Published online 4 June 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 2 **1**
https://doi.org/10.1093/nargab/lqab042

Known sequence features can explain half of all human gene ends

Aleksei Shkurin[1,2] and Timothy R. Hughes[1,2,*]

[1]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada and [2]Terrence Donnelly Centre for Cellular & Biomolecular Research, Toronto, ON M5S 3E1, Canada

Received February 18, 2021; Revised April 14, 2021; Editorial Decision April 19, 2021; Accepted May 10, 2021

ABSTRACT

Cleavage and polyadenylation (CPA) sites define eukaryotic gene ends. CPA sites are associated with five key sequence recognition elements: the upstream UGUA, the polyadenylation signal (PAS), and U-rich sequences; the CA/UA dinucleotide where cleavage occurs; and GU-rich downstream elements (DSEs). Currently, it is not clear whether these sequences are sufficient to delineate CPA sites. Additionally, numerous other sequences and factors have been described, often in the context of promoting alternative CPA sites and preventing cryptic CPA site usage. Here, we dissect the contributions of individual sequence features to CPA using standard discriminative models. We show that models comprised only of the five primary CPA sequence features give highest probability scores to constitutive CPA sites at the ends of coding genes, relative to the entire pre-mRNA sequence, for 41% of all human genes. U1-hybridizing sequences provide a small boost in performance. The addition of all known RBP RNA binding motifs to the model, however, increases this figure to 49%, and suggests an involvement of both known and suspected CPA regulators as well as potential new factors in delineating constitutive CPA sites. To our knowledge, this high effectiveness of established features to predict human gene ends has not previously been documented.

INTRODUCTION

Cleavage and polyadenylation (CPA) is the process of cleaving precursor mRNA and adding a string of adenine (A) nucleotides to the 3′-end of a primary RNA transcript (1,2). In human, CPA is mediated by four main protein complexes (the 'core' CPA machinery) that recognize five *cis*-acting RNA elements in the pre-mRNA (3) (Figure 1). First, one or more instances of UGUA are usually found up to 100 nt upstream of the CPA site. The UGUA elements are recognized by NUDT21/CFIm25, a subunit of Cleavage Factor Im (CFIm) (2,4,5). Second, the polyadenylation signal (PAS), typically either AAUAAA or AUUAAA (or ∼11 minor variants), is found around 30 nt upstream the CPA site (6,7). The PAS is the best known of the sequence signals, as it is found in the majority of known human CPA sites (8,9). It is recognized by the CPSF (Cleavage and Polyadenylation Specificity Factor) subunit WDR33, likely in conjunction with CPSF4/CPSF30 and CPSF1/CPSF160 (10,11). The endonuclease subunit CPSF3/CPSF73 mediates the cleavage, with the cleavage site usually preceded by a CA or UA dinucleotide (2). Poly-U sequences, preferred by CPSF4/CPSF30 (12), are also often found surrounding the cleavage site (13), and sometimes further upstream (14). Finally, degenerate U- and GU-rich downstream elements (DSEs) are often found starting ∼20 nucleotides downstream of CPA sites (15). These elements are recognized by CSTF2/CstF-64, the RRM-containing subunit of the Cleavage Stimulation Factor complex (16). Consistent with the fundamental importance of CPA, its misregulation is associated with a wide range of genetic disorders. For example, a mutation within the PAS of FOXP3 (AATAAA to AATGAA) leads to immunodysregulation polyendocrinopathy (17), while mutation of AATAAA to AATACA in TP53 increases susceptibility to cancers including cutaneous basal cell carcinoma, prostate cancer, glioma and colorectal adenoma (18). Similarly, misregulation of U-rich upstream elements was associated with conditions affecting inflammatory hypercoagulation and tumor invasion (19).

