Species-specific factors mediate extensive heterogeneity of mRNA 3′ ends in yeasts

Zarmik Moqtaderi[1], Joseph V. Geisberg[1], Yi Jin, Xiaochun Fan[2], and Kevin Struhl[3]

Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115

Contributed by Kevin Struhl, May 22, 2013 (sent for review April 30, 2013)

Most eukaryotic genes express mRNAs with alternative polyadenylation sites at their 3′ ends. Here we show that polyadenylated 3′ termini in three yeast species (*Saccharomyces cerevisiae*, *Kluyveromyces lactis*, and *Debaryomyces hansenii*) are remarkably heterogeneous. Instead of a few discrete 3′ ends, the average yeast gene has an "end zone," a >200 bp window with >60 distinct poly(A) sites, the most used of which represents only 20% of the mRNA molecules. The pattern of polyadenylation within this zone varies across species, with *D. hansenii* possessing a higher focus on a single dominant point closer to the ORF terminus. Some polyadenylation occurs within mRNA coding regions with a strong bias toward the promoter. The polyadenylation pattern is determined by a highly degenerate sequence over a broad region and by a local sequence that relies on A residues after the cleavage point. Many dominant poly(A) sites are predicted to adopt a common secondary structure that may be recognized by the cleavage/polyadenylation machinery. We suggest that the end zone reflects a region permissive for polyadenylation, within which cleavage occurs preferentially at the A-rich sequence. In *S. cerevisiae* strains, *D. hansenii* genes adopt the *S. cerevisiae* polyadenylation profile, indicating that the polyadenylation pattern is mediated primarily by species-specific factors.

3′ end formation | transcription termination | evolution | gene expression | mRNA processing

The 3′ ends of eukaryotic mRNAs are generated by cleavage of the nascent transcript and the addition of A residues, resulting in a polyadenylated tail. In a wide variety of eukaryotic organisms, most genes express mRNAs with alternative polyadenylation sites (1–7). In metazoans, poly(A) site selection is determined primarily by an AAUAAA motif in the RNA 10–30 nt upstream of the polyadenylation site (8–10). A less-conserved U-rich element downstream of the cleavage site also plays a role in poly(A) site selection (9–11). In contrast, sequence preferences for poly(A) site selection in the yeast *Saccharomyces cerevisiae* are considerably less strict, and several degenerate sequence determinants have been identified: an AU-rich efficiency element located a variable distance upstream of the poly(A) site, an A-rich positioning element 10–30 nt upstream of the cleavage site, and two U-rich sequences, one on each side of the cleavage site (10–12). The 3′ end formation near these sites is accomplished by an apparatus of ∼20 proteins organized into subcomplexes including cleavage and polyadenylation factor (CPF), cleavage factor IA (CF1A), and CF1B (11, 13). However, the precise sequence requirements for polyadenylation are poorly understood, with respect to both why it occurs selectively downstream of the protein-coding region and how specific sites are chosen.

Although the sequence determinants for polyadenylation differ between mammals and *S. cerevisiae*, it is unknown to what extent the specificity of polyadenylation is conserved. In general, some sequence-dependent processes are conserved, whereas others are not. For instance, there are many examples of transcription factors with indistinguishable DNA-binding specificities and homologous DNA-binding domains in a wide variety of eukaryotic species. Within core promoters, TATA-binding protein (TBP) recognition of the TATA element is functionally conserved from yeast to human, but the Initiator and other downstream promoter elements are not.

Even in cases of conserved sequence-dependent processes, species-specific factors can interpret DNA sequences in different manners. For example, although TBP is the major sequence-specific component of structurally homologous preinitiation complexes, the pattern of mRNA initiation sites varies among species. Similarly, although the general pattern of nucleosome positioning is strongly conserved across yeast species, there are species-specific differences in nucleosome spacing and other parameters (14). As determined by analysis of foreign yeast DNA (on artificial chromosomes) in *S. cerevisiae*, most (but not all) aspects of nucleosome positioning are due to species-specific factors (15). Thus, it is impossible to predict whether the sequences necessary for polyadenylation are conserved across yeast species or to predict whether polyadenylation factors are species-specific for determining polyadenylation sites.

Here, we use direct RNA sequencing to identify polyadenylation sites in three yeast species. In all cases, polyadenylation sites downstream of ORFs are remarkably heterogeneous, and a very small minority of poly(A) sites are located near promoters. In all species tested, the polyadenylation pattern is determined by a highly degenerate sequence over a broad region and by a previously unknown local element that relies on A residues after the cleavage point. In addition, many dominant poly(A) sites are predicted to adopt a common secondary structure that is much less frequently observed among poly(A) sites in general. Despite these overall similarities, the pattern of poly(A) sites differs among these yeast species. Using a functional evolutionary approach involving expression of heterologous genomic regions in *S. cerevisiae* (15), we demonstrate that polyadenylation factors (and not the underlying sequences) are primarily responsible for the species-specific pattern of poly(A) sites.

Results

Numerous Polyadenylation Sites Occur Within a Wide "End Zone." We used direct RNA sequencing (16) to determine the 3′ ends of polyadenylated RNA isolated from three yeast species growing exponentially in rich medium. In this technique, the poly(A) tail of an individual mRNA molecule is hybridized to an immobilized oligo-dT primer to direct sequencing of the RNA without the need for intermediate amplification or cDNA synthesis. This method obviates biases introduced by PCR amplification, and hence can reveal minor species that would be undetectable by conventional strategies. For each experiment, we generated ∼8 million mapped

Author contributions: Z.M., J.V.G., Y.J., X.F., and K.S. designed research; Z.M., J.V.G., Y.J., and X.F. performed research; Z.M. contributed new reagents/analytic tools; Z.M., J.V.G., Y.J., and K.S. analyzed data; and Z.M., J.V.G., and K.S. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The sequences reported in this paper have been deposited in the NCBI GEO database (accession no. GSE47661).

[1]Z.M. and J.V.G. contributed equally to this work.

[2]Present address: Biochemical Sciences and Engineering, Central Research and Development, E. I. du Pont de Nemours and Company, Wilmington, DE 19880.

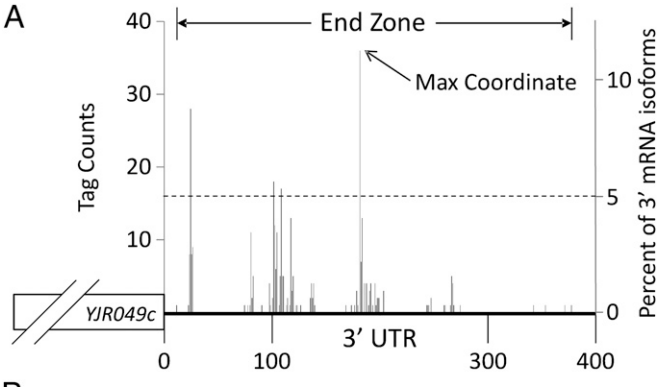
[3]To whom correspondence should be addressed. E-mail: kevin@hms.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1309384110/-/DCSupplemental.

sequence reads. A similarly sized dataset for *S. cerevisiae* RNAs has been obtained previously (17), but not analyzed in the manner described below. This approach measures the steady-state levels of the many individual mRNA isoforms, which result from both their originally produced levels and from their stabilities in vivo.

Although previous studies have generally described a small number of distinct poly(A) sites per gene (18), we observe a striking heterogeneity of mRNA 3′ termini at virtually all *S. cerevisiae* genes. The genome-wide median is 62 distinct polyadenylated 3′ termini per protein-coding gene; these occur over an unexpectedly wide (median 269 nt) region in the 3′ untranslated region (3′UTR) of the gene (Fig. 1*A*). We have termed this window the end zone.

Several observations indicate that the striking heterogeneity reflects real poly(A) sites that are used in vivo and is not an experimental artifact. First, dominant RNA poly(A) species correlate well with termination sites defined in previous work using older, less sensitive technologies (19). Second, poly(A) sites in coding regions occur at a much lower frequency than in 3′UTRs, and when observed they cluster near the 5′ ends of the mRNAs (see below). Third, because the RNA is sequenced directly from individual molecules without amplification, the minor species are not artifacts of any amplification step. Fourth, many preferred termination sites are surrounded by lesser used sites that occur at comparable levels both upstream and downstream of the preferred site. This is inconsistent with directional artifacts due to systematic errors, such as sequencing erroneously initiating from the penultimate nucleotide instead of the last one before the poly(A) tail. Fifth, the naturally occurring 14-A stretch in the *TFC1* locus is insufficient to generate internally primed sequence reads, and we observe essentially no sequence reads just upstream of other poly(dA) stretches that occur in the genome. Lastly, as will be described elsewhere, different mRNA isoforms of the same gene have different half-lives, indicating that they are of biological, as opposed to experimental, origin. Thus, the extreme diversity of mature 3′ ends observed here reflects true promiscuity of the poly(A) machinery within the end zone.



B

	S. cerevisiae	*K. lactis*	*D. hansenii*
Width of end zone	269	260	280
Most proximal coordinate (position vs ORF stop)	36	40	24
Most distal coordinate (position vs ORF stop)	320	323	322
Number of major coordinates each accounting for at least 10% of total	2	2	2
Most-used coordinate (position vs ORF stop)	121	121	88
Percentage of overall signal found at most-used coordinate	20.2	23.3	29.6
Number of coordinates with ≥5% of the signal of the most-used coordinate	18	14	10
Region in bp spanned by these "5% of max" coordinates	79	60	42

Fig. 1. Schematic representation of a typical protein-coding gene and end zone statistics. (*A*) Poly(A) site distribution of a representative locus, *YJR049C*. Reads initiating from each coordinate are shown as a percentage of all reads (right axis) and in absolute magnitude (left axis). (*B*) Properties of end zones in three yeast species. Median values are shown for all protein-coding genes with at least 100 total sequence reads in the 400-bp region immediately downstream of the ORF. The analysis excludes genes whose 5′ and 3′ neighbors on the same strand are less than 500 bp away.

End Zone Properties Differ in Three Yeast Species. The *S. cerevisiae* end zone has a median length of 269 and ranges from 21 to 399. The median most highly preferred position for polyadenylation is at 121 past the stop codon (Fig. 1*B*), but it can occur anywhere within the window. The median most proximal polyadenylated RNA isoform in the 3′UTR is at 36 after the translation stop, and the median most distal poly(A) site occurs at 320 after the translation stop. On average, there are two major polyadenylated isoforms each representing at least 10% of the overall total. The most frequently used poly(A) site in the 3′UTR typically accounts for just 20% of the RNA derived from sites within 400 bp after the ORF end. Thus, >50% of the 3′ ends for a typical *S. cerevisiae* gene arise from numerous minor poly(A) sites over the entire end zone. *Kluyveromyces lactis* exhibits a polyadenylation profile largely similar to that of *S. cerevisiae*, with an end zone of comparable width and a median 47 different poly(A) sites. In contrast, the more evolutionarily distant yeast *Debaryomyces hansenii* has a distinct profile consisting of end zones beginning closer to the ORF and a higher concentration of signal at the most preferred coordinate (median 29.6% of the total, positioned 88 bp downstream of the translational stop) (Fig. 1*B*). In *D. hansenii*, the span of moderately used sites (≥5% of the most preferred site in a given gene) is appreciably narrower than in *S. cerevisiae* (42 vs. 79 nt).

