Title

Poly(A) code analyses reveal key determinants for tissue-specific mRNA alternative polyadenylation

Journal

RNA, 22(6)

ISSN

1355-8382

Authors

Weng, Lingjie

Li, Yi

Xie, Xiaohui

et al.

Publication Date

2016-06-01

DOI

10.1261/rna.055681.115

Peer reviewed

Poly(A) code analyses reveal key determinants for tissue-specific mRNA alternative polyadenylation

LINGJIE WENG,[1,2,3,4] YI LI,[2,3,4] XIAOHUI XIE,[2,3] and YONGSHENG SHI[1]

[1]Department of Microbiology and Molecular Genetics, School of Medicine, University of California, Irvine, Irvine, California 92697, USA
[2]Institute for Genomics and Bioinformatics,[3]Department of Computer Science, University of California, Irvine, Irvine, California 92697, USA

ABSTRACT

mRNA alternative polyadenylation (APA) is a critical mechanism for post-transcriptional gene regulation and is often regulated in a tissue- and/or developmental stage-specific manner. An ultimate goal for the APA field has been to be able to computationally predict APA profiles under different physiological or pathological conditions. As a first step toward this goal, we have assembled a poly(A) code for predicting tissue-specific poly(A) sites (PASs). Based on a compendium of over 600 features that have known or potential roles in PAS selection, we have generated and refined a machine-learning algorithm using multiple high-throughput sequencing-based data sets of tissue-specific and constitutive PASs. This code can predict tissue-specific PASs with >85% accuracy. Importantly, by analyzing the prediction performance based on different RNA features, we found that PAS context, including the distance between alternative PASs and the relative position of a PAS within the gene, is a key feature for determining the susceptibility of a PAS to tissue-specific regulation. Our poly(A) code provides a useful tool for not only predicting tissue-specific APA regulation, but also for studying its underlying molecular mechanisms.

Keywords: mRNA 3′ processing; tissue specificity; post-transcriptional gene regulation; machine learning

INTRODUCTION

The 3′ ends of most eukaryotic mRNAs are formed by an endonucleolytic cleavage and subsequent polyadenylation (Colgan and Manley 1997; Zhao et al. 1999). Recent global studies revealed that ~70% of eukaryotic genes produce multiple RNA isoforms through the usage of different cleavage/polyadenylation sites (PASs), a phenomenon called alternative polyadenylation (APA) (Di Giammartino et al. 2011; Shi 2012; Elkon et al. 2013; Tian and Manley 2013). APA has been increasingly recognized as a crucial mechanism for eukaryotic gene regulation. Usage of alternative PASs located in distinct terminal exons often leads to the production of mRNAs that encode proteins with related, distinct, or even opposing functions. On the other hand, APA involving PASs in the same terminal exon produces mRNAs that share the same coding region, but differ in their 3′ untranslated regions (UTRs). The diverse 3′ UTRs generated by APA may confer different stability, translation efficiency, or subcellular localization to the mRNA isoforms (Mayr 2015). Importantly, APA is highly regulated in development and in a tissue-specific manner, and deregulation of APA has been implicated in a wide range of human diseases, including cancer and neuromuscular disorders (Di Giammartino et al. 2011; Shi 2012). Given the functional importance of APA, it is critical to understand how PAS selection is regulated.

Both *cis* elements and *trans*-acting factors can influence PAS selection (Di Giammartino et al. 2011; Shi 2012; Elkon et al. 2013; Tian and Manley 2013). First, alternative PASs within a transcript often have different intrinsic strengths, which refer to the efficiencies by which these sites are recognized and processed by the core mRNA 3′ processing machinery (Shi 2012). PASs whose sequences are more consistent with the consensus PAS sequences tend to be stronger sites, thus more likely to be selected (Takagaki et al. 1996; Yao et al. 2012; Lackford et al. 2014). Second, the efficiency of 3′ processing at any given site is also affected by the overall activity levels of the mRNA 3′ processing machinery. For example, changes in the protein levels of 3′ processing factors often lead to APA changes in a specific subset of genes (Takagaki et al. 1996; Yao et al. 2012; Lackford et al. 2014; Li et al. 2015). Third, regulatory factors, including RNA binding proteins, may promote or inhibit PAS recognition by the core 3′ processing machinery (Shi and Manley 2015).

[4]**These authors contributed equally to this work.**
Corresponding authors: yongshes@uci.edu, xhx@ics.uci.edu
Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.055681.115.

Finally, the sequence context of alternative PASs may also influence APA outcomes. Promoter-proximal PASs have a natural advantage over the distal sites as they are transcribed earlier. The magnitude of such advantage can be influenced by the transcript structure (such as the distance between the proximal and distal PAS) and the transcription elongation rate (Davis and Shi 2014). The combinatorial effects of all the aforementioned factors and the stabilities of the mRNA isoforms determine the steady-state APA landscape. Given the highly complex interplay between these factors, computational approaches are essential for integrating the effects of these factors in order to fully elucidate the regulatory mechanisms of APA.

One of the ultimate goals of the APA field is to be able to computationally predict developmental stage- or tissue-specific APA outcomes under physiological or pathological conditions. Surprisingly, however, research progress in this area has been limited. Most previous studies have focused on the characterization and prediction of PASs in various genomes. Based on known and computationally identified PAS-associated *cis* elements, a number of approaches have been proposed for predicting PAS, including hidden Markov model or support vector machines (Tabaska and Zhang 1999; Cheng et al. 2006; Akhtar et al. 2010; Kalkatawi et al. 2012; Hafez et al. 2013; Xie et al. 2013). Although these approaches have achieved considerable success in predicting PAS with canonical features, they are not suitable for distinguishing constitutive and tissue-specific PASs. As RNA sequences within the core PASs alone are not sufficient to account for tissue-specific APA regulation, more sophisticated algorithms that incorporate additional features are necessary for deciphering the rules for PAS selection. In this report, we have developed the first poly(A) code, a machine-learning algorithm that uses a compendium of over 600 RNA features to accurately distinguish constitutive and tissue-specific PASs. Additionally, using this approach, we have identified novel molecular features important for tissue-specific APA regulation.