Despite this detailed knowledge, the precise RNA sequence cues that determine actual CPA sites remain a topic of active research. Collectively, an exact match to all of the sequence features above has the potential for relatively high specificity, such that only one or a few sites would be expected in random sequence of the average size of a human pre-mRNA (23 kb) (see Materials and Methods for estimates). In reality, however, CPA sites are heterogeneous, with each containing a different assembly of sequence features; combinations of only a subset of the CPA sequence elements would occur much more often. Indeed, 'cryptic' CPA sites, which would lead to truncated transcripts, ap-

*To whom correspondence should be addressed. Tel: +1 416 859 1492; Email: t.hughes@utoronto.ca

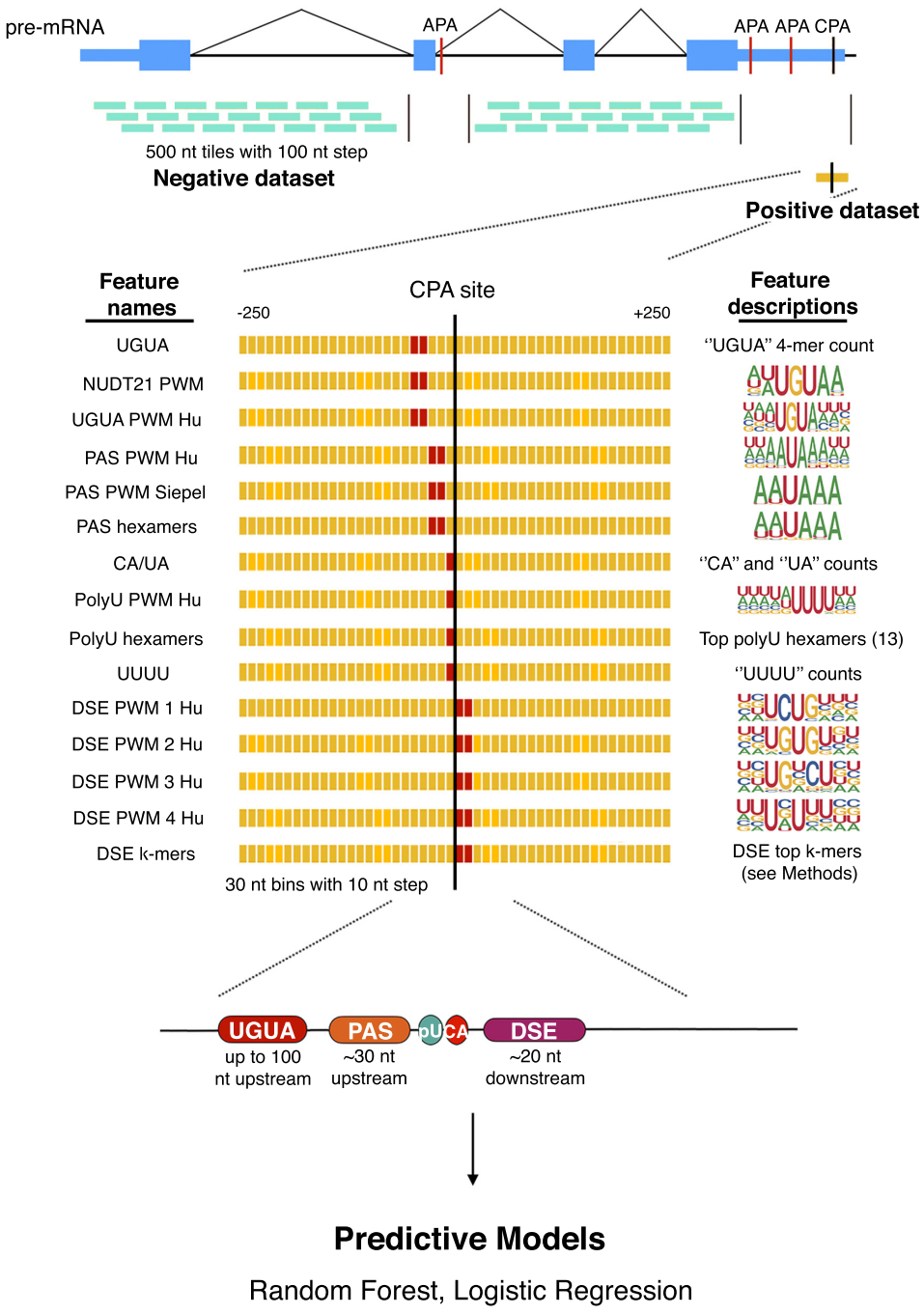

Figure 1. Schematic of the baseline model pipeline. CPA sites from PolyA_DB 3 (37) were processed to identify one constitutive CPA site per gene; 500 nt sequences surrounding these sites are used as the positive dataset. The negative dataset consists of 500 nt portions of genes not overlapping with any annotated CPA sites. The feature matrix is calculated using PWMs scores or *k*-mer counts within tiling windows. Red positions within the grid indicate the expected positions of each corresponding element, as illustrated below the grid. AUROC and AUPRC are used as evaluation metrics.

pear to be widely distributed, and at least one mechanism is known to suppress usage of these sites (the U1 snRNP) (20–23).

Metazoans appear to have taken advantage of this flexibility in CPA, with most genes containing multiple CPA sites that produce functional transcript isoforms differing in their terminal exons or 3′ UTR length (3,24,25), thus impacting the protein sequence and/or regulation of the transcript. Alternative CPA sites are often tissue-specific (24,26–28) and presumably there are specific mechanisms that dictate their usage. Indeed, there are several examples where corresponding regulators have been identified. For example, the neuronal RBP Nova acts as an inhibitor when binding close to CPA sites, and as an enhancer when binding distant from CPA (29), while neural Hu proteins inhibit CPA sites with U-rich elements (30).

A series of previous computational analyses have sought to predict CPA sites from RNA sequence. CPA is a tractable computational problem in which the goal is to find patterns of sequence features that discriminate actual CPA sites from the remainder of the gene. Sequence elements and their positions can be described as a vector that is compatible with probabilistic inference methods. Early efforts to computationally predict human CPA (31,32) used motif thresholding and quadratic discriminant analysis, respectively, to show that the sequence determinants known at the time (PAS and DSE) have significant classification ability (e.g. 79% accuracy (32)). More recently, CPA site prediction has increasingly used machine learning with larger feature sets encompassing both k-mers and known RNA binding motifs for proteins. Better overall statistical performance has been reported, but the data sets and evaluation criteria employed varied dramatically, complicating direct comparisons among studies. Xie *et al*. (33) used 3-mers derived from top variants of the PAS as input into an HMM-SVM, and reported accuracy of 85%. Hafez *et al*. (34) used an SVM trained on ±100 nt around CPA sites to obtain an area under the receiver operating curve (AUROC) of 0.996, but employed only the terminal exon sequences as negatives. Leung *et al*. (35) used both 'hand-crafted' feature vectors (composed of RBP RNA binding motifs and k-mers) and k-mer sets learned directly from the sequence, employing Convolutional Neural Networks to directly learn alternative polyadenylation patterns, and reported AUROC of 0.97 (hand-crafted) and 0.98 (*k*-mers learned directly) at the task of discriminating CPA sites from neighbouring genomic sequence. None of these papers explicitly report CPA site predictions genome-wide, and do not address why CPA does not occur elsewhere in the primary transcript, which is typically many times longer than the terminal exon. AUROCs in this range would be expected to predict many CPA sites per human gene, on average (AUROC of 0.99 would be roughly equivalent to 1 out of 50 randomly selected sequences scoring as a false positive).

Overall, several critical issues remain unresolved. First, none of the previous studies addressed whether the five well-established sequence features can indeed specify known constitutive CPA sites relative to all non-CPA sequence within primary transcripts. Second, it is difficult to compare the results of previous analyses because different sets of sequences and evaluation metrics were used. Third, linking k-mers to biological mechanisms can be challenging. For example, Hafez *et al.* (34), which used k-mers as features to generate a model with good predictive ability, provided very limited mechanistic explanation (primarily a sequence logo reflecting the general pattern of the most informative sequences relative to CPA, which resemble known regulatory elements). Fourth, until the recent availability of large 3′-end seq datasets, many studies used reference databases that filter out potential CPA sites lacking the established PAS sequence, thus introducing circularity.

