al. 2012. Analysis of *C. elegans* intestinal gene expression and polyadenylation by fluorescence-activated nuclei sorting and 3′-end-seq. *Nucleic Acids Res* **40:** 6304–6318.

Hong LZ, Li J, Schmidt-Kuntzel A, Warren WC, Barsh GS. 2011. Digital gene expression for non-model organisms. *Genome Res* **21:** 1905–1915.

Jan CH, Friedman RC, Ruby JG, Bartel DP. 2011. Formation, regulation and evolution of *Caenorhabditis elegans* 3′UTRs. *Nature* **469:** 97–101.

Janicke A, Vancuylenberg J, Boag PR, Traven A, Beilharz TH. 2012. ePAT: a simple method to tag adenylated RNA to measure poly (A)-tail length and other 3′ RACE applications. *RNA* **18:** 1289–1295.

Ji Z, Tian B. 2009. Reprogramming of 3′ untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PLoS One* **4:** e8419.

Kapranov P, Ozsolak F, Kim SW, Foissac S, Lipson D, Hart C, Roels S, Borel C, Antonarakis SE, Monaghan AP, et al. 2010. New class of gene-termini-associated human RNAs suggests a novel RNA copying mechanism. *Nature* **466:** 642–646.

Lackner DH, Beilharz TH, Marguerat S, Mata J, Watt S, Schubert F, Preiss T, Bahler J. 2007. A network of multiple regulatory layers shapes gene expression in fission yeast. *Mol Cell* **26:** 145–155.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359.

Lee MC, Jänicke A, Beilharz TH. 2014. Using Klenow-mediated extension to measure poly(A)-tail length and position in the transcriptome. *Methods Mol Biol* **1125:** 25–42.

Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein MR, Khivansara V, et al. 2010. The landscape of *C. elegans* 3′UTRs. *Science* **329:** 432–435.

Mayr C, Bartel DP. 2009. Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138:** 673–684.

Minvielle-Sebastia L, Beyer K, Krecic AM, Hector RE, Swanson MS, Keller W. 1998. Control of cleavage site selection during mRNA 3′ end formation by a yeast hnRNP. *EMBO J* **17:** 7454–7468.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320:** 1344–1349.

Nookaew I, Papini M, Pornputtpong N, Scalcinati G, Fagerberg L, Uhlén M, Nielsen J. 2012. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*. *Nucleic Acids Res* **40:** 10084–10097.

Ortiz-Zapater E, Pineda D, Martinez-Bosch N, Fernandez-Miranda G, Iglesias M, Alameda F, Moreno M, Eliscovich C, Eyras E, Real FX, et al. 2012. Key contribution of CPEB4-mediated translational control to cancer progression. *Nat Med* **18:** 83–90.

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.

Pak C, Garshasbi M, Kahrizi K, Gross C, Apponi LH, Noto JJ, Kelly SM, Leung SW, Tzschach A, Behjati F, et al. 2011. Mutation of the conserved polyadenosine RNA binding protein, ZC3H14/dNab2, impairs neural function in *Drosophila* and humans. *Proc Natl Acad Sci* **108:** 12390–12395.

Panepinto JC, Heinz E, Traven A. 2013. The cellular roles of Ccr4-NOT in model and pathogenic fungi—implications for fungal virulence. *Front Genet* **4:** 302.

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11:** R25.

Ruzanov P, Riddle DL. 2010. Deep SAGE analysis of the *Caenorhabditis elegans* transcriptome. *Nucleic Acids Res* **38:** 3252–3262.

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.

Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17:** 761–772.

Slomovic S, Fremder E, Staals RH, Pruijn GJ, Schuster G. 2010. Addition of poly(A) and poly(A)-rich tails during RNA degradation in the cytoplasm of human cells. *Proc Natl Acad Sci* **107:** 7407–7412.

Smyth GK. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3:** Article3.

Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP. 2014. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508:** 66–71.

Sun M, Schwalb B, Pirkl N, Maier KC, Schenk A, Failmezger H, Tresch A, Cramer P. 2013. Global analysis of eukaryotic mRNA degradation reveals Xrn1-dependent buffering of transcript levels. *Mol Cell* **52:** 52–62.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2012. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14:** 178–192.

Tucker M, Valencia-Sanchez MA, Staples RR, Chen J, Denis CL, Parker R. 2001. The transcription factor associated Ccr4 and Caf1 proteins are components of the major cytoplasmic mRNA deadenylase in *Saccharomyces cerevisiae*. *Cell* **104:** 377–386.

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.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. 1995. Serial analysis of gene expression. *Science* **270:** 484–487.

Wei W, Pelechano V, Jarvelin AI, Steinmetz LM. 2011. Functional consequences of bidirectional promoters. *Trends Genet* **27:** 267–276.

Wilkening S, Pelechano V, Jarvelin AI, Tekkedil MM, Anders S, Benes V, Steinmetz LM. 2013. An efficient method for genome-wide polyadenylation site mapping and RNA quantification. *Nucleic Acids Res* **41:** e65.

Wu ZJ, Meyer CA, Choudhury S, Shipitsin M, Maruyama R, Bessarabova M, Nikolskaya T, Sukumar S, Schwartzman A, Liu JS, et al. 2010. Gene expression profiling of human breast tissue samples using SAGE-Seq. *Genome Research* **20:** 1730–1739.

Wyers F, Rougemaille M, Badis G, Rousselle JC, Dufour ME, Boulay J, Regnault B, Devaux F, Namane A, Seraphin B, et al. 2005. Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121:** 725–737.

Yoon OK, Brem RB. 2010. Noncanonical transcript forms in yeast and their regulation during environmental stress. *RNA* **16:** 1256–1267.

PAT-seq: a method to study the integration of 3′-UTR dynamics with gene expression in the eukaryotic transcriptome

Paul F. Harrison, David R. Powell, Jennifer L. Clancy, et al.

RNA 2015 21: 1502-1510 originally published online June 19, 2015
Access the most recent version at doi:10.1261/rna.048355.114