RESULTS

Identify tissue-specific PASs

mRNA 3′-end sequencing not only maps PASs on the transcriptome level, but quantifies the relative abundance of each mRNA isoform, making it the ideal approach for APA analysis (Shi 2012). Several 3′-end sequencing data sets have recently been published that include different mammalian tissues with variable coverage and depths (Derti et al. 2012; Lin et al. 2012; Lianoglou et al. 2013). For our initial modeling, we chose to use the data set generated by the Mayr laboratory that covers seven human tissues, including naive B cells, brain, breast, embryonic stem (ES) cells, ovary, skeletal muscle, and testis (Lianoglou et al. 2013). We used the "cleaned alignment" version of their 3′ seq data set, in which only uniquely mapped reads were kept and internal priming or spurious antisense reads were computationally removed. As sequencing depths vary for different tissues, we normalized the read counts by sequencing depth (counts per million) in each tissue. For each PAS located in an annotated gene, we calculated its usage frequency (ϕ) by dividing the read counts for this PAS by the total read counts for the gene. To remove extremely infrequently used PASs, we filtered out PASs whose ϕ values are <5% in all tissues. In total, 18,494 PASs covering 8202 genes met this criterion and were used for subsequent analyses.

For our analyses of tissue-specific PAS selection, we have focused on genes that have multiple PASs and are expressed in multiple tissues. To identify PASs that showed tissue-specific usage, we calculated the Shannon entropy scores (H) of the ϕ values of each PAS across the seven tissues, which ranges from zero to $\log_2(N)$, where N is the number of tissues. Shannon entropy measures the uniformity of data across different samples and has been widely used in gene expression analyses (Schug et al. 2005). The original Shannon entropy scores are only suitable for identifying PASs that are highly used in one or a few tissues, but fail to detect those that are suppressed in one or a few tissues. In order to efficiently identify all types of tissue-specific PASs, we applied Tukey biweight (T_{bw}) to obtain an adjusted entropy score (H', see Materials and Methods for details) (Grant et al. 2005). PASs with H' values close to zero are selected or avoided in a single tissue, and PASs with high H' are more broadly used in different tissues. We have designated 2276 PASs as tissue-specific ($H' < 1.8$ and $H < 2.2$, red area) and 3903 PAS as constitutive ($H' > 2.2$ and $H > 2.7$, green area) (Fig. 1A). These thresholds were chosen to obtain sufficient numbers of both tissue-specific and constitutive PASs while maintaining data quality (i.e., the tissue-specific and constitutive PAS groups are sufficiently distinct). For all tissue-specific PASs, we next identified the specific tissue(s) that have significantly different ϕ values by using an outlier detection method called ROKU (see Materials and Methods for details). The ROKU method is based on Akaike's information criterion (AIC) (Kadota et al. 2006), and it can handle various situations, including the "up-type" in which a PAS is highly used in a single or small number of tissues, the "down-type" in which a PAS is depleted in a single or a few tissues, and the "mix-type" in which a PAS is highly used in some tissues and depleted in others. For example, the proximal and intermediate PASs in the *Mgea5* gene have H' values of 0.98 and 1.01, and their ϕ values are significantly different in testis and B cells, respectively (Fig. 1B). Thus these sites were designated as testis- and B cell-specific PASs.

Characteristics of tissue-specific PASs

We next characterized the tissue-specific PASs in detail and made several interesting observations. First, different tissues showed drastically different levels of tissue-specific PAS selection (Fig. 2A). Of the seven tissues in this data set, the breast

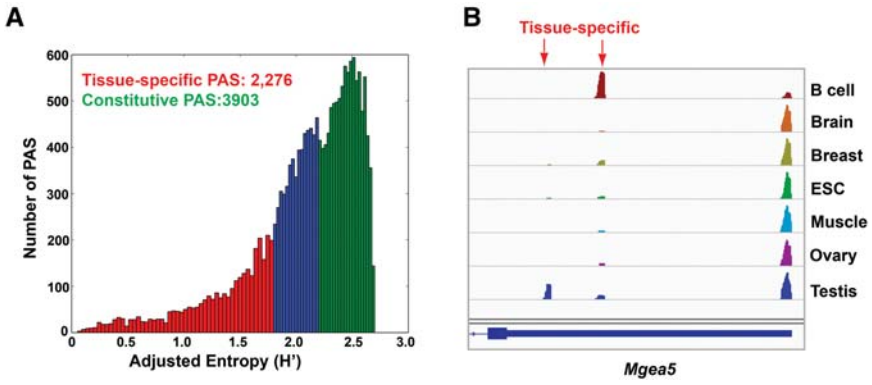

FIGURE 1. Identification of tissue-specific poly(A) sites. (*A*) A histogram of PASs at each adjusted entropy score (H'). The red area corresponds to tissue-specific PASs ($H' < 1.8$) and green to constitutive ($H' > 2.2$). (*B*) PAS usage frequency for the *Mgea5* gene as determined by 3′ seq. The proximal and intermediate PASs were identified as tissue-specific for testis and B cells, respectively.

and ovary had the lowest number of tissue-specific PAS usage. On the other hand, 493 and 423 PASs showed tissue-specific usage in naïve B cells and testis, respectively. Surprisingly, brain and ESC have intermediate levels of tissue-specific PASs (Fig. 2A). Second, we have carried out gene ontology (GO) analyses of the genes that contain tissue-specific and constitutive PAS in this data set. Tissue-specific PAS-containing genes showed significant overrepresentation of genes that function in protein modification, especially phosphorylation, and intracellular trafficking (Fig. 2B). In contrast, constitutive PAS-containing genes are most enriched with genes that function in nucleic acid metabolism and mRNA processing (Fig. 2B). Third, we have examined the relative positions of tissue-specific PASs within their host genes and found that the proximal (blue bars) and intermediate (green bars) PASs are more likely to be regulated in a tissue-specific manner than distal sites (brown bars) (compare constitutive and tissue-specific columns, Fig. 2C). Fourth, we have compared the sequence conservation levels based on phastCons scores of the 200 nucleotides (nt) region spanning the cleavage sites of tissue-specific or constitutive PASs (Siepel et al. 2005). Constitutive alternative PASs showed similar conservation levels (Fig. 2D, bottom panel). In contrast, tissue-specific PASs tend to be less conserved than constitutive sites from the same genes regardless of their relative positions (Fig. 2D, top and middle panels). Finally, we compared the nucleotide composition between tissue-specific and constitutive PASs. Constitutive PASs contain a highly A-rich region centered around −20 nt and a U-rich region at approximately +25 nt (relative to the cleavage sites) (Supplemental Fig. 1A). Interestingly, although tissue-specific PASs also contain an A-rich peak upstream of the cleavage sites, there is a significant decrease in A-levels near −10 nt and a concomitant increase of Us in the same region (Supplemental Fig. S1B, marked by arrows). Consistently, similar patterns were also observed in the nucleotide composition of B cell- and testis-specific PASs (Supplemental Fig. S1C,D). Together, these results suggest that tissue-specific PASs differ from constitutive sites in multiple aspects and such information may be harnessed for building a computational model for distinguishing tissue-specific and constitutive PASs.

Assemble the poly(A) code and define key features for tissue-specific PAS prediction

Using the tissue-specific and constitutive PAS data sets that we have compiled, we next set out to assemble a machine-learning algorithm for predicting whether any given PAS is used in a tissue-specific or constitutive manner. To this end, we have compiled a compendium of 658 RNA features, covering major parameters that are known to impact PAS selection or have the potential to do so (the full list is included in Supplemental Table S1, see Materials and Methods for details). They include the core PAS *cis* elements, known and potential regulatory motifs within and adjacent to PASs, sequence conservation levels, secondary structures, nucleosome occupancy, and the PAS context. Based on these features, we then carried out supervised classification to distinguish between tissue-specific and constitutive PASs using various machine-learning algorithms.

As mentioned earlier, we have identified 2276 tissue-specific PASs and 3903 constitutive PASs. To equalize the number of PASs, we randomly sampled 2276 samples from the pool of 3903 constitutive PASs. Given the relatively limited data, we used a 10-fold cross validation procedure. We held out 552 samples for model testing, and used the remaining PASs with 10-fold cross validation to learn the optimal hyper parameters for each prediction model. We used both accuracy and the area under the ROC curve (AUC) to evaluate the classification performance. Accuracy is the proportion of true results, including both true positive and true negatives. AUC is determined by true-positive rate (TPR) and false-positive rate (FPR).

We applied several classifier models, including logistic regression (LR), linear support vector machine with regularization (LSVM), SVM with WD-kernel (WDSVM), and adaptive boosting (AdaBoost). The WDSVM model extracts sequences around PAS and uses WD-kernel to compute similarities between two sequences while taking positional information into account (Hafez et al. 2013). In comparison, AdaBoost builds a sequence of weak base estimators and combines them in a weighted manner by iteratively refining the weight assigned to each sample until optimal performance is achieved (Freund and Schapire 1997). As shown in Figure 3A, the *cis* element-based WDSVM model performed poorly (AUC = 0.64), suggesting that PAS sequences

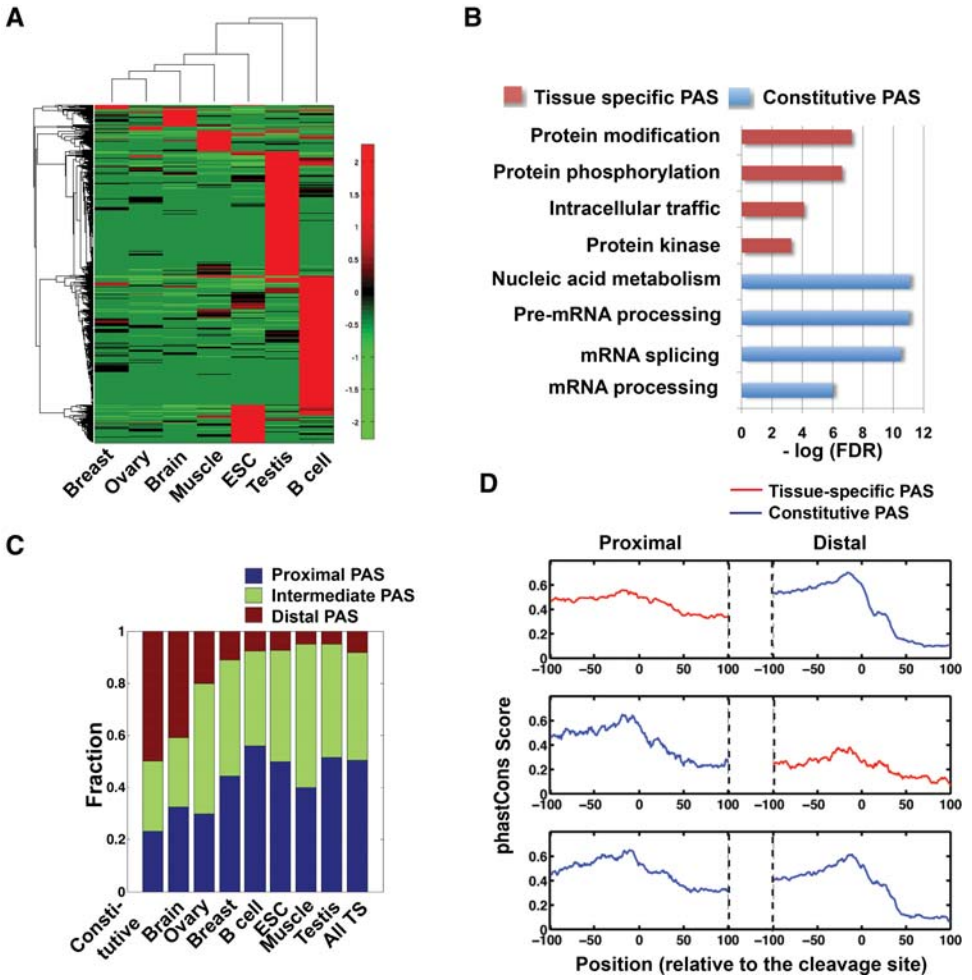

FIGURE 2. Functional characterization of tissue-specific poly(A) sites. (*A*) A heat map of standardized usage frequency for all tissue-specific PASs in all seven tissues. (*B*) Gene ontology analyses of tissue-specific and constitutive PASs. (*C*) The proportions of proximal, intermediate, and distal PASs in constitutive PASs, or PASs that are specific to each tissue, or all tissue-specific (TS) PASs. (*D*) The average phastCons scores for tissue-specific (red lines) and the corresponding constitutive PASs (blue lines) from the same genes.

are not sufficient for predicting tissue-specific APA. The two linear models with our assembled RNA features as input yielded better performances (AUC $\geq$ 0.8). The best performance was achieved with AdaBoost (accuracy = 85.3%, AUC = 0.92). In addition to the general tissue-specific PASs, we have also applied the same algorithm and feature set to predicting PASs that are preferentially used or avoided in B cells and testis, the only two tissues with a sufficiently large number of tissue-specific PASs in our data set. As shown in Figure 3B, our poly(A) code achieved similar accuracy levels in predicting B cell- or testis-specific PASs as it did for general tissue-specific PASs. These results demonstrate that our AdaBoost classification model outperformed the linear models, and can be used to distinguish between tissue-specific and constitutive PASs with high accuracy.

We next wanted to use our poly(A) code to define the key feature(s) for the prediction performance. To this end, we carried out predictions with the same AdaBoost classifier model using individual features and compared their prediction performances. As shown in Figure 3B, these individual features led to variable prediction accuracy levels. For example, PAS context alone gave a prediction accuracy of over 80%. In sharp contrast, RNA secondary structure yielded the lowest accuracy of ~50%, similar to random selection. Based on this analysis, the top four features for predicting tissue-specific PASs are PAS context, 6-mer motifs, conservation levels, and PAS signal sequence (Fig. 3B).

To determine the robustness of our model, we have tested it on another RNA 3′ sequencing data set. This data set was generated using the Helicos Direct RNA sequencing platform and covers five human tissues, including breast, colon, kidney, liver, and lung (Lin et al. 2012). Using this data set, our model gave a prediction accuracy of 78.6% (AUC = 0.83, Supplemental Fig. S2). Furthermore, PAS context was again one of the top features for the prediction performance (73.5% accuracy, AUC = 0.77, Supplemental Fig. S2). Therefore, our poly(A) code gave comparable performance on two independently generated data sets and consistently

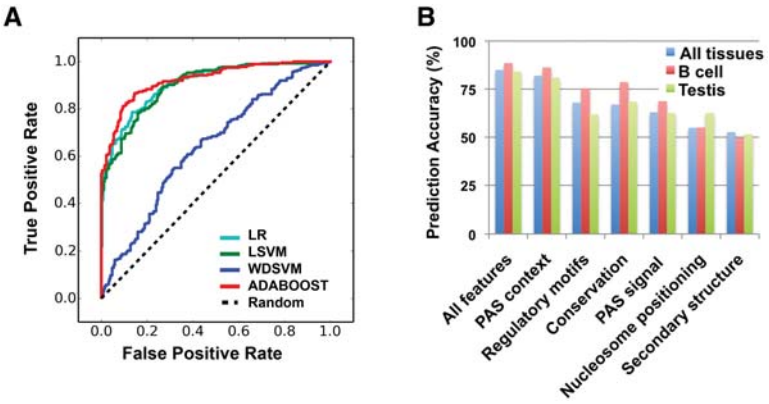

FIGURE 3. Assembling and testing the poly(A) code. (*A*) ROC curves for random selection and the four models tested, including linear regression (LR), linear SVM (LSVM), SVM with WD kernel (WDSVM), and adaptive boosting (Adaboost). Prediction accuracy and the area under the curve (AUC) are listed. (*B*) The accuracy of predicting all tissue-specific (blue columns), or B cell-specific (red columns), or testis-specific PASs (green columns) using different feature sets as listed on the *x*-axis.

identified PAS context as a key feature for distinguishing tissue-specific and constitutive PASs.

APA is subject to regulation by *trans*-acting factors, including the core mRNA 3′ processing factors and various RNA-binding proteins (Shi and Manley 2015). As these factors are often expressed at different levels in different tissues, they are expected to impact tissue-specific APA profiles. To explore this, we first compared the expression levels (based on the normalized 3′ sequencing read counts) of 106 *trans*-acting factors in all seven tissues. They include 26 core 3′ processing factors and 80 RNA-binding proteins whose RNA-binding specificities have been characterized (Shi et al. 2009; Ray et al. 2013). Indeed, great variations were observed in their mRNA levels (Supplemental Fig. S3).

Next we wanted to determine whether the expression patterns of *trans*-acting factors could be used for predicting tissue-specific APA. For this purpose, we implemented a machine-learning model for predicting whether a tissue-specific PAS is utilized in a specific tissue or not. To train and test this model, we used the expression levels of the 106 *trans*-acting factors together with the original 658 features as input without specifying the tissue source. Note that here we had to use a distinct machine-learning model to study the effect of *trans*-acting factors, because our original model was designed for distinguishing between tissue-specific and constitutive PASs as two general groups. In contrast, this new model has to be able to predict PAS usage in specific tissues. To train the new model, we determined the usage of each of the 2276 tissue-specific sites in each of the seven tissues, labeling a site as either positive if the PAS is utilized in that tissue, or as negative if the PAS is not used. The entire data set, consisting of 15,932 samples, was then separated into a training data set (3000 positive and 10,000 negative PASs) and a test data set (762 positive and 2170 negative PASs). We applied AdaBoost to train our model using either the original 658 features or these features plus the expression levels of all *trans*-acting factors. We found that the model using the original 658 features performed poorly (AUC = 0.359). Adding the *trans*-acting factor levels to the feature set led to significantly higher prediction accuracy (AUC = 0.807). These results suggest that the expression pattern of *trans*-acting factors is a useful feature for predicting tissue-specific APA.

PAS context is a major determinant of regulated APA

Our poly(A) code analyses revealed that PAS context is a key feature for distinguishing tissue-specific and constitutive PASs (Fig. 2B; Supplemental Fig. S2). PAS context is a composite feature that includes the relative position of a PAS within the gene, the distance from neighboring PASs, the distance to neighboring splicing sites, and the distance to the stop codon (see Materials and Methods for details). One of the main features in PAS context is the distance between a PAS and its nearest neighboring PAS. We next compared the distance between tissue-specific or constitutive PASs and their nearest neighboring sites. As shown in Figure 4A, the median distance between constitutive PASs and their closest neighboring PASs is ~680 nt. In contrast, tissue-specific PASs are 1~1.4 kb away from their nearest neighboring PASs, significantly greater than that of constitutive PAS ($P = 3.4 \times 10^{-4}$–3.4×10^{-54}, K–S test) (Fig. 4A). Among the seven tissues, ESC-specific PASs are the farthest from their neighboring sites with a median distance of ~1.4 kb. These results suggest that tissue-specific PASs tend to be located farther away from their neighboring PAS compared to constitutive PASs. Consistent with this conclusion, when all tissue-specific PASs were considered together, we detected a positive correlation between the distance of neighboring PASs and the percentage of tissue-specific PASs ($r^2 = 0.56$) (Fig. 4B).

As tissue-specific PAS selection is mediated, at least in part, by *trans*-acting regulatory proteins, we next tested whether the distance to the nearest neighboring PAS also plays a role in APA control by known regulators. Recent studies have characterized the APA regulation mediated by a number of mRNA 3′ processing factors, including CPSF, CstF, and CFIm (Martin et al. 2012; Yao et al. 2012; Lackford et al. 2014; Li et al. 2015). Although these factors are believed to play a general role in mRNA 3′ processing, depletion of these factors led to APA changes of a specific set of transcripts and it remains unclear why certain APA events are more sensitive to the protein levels of core mRNA 3′ processing factors. Since our analyses suggest that longer distance between alternative PASs is a strong predictor of tissue-specific APA

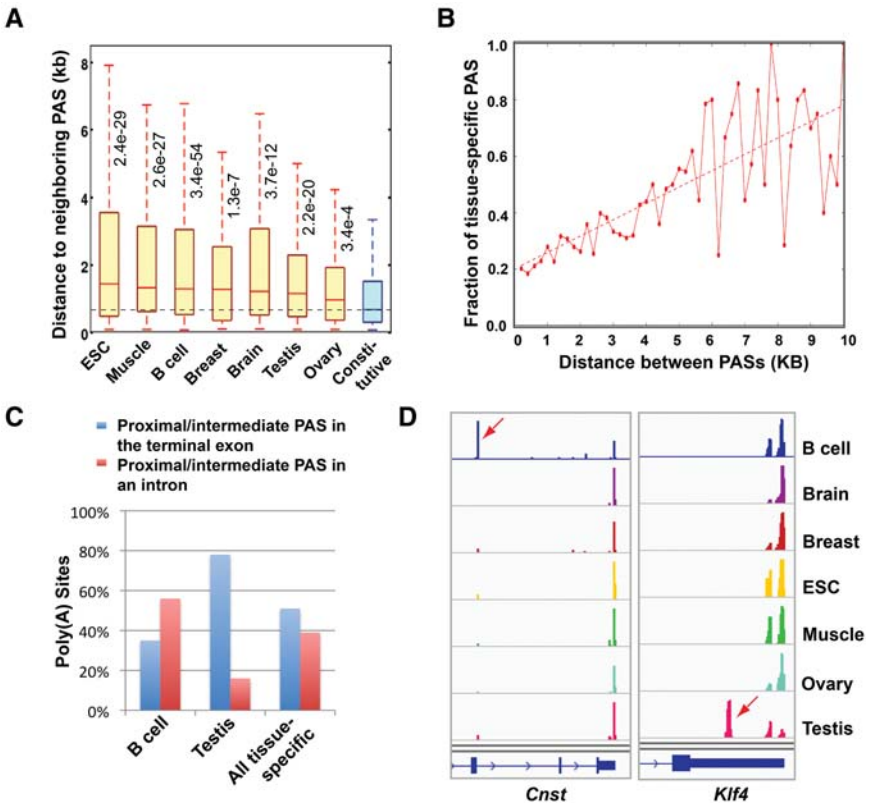

FIGURE 4. Key features that distinguish constitutive and tissue-specific APA events. (*A*) A box plot showing the distribution of the distance between neighboring PASs for tissue-specific and constitutive PASs. The *P*-values (K–S test) are listed for the difference between tissue-specific and constitutive PASs. (*B*) The relationship between *x* (the distance between neighboring PASs) and *y* (the percentage of tissue-specific PAS when they are located at that distance from its neighboring PAS). $r^2 = 0.56$. (*C*) Percentage of proximal/intermediate PASs in the terminal exon (blue bars) or proximal/intermediate PASs in an intron (red bars) in B cell-, testis-, or all tissue-specific PASs. (*D*) Integrated Genome Viewer (IGV) tracks of 3′ sequencing results for *Cnst* and *Klf4* genes in seven tissues.

regulation, we compared the distance between alternative PASs that are regulated by specific 3′ processing factors and those that are not affected. Remarkably, the mean distance between alternative PASs that are regulated by CFIm68 (a subunit of the CFIm complex), Fip1 (a subunit of the CPSF complex), and CstF64/τ (subunits of the CstF complex) are all significantly longer than those of nontargets (Supplemental Fig. S4). Together, these results strongly suggest that alternative PASs that are far from their neighboring sites are more likely to be regulated. Therefore, the distance between alternative PASs is a key determinant for the susceptibility of APA events to regulation.

The relative position of a PAS within the gene is another important feature of the PAS context. When we compared the relative positions of B cell-, testis-, and all tissue-specific PASs, we found some interesting differences. Over half (56%) of B cell-specific PASs are proximal or intermediate PASs found within an intron, and usage of these PASs are expected to produce an mRNA that encodes a distinct protein (Fig. 4C). For example, *Cnst* transcripts are primarily polyadenylated at the annotated PAS within the terminal exon to produce full-length mRNAs in most tissues (Fig. 4D, left panel). However, in B cells, a PAS within the penultimate intron is preferentially used, and the resultant mRNA is predicted to encode a different protein. In contrast, 78% of testis-specific PASs are proximal or intermediate PASs within the annotated terminal exon (Fig. 4C). Thus, similar to the *Klf4* gene shown in Figure 4D and the *Mgea5* gene shown in Figure 1B, the majority of testis-specific APA events lead to shorter 3′ UTRs without affecting the protein coding regions. These results revealed that the PAS context significantly differ for B cell- and testis-specific PASs and indicated that APA may be regulated by different mechanisms in these tissues. Together, these results suggest that PAS context, including the distance between alternative PASs and the relative positions of alternative PASs, is a critical feature for determining tissue-specific APA profiles.

DISCUSSION

In this study, we first identified thousands of tissue-specific PASs based on multiple published RNA 3′ sequencing data sets. Using these data sets and over 600 RNA features, we trained and tested various machine-learning algorithms for predicting tissue-specific PASs and achieved ~85% accuracy using an AdaBoost algorithm. Finally, we compared the contribution of different RNA features to the prediction performance of our model and identified key features, including PAS context, for determining tissue-specific APA regulation. These results represent an important step in deciphering the "poly(A) code" and their implications are discussed below.

Computational prediction of APA has been a major goal for the mRNA 3′ processing field, but surprisingly little progress has been made. Previous studies have mainly focused on PAS prediction. For example, earlier studies identified a number of *cis* elements using the EST-based PAS database, and position weight matrices of these *cis* elements were then used as input features for a hidden Markov model or support vector machine (SVM) to predict PASs (Tabaska and Zhang 1999; Cheng et al. 2006; Akhtar et al. 2010; Kalkatawi et al. 2012; Xie et al. 2013). A recent study built an SVM with string kernels based on high throughput sequencing data and improved the PAS prediction performance (Hafez et al. 2013). Although these studies represent

important advances in the field, they also have multiple limitations. For examples, all previous models used *cis* elements near PAS as the sole or main feature. However, only several *cis* elements have been experimentally shown to contribute to PAS definition, and our current understanding of the relationship between RNA sequence and PAS strength remains rudimentary (Tian and Graber 2012). In addition, similar to alternative splicing (Barash et al. 2010), other features are likely important for determining APA outcomes, but were not considered in earlier studies. Finally, previous studies used an EST-based PAS data set or a very limited high throughput-sequencing data set that only contained several tissue types (Tabaska and Zhang 1999; Cheng et al. 2006; Akhtar et al. 2010; Kalkatawi et al. 2012; Hafez et al. 2013; Xie et al. 2013). As a result, although these earlier models achieved reasonable performance in distinguishing constitutive PASs from background genomic or UTR sequences, they are not suitable for predicting tissue-specific PASs.

In this study, we took several steps to overcome the limitations associated with earlier studies. First, we have trained and tested our computational models using multiple high-throughput sequencing data sets of tissue-specific PASs that had higher coverage of different tissue types (Lin et al. 2012; Lianoglou et al. 2013). Indeed our model achieved higher prediction accuracy with the Mayr laboratory data set, which contains seven tissues, than with the John laboratory data set with only five tissues, further highlighting the importance of data coverage. To further improve the performance of the poly(A) code, it will be important for future studies to generate high-quality and comprehensive 3′ sequencing data from normal tissues as well as patient samples. Second, we have tested a number of different machine-learning algorithms, including those used in previous studies. Our results demonstrated that AdaBoost, which has not been applied to APA analyses before, outperformed all of the previously used algorithms (Fig. 3A). Finally, we have compiled over 600 features and used them for our modeling. Our data suggest that, instead of the known *cis* elements within the PAS regions, PAS context was the most informative feature in distinguishing tissue-specific and constitutive PASs (Fig. 3B). Adding the expression levels of *trans*-acting factors to the feature list can further improve the prediction accuracy of the poly(A) code. Therefore, to improve the performance of the poly(A) code, it will be critical to further expand and refine the RNA feature repertoire.

Our poly(A) code predicts tissue-specific PASs based on over 600 features. Through our computational modeling, we have identified the PAS context as the most informative feature for distinguishing tissue-specific and constitutive PASs (Fig. 3B). As mentioned earlier, PAS context is a composite feature that includes APA types and the distance from neighboring splice sites and PASs. Importantly, our results suggest that there is a positive correlation between the distance between neighboring PASs and the probability of tissue-specific APA (Fig. 4B). This is due, at least in part, to the fact that the alternative PASs that are located far from each other are more likely to be regulated by *trans*-acting factors such as the core mRNA 3′ processing factors (Supplemental Fig. S2). This is consistent with a number of published studies. For example, we first provided evidence that the distance between neighboring PASs plays an important role in determining the mode of regulation by an APA regulator Fip1 (Lackford et al. 2014). This conclusion was further substantiated by a large-scale analysis showing that the APA events targeted by various regulators tend to have longer distances between neighboring PASs (Li et al. 2015). Although the underlying mechanism for this phenomenon is still poorly understood, the effect of distance between neighboring PASs on APA could be, at least in part, due to the fact that promoter-proximal PASs have an intrinsic advantage over the distal sites as they are transcribed earlier and thus have more time to be recognized by the 3′ processing machinery (Shi 2012). The magnitude of such an advantage is determined by the interval between the times when these PASs are transcribed. Longer distance between neighboring PASs would give the proximal sites a greater advantage and, in turn, more room for regulation.