Here, we dissect the contributions of diverse RNA sequence features to CPA site discrimination, with the goals of simultaneously increasing performance in a realistic test framework (i.e. with a large excess of negatives derived from real genic sequences), and deriving a set of minimal features that are sufficient to obtain high performance. We find that standard supervised learning approaches (Random Forests and Logistic Regression), employing a small number of established features represented as either classical position weight matrix (PWM) motifs or a handful of short *k*-mers, are surprisingly effective at identifying constitutive CPA sites at the ends of human genes. Addition of hundreds of diverse sequence features to the model (U1 binding sites, and all known RBP RNA binding motifs) improves the model by ∼20%, such that the constitutive CPA site is the highest scoring sequence window in half of all human genes. Thus, while CPA is potentially controlled by a large number of protein factors, the core CPA machinery alone plays a major role in defining human gene structures.

MATERIALS AND METHODS

Initial calculation of specificity of known CPA sequence features in random sequence

The probability of observing UGUA within a 100 base window is 0.39 ($100/4^4$). The probability of observing any of the 13 variants of the PAS in a 20 base window is 0.063 ($13*20/4^6$), if they are weighted equally, or 0.022, if they are weighted by their probability at CPA sites (as a proxy for activity). The probability of observing UUUU in a 30 base window is 0.11 ($30/4^4$). The probability of observing a CA or UA dinucleotide is 0.13. The probability of observing G or U for eight consecutive bases (taken as a strong DSE (36) within a 40 base window is roughly estimated as 0.156 ($40/2^8$). The expected frequency of all five features occurring surrounding any base is $0.39 \times 0.063 \times 0.11 \times 0.13 \times 0.156 = 0.000055$, or one every 18 kb if the 13 PAS variants are weighted equally (every 52 kb with the PAS sites weighted). We note that the low G/C content of the human genome would make these A/U-rich sequences more likely.

Data sources

Human CPA site annotations were obtained from PolyA_DB 3 (37). To select constitutive CPA sites, we developed a significance metric by multiplying percentage of samples expressed (PSE) and mean reads per million (RPM) scores provided by PolyA_DB 3. Sites that were selected for positive training dataset are those that have highest PSExRPM score, and located in (or at the end of) the 3′UTR of the longest isoform of the corresponding UCSC gene on the table browser. We retrieved ±250 nt of sequence around each constitutive CPA site from UCSC (hg19).

To create a negative dataset for training the classifiers, we tiled the same longest UCSC isoforms genes into 500 nt windows with 100 nt steps, and removed those that overlap any CPA sites in PolyA_DB 3. For training, we then randomly subset negative sequences to obtain a 30:1 ratio of negative to positive data.

PWMs used for this study were taken from CISBP-RNA database (38) and ENCORE (39). We derived the 'Hexamer' PAS PWM by weighting each hexamer in (8) according to its counts in the constitutive CPA sites described above. DSE PWMs were obtained from (13). To score U1 sites we used the 5′SS MaxEntScan score (40) and RNAhybrid (41). The 'Siepel PAS' site is from (42). The 'DSE *k*-

mers' feature is calculated as the count of all instances of 'G', 'U', 'GG', 'GU', 'UG' and 'UU'. The 'UGUA' feature is a PWM representing this single sequence.

Feature matrices

To calculate feature matrices, we break each individual 500 nt sequence from positive and negative data into bins of size 30 nt with a 10 nt step. Next, for each of the PWMs in each 30 nt bin, we calculate the maximum log(odds) score in each of those bins, and convert these predicted energy scores to predicted affinity. For MaxEntScore and RNAhybrid, we also select the maximum score per each bin (leaving MaxEntScore in log domain).

Machine learning

Random Forest and Logistic Regression classifiers were trained in Python (v3.5.1) using scikit-learn library version 0.22.2. Random Forest training used RandomForestClassifier with 30 000 trees, a minimum sample split of 5, class weight 'balanced'. For the baseline Logistic Regression, we selected 'l1' penalty, regularisation strength of 0.1, tolerance of 0.01, 'saga' solver, and 'balanced' class weight. For the constitutive vs cryptic Logistic Regression model, use same parameters, except regularisation strength of 0.0018 (we examined a series of regularization parameters (lambda values) and identified a value after which there was a rapid decline in performance).

RESULTS

Compilation of CPA data and 'core' CPA motifs

We began by organizing a system to computationally interrogate the contributions of the five established *cis*-acting RNA elements, and their positions relative to the cleavage site (Figure 1 shows a schematic). This system is comprised of four basic components: (i) a dataset of CPA sites (positives), and non-CPA sites (negatives); (ii) motif models (i.e. PWMs), individual k-mers, and other scores that represent predicted affinity of RBPs to any given sequence, and scores obtained from these models for tiled sequence windows relative to the CPA sites; (iii) algorithms that input the RNA binding motif scores for each tiling window as features, and output both a probability that reflects confidence that any given example is a CPA site, as well as information about the relative importance of the individual features and (iv) a testing regime, which quantifies the predictive ability of each algorithm using several criteria. Implementing each of these components involves numerous choices. In each case, we sought to minimize bias and circularity, and to achieve results that are mechanistically interpretable, in order to be biologically meaningful, and as simple as possible, to avoid overfitting and ambiguity.

For the dataset of CPA sites ('positives'), we employed PolyA_DB 3, which is based on 3′-READS (a 3′-end sequencing method) applied to a panel of cell lines and also to mixed tissue (37). This database includes 58 676 CPA sites, each associated with values including mean RPKM and PSE (Percentage Samples Expressed). Because our initial goal was to characterize contributions of the core machinery to CPA, and the core machinery is presumably constitutive, we selected our initial set of positives as 15 848 CPA sites (allowing only one per UCSC gene) with both high RPKM and PSE scores that overlap or flank 3′UTRs (see Materials and Methods). We refer to these as 'constitutive' CPA sites. We excluded all other CPA sites annotated in PolyA_DB 3, as they represent alternative CPA sites. We used a 500 nt window to represent each CPA site (–250 to +250). We generated 'negative' sequences (i.e. those that are not CPA sites) by first collecting all 500-nt tiling windows (with offset 100 nt) in the sense strand of genes with constitutive CPA sites, and then removing windows that overlap more than 10% with any CPA site in PolyA_DB 3. This process generated 15,352,546 negative examples; generally, only a randomly selected subset of ∼100 000 were used for training the models, and ∼400,000 employed for testing. We generated training and testing sets by splitting the chromosomes (Chromosomes 1–14 are used for training, and 15–22 and X are used for testing).

We compiled RNA binding motif models from diverse sources (shown in Figure 1; see Materials and Methods for details) in order to calculate features in the models. We included multiple representations for each component of the core CPA machinery (e.g. several PAS signals have been described, and to our knowledge it is unknown which best reflects binding of CPSF, or whether a single motif is sufficient). It is unclear whether motif models learned from CPA sites accurately represent the full sequence preferences of these proteins; therefore, where possible, we used RNA binding motifs derived from data collected without using knowledge of established CPA sequences (e.g. from *in vitro* assays such as RNAcompete (43)). We note that some of the motifs have been learned from human mRNA sequences present in the training data, which could lead to circularity, but we also note that the motif representations are simple (and thus less likely to be overfitted). To generate a feature vector for each CPA site (positive or negative), we scored windows of length 30 bases, in 10 base tiling steps, over each 500 base sequence. The motif scores were represented in linear domain (i.e. 10^log(odds)), which we reasoned would reflect relative preference (i.e. relative Ka). We used the max score for each window. Thus, with 47 windows and 15 different representations of the elements (Figure 1), there are initially 705 features comprising the feature vector for each example sequence.

For the learning algorithms, we employed two commonly used implementations of different machine learning strategies. The first, Random Forests (RF), employs a large set of decision trees, which has the advantage that it inherently captures logic relationships and is thought to be less prone to overfitting because it uses an ensemble of decorrelated classifiers. It can also be used to obtain importance scores for each feature. The second, Logistic Regression (LR) with L1 regularization, does not inherently capture logical relationships, but it has the advantage that the feature selection through Lasso regularization tends to collapse redundant features, giving zero weight to noninformative features. The directional weights (coefficients) given to each feature are easily obtained.

Performance of prediction methods confirms critical importance of PAS and DSE

To evaluate the performance of the models, we report both AUROC (Area under the Receiver Operating Curve) and AUPRC (area under the Precision Recall Curve), with a ∼30-fold excess of negatives to positives (which represents a large excess, but avoids unwieldy run times). Figure 2 shows these values for six different variants of the predictors, which together with the feature importance scores and weights (Figure 3) enable dissection of the contributions of the core machinery. The first two models are the initial RF and LR models, with 705 features per example; we refer to the 705 feature RF model as the 'baseline' model. These are the best performing models (Figure 2), and RF and LR are in reasonable agreement regarding which features are important (Figure 3). Strikingly, the AUROC values obtained (0.98 and 0.97, respectively) are similar or better to those reported in previous studies that employed more complex and less interpretable models (33–35). The models largely confirmed the regions where the CPA factors are known to act, with the PAS most critical 20–30 bases upstream of the CPA site, and the DSE at 10–20 bases downstream. The feature importance scores for the models differ somewhat, presumably because only LR involves a regularization step, in which the number of weighted features is minimized, but overall the two models are consistent. The CA/UA dinucleotide appears to be dispensable to both models, perhaps because it occurs frequently at random and controls precise local placement of the CPA site, which was not considered here.

The third and fourth curves (Top10 LR and Top10 RF) in Figure 2 show the performance of RF and LR models encompassing the ten features with highest weights in the baseline LR model (all reside in the PAS and DSE in Figure 3B; indicated with black boxes and white circles). These models are nearly as effective as the full 705 features, albeit with a ∼10% decrease in AUPRC. Upon further simplification, however, the models are deeply compromised: the fifth and sixth curves in Figure 2, 'PAS/DSE', are derived using the maximum scores of the three PAS features and seven DSE features within the bins where they are boxed in Figure 3B, respectively. Intriguingly, four very different representations of the DSE all appear to be important, at the same positions: collapsing them into a single value (by taking the maximum score of any of them per bin) results in substantial decline in performance. We speculate that the interaction of CPSF and CstF with each individual sequence and with each other may be more complex than what can be captured by PWM motif models (see Discussion).

Model predictions on complete pre-mRNAs and pathogenic mutations

We next asked how well the baseline model specifically predicts the 3′ ends of genes. To do this, we obtained the primary transcripts from UCSC for the 15 848 human protein-coding genes that had a 'constitutive' CPA site, as defined above, and analyzed all possible 500-base tiling windows within the pre-mRNA, i.e. starting at every base. Examples are shown in Figure 4A and B, illustrating that there is little bias in position within the gene; the probability distributions for the constitutive CPA sites and all negative sequences are shown in Figure 4C. Strikingly, the baseline model assigned the highest probability to the 'constitutive' CPA site for 41% of all genes. The significance of this outcome is discussed below (see Discussion), but we believe it is a higher figure than would have been anticipated, and we take it to confirm that the core CPA machinery plays a major role in determining not only CPA sites but also gene ends. As expected, short genes are more likely to have the highest scoring sequence at the end of the gene, because the probability of encountering a cryptic site at random would be lower (Figure 4D).

To assess the ability of the baseline model to forecast the consequences of mutation events, we used FOXP3 and TP53 CPA sequences and compared the prediction scores with and without known PAS mutations (17,18). For FOXP3, the original CPA site was assigned a probability of 0.67, and the mutated sequence 0.04. In case of TP53, the performance changed from 0.73 to 0.34. These results show that the model is sensitive to known pathogenic mutations and further demonstrate its ability to identify functional CPA sites.

Predicting modifiers of CPA beyond the baseline model

We next asked whether inclusion of additional features in the model would aid in discriminating constitutive from cryptic CPA sites. We defined cryptic sites as those that the baseline model assigned probability scores higher than the overall average for constitutive CPA sites ($D > 0.8$), and that did not overlap with any known CPA sites in PolyA _DB 3, therefore excluding the possibility that these sequences are alternative CPA sites. Among the 15 848 genes, there were 58 271 such sites. We formulated the problