Cell



# Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation

Fatih Ozsolak,[1,*] Philipp Kapranov,[1] Sylvain Foissac,[2] Sang Woo Kim,[3] Elane Fishilevich,[3] A. Paula Monaghan,[4] Bino John,[3] and Patrice M. Milos[1,*]
[1]Helicos BioSciences Corporation, Cambridge, MA 02139, USA
[2]Integromics, Madrid 28760, Spain
[3]Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15260, USA
[4]Department of Neurobiology, University of Pittsburgh, Pittsburgh, PA 15260, USA
*Correspondence: fatihozsolak@gmail.com (F.O.), pmilos@helicosbio.com (P.M.M.)
DOI 10.1016/j.cell.2010.11.020

## SUMMARY

The emerging discoveries on the link between polyadenylation and disease states underline the need to fully characterize genome-wide polyadenylation states. Here, we report comprehensive maps of global polyadenylation events in human and yeast generated using refinements to the Direct RNA Sequencing technology. This direct approach provides a quantitative view of genome-wide polyadenylation states in a strand-specific manner and requires only attomole RNA quantities. The polyadenylation profiles revealed an abundance of unannotated polyadenylation sites, alternative polyadenylation patterns, and regulatory element-associated poly(A)$^+$ RNAs. We observed differences in sequence composition surrounding canonical and noncanonical human polyadenylation sites, suggesting novel noncoding RNA-specific polyadenylation mechanisms in humans. Furthermore, we observed the correlation level between sense and antisense transcripts to depend on gene expression levels, supporting the view that overlapping transcription from opposite strands may play a regulatory role. Our data provide a comprehensive view of the polyadenylation state and overlapping transcription.

## INTRODUCTION

The known regulatory role of 3′ untranslated regions (3′UTRs) and poly(A) tails in mRNA localization, stability, and translation (reviewed by Andreassi and Riccio, 2009), and polyadenylation regulation defects leading to human diseases such as oculopharyngeal muscular dystrophy, thalassemias, thrombophilia, and IPEX syndrome (Bennett et al., 2001; Brais et al., 1998; Gehring et al., 2001; Higgs et al., 1983; Lin et al., 1998; Orkin et al., 1985) underscores the need to fully characterize polyadenylation sites and mechanisms. Our knowledge in this area primarily originates from expressed sequence tag (EST) databases and predictions relying on polyadenylation-associated motif elements (Graber et al., 2002; Lutz, 2008; Tian et al., 2005). EST databases are valuable but insufficient for in-depth mapping of polyadenylation sites due to data quality problems, such as low numbers of full-length ESTs, chimeric sequences (due to cDNA template switching; Cocquet et al., 2006), internal cDNA priming events leading to cloning of incomplete transcripts, and low-quality sequences at the ends of ESTs (Zhang et al., 2005a, 2005b). For applications requiring identification of polyadenylation site usage frequency changes across biological conditions, EST databases, motif searches, and classic polyadenylation site mapping approaches (Slomovic et al., 2008), such as RACE, RT-PCR, and nuclease sensitivity assays, do not provide the required simplicity, sensitivity, depth, and quantitative genome-wide view. Annotation of the 3′ ends of yeast genes were attempted previously with RNA-seq (Nagalakshmi et al., 2008) and microarray-based (David et al., 2006) approaches, but these studies did not have sufficient resolution to map individual cleavage sites for polyadenylation. Furthermore, despite much interest devoted to overlapping transcription, we still do not have a complete understanding of sense/antisense transcription (reviewed by Faghihi and Wahlestedt, 2009). To date, our knowledge in this area comes from methods relying on reverse transcription that suffers from spurious second-strand cDNA products (Gubler, 1987; Spiegelman et al., 1970), complicating analyses requiring unambiguous determination of RNA strand. Although methods have recently been developed that preserve the RNA strand information through RNA-level modifications, such as bisulfite treatment or RNA-level adaptor ligation (He et al., 2008; Mamanova et al., 2010), these still rely on cDNA synthesis, ligation, and amplification steps that may introduce artifacts and complicate the quantitation of various RNA species.

To avoid the known biases and artifacts introduced to RNA measurements during reverse transcription (Cocquet et al., 2006; Liu and Graber, 2006; Mamanova et al., 2010; Wu et al., 2008) or other sample manipulation steps, we recently developed the direct RNA sequencing (DRS) technology (Ozsolak et al., 2009). DRS sequences RNA molecules in a massively parallel manner without its prior conversion to cDNA or the need for biasing ligation or amplification steps. Since this proof-of-concept study, we have improved and adapted DRS

for use with the Helicos Genetic Analysis System. DRS produces alignable reads up to 55 nt (mean read length, 33–34 nt). Unlike other RNA analysis approaches, which require multiple nucleic acid manipulation steps, DRS only requires polyadenylated and 3′ blocked RNA templates for sequencing.

We here applied DRS to generate a comprehensive and high-resolution map of polyadenylation sites of human and yeast transcripts. Using multiple independent approaches, we validated our findings and demonstrated the usefulness of the approach to identify alternative polyadenylation events. We observed many unannotated polyadenylation sites and novel RNA species associated with open chromatin sites that may function to regulate gene expression. We also observed that sequence and motif contexts surrounding novel intergenic and genic sense/antisense polyadenylation sites away from 3′ ends of known genes exhibit significant differences than sequence and motif contexts surrounding polyadenylation sites near known gene 3′ ends. This observation suggests alternative mechanisms and/or purposes of RNA polyadenylation. In addition, we have examined overlapping transcription patterns of poly(A)$^+$ transcripts. Between the steady-state quantities of sense and antisense transcripts, we observed a complex correlation pattern that depends on gene expression levels.

## RESULTS

### Mapping Global 3′ Polyadenylation Sites with DRS

To determine polyadenylation locations, 200–300 picograms of human liver and yeast poly(A)$^+$ RNAs, and 3 ng of human brain total RNA blocked at their 3′ ends were used per sequencing channel. Given that poly(A)$^+$ RNA species already contain a natural poly(A) tail, additional polyadenylation was not needed. After the capture of poly(A)$^+$ RNA species on poly(dT)-coated flow cell surfaces by hybridization, a ''fill'' step with natural dTTP and a ''lock'' step with fluorescently labeled proprietary Virtual Terminator (VT)-A, -C, and -G nucleotides were performed. These steps correct for any misalignments that may be present in poly (A/T) duplexes and ensure that the sequencing starts in the template rather than the poly(A) tail. After the completion of fill and lock steps, DRS was initiated. The 5′ ends of DRS reads signify cleavage locations. The resolution for identification of the polyadenylation cleavage nucleotide is dependent on fill and lock efficiency and the ability of the sequencing reaction to start immediately upstream of the poly(A) tails. We measured this efficiency using polyadenylated oligoribonucleotides and determined the resolution to be ±2 nt (see Figure S1A available online). To determine whether our results might have been negatively affected by potential internal priming events, we performed experiments to observe the sequencing behavior of templates containing internal poly(A) stretches with 3′ noncomplementary overhangs and examined the fraction of polyadenylation regions containing downstream poly(A)-rich regions. We observed rare or no occurrence of internal priming events (Table S1 and Extended Experimental Procedures). Thus, the technology is capable of mapping the extensive 3′ end heterogeneity we and others (Iseli et al., 2002; Lopez et al., 2006; Muro et al., 2008) observed in the majority of yeast and human genes in a genome-wide manner and at nucleotide resolution (Figures 1A–1D).

### Genome-wide 3′ Polyadenylation State in Yeast

We obtained 7,036,730 DRS reads uniquely aligned to the yeast genome, each read representing a polyadenylation site of an independent transcript, to deduce the yeast polyadenylation landscape (Table S2). To verify our findings, we compared the polyadenylation sites identified here to the sites identified previously for 11 yeast genes using classic approaches, observing high overlap (Figure 1B). Because of its higher resolution, DRS found the frequently used cleavage locations reported previously and other generally lower-frequency cleavage positions (Figure 1A and Figure S1B). In addition, DRS data agreed well with the polyadenylation sites mapped previously for ten genes and seven snoRNAs using PCR amplification of 3′ transcript ends in a manner that preserves the variability in the 3′ ends, followed by high-throughput DNA sequencing of the RT-PCR products (Ozsolak et al., 2009). Furthermore, we validated four previously unannotated intergenic and genic polyadenylation locations using cloning and RACE approaches (Figure S1C and Table S3). We also compared DRS reads to the 60,218 3′ end tags, which constitute ~0.2% of RNA-seq reads, are analogous to DRS reads and mark yeast polyadenylation sites (Nagalakshmi et al., 2008), observing 53,849 (89.4%) of end tags to be within 5 nt of DRS read start locations. The difference observed in the remaining ~10% may be due to differences in the resolution of both methods, different yeast strains and RNA preparation approaches used in both studies.

The median length of the 3′UTRs of 5759 yeast open reading frames (ORFs) was 166 nt (Figure 2A and Table 1). With the number of reads and depth we generated for this study, we observed that 72.1% of the yeast genes exhibited polyadenylation locations separated by at least 50 nt, and frequently more, and thus have multiple polyadenylation sites. The higher levels observed here relative to the 10%–15% level reported previously (Nagalakshmi et al., 2008) may be due to the higher resolution of the approach presented here and the higher number of transcripts analyzed. Similar to previous reports (Nagalakshmi et al., 2008), we observed 14% of genes to be orientated in tail-to-tail orientation and have overlapping 3′ ends (see below). Fourteen percent of yeast DRS reads mapped to regions within the yeast ORFs either in exons or introns (Table 1). Intronic polyadenylation sites are possibly due to a dynamic interplay between splicing and polyadenylation (Tian et al., 2007) and may represent transcripts encoding shorter proteins.

10.6% of yeast DRS reads did not map downstream of annotated yeast 3′ ends or within the ORFs. To examine the degree of association of yeast poly(A)$^+$ transcripts with regulatory regions, we took advantage of the regulatory protein binding sites defined recently by DNase I hypersensitive site (DHS) mapping (Hesselberth et al., 2009). We observed a significant enrichment of divergent transcripts (e.g., transcribed away from DHSs) in regions that are in proximity to intergenic DHSs ($p = 8.041e\text{-}07$, nonparametrical two-sample Kolmogorov-Smirnov test) (Figure S2).

### Genome-wide 3′ Polyadenylation State in Humans

A total of 11,882,580 uniquely mapping reads were obtained from human liver poly(A)$^+$ RNA, of which 1,322,970 were derived from mitochondria and 2,570 reads from rRNA. This is consistent with the observations that human mitochondrial transcripts and

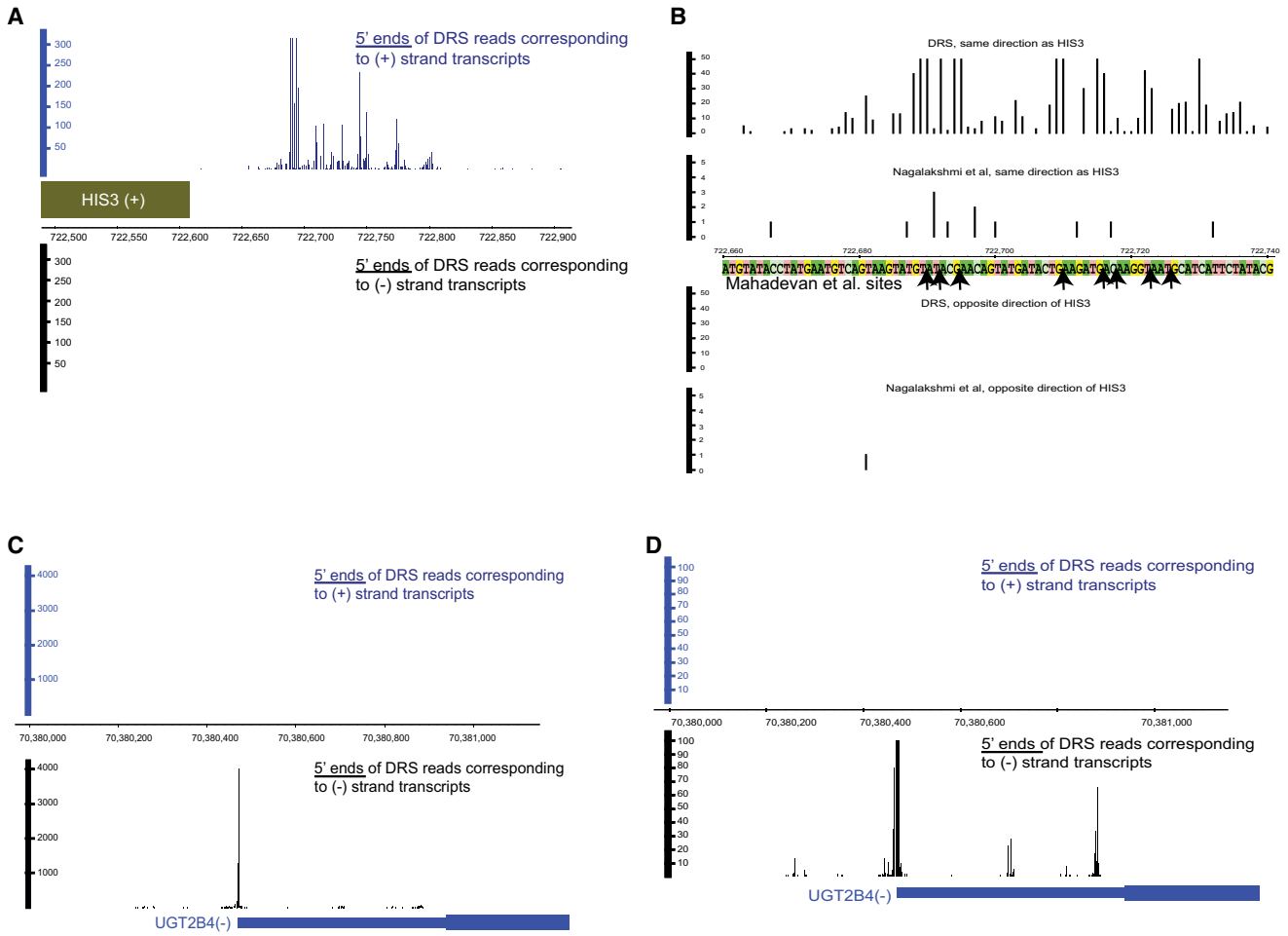


**Figure 1. Polyadenylation Site Detection in Yeast and Human**

(A) The blue and black panels show the DRS reads emanating from transcripts in the + and – direction, respectively. The major peaks in the blue panel correspond to the 13 polyadenylation sites at locations 722690, 722692, 722695, 722710, 722716, 722718, 722723, 722726, 722746, 722750, 722752, 722775, and 722777 previously identified for HIS3 (Mahadevan et al., 1997) using 3′ RACE-PCR.

(B) Zoomed-in view of (A). *y* axis was reduced from 0–300 scale to 0–50. *x* axis was reduced from 722,500–722,900 scale to 722,660–722,740. All "end tags" identified by Nagalakshmi et al. (2008) in this region are also shown (*y* axis for these tags is on the scale of 0–5). Arrows mark the sites identified by Mahadevan et al. (1997) in the region shown.

(C and D) Overview (B) and a zoomed-in view (C) of reads mapping to *UGT2B4* 3′ annotated ends. Multiple potential polyadenylation sites are evident in panel C (see also Figure S2 and Table S1).

a fraction of rRNAs are polyadenylated, perhaps for the purposes of degradation (Nagaike et al., 2005; Slomovic et al., 2010); 56.1% of DRS reads overlapped with 19,871 of 28,858 polyadenylation sites previously annotated using EST databases and motif searches (Zhang et al., 2005a). The differences observed may be due to the single tissue examined here, whereas EST database searches include data from multiple tissue types. More than half (55.7%) of liver DRS reads emerged from within 10 nt of annotated 3′ ends of UCSC Genes (Figure 2B, Figure S3A, and Table 1). The remaining 44.3% of the reads represent either novel RNAs or alternative polyadenylation sites of known mRNAs. Although estimation of the extent of noncoding transcription based on this data is difficult because the full structures of transcripts represented by the DRS reads are not known, at the very least, 9% of reads are located in intergenic regions that are at least 5 kb away from known genes and thus likely to represent novel RNAs; 37% of intergenic reads in humans are within 5 kb of known transcripts, and 42% are within 10 kb. Thus, a considerable fraction of intergenic reads are in proximity to known genes (van Bakel et al., 2010). An additional 14.7% of reads fall within introns on either strand. Polyadenylation events near the 3′ ends of known genes tend to happen more frequently in 3′UTR regions rather than the region immediately downstream of the 3′ ends of genes (Table 1 and Figure S3A). This may be caused by degradation intermediates of prematurely terminated transcripts, or the 3′ end annotations generated from EST databases favoring more downstream polyadenylation locations over upstream ones due to concerns such as incomplete cDNA clones and sequences, and thus, underrepresenting the diversity of polyadenylation sites.

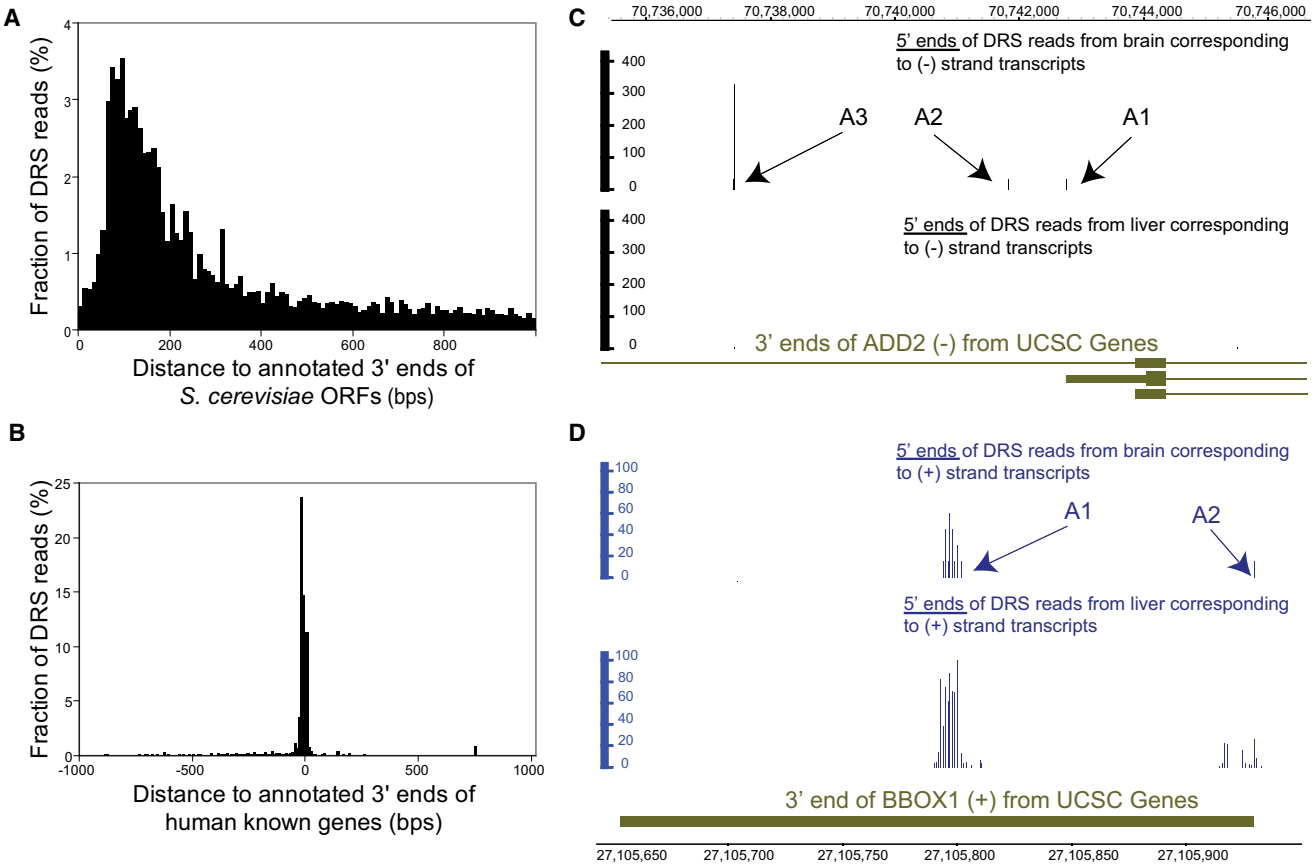


**Figure 2. Characteristics of Polyadenylation Sites in Yeast and Human**

(A and B) Y-axes indicate the fraction of DRS reads aligning at x-distances (in 10 bp bins) relative to the annotated 3′ ends of yeast ORFs (A) and the annotated 3′ ends of human UCSC genes (B).

(C and D) *ADD2* (C) and *BBOX1* (D) polyadenylation sites in human liver and brain. The polyadenylation sites identified (indicated as A1, A2 ,and A3) for both genes agree well with previous findings (Costessi et al., 2006; Rigault et al., 2006) (see also Figure S3 and Table S3).

To exemplify the ability of DRS to identify alternative polyadenylation events, we profiled human brain total RNA. *ADD2* mRNAs were found to have one major and additional minor polyadenylation sites in brain but none in liver (Figure 2C), as reported previously (Costessi et al., 2006). In addition, in concordance with previous results (Rigault et al., 2006), we observed two polyadenylation sites for *BBOX1* and a higher quantity of the "short" versus the "long" transcript in both tissues (Figure 2D).

## Sense/Antisense Poly(A)$^+$ Transcripts in Yeast and Human

DRS can not only pinpoint the sites of sense and antisense transcription, but also provide quantification of such transcripts without biases introduced by steps such as ligation, amplification, and other manipulations. Of the 5769 annotated yeast ORFs, at least 3492 (60.5%) had an antisense transcript, as evidenced by at least 10 antisense reads within the annotated ORFs (Figures S3B and S3C). These antisense reads compose 9.2% of the total DRS reads. When we considered the ambiguity in yeast 3′ end annotations and included regions 200 nt downstream of the 3′ annotation, the fraction of antisense reads increased to 41.2% and the ORFs with antisense transcripts increased to 4641 (80.4%), in part due to the genes with overlapping 3′ ends.

In the human liver RNA, at least 19,680/65,260 (30.2%) of all annotated transcripts were found to have antisense transcription as defined by at least 10 antisense reads either in exons or introns (Figures S3D and S3E). Although prevalent, the antisense transcription is still a minority in terms of transcript abundance: ~8% of all reads that overlap an annotated transcript are antisense to it. This number is similar to the 11% reported previously (He et al., 2008). Importantly, these numbers were obtained from poly(A)$^+$ RNA and do not represent the extent of poly(A)$^-$ antisense transcription (Dutrow et al., 2008; Kiyosawa et al., 2005).

## Quantification of Sense/Antisense Poly(A)$^+$ Transcriptome

We then explored the correlation between the quantities of sense and antisense transcripts. This analysis was attempted to observe the relationship between sense and antisense transcripts encoded by the same genomic region, given the presence of certain biological constraints such as transcription in both directions in a locus and pathways degrading

**Table 1. Distribution of Yeast and Human Liver Reads across Genomic Regions**

| Human | 5′UTR | 3′UTR | CDS | Introns | Transcripts | ±200 nt of 5′ Ends | ±200 nt of 3′ Ends | ±10 nt of 3′ Ends |
|---|---|---|---|---|---|---|---|---|
| Sense | 6.46 | 79.38 | 1.02 | 8.8 | 83.94 | 0.59 | 71.32 | 55.7 |
| Antisense | 0.18 | 2.1 | 0.23 | 5.86 | 7.98 | 0.1 | 2.96 | 1.12 |
| Yeast | CDS | Introns | Transcripts | | ±1000 nt of 3′ Ends of ORFs | | | |
| Sense | 4.68 | 0.19 | 4.86 | | 91.36 | | | |
| Antisense | 9.16 | 0.04 | 9.19 | | 53.04 | | | |

The numbers indicate percentages of uniquely aligned yeast and human DRS reads (Table S2) as provided by the SeqSolve software (Integromics). The categories shown are not exclusive, and each proportion was computed independently. Hence, proportions are not expected to add up to 100%. The relatively high percentage of reads in the category of antisense yeast reads within 1000 nt of 3′ ORF ends is due to ~2000 yeast ORFs whose 3′ ends are close to each other. CDS: coding sequence, UTR: untranslated region, ORF: open reading frame, Transcripts: within annotated gene boundaries (see also Figure S1 and Table S2).

complementary RNA species, such as microRNA or similar pathways in human. The distribution of the sense and antisense counts for yeast and human did not represent the normal distribution (Shapiro-Wilk test, $p < 0.0001$), even after converting the values into log space. Thus, we used the nonparametric Spearman correlation for this analysis based on the raw (non-log converted) values of sense and antisense expression levels of annotated genes. We separated the annotated genes into four quartiles according to their sense expression levels (Table 2). We did find a weak, but significant (see below), negative correlation between the levels of sense and antisense polyadenylated RNAs in the top quartile (Q1). The correlation became progressively more positive as the levels of the sense transcripts decreased, as exemplified by the positive correlation for the bottom fourth and third quartile of expression for the yeast and human samples, respectively. Because the expression levels of transcripts that do not overlap in the genome could also correlate and the negative correlations obtained for the high expressors could be influenced by the extreme values, we introduced a permutation test whereby pairing of sense-antisense values for each gene was reassigned: for each annotated gene, the sense value was kept the same, the antisense value was randomly chosen from another gene, and the Spearman correlation was calculated. This test shows that all (even the lowest) correlations found between the real sense and antisense reads counts are indeed highly significant ($p < 0.001$). Similar trends were observed when converging genes in yeast were omitted from the analyses (Table S4).

## Sequence Structure Surrounding Polyadenylation Sites

Having generated an in-depth view of polyadenylation cleavage locations, we examined the sequence patterns potentially governing transcription termination and polyadenylation. We first performed a de novo search for motifs near human polyadenylation locations and detected three novel motifs and the canonical signal (Figure 3). For this analysis, we used confident polyadenylation sites we defined using a clustering approach and supported by multiple reads (Figure S4, Table S5, and Extended Experimental Procedures). We identified a novel TTTTTTTTT motif ($e = 10^{-158}$) (Figure 3A) and an AAWAAA motif closely resembling the canonical AWTAAA signal ($e = 10^{-112}$) (Figure 3C) upstream of the polyadenylation sites (Zhao et al., 1999). We examined the distribution of these motifs across five polyadenylation site categories (C1-5) generated depending on site orientation (e.g., sense or antisense) and proximity relative to known 3′ ends of genes (Figure 3 and Experimental Procedures). Just like the canonical AWTAAA signal (Figure 3D and Table S6), TTTTTTTTT occurs in a highly position-specific manner ~21 nt upstream of the polyadenylation site (Figure 3B), suggesting that these motifs are mechanistically important for

**Table 2. Spearman Correlation Coefficients between Sense and Antisense Transcript Levels**

| Yeast | | | | |
|---|---|---|---|---|
| | Q1 | Q2 | Q3 | Q4 |
| Actual correlation | −0.11 | 0 | −0.01 | 0.36 |
| 1000 permutations, minimum | −2.39E-05 | −5.55E-05 | −3.79E-05 | −6.78E-05 |
| 1000 permutations, maximum | 9.05E-05 | 7.01E-05 | 8.47E-05 | 7.84E-05 |
| Human Liver | | | | |
| | Q1 | Q2 | Q3 | |
| Actual Correlation | −0.11 | 0.02 | 0.12 | |
| 1000 permutations, minimum | −9.59E-05 | −3.40E-05 | −9.00E-05 | |
| 1000 permutations, maximum | 9.25E-05 | 9.80E-06 | 5.69E-07 | |

Q1–4 indicates quartiles, with Q1 indicating the genes with highest sense expression values. For the human liver sample, we performed the analysis only for the top three quartiles since genes with zero expression level dominated the fourth quartile. The minimum and maximum correlation coefficients obtained after 1000 permutations were reported (thus $p < 0.001$). Similar trends were observed for yeast after the removal of potentially overlapping transcripts (see also Table S4).

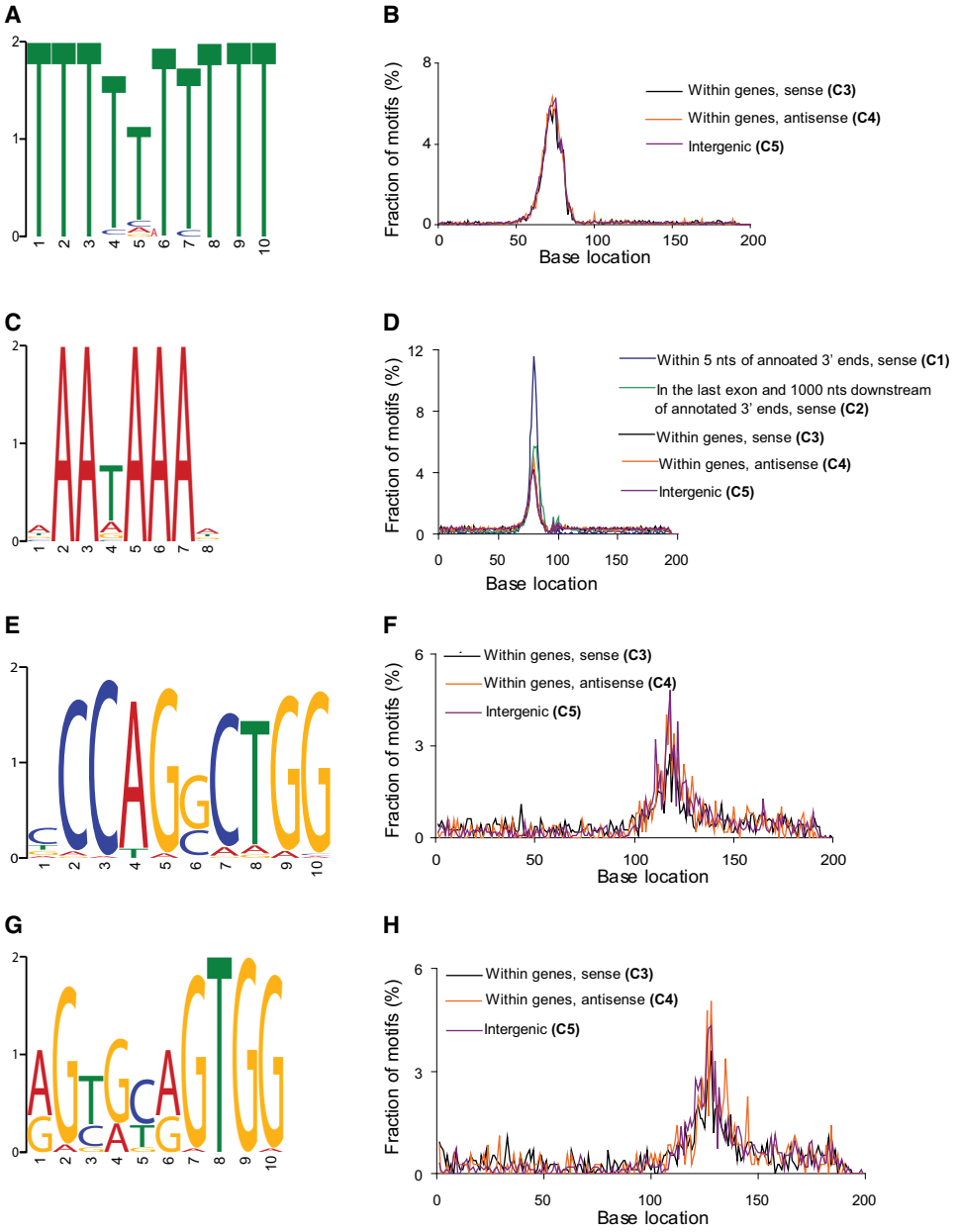


**Figure 3. Polyadenylation Motif Analyses**

Panels (A), (C), (E), and (G) indicate human motif elements identified. TTTTTTTTT (B), AWTAAA (D), CCAGSCTGG (F), and RGYRYRGTGG (H) distance distribution are shown in respective panels. Human categories were defined as sites that are within 5 nucleotides of annotated 3′ ends of known human genes in sense orientation (category 1), in the last exon and 1 kb downstream of annotated 3′ ends of human known genes in sense orientation (category 2), located anywhere within the transcripts in sense orientation except in categories 1 and 2 (category 3), antisense to genes (category 4) and in intergenic regions (category 5). In distance plots, *y* axis indicates the fraction of motifs (in percentages) at x-distances relative to the polyadenylation location (at base location 101) in each category. X-distances were calculated between the polyadenylation location identified with DRS and the first base immediately before the motif element. In panels B, F, and H, only the categories 3, 4, and 5 representing genic and intergenic sites were shown, because less than 10% (250–350) of these motifs were in categories 1 and 2, and not in sufficient numbers to be plotted in the graphs. Absolute numbers of motif counts for these latter three panels across all five human categories were provided in Figures S6A–S6C (see also Figure S4 and Table S5).

polyadenylation. However, the TTTTTTTTT motif is largely present in the genic and intergenic regions (C3-5 in Figure 3), unlike the canonical motif which is largely present near the annotated 3′ ends of genes (C1-2).