Polyadenylation Sites Outside of 3′UTRs Are Biased to the Beginning of the Coding Region. As mentioned above, a small amount of polyadenylation occurs within coding regions, at a level averaging 4% of that observed in the 3′UTR in steady state. Polyadenylation can also occur outside of 3′UTRs in other species, and cryptic poly(A) sites in introns can be suppressed by U1 snRNP, which also plays a role in mRNA isoform selection (20, 21). Strikingly, polyadenylation within the first 15 nucleotides of the ORF occurs fivefold more frequently than within the rest of the coding region (Fig. 2*A*). In the 3′UTRs of all three yeasts, positions used for polyadenylation are more widely distributed, even though in *D. hansenii* the distribution of sites is skewed closer to the 3′ end of the ORF (Fig. 2*B*). We also observe low-level polyadenylation of snoRNAs and tRNAs, which is suspected to be generated by the TRAMP complex that adds short oligo(A) sequences to abortive RNAs and targets them for degradation (22–24). In this regard, sequence preferences surrounding poly(A) sites at tRNAs are very different from those surrounding poly(A) sites at 3′UTRs (see below and Fig. S1). However, poly(A) sites at the beginning of the ORF and at snoRNAs have similar sequence preferences to those at 3′UTRs (Fig. S1), suggesting that they may arise from the standard polyadenylation machinery.

Sequence Preferences Around Major Polyadenylation Sites. To identify sequence determinants that contribute to poly(A) site selection, we analyzed nucleotide frequencies over the 200-nt sequence centered on the single-most preferred cleavage/polyadenylation position for every protein-coding gene (Fig. 3 *A*–*C*; frequencies of all possible dinucleotides are shown in Fig. S2). In *S. cerevisiae*, the 100 nt region upstream of the cleavage site is slightly enriched in A and U residues relative to the region downstream. In addition, there is an 11-nt A-rich stretch between –24 and –13 relative to the poly(A) site, and U-rich segments upstream (–11 to –1) and downstream of the cleavage site (+7 to +21). These correspond fairly well with the previously identified efficiency, positioning, and U-rich elements (25).

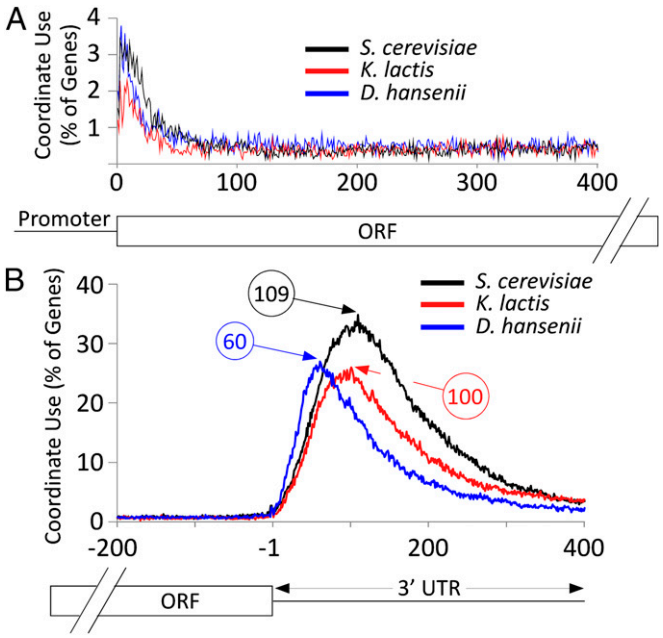


Fig. 2. Polyadenylation in 5′ ends of ORFs and in 3′UTRs of yeast. (*A*) Percentage of protein-coding genes with any polyadenylation at the indicated coordinates relative to the ORF start. (*B*) Polyadenylation in 3′UTRs, with the most frequently occurring poly(A) sites circled. In this analysis, a position is counted only once for each gene regardless of the number of reads initiating there. To avoid confusion by any possible read-through transcripts, we performed this analysis using only that subset of protein-coding genes separated by at least 500 nucleotides from any 5′-flanking gene on the same strand.

Unexpectedly, the second nucleotide downstream of the cleavage site is very strongly enriched for an A residue, and this enrichment for A continues for an additional three nucleotides. More importantly, these A residues at +2 to +5 are the most critical determinant of relative poly(A) site strength. Although sequence elements more distant from the cleavage site are essentially unchanged between the most preferred and the 10th most preferred cleavage sites for individual genes, the most preferred sites have a much higher frequency of these A residues (Fig. 3*D*). As independent confirmation of the functional importance of these A residues, exceptionally dominant poly(A) sites (accounting for >40% of a gene's mRNA molecules, which occurs in ~300 genes) almost invariably possess an A at +2, and are significantly more enriched for an A at +3 and +4 (Fig. S3). In addition, at such dominant poly(A) sites, there is also a significant ($P < 10^{-3}$) enrichment for U residues immediately upstream of the cleavage site.

Despite their considerable evolutionary distance from *S. cerevisiae*, *K. lactis* and *D. hansenii* possess remarkably similar general nucleotide preference patterns in the vicinity of dominant poly(A) sites (Fig. 3 *B* and *C*). All three yeasts share an extreme preference for an A base at the +2 position downstream of the cleavage site. The presence of an A at the +2 position and to a lesser extent at the +3 and +4 positions is a primary determinant of the strength of a particular poly(A) site. As in *S. cerevisiae*, the most dominant poly (A) sites almost uniformly possess an A in the +2 position. Although the global similarities among the consensus motifs in the three yeasts are striking, some subtle differences are evident upon close inspection, especially close to the cleavage position (Fig. S4). Unlike *K. lactis*, in which sequence composition around the cleavage site is nearly identical to that in *S. cerevisiae*, *D. hansenii* shows enrichment in A residues and a drop in G/C use immediately upstream of its cleavage sites, consistent with its greater evolutionary divergence from *S. cerevisiae*.

Motif Scores Predict Preferential Polyadenylation in 3′UTRs. Given that the polyadenylation machinery operates over wide zones with different sequences and that the few critical A residues have limited information content, why does polyadenylation not occur more frequently within the ORF? To address this issue, we generated a position weight matrix of a 51 nt region centered on the most dominant poly(A) sites (Fig. S5). Based on this matrix, we obtained poly(A) motif scores for all 51 nt regions over the entire genome. Importantly, motif scores are much higher in the 3′UTR than in the ORF, with the highest scores occurring >50 nt after the stop codon (Fig. 3*E*). This observation explains why poly(A) sites are much less frequent in ORFs and why end zones typically do not begin immediately after the end of the coding sequence. In accord with this suggestion, in the 46 ORFs whose end zones start particularly close to (within 30 nt of) the ORF end, the motif scores are unusually high in the latter part of the ORF (Fig. 3*F*). Conversely, high motif scores are shifted further downstream in the subset of genes whose end zones start most distal from the ORF. Lastly, the motif score is relatively high in the 5′UTR and decreases precipitously after the ATG, consistent with the observation that poly(A) sites are most prevalent in the first few nucleotides of the ORF.

Structural Motif Associated Preferentially with Dominant Polyadenylation Sites. To determine whether the sequence patterns near favored poly(A) sites are reflective of a structural motif recognized by the polyadenylation machinery, we used mfold (26) to model the RNA secondary structure in the vicinity of dominant and weak sites. Interestingly, many of the most dominant poly(A) sites are predicted to adopt a similar structure in which the poly(A) addition site is within a double-stranded stem and is immediately adjacent to a single-stranded region (Fig. 3*G*). Weaker poly(A) sites are predicted significantly less often to conform to this structural motif ($P = 4.3 \times 10^{-6}$), and they do so at a frequency similar to or slightly above the background occurrence of the structure in 3′UTRs. At weak sites that do form the structure, the double-stranded stem encompassing the poly(A) site is significantly less stable than at strong sites (average ΔG –5.7 vs. –7.4, $P = 5.7 \times 10^{-3}$). (Fig. 3*H*). Although these results clearly indicate a structural component to poly(A) site selection, the specific structures involved in this process remain to be determined.

Transplanted Foreign Yeast DNA Assumes the Polyadenylation Profile of the Host Strain. The subtle distinctions in poly(A) site selection among the yeast species make it possible to test whether RNA sequence or *trans*-acting factors are the major determinant of cleavage specificity. Specifically, we compared the polyadenylation pattern of large chromosomal segments from *D. hansenii* that had been introduced into *S. cerevisiae* (15) with the pattern in the native organism. Strikingly, transplanted *D. hansenii* loci exhibit a marked change from their native polyadenylation patterns to a more *S. cerevisiae*–like profile (Fig. 4*A*). Specifically, the distribution of significant cleavage sites within their end zones appears markedly wider, with signal distributed over several major sites as opposed to a single dominant one (Fig. 4 *B* and *C*). Overall, native *D. hansenii* poly(A) sites are poorly correlated with poly(A) sites at the same genes transplanted into a *S. cerevisiae* context ($R = 0.41$ vs. $R = 0.95$ for two biological replicates of *S. cerevisiae*) (Fig. 4 *D* and *E*). Thus, although the sequence patterns around preferred sites are globally similar in these two species, this overall similarity is insufficient to specify the same polyadenylation pattern. Instead, factors must act differently in each species to produce the poly(A) site distribution.

Discussion

Species-Specific Patterns of Polyadenylation Are Determined Primarily by Polyadenylation Factors. The polyadenylation patterns of *S. cerevisiae*, *K. lactis*, and *D. hansenii* are broadly similar. In these yeast species, polyadenylation sites within 3′UTRs are located

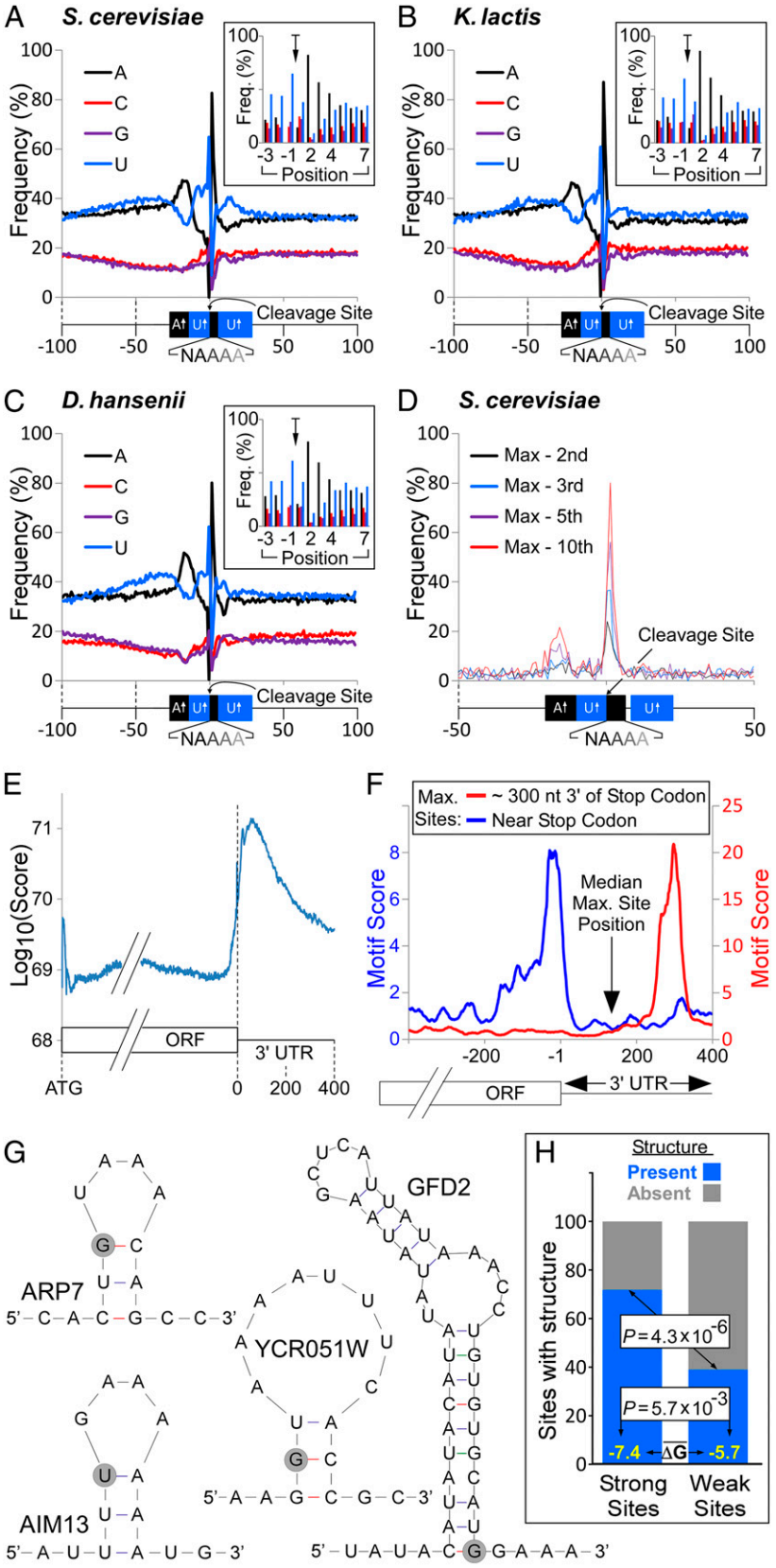


Fig. 3. Distinct sequence motifs govern efficient mRNA 3′ end formation in fungi. (*A*–*C*) Nucleotide frequency distributions derived from the alignment of thousands of sequences (±100 nt relative to favored cleavage sites) are depicted for *S. cerevisiae* (*A*), *D. hansenii*, (*B*) and *K. lactis* (*C*). *Inset* graphs present close-up views of nucleotide compositions in the immediate vicinities of the cleavage sites (downward arrows). Locations of upstream A-rich (black boxes), U-rich (blue boxes), downstream U-rich (blue boxes), and downstream A residues (relative importance indicated by darkness) are depicted below the *x*-axis. (*D*) A-rich sequences immediately downstream of cleavage sites are the primary determinants of cleavage efficiency in 3′ UTRs of Pol II–transcribed genes. Position-specific difference plots were constructed for *S. cerevisiae* between max use sites and lesser quality sites as indicated. Euclidean distance (see *Materials and Methods* for a detailed explanation), which represents a direct measure of the extent of change in nt composition (and hence importance), is plotted for a 100 nt window encompassing the cleavage site. (*E*) Position weight matrix (Fig. S3) motif plots predict low likelihood of mRNA 3′end formation in ORFs. Log_{10} of median scores at each position across the genome are displayed aligned to either the ATG (left part) or stop codon (right part). (*F*) Motif scores of maximal use sites clustered within 30 nt of translational stops predicts an ORF-proximal polyadenylation pattern. (*G*) Representative cleavage site structures, with poly(A) addition sites highlighted in gray circle. (*H*) Presence or absence of structures in 100 strong and weak cleavage sites. *P* values demonstrate that structures are disproportionately present and more stable (lower ΔG values) in strong sites.

over end zones that are over 200 nucleotides long. There is a remarkable heterogeneity of poly(A) sites within the end zone, and the most used sites within a gene typically account for a minority of the overall mRNA population. In addition, the small minority of poly(A) sites located outside 3′UTRs are biased to the beginning of the ORFs. Lastly, the sequence preferences around the dominant

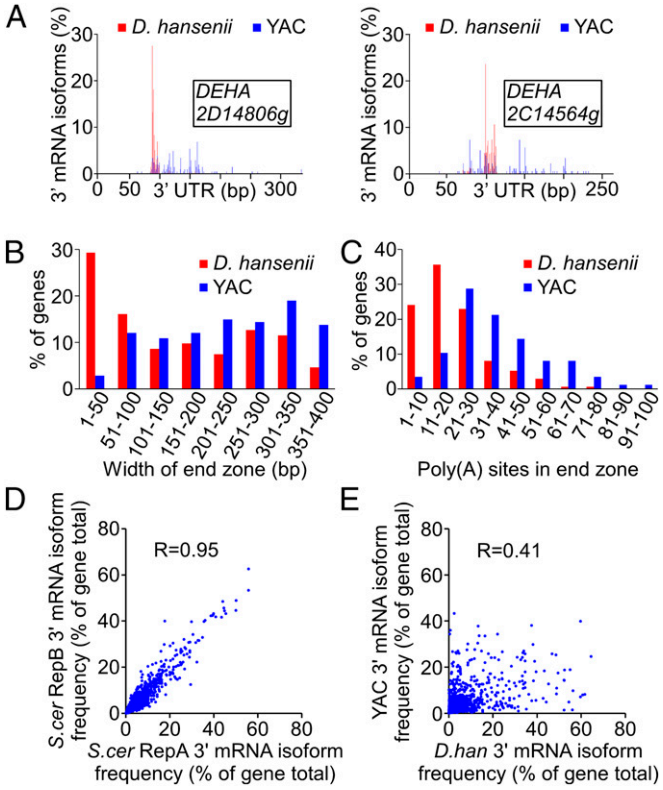


Fig. 4. Characterization of mRNA 3′ ends in YAC-containing *S. cerevisiae* strains with large genomic regions from *D. hansenii*. (*A*) Two examples of different 3′ ends heterogeneity of *D. hansenii* mRNAs in native species or YAC strains. Red shows pattern of native mRNA 3′ ends in *D. hansenii*; blue indicates 3′ ends from the same sequence on a YAC in *S. cerevisiae*. (*B*) The distribution of width of end zone from all *D. hansenii* 3′UTRs on YACs and the same UTRs in *D. hansenii*. End zone here was defined as the span between the 5′-most and 3′-most coordinates with 2% signal of the most preferred poly(A) site within the 400 nt 3′UTR. (*C*) The distribution of number of poly(A) sites within end zones defined in *B*. Only sites with at least 2% signal of the most preferred site within the same end zone were included in the analysis. (*D*) A scatter plot was used to compare percent frequency of 3′ end utilization within the 400 nt 3′UTR of *S. cerevisiae* ORFs on chromosome III from two independent datasets. mRNA 3′ ends with at least three raw reads were included in the graph. R-value indicates the correlation coefficient between the two datasets. (*E*) Frequency of 3′ end utilization within 3′UTRs on YACs was compared with native utilization frequency for the same regions, as in *D*.

poly(A) sites are extremely similar, although not identical, in the three yeast species.

Despite these broad similarities, the polyadenylation profiles show species specificity. Most significantly, in comparison with *S. cerevisiae*, *D. hansenii* end zones begin closer to the ORF, have stronger utilization of the most preferred site, and have a more narrow span of moderately used sites. Using a functional evolutionary approach first developed to analyze the basis of species specificity of nucleosome positioning (15), we show that the polyadenylation profiles of *D. hansenii* genes in *S. cerevisiae* are significantly different from those in the endogenous *D. hansenii*, indicating differential utilization of the RNA sequences in these two species. Furthermore, the overall profile of *D. hansenii* genes in *S. cerevisiae* resembles the overall profile of endogenous *S. cerevisiae* genes, indicating that species-specific factors, not sequence, are primarily responsible for the species-specific profiles.

Although the mechanistic basis of species specificity remains to be determined, the intrinsic specificity of RNA cleavage is unlikely to be a major determinant. This is because (*i*) the species-specific differences in the pattern relate to the nature and location of the end zone and (*ii*) the nucleotide preferences are extremely similar among the three species. In accord with this suggestion, the overall *D. hansenii* profile becomes *S. cerevisiae*–like when the genes are introduced into *S. cerevisiae*, and it is inconceivable that this could reflect fortuitous *S. cerevisiae*–favored sequences at common locations with *D. hansenii* genes. It is likely that the factors that mediate species specificity are involved primarily in interpreting the sequences that determine the end zone (see below), although we cannot exclude the possibility they might affect mRNA stability.

Selection of Polyadenylation Sites. Our results suggest a two-stage process of polyadenylation. In the first stage, a wide region of the 3′ UTR is permissive to polyadenylation by virtue of its overall sequence (largely AU richness). In the second stage, the relative utilization of sites within this region is determined by how well they conform to sequence and structural preferences surrounding the cleavage site, particularly the A residues between +2 and +5.

For the first stage, there are DNA-based and RNA-based models, and we suggest the following possibilities. In a DNA-based model, the large permissive region may correspond to sequences that cumulatively and progressively slow the rate of RNA Polymerase II (Pol II) elongation, and CPFs require increased Pol II dwell time to carry out their functions. In support of this view, induced pausing of Pol II by G-rich sequences in vitro results in increased polyadenylation of an upstream site (27), and reducing the in vivo Pol II elongation rate (*rpb2-10* mutation or 6-azauracil) causes a pronounced defect in processivity, in which many Pol II molecules fail to traverse the entire gene (28). As elongating Pol II is intimately associated with the DNA template to preclude spontaneous dissociation, a decrease in processivity very likely involves a cleavage step that leads to Pol II termination and dissociation. Premature transcript cleavage due to a processivity defect may result in polyadenylation within coding regions. In an RNA-based model, the AU-rich region forms distinct types of structures that favor polyadenylation (or inhibits structures that disfavor polyadenylation). Alternatively, the polyadenylation machinery might start the process by recognizing transcriptional elongation complexes with less stable RNA–DNA duplexes with AU-rich sequences and/or complexes with AU-rich RNA sequences extruding from the elongating complex. By any of these models, *D. hansenii* and *S. cerevisiae* factors would differentially read the same DNA or RNA sequence.

In the second stage, specific cleavage/polyadenylation sites within the permissive region are favored by virtue of the immediately adjacent sequence composition and secondary structure. One possibility is that such favored motifs preferentially interact with the specificity factor Cft1 (29). However, consistent with the relatively low information content of the motif surrounding the poly(A) site, the specificity for particular sites is only modest. Instead, the extreme diversity of mature 3′ ends reflects considerable promiscuity of the poly(A) machinery within end zones. After the completion of the present work, extensive poly(A) site diversity was observed by a different method in *S. cerevisiae* (30) and was also reported in *Arabidopsis thaliana* (31), suggesting that such promiscuity of the poly(A) machinery may be widespread in eukaryotes.

It is possible that much of the extreme heterogeneity of poly (A) sites merely reflects biological noise (32); that is, biochemical events that occur in vivo as a result of a low-specificity process that has not been subjected to evolutionary optimization. However, cells can alter the distribution of poly (A) sites in response to particular growth conditions (33), perhaps by changing the expression of specific CF1A components influencing site preference (34, 35). In addition, the half-lives of various minor species within a given gene differ considerably, suggesting a potential biological role. Thus, it seems likely that at least some of the heterogeneity in poly(A) sites

could have functional consequences for RNA stability and posttranscriptional events.

Materials and Methods

Yeast Strains. Yeast strain JGY2000, a derivative of BY4741 (ATCC 4040002; *MATa*, *his3Δ0, leu2Δ0, met15Δ0, ura3Δ0, rpb1::RPB1–FRB, rpl13::RPL13–FK512*) was grown in YPD medium to $OD_{600} = 0.6$. *S. cerevisiae* AB1380, *D. hansenii* (NCYC 2572), *K. lactis* (CLIB 209), and *S. cerevisiae* YAC6 and YAC7 strains harboring Yeast Artificial Chromosomes with *D. hansenii* sequences (15) were grown in rich medium to midlog phase. Cells were harvested and direct RNA sequencing was performed as described (10). Results from the AB1380 strain were essentially identical to those obtained from JGY2000.

Sequencing. Raw reads were filtered and aligned to the *S. cerevisiae* (2008 SGD Version 61), *D. hansenii* (Release 4–2/2012; www.genolevures.org/download/derived_files/CSV/Deha.csv) or *K. lactis* (Release 4–2/2012; www.genolevures.org/download/derived_files/CSV/Klla.csv) genome sequences. Reads whose 5′ ends initiated with a T-residue (corresponding to A at the 3′ ends of transcripts) were excluded from analysis, as sequences stemming from cleavage events immediately downstream of A-residues cannot be distinguished from sequences originating from extension of incompletely filled-in poly(A) tails hybridized to poly (dT)-coated surfaces (17). Approximately 7.5–8.5 million *S. cerevisiae*, 3.1 million *D. hansenii*, and 3.4 million *K. lactis* aligned reads were used to construct gene-specific end zones and to perform various statistical analyses.

Data Analysis. Nucleotide frequencies were computed by using custom-written Python scripts that are available upon request. For determination of patterns surrounding preferred polyadenylation sites, we limited the analysis to genes for which at least 100 sequence reads initiated in the 400 bp region downstream of the ORF. In *S. cerevisiae*, ∼5,000 genes met this criterion. Sequence motifs were also identified by using MEME (36). Motif scoring was performed using custom-written Python scripts.

For determination of cleavage frequency within ORFs, we restricted our analysis to ORFs at least 500 bp downstream of any 5′ neighboring genes transcribed on the same strand. For Fig. 3*E*, we calculated position-specific composition changes at each nucleotide n by measuring the Euclidean distance D_n between nucleotide frequencies using the following formula:

$$D_n = \sqrt{\left(f_A^{Max} - f_A^k\right)^2 + \left(f_C^{Max} - f_C^k\right)^2 + \left(f_G^{Max} - f_G^k\right)^2 + \left(f_U^{Max} - f_U^k\right)^2} = \sqrt{\sum_{i=A,C,G,U} \left(f_i^{Max} - f_i^k\right)^2}$$

where f_i^{Max} is the frequency of base i (i = {A,C,G,U}) at maximally preferred coordinates and f_i^k is the frequency of i at k preferred sites—in this case, the second best, third best, fifth best, and 10th best sites. Fig. 3*E* in essence represents the plot of D_n versus position in the vicinity of the cleavage sites.

Significance of enrichment of the U-rich stretch (–11 to –1) in the dominant sites (>40% of overall 3′UTR signal) relative to the weak sites (2% of signal) was calculated using a Student t test.

RNA secondary structures were modeled and ΔG stability values were calculated using mfold (Version 3.5) (26), with maximum distances between base pairs constrained to either 200 or 50 nt. Cleavage sites were judged to contain a structural motif if two consecutive bases at positions –1/+1, –2/–1, or –3/–2 were double stranded and positions +2 and/or +3 were in single-stranded conformation. Statistical significance of overrepresentation of strong cleavage sites relative to their weak counterparts was determined with a Fisher's Exact test, whereas the P value for the mean ΔG difference was calculated with a Student's test.

ACKNOWLEDGMENTS. The authors thank Fatih Ozsolak for arranging direct RNA sequencing at Helicos BioSciences. This work was supported by Grant GM30186 (to K.S.) from the National Institutes of Health.

1. Tian B, Hu J, Zhang H, Lutz CS (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* 33(1):201–212.
2. Jan CH, Friedman RC, Ruby JG, Bartel DP (2011) Formation, regulation and evolution of Caenorhabditis elegans 3′UTRs. *Nature* 469(7328):97–101.
3. Wu X, et al. (2011) Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. *Proc Natl Acad Sci USA* 108(30):12533–12538.
4. Derti A, et al. (2012) A quantitative atlas of polyadenylation in five mammals. *Genome Res* 22(6):1173–1183.
5. Smibert P, et al. (2012) Global patterns of tissue-specific alternative polyadenylation in Drosophila. *Cell Rep* 1(3):277–289.
6. Di Giammartino DC, Nishida K, Manley JL (2011) Mechanisms and consequences of alternative polyadenylation. *Mol Cell* 43(6):853–866.
7. Tian B, Manley JL (2013) Alternative cleavage and polyadenylation: The long and short of it. *Trends Biochem Sci* 38(6):312–320.
8. Proudfoot NJ, Brownlee GG (1976) 3′ non-coding region sequences in eukaryotic messenger RNA. *Nature* 263(5574):211–214.
9. Colgan DF, Manley JL (1997) Mechanism and regulation of mRNA polyadenylation. *Genes Dev* 11(21):2755–2766.
10. Zhao J, Hyman L, 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(2):405–445.
11. Mandel CR, Bai Y, Tong L (2008) Protein factors in pre-mRNA 3′-end processing. *Cell Mol Life Sci* 65(7-8):1099–1122.
12. Zaret KS, Sherman F (1982) DNA sequence required for efficient transcription termination in yeast. *Cell* 28(3):563–573.
13. Proudfoot N, O'Sullivan J (2002) Polyadenylation: A tail of two complexes. *Curr Biol* 12(24):R855–R857.
14. Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ (2010) The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol* 8(7):e1000414.
15. Hughes AL, Jin Y, Rando OJ, Struhl K (2012) A functional evolutionary approach to identify determinants of nucleosome positioning: A unifying model for establishing the genome-wide pattern. *Mol Cell* 48(1):5–15.
16. Ozsolak F, et al. (2009) Direct RNA sequencing. *Nature* 461(7265):814–818.
17. Ozsolak F, et al. (2010) Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* 143(6):1018–1029.
18. Nagalakshmi U, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320(5881):1344–1349.
19. David L, et al. (2006) A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci USA* 103(14):5320–5325.
20. Kaida D, et al. (2010) U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* 468(7324):664–668.
21. Berg MG, et al. (2012) U1 snRNP determines mRNA length and regulates isoform expression. *Cell* 150(1):53–64.
22. Kadaba S, et al. (2004) Nuclear surveillance and degradation of hypomodified initiator tRNAMet in S. cerevisiae. *Genes Dev* 18(11):1227–1240.
23. Grzechnik P, Kufel J (2008) Polyadenylation linked to transcription termination directs the processing of snoRNA precursors in yeast. *Mol Cell* 32(2):247–258.
24. Wlotzka W, Kudla G, Granneman S, Tollervey D (2011) The nuclear RNA polymerase II surveillance system targets polymerase III transcripts. *EMBO J* 30(9):1790–1803.
25. Tian B, Graber JH (2012) Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip Rev RNA* 3(3):385–396.
26. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31(13):3406–3415.
27. Yonaha M, Proudfoot NJ (1999) Specific transcriptional pausing activates polyadenylation in a coupled in vitro system. *Mol Cell* 3(5):593–600.
28. Mason PB, Struhl K (2005) Distinction and relationship between elongation rate and processivity of RNA polymerase II *in vivo*. *Mol Cell* 17(6):831–840.
29. Dichtl B, et al. (2002) Yhh1p/Cft1p directly links poly(A) site recognition and RNA polymerase II transcription termination. *EMBO J* 21(15):4125–4135.
30. Pelechano V, Wei W, Steinmetz LM (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* 497(7447):127–131.
31. Sherstnev A, et al. (2012) Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation. *Nat Struct Mol Biol* 19(8):845–852.
32. Struhl K (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol* 14(2):103–105.
33. Hoopes BC, Bowers GD, DiVisconte MJ (2000) The two Saccharomyces cerevisiae SUA7 (TFIIB) transcripts differ at the 3′-end and respond differently to stress. *Nucleic Acids Res* 28(22):4435–4443.
34. Bucheli ME, He X, Kaplan CD, Moore CL, Buratowski S (2007) Polyadenylation site choice in yeast is affected by competition between Npl3 and polyadenylation factor CFI. *RNA* 13(10):1756–1764.
35. Kim Guisbert KS, Li H, Guthrie C (2007) Alternative 3′ pre-mRNA processing in Saccharomyces cerevisiae is modulated by Nab4/Hrp1 in vivo. *PLoS Biol* 5(1):e6.
36. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2:28–36.