In addition to the distance between neighboring PASs, the relative position of a PAS with the gene is another important feature in PAS context. Our analyses revealed interesting differences in APA types in different tissues. For example, naïve B cells display higher usage of many intronic PASs, resulting in the production of distinct protein product (Fig. 4C,D). In keeping with this, usage of intronic PASs in B cells has previously been reported for the IgM transcripts (Alt et al. 1980; Early et al. 1980; Rogers et al. 1980). In contrast, the majority of testis-specific PASs are proximal/intermediate PASs within the terminal exon, leading to the production of mRNAs that encode the same proteins but have shorter 3′ UTRs (Fig. 4C,D). Therefore, given the different PAS context, tissue-specific APA profiles have distinct functional impact and are likely regulated by different mechanisms. For example, APA events involving intronic PASs, such as many B cell-specific PASs, are more likely to be regulated by both 3′ processing factors as well as splicing factors, such as U1 snRNP (Kaida et al. 2010; Berg et al. 2012; Li et al. 2015).

In summary, we have assembled a poly(A) code for accurately distinguishing tissue-specific PASs from constitutive sites based on over 600 RNA features. Our results provided evidence that computational modeling is a useful tool not only for predicting tissue-specific APA patterns, but also for studying the underlying regulatory mechanisms.

MATERIALS AND METHODS

RNA 3′ sequencing data sets used in this study were downloaded from the NCBI Sequence Read Archive (SRP029953) (Lianoglou et al. 2013) or from http://johnlab.org/xpad/ (Lin et al. 2012). All the collected PASs, associated features and source code for prediction models are available at https://github.com/uci-cbcl/polyAcode.

Identify tissue-specific and constitutive PASs

To identify tissue-specific and constitutive PASs, we first calculated the usage frequency (ϕ) for each PAS. Specifically, we normalized the read counts by sequencing depth (counts per million) in each tissue. We then calculated ϕ for each PAS located in an annotated gene by dividing the read counts for this PAS by the total read counts for the gene. We filtered extremely infrequently used PASs, whose ϕ values are <5% in all tissues. Next, we normalized ϕ values across seven different tissues to obtain a probability vector and calculated the Shannon entropy scores (H) as the measure of tissue-specific usage:

$$H = -\sum_{t=1}^{7} p_t \log_2(p_t),$$

where $p\phi_t$ is the usage frequency of the PAS in tissue type t. We introduced one-step Tukey biweight (T_{bw}) to adjust the original, $\phi'_t = |\phi_t - T_{bw}|$ (additional details on Tukey biweight calculation can be found in Supplemental Methods). We then calculated the adjusted H' based on ϕ'_t using the same formula. Finally, with stringent thresholds, we have designated 2276 PASs as tissue-specific ($H' < 1.8$) and 3903 PASs as constitutive ($H' > 2.2$).

To identify the tissues in which these PASs showed significantly different usage frequency, we applied ROKU (Kadota et al. 2006). We first normalized the ϕ value of each PAS across different tissues to be zero mean and unit variance, and ranked all tissues by the normalized values in the order of increasing magnitude. For each PAS, we considered three possible types of tissue-specific patterns: (i) "up-type": a PAS is highly utilized in one or a small number of tissues compared to the rest; (ii) "down-type": a PAS is depleted in one or a few tissues; (iii) "mix-type": neither (i) nor (ii). We enumerated all possible outlier patterns with different combinations of tissues. For each combination, we computed a statistic:

$$U = \frac{1}{2}\mathrm{AIC} = n \log \delta + \sqrt{2} \times s \times \frac{\log n!}{n},$$

where s and n denote the number of outlier and nonoutlier tissues, respectively, and δ denotes the standard deviation of the scores assigned with n nonoutlier tissues (Kadota et al. 2003). The first term in the statistic measures the variance of nonoutliers, while the second term measures the model complexity. The combination with the lowest U was assigned to be the best tissue-specific pattern.

Compendium of putative regulatory features

We assembled a compendium of 658 RNA features, covering all major parameters known or with great potential to influence poly(A) regulation, including motifs of known polyadenylation regulator, unknown but potential motifs, evolutionary conservation level, secondary structure information, nucleosome positioning, and features describing transcript structures. For *trans*-acting factors, normalized 3′ seq read counts for 106 factors were used to represent their expression levels. The 764 new features were jointly standardized by subtracting by mean and dividing by standard deviation. The complete list of features is available at https://github.com/uci-cbcl/polyAcode/blob/master/features_annotation.xls.

Prediction models

We applied several classifier models to distinguish tissue-specific and constitutive PASs based on the 658 RNA features and the sequence context of each PAS, including logistic regression (LR), linear support vector machine (LSVM) with regularization, SVM with WD-kernel (WDSVM), and Adaboost (adaptive boosting). Adaboost achieved the best performance and was used for downstream analysis. It is a boosting algorithm and has been shown to solve many classification problems (Freund and Schapire 1997). Specifically, Adaboost fits a sequence of weak learners, such as small decision trees, on weighted training data. The predictions from all of the weak learners are combined through a weighted majority vote, that weak classifiers with lower classification errors usually have higher weights, to produce the final prediction. All the prediction models were implemented using Python and Scikit-learn (http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html). The WDSVM model was implemented using shogun (http://www.shogun-toolbox.org/).

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank Dr. Klemens Hertel for his critical input. This study was supported by grants from the National Institutes of Health (GM090056 and CA177651), the American Cancer Society (RSG-12-186) to Y.S., and the National Institutes of Health (HG006870) to X.X.

Received December 17, 2015; accepted February 22, 2016.

REFERENCES

Akhtar MN, Bukhari SA, Fazal Z, Qamar R, Shahmuradov IA. 2010. POLYAR, a new computer program for prediction of poly(A) sites in human sequences. *BMC Genomics* **11:** 646.

Alt FW, Bothwell AL, Knapp M, Siden E, Mather E, Koshland M, Baltimore D. 1980. Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3′ ends. *Cell* **20:** 293–301.

Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ. 2010. Deciphering the splicing code. *Nature* **465:** 53–59.

Berg MG, Singh LN, Younis I, Liu Q, Pinto AM, Kaida D, Zhang Z, Cho S, Sherrill-Mix S, Wan L, et al. 2012. U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150:** 53–64.

Cheng Y, Miura RM, Tian B. 2006. Prediction of mRNA polyadenylation sites by support vector machine. *Bioinformatics* **22:** 2320–2325.

Colgan DF, Manley JL. 1997. Mechanism and regulation of mRNA polyadenylation. *Genes Dev* **11:** 2755–2766.

Davis R, Shi Y. 2014. The polyadenylation code: a unified model for the regulation of mRNA alternative polyadenylation. *J Zhejiang Univ Sci B* **15:** 429–437.

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22:** 1173–1183.

Di Giammartino DC, Nishida K, Manley JL. 2011. Mechanisms and consequences of alternative polyadenylation. *Mol Cell* **43:** 853–866.

Early P, Rogers J, Davis M, Calame K, Bond M, Wall R, Hood L. 1980. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* **20:** 313–319.

Elkon R, Ugalde AP, Agami R. 2013. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat Rev Genet* **14:** 496–506.

Freund Y, Schapire RE. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* **55:** 119–139.

Grant GR, Liu J, Stoeckert CJ Jr. 2005. A practical false discovery rate approach to identifying patterns of differential expression in microarray data. *Bioinformatics* **21:** 2684–2690.

Hafez D, Ni T, Mukherjee S, Zhu J, Ohler U. 2013. Genome-wide identification and predictive modeling of tissue-specific alternative polyadenylation. *Bioinformatics* **29:** i108–i116.

Kadota K, Nishimura S, Bono H, Nakamura S, Hayashizaki Y, Okazaki Y, Takahashi K. 2003. Detection of genes with tissue-specific expression patterns using Akaike's information criterion procedure. *Physiol Genomics* **12:** 251–259.

Kadota K, Ye J, Nakai Y, Terada T, Shimizu K. 2006. ROKU: a novel method for identification of tissue-specific genes. *BMC Bioinformatics* **7:** 294.

Kaida D, Berg MG, Younis I, Kasim M, Singh LN, Wan L, Dreyfuss G. 2010. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468:** 664–668.

Kalkatawi M, Rangkuti F, Schramm M, Jankovic BR, Kamau A, Chowdhary R, Archer JA, Bajic VB. 2012. Dragon PolyA Spotter: predictor of poly(A) motifs within human genomic DNA sequences. *Bioinformatics* **28:** 127–129.

Lackford B, Yao C, Charles GM, Weng L, Zheng X, Choi EA, Xie X, Wan J, Xing Y, Freudenberg JM, et al. 2014. Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal. *EMBO J* **33:** 878–889.

Li W, You B, Hoque M, Zheng D, Luo W, Ji Z, Park JY, Gunderson SI, Kalsotra A, Manley JL, et al. 2015. Systematic profiling of poly(A)$^+$ transcripts modulated by core 3′ end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. *PLoS Genet* **11:** e1005166.

Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. 2013. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev* **27:** 2380–2396.

Lin Y, Li Z, Ozsolak F, Kim SW, Arango-Argoty G, Liu TT, Tenenbaum SA, Bailey T, Monaghan AP, Milos PM, et al. 2012. An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Res* **40:** 8460–8471.

Martin G, Gruber AR, Keller W, Zavolan M. 2012. Genome-wide analysis of pre-mRNA 3′ end processing reveals a decisive role of human cleavage factor I in the regulation of 3′ UTR length. *Cell Rep* **1:** 753–763.

Mayr C. 2015. Evolution and biological roles of alternative 3′UTRs. *Trends Cell Biol* **26:** 227–237.

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499:** 172–177.

Rogers J, Early P, Carter C, Calame K, Bond M, Hood L, Wall R. 1980. Two mRNAs with different 3′ ends encode membrane-bound and secreted forms of immunoglobulin mu chain. *Cell* **20:** 303–312.

Schug J, Schuller WP, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ Jr. 2005. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol* **6:** R33.

Shi Y. 2012. Alternative polyadenylation: new insights from global analyses. *RNA* **18:** 2105–2117.

Shi Y, Manley JL. 2015. The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site. *Genes Dev* **29:** 889–897.

Shi Y, Di Giammartino DC, Taylor D, Sarkeshik A, Rice WJ, Yates JR III, Frank J, Manley JL. 2009. Molecular architecture of the human pre-mRNA 3′ processing complex. *Mol Cell* **33:** 365–376.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050.

Tabaska JE, Zhang MQ. 1999. Detection of polyadenylation signals in human DNA sequences. *Gene* **231:** 77–86.

Takagaki Y, Seipelt RL, Peterson ML, Manley JL. 1996. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. *Cell* **87:** 941–952.

Tian B, Graber JH. 2012. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip Rev RNA* **3:** 385–396.

Tian B, Manley JL. 2013. Alternative cleavage and polyadenylation: the long and short of it. *Trends Biochem Sci* **38:** 312–320.

Xie B, Jankovic BR, Bajic VB, Song L, Gao X. 2013. Poly(A) motif prediction using spectral latent features from human DNA sequences. *Bioinformatics* **29:** i316–325.

Yao C, Biesinger J, Wan J, Weng L, Xing Y, Xie X, Shi Y. 2012. Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation. *Proc Natl Acad Sci* **109:** 18773–18778.

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:** 405–445.

Poly(A) code analyses reveal key determinants for tissue-specific mRNA alternative polyadenylation

Lingjie Weng, Yi Li, Xiaohui Xie, et al.

RNA published online April 19, 2016