We also detected a novel palindromic sequence, CCAGSCTGG ($e = 10^{-33}$) (Figure 3E), downstream of the polyadenylation sites that manifests a strong position-specific pattern (Figure 3F). Further analysis using less stringent motif scans led

to the identification of RGYRYRGTGG (Figure 3G) that co-occur (p = ~0) with the CCAGSCTGG motif at a frequency of ~45%, and localizes ~31 nt downstream from the polyadenylation location (Figure 3H). Notably, we found that CCAGSCTGG and RGYRYRGTGG also strongly co-occur with the TTTTTTTTT motif (p = ~0) in the intergenic and genic regions (C3-5), whereas these motifs does not co-occur and anticorrelate with the canonical AWTAAA localization (p = ~0). The pervasive presence of the TTTTTTTTT motif in the novel genic and intergenic polyadenylation sites, its similarity to the AWTAA signal with respect to its positional preference, its anticorrelation to AWTAAA localization, and its co-occurrence with the CCAGSCTGG and RGYRYRGTGG motifs are intriguing and may point to uncharacterized polyadenylation mechanism(s) in humans. We applied similar approaches to yeast, detecting no additional motifs beyond the previously characterized positioning (PE, AAWAAA) and upstream efficiency (EE, TAYRTA) elements (Zhao et al., 1999). The general positioning of the upstream PE motif (Figure S5A) were closer to the cleavage site than the localization of the EE motif (Figure S5B), as expected (Zhao et al., 1999).

We then examined the nucleotide composition around the polyadenylation cleavage locations in each group. We observed a difference in the profiles of nucleotide frequency distributions surrounding human cleavage sites in regions near 3′ known gene ends (C1-2) and in genic and intergenic regions (C3-5, Figure 4). As expected, the categories 1 and 2 had the T-rich downstream sequence element (DSE) 20-30 bases downstream of the polyadenylation sites and A-rich sequences upstream (Zhao et al., 1999). On the other hand, the nucleotide profiles around the sites in the categories 3–5 were different and similar to the yeast sites (Figures S5C–S5F) with the pronounced upstream T-rich sequences, in line with the TTTTTTTTT motif identified in the upstream regions above (Figure 4). The presence of a T-rich polyadenylation enhancer sequence element upstream of the AATAAA motif is common among viruses and has been previously found in a few human genes (Bhat and Wold, 1987; Moreira et al., 1995). However, the T-rich pattern observed here is immediately upstream of the sense/antisense genic and intergenic cleavage sites, and therefore represents a different and novel observation. This latter similarity at the yeast and human nucleotide profiles prompted us to examine yeast motif presence in humans. Interestingly, we observed an enrichment of the yeast EE motif immediately upstream of the human cleavage sites in categories 3–5, but not in categories 1 and 2 (Figure 5). The yeast EE motif however does not co-occur with the novel CCAGSCTGG, RGYRYRGTGG, and TTTTTTTTT motifs identified above, and thus may be present in an independent subset of genic and intergenic sites. This latter finding may point to the existence of another, perhaps yeast-like, polyadenylation sequence structure in a subset of human polyadenylation sites.

## DISCUSSION

This study presents genome-wide polyadenylation maps that incorporate the accuracy of a high-throughput sequencing-based methodology and true strand-specificity. Other sequencing-based polyadenylation mapping approaches have recently become available (Mangone et al., 2010; Yoon and Brem, 2010). Compared to these approaches, the DRS-based approach is in quantitative nature, free of reverse transcription and ligation artifacts, and requires only minute RNA quantities. The nucleotide resolution of the approach is similar to other classic methods of polyadenylation site mapping. However, just like these other approaches, the DRS-based approach cannot truly differentiate cases where the template cleavage may occur right after an A-residue. Such sites may cause the resolution of the approach to elevate from its current level of ±2 nt. Because sequencing technologies available or in development today, including DRS, do not provide the full transcript sequence, it is not possible to know the sequence of the entire RNA molecule represented by each read by any sequencing technology. It is therefore possible that the reads found around the annotated transcriptional start and polyadenylation sites may partly represent short poly(A)$^+$ RNAs previously found to be associated with the gene termini (Kapranov et al., 2007a, 2010; Affymetrix ENCODE Transcriptome Project, 2009). A fraction of reads found around the annotated polyadenylation site of known messages may not represent the annotated form, but other isoforms or correspond to other overlapping transcripts that share the same polyadenylation region. Furthermore, polyadenylation sites observed downstream of annotated 3′ ends may represent alternative polyadenylation events or transcription termination products (Kim et al., 2004; Teixeira et al., 2004; West et al., 2004).

Our results show that most yeast and human transcripts have yet uncharacterized polyadenylation sites. This dataset, along with additional biological replicates and data from different cell types and states, will allow empirical annotation of such sites and provide the substrate for biological experimentation examining changes in these sites. The enrichment of reads in yeast intergenic functional transcription factor-binding sites and DHSs suggests that these potential regulatory regions may indeed encode for RNAs. The presence of RNAs from a subset of potential mammalian enhancers (eRNAs) and open chromatin regions has recently been described (De Santa et al., 2010; Kim et al., 2010; van Bakel et al., 2010). Unlike the report by Kim et al., (2010), which found eRNAs to lack poly(A) tails, our results indicate the potential existence of poly(A)-tail containing RNAs associated with regulatory elements in yeast. We speculate that these regulatory region-associated reads may represent a recently described class of polyadenylated noncoding RNAs that regulate gene expression (Bumgarner et al., 2009; Orom et al., 2010). They may also represent divergent transcription events from unannotated promoters (Neil et al., 2009; Seila et al., 2008; Xu et al., 2009). Alternatively, given the likely association of RNA polymerase II with the transcriptional factors binding to these regions, these RNAs may emerge from transcriptional noise events postulated to occur (Struhl, 2007). The lack of comprehensive transcription factor-binding site and enhancer maps in humans prevented us from examining such RNAs in our human studies. However, the relatively high fraction of intergenic DRS reads obtained in the human samples suggest that at least a fraction of these reads may emerge from enhancers. Further studies are needed to delineate the functions, if any, of these RNAs and how they may be contributing to regulatory function.

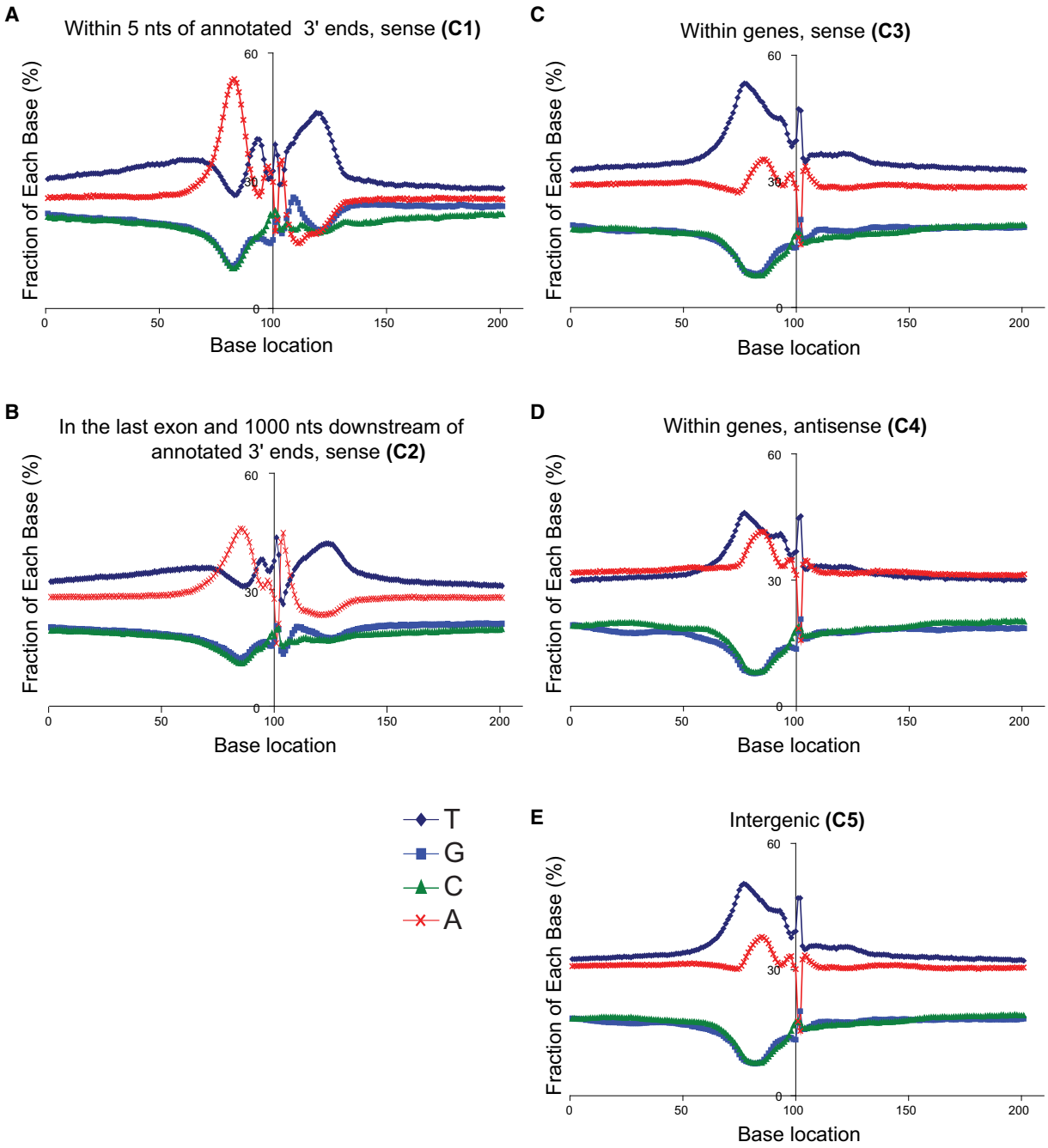


**Figure 4. The Nucleotide Composition Surrounding Polyadenylation Cleavage Locations in Humans**

(A–E) Category descriptions were provided in Figure 3. *y* axis indicates the nucleotide composition (in percentages) at x-locations relative to the cleavage positions (at base location 101). Dark blue (diamond), blue (rectangle), green (triangle), and red (cross) lines indicate T, G, C, and A nucleotides, respectively. Polyadenylation locations in C3-5 differ from those in C1-2, and exhibit elevated T and A content in 40–50 nt upstream of polyadenylation cleavage positions (see also Figure S5 and Table S6).

Our observation of novel polyadenylation patterns including novel co-occurring motifs (CCAGSCTGG, RGYRYRGTGG, and TTTTTTTTT) and enrichment of T-rich and yeast EE motif sequences near sites corresponding to noncoding transcript categories (antisense, sense genic, and intergenic) compared to sites in proximity to the 3′ ends of known genes suggest interesting possibilities for human polyadenylation. Particularly, the anticorrelation we observed between the localizations of the three novel motifs above and the canonical AWTAAA suggests alternative and yet to be characterized mechanisms of

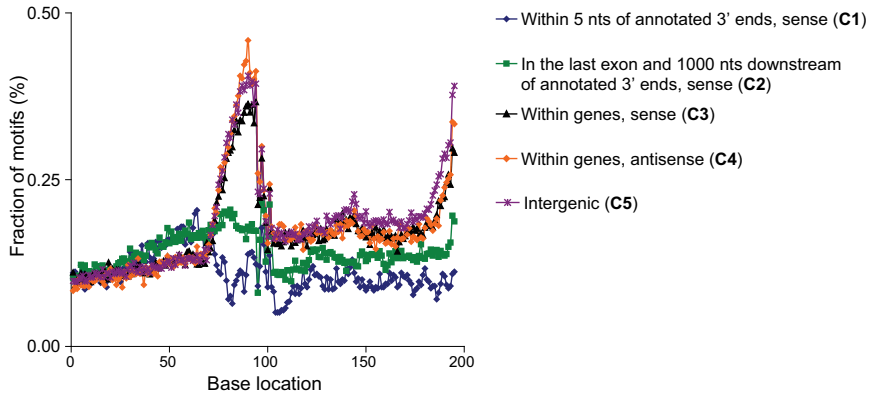


**Figure 5. Distance Distribution of Yeast EE (TAYRTA) Motif across Human Categories**

*y* axis indicates the fraction of motifs (in percentages) at x-distances relative to the cleavage positions (at base location 101) in each category. X-distances were calculated between the cleavage location identified with DRS and the first base immediately before the motif element. Human category descriptions were provided in Figure 3 legend. The enrichment of the EE motif immediately upstream of the cleavage sites in human categories 3, 4, and 5, but not in categories 1 and 2, is in parallel to the upstream human T-enrichment pattern shown in Figure 4 (see also Figure S6).

transcription termination, cleavage, and polyadenylation. Given that RNAs in these regions are likely to be noncoding, perhaps alternative modes of polyadenylation exist for noncoding RNAs. These three novel motifs are present in a relatively small fraction of polyadenylation sites and cleavage events (Table S6C). This may partly be explained by the relatively low fraction of polyadenylated noncoding RNAs relative to mRNAs of protein-coding genes in terms of mass. Combined with the recent observation that even very low abundance noncoding RNAs, as low as four copies per cell, can regulate target genes (Wang et al., 2008), these new motifs may be specific to such a subset of noncoding RNAs. Further in-depth de novo motif analyses in these novel regions and the identification of the components of this potential alternative polyadenylation machinery would open a number of conceptual and experimental possibilities. First, we may learn more about the RNAs they process, which may include various species (Buratowski, 2008) such as promoter-associated RNAs (Core et al., 2008; Neil et al., 2009; Seila et al., 2008), cryptic unstable RNAs (Preker et al., 2008; Wyers et al., 2005), long intergenic noncoding RNAs (Guttman et al., 2009), and polyadenylated RNAs resulting from degradation events (Slomovic et al., 2010). Second, we may get a more mechanistic understanding of polyadenylation and its connection with other cellular processes. For instance, the CCAGSCTGG palindromic motif identified here is a candidate binding site for human topoisomerase II (topo-II) (Spitzner and Muller, 1988). Topo-II is part of the RNA polymerase II holoenzyme and relaxes the superhelical tension that accumulates during transcription elongation (Mondal and Parvin, 2001). Perhaps the presence of such a motif downstream of polyadenylation sites is to ensure that transcriptional superhelical tension does not extend beyond the boundaries of the transcripts and thus do not disturb downstream regions.

In line with previous studies (He et al., 2008; Kapranov et al., 2007b), we observed that antisense transcription is prevalent in the yeast and human genomes and that the quantities of steady-state levels of sense and antisense transcripts occupying the same genomic space can negatively correlate with each other. Our results indicate a complex picture where the highly expressed genes in the top quartile tend to negatively correlate with the expression of antisense transcripts. On the other hand, the genes in the bottom quartile show a positive correlation between the sense and antisense transcription. Although both results are significant, the former effect is relatively small and similar to what has been detected previously (Chen et al., 2005), whereas the latter effect is the strongest (at least in yeast) and is similar to the results obtained in *Schizosaccharomyces pombe* (Dutrow et al., 2008) and mouse (Katayama et al., 2005), where positive correlation was found. In view of these results, it is perhaps not surprising that the correlation of sense and antisense transcripts has remained a controversial issue as often both were found to be positively correlated (Kapranov et al., 2007b). The relatively low negative correlation values most likely reflect the fact the overlapping positioning in the genome is only one of many ways to regulate stable levels of polydenylated RNAs species. It is however tempting to speculate that in highly expressed genes, the physical interference of converging RNA polymerase complexes could exert a dominant effect, whereas this possibility may be less of a factor in the genes with lower transcriptional activity. In the latter cases, other factors, such as chromatin accessibility that could permit transcription from both strands could be a larger determining factor. To what extent the observed negative correlation is due to sense/antisense transcripts occupying the same genomic space and/or other transcriptional control mechanisms needs further exploration.

This study represents the first step for the adaptation of the direct RNA sequencing technology to decipher the genome and its functions. Future studies will focus on the functional characterization of novel poly(A)$^+$ regulatory region-associated RNAs, antisense transcripts, and polyadenylation sites identified in this study, and the adaptation of DRS for other existing and novel RNA applications.

## EXPERIMENTAL PROCEDURES

### Sample Preparation for DRS

Yeast (*Saccharomyces cerevisiae*) and human liver poly(A)$^+$ RNAs were obtained from Clontech, CA (USA). Human brain total RNA was from Ambion. The 3′ blocking reaction was performed with poly(A) tailing kit (Ambion, TX, USA) and 3′deoxyATP (Jena Biosciences, Germany), incubating the reaction mixture at 37°C for 30 min. The blocked RNA was hybridized to flow cell

surfaces for sequencing with DRS without additional cleaning steps (Ozsolak et al., 2009).

### Data Analysis

Raw DRS reads were filtered using a suite of Helicos tools available at http://open.helicosbio.com/mwiki/index.php/Releases and described at http://open.helicosbio.com/helisphere_user_guide/. Alignments were conducted with indexDPgenomic available on the Helicos website (http://open.helicosbio.com/mwiki/index.php/Releases). For the genomic alignments, reads were aligned to the yeast SGD/sacCer2 and human NCBIv36 version of the genome supplemented with the complete ribosomal repeat unit (GenBank accession number U13369.1). Reads with a minimal length of 25 nt and alignment score of 4.3 and above were allowed. Aligned reads were further filtered for reads having a unique best alignment score. Total raw per base error rate was 4%–5%, dominated by missing base errors (2%–3%).

Downstream analysis was performed with the SeqSolve NGS software (Integromics, Spain). Annotated yeast or human transcriptome was defined as either the SGD Genes from *Saccharomyces* Genome Database track or UCSC Genes track on the UCSC Genome Browser. Counts within each annotation were derived from either the sense or antisense strand using the positions of the 5′ ends of reads aligned to the appropriate strand. Yeast median UTR length was calculated by taking the median of the distances between the annotated 3′ end locations of yeast ORFs and the reads that map in the sense orientation and within 1000 bp downstream of ORF 3′ ends.

For the sequence composition surrounding polyadenylation cleavage site analysis, the 5′ ends of reads representing the 3′ cleavage sites were grouped based on overlap with the genomic annotation, as described in Figure 3 and Figure S5. Mitochondrial reads were not used for the sequence analysis. These categories for human were (1) sense cleavage locations that are within 5 bases of annotated 3′ ends, (2) sense cleavage locations that are not in category #1 and are in the last exons or 1 kb downstream of the annotated 3′ ends, (3) sense cleavage locations that are not in categories 1 and 2 and are within annotated genes, (4) antisense cleavage locations that are within annotated genes, and (5) intergenic cleavage locations that are not in categories 1–4. The categories for yeast were (1) sense cleavage locations that are located within 200 bp downstream of the annotated 3′ ends of yeast ORFs, (2) sense cleavage locations that are not within category 1 and are within bodies of ORFs, (3) antisense cleavage locations that are not within category 1 and are within bodies of ORFs, and (4) intergenic cleavage locations that are not in categories 1, 2, and 3, and are at least 1 kb away from the 3′ ends of yeast ORFs. Reads in each category were then collapsed according to their unique 5′ ends representing unique polyadenylation cleavage locations. Sequences 100 bases on each side of each collapsed locations were analyzed as described in the text.

### Detection of Novel Motifs

To investigate the presence of new sequence motifs, upstream and downstream genomic sequences (50 bases) of novel polyadenylation sites (Figure S4) were scanned independently using MEME (Bailey et al., 2006). To reduce the occurrence of spurious motifs, motif searches were performed using a highly stringent E-value ($10^{-25}$) threshold, based on a nonredundant set of 1000 sequences that were sampled uniformly from the complete set of upstream/downstream sequences. The threshold ($10^{-25}$) was used because even when sites across each chromosome was separately analyzed (24 control experiments) to rule out dataset artifacts, the three human motifs were consistently detected. The various motif variants were manually inspected to select a single motif for display representation. For additional validations of the motifs, the up/downstream occurrences and co-occurrences were analyzed. Total occurrences of motifs in up/downstream sequences were determined by searching for all short strings that matched (>90%) the position-specific scoring (log-odds) matrix profile of the motifs detected by MEME. To test the statistical significance of co-occurrence between two motifs, hypergeometric tests (Lee et al., 2007) were performed based on the total number of occurrences of the two motifs in the complete set of nonredundant sequences. Because only four motifs were compared (six comparisons) to each other for co-occurrence analysis, and because the reported p values are close to zero, the Bonferroni correction factor of 6 was not used.

## ACCESSION NUMBERS

Sequencing datasets described in this study have been deposited at the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), accession no SRA012232. The datasets are also available as wiggle files at the Helicos open access website (HeliSphere, http://open.helicosbio.com/) along with yeast and human polyadenylation sites defined in this study.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, six figures, and six tables and can be found with this article online at doi:10.1016/j.cell.2010.11.020.

## ACKNOWLEDGMENTS

We thank our Helicos and Integromics colleagues for technical assistance and discussions. This work was supported by the National Human Genome Research Institute (grant R01-HG005230 to F.O. and P.M.M.). B.J. is supported by the National Institutes of Health (grant GM079756) and the American Cancer Society (grant RSG0905401), A.P.M. is supported by the National Institutes of Health (grant MH60774), and S.F. is supported by the Spanish Ministry of Science and Innovation—FEDER (CDTI loan IDI-20091293). F.O., P.K., and P.M.M. are employees of Helicos BioSciences Corporation. S.F. is an employee of Integromics.

Received: May 26, 2010
Revised: September 28, 2010
Accepted: November 9, 2010
Published: December 9, 2010

## REFERENCES

Andreassi, C., and Riccio, A. (2009). To localize or not to localize: mRNA fate is in 3′UTR ends. Trends Cell Biol. *19*, 465–474.

Bailey, T.L., Williams, N., Misleh, C., and Li, W.W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. *34*, W369–W373.

Bennett, C.L., Brunkow, M.E., Ramsdell, F., O'Briant, K.C., Zhu, Q., Fuleihan, R.L., Shigeoka, A.O., Ochs, H.D., and Chance, P.F. (2001). A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA→AAUGAA) leads to the IPEX syndrome. Immunogenetics *53*, 435–439.

Bhat, B.M., and Wold, W.S. (1987). A small deletion distant from a splice or polyadenylation site dramatically alters pre-mRNA processing in region E3 of adenovirus. J. Virol. *61*, 3938–3945.

Brais, B., Bouchard, J.P., Xie, Y.G., Rochefort, D.L., Chretien, N., Tome, F.M., Lafreniere, R.G., Rommens, J.M., Uyama, E., Nohira, O., et al. (1998). Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nat. Genet. *18*, 164–167.

Bumgarner, S.L., Dowell, R.D., Grisafi, P., Gifford, D.K., and Fink, G.R. (2009). Toggle involving cis-interfering noncoding RNAs controls variegated gene expression in yeast. Proc. Natl. Acad. Sci. USA *106*, 18321–18326.

Buratowski, S. (2008). Transcription: gene expression–where to start? Science *322*, 1804–1805.

Chen, J., Sun, M., Hurst, L.D., Carmichael, G.G., and Rowley, J.D. (2005). Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. Trends Genet. *21*, 326–329.

Cocquet, J., Chong, A., Zhang, G., and Veitia, R.A. (2006). Reverse transcriptase template switching and false alternative transcripts. Genomics *88*, 127–131.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848.

Costessi, L., Devescovi, G., Baralle, F.E., and Muro, A.F. (2006). Brain-specific promoter and polyadenylation sites of the beta-adducin pre-mRNA generate an unusually long 3′-UTR. Nucleic Acids Res. *34*, 243–253.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. (2006). A high-resolution map of transcription in the yeast genome. Proc. Natl. Acad. Sci. USA *103*, 5320–5325.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. *8*, e1000384.

Dutrow, N., Nix, D.A., Holt, D., Milash, B., Dalley, B., Westbroek, E., Parnell, T.J., and Cairns, B.R. (2008). Dynamic transcriptome of *Schizosaccharomyces pombe* shown by RNA-DNA hybrid mapping. Nat. Genet. *40*, 977–986.

Faghihi, M.A., and Wahlestedt, C. (2009). Regulatory roles of natural antisense transcripts. Nat. Rev. Mol. Cell Biol. *10*, 637–643.

Gehring, N.H., Frede, U., Neu-Yilik, G., Hundsdoerfer, P., Vetter, B., Hentze, M.W., and Kulozik, A.E. (2001). Increased efficiency of mRNA 3′ end formation: a new genetic mechanism contributing to hereditary thrombophilia. Nat. Genet. *28*, 389–392.

Graber, J.H., McAllister, G.D., and Smith, T.F. (2002). Probabilistic prediction of *Saccharomyces cerevisiae* mRNA 3′-processing sites. Nucleic Acids Res. *30*, 1851–1858.

Gubler, U. (1987). Second-strand cDNA synthesis: mRNA fragments as primers. Methods Enzymol. *152*, 330–335.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

He, Y., Vogelstein, B., Velculescu, V.E., Papadopoulos, N., and Kinzler, K.W. (2008). The antisense transcriptomes of human cells. Science *322*, 1855–1857.

Hesselberth, J.R., Chen, X., Zhang, Z., Sabo, P.J., Sandstrom, R., Reynolds, A.P., Thurman, R.E., Neph, S., Kuehn, M.S., Noble, W.S., et al. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat. Methods *6*, 283–289.

Higgs, D.R., Goodbourn, S.E., Lamb, J., Clegg, J.B., Weatherall, D.J., and Proudfoot, N.J. (1983). Alpha-thalassaemia caused by a polyadenylation signal mutation. Nature *306*, 398–400.

Iseli, C., Stevenson, B.J., de Souza, S.J., Samaia, H.B., Camargo, A.A., Buetow, K.H., Strausberg, R.L., Simpson, A.J., Bucher, P., and Jongeneel, C.V. (2002). Long-range heterogeneity at the 3′ ends of human mRNAs. Genome Res. *12*, 1068–1074.

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermuller, J., Hofacker, I.L., et al. (2007a). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science *316*, 1484–1488.

Kapranov, P., Willingham, A.T., and Gingeras, T.R. (2007b). Genome-wide transcription and the implications for genomic organization. Nat. Rev. Genet. *8*, 413–423.

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

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., et al. (2005). Antisense transcription in the mammalian transcriptome. Science *309*, 1564–1566.

Kim, M., Krogan, N.J., Vasiljeva, L., Rando, O.J., Nedea, E., Greenblatt, J.F., and Buratowski, S. (2004). The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. Nature *432*, 517–522.

Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182–187.

Kiyosawa, H., Mise, N., Iwase, S., Hayashizaki, Y., and Abe, K. (2005). Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. Genome Res. *15*, 463–474.

Lee, J., Li, Z., Brower-Sinning, R., and John, B. (2007). Regulatory circuit of human microRNA biogenesis. PLoS Comput. Biol. *3*, e67.

Lin, C.L., Bristol, L.A., Jin, L., Dykes-Hoberg, M., Crawford, T., Clawson, L., and Rothstein, J.D. (1998). Aberrant RNA processing in a neurodegenerative disease: the cause for absent EAAT2, a glutamate transporter, in amyotrophic lateral sclerosis. Neuron *20*, 589–602.

Liu, D., and Graber, J.H. (2006). Quantitative comparison of EST libraries requires compensation for systematic biases in cDNA generation. BMC Bioinformatics *7*, 77.

Lopez, F., Granjeaud, S., Ara, T., Ghattas, B., and Gautheret, D. (2006). The disparate nature of "intergenic" polyadenylation sites. RNA *12*, 1794–1801.

Lutz, C.S. (2008). Alternative polyadenylation: a twist on mRNA 3′ end formation. ACS Chem. Biol. *3*, 609–617.

Mahadevan, S., Raghunand, T.R., Panicker, S., and Struhl, K. (1997). Characterisation of 3′ end formation of the yeast HIS3 mRNA. Gene *190*, 69–76.

Mamanova, L., Andrews, R.M., James, K.D., Sheridan, E.M., Ellis, P.D., Langford, C.F., Ost, T.W., Collins, J.E., and Turner, D.J. (2010). FRT-seq: amplification-free, strand-specific transcriptome sequencing. Nat. Methods *7*, 130–132.

Mangone, M., Manoharan, A.P., Thierry-Mieg, D., Thierry-Mieg, J., Han, T., Mackowiak, S.D., Mis, E., Zegar, C., Gutwein, M.R., Khivansara, V., et al. (2010). The landscape of *C. elegans* 3′UTRs. Science *329*, 432–435.

Mondal, N., and Parvin, J.D. (2001). DNA topoisomerase IIalpha is required for RNA polymerase II transcription on chromatin templates. Nature *413*, 435–438.

Moreira, A., Wollerton, M., Monks, J., and Proudfoot, N.J. (1995). Upstream sequence elements enhance poly(A) site efficiency of the C2 complement gene and are phylogenetically conserved. EMBO J. *14*, 3809–3819.

Muro, E.M., Herrington, R., Janmohamed, S., Frelin, C., Andrade-Navarro, M.A., and Iscove, N.N. (2008). Identification of gene 3′ ends by automated EST cluster analysis. Proc. Natl. Acad. Sci. USA *105*, 20286–20290.

Nagaike, T., Suzuki, T., Katoh, T., and Ueda, T. (2005). Human mitochondrial mRNAs are stabilized with polyadenylation regulated by mitochondria-specific poly(A) polymerase and polynucleotide phosphorylase. J. Biol. Chem. *280*, 19721–19727.

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

Neil, H., Malabat, C., d'Aubenton-Carafa, Y., Xu, Z., Steinmetz, L.M., and Jacquier, A. (2009). Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature *457*, 1038–1042.

Orkin, S.H., Cheng, T.C., Antonarakis, S.E., and Kazazian, H.H., Jr. (1985). Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene. EMBO J. *4*, 453–456.

Orom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., et al. (2010). Long noncoding RNAs with enhancer-like function in human cells. Cell *143*, 46–58.

Ozsolak, F., Platt, A.R., Jones, D.R., Reifenberger, J.G., Sass, L.E., McInerney, P., Thompson, J.F., Bowers, J., Jarosz, M., and Milos, P.M. (2009). Direct RNA sequencing. Nature *461*, 814–818.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science *322*, 1851–1854.

Affymetrix ENCODE Transcriptome Project; Cold Spring Harbor Laboratory ENCODE Transcriptome Project. (2009). Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. Nature *457*, 1028–1032.

Rigault, C., Le Borgne, F., and Demarquoy, J. (2006). Genomic structure, alternative maturation and tissue expression of the human BBOX1 gene. Biochim. Biophys. Acta *1761*, 1469–1481.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849–1851.

Slomovic, S., Portnoy, V., and Schuster, G. (2008). Detection and characterization of polyadenylated RNA in Eukarya, Bacteria, Archaea, and organelles. Methods Enzymol. *447*, 501–520.

Slomovic, S., Fremder, E., Staals, R.H., Pruijn, G.J., and 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. USA *107*, 7407–7412.

Spiegelman, S., Burny, A., Das, M.R., Keydar, J., Schlom, J., Travnicek, M., and Watson, K. (1970). DNA-directed DNA polymerase activity in oncogenic RNA viruses. Nature *227*, 1029–1031.

Spitzner, J.R., and Muller, M.T. (1988). A consensus sequence for cleavage by vertebrate DNA topoisomerase II. Nucleic Acids Res. *16*, 5533–5556.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat. Struct. Mol. Biol. *14*, 103–105.

Teixeira, A., Tahiri-Alaoui, A., West, S., Thomas, B., Ramadass, A., Martianov, I., Dye, M., James, W., Proudfoot, N.J., and Akoulitchev, A. (2004). Autocatalytic RNA cleavage in the human beta-globin pre-mRNA promotes transcription termination. Nature *432*, 526–530.

Tian, B., Hu, J., Zhang, H., and Lutz, C.S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res. *33*, 201–212.

Tian, B., Pan, Z., and Lee, J.Y. (2007). Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. Genome Res. *17*, 156–165.

van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2010). Most "dark matter" transcripts are associated with known genes. PLoS Biol. *8*, e1000371.

Wang, X., Arai, S., Song, X., Reichart, D., Du, K., Pascual, G., Tempst, P., Rosenfeld, M.G., Glass, C.K., and Kurokawa, R. (2008). Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. Nature *454*, 126–130.

West, S., Gromak, N., and Proudfoot, N.J. (2004). Human 5′–> 3′ exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. Nature *432*, 522–525.

Wu, J.Q., Du, J., Rozowsky, J., Zhang, Z., Urban, A.E., Euskirchen, G., Weissman, S., Gerstein, M., and Snyder, M. (2008). Systematic analysis of transcribed loci in ENCODE regions using RACE sequencing reveals extensive transcription in the human genome. Genome Biol. *9*, R3.

Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.C., Dufour, M.E., 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.

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Munster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L.M. (2009). Bidirectional promoters generate pervasive transcription in yeast. Nature *457*, 1033–1037.

Yoon, O.K., and Brem, R.B. (2010). Noncanonical transcript forms in yeast and their regulation during environmental stress. RNA *16*, 1256–1267.

Zhang, H., Hu, J., Recce, M., and Tian, B. (2005a). PolyA_DB: a database for mammalian mRNA polyadenylation. Nucleic Acids Res. *33*, D116–D120.

Zhang, H., Lee, J.Y., and Tian, B. (2005b). Biased alternative polyadenylation in human tissues. Genome Biol. *6*, R100.

Zhao, J., Hyman, L., and Moore, C. (1999). Formation of mRNA 3′ ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol. Mol. Biol. Rev. *63*, 405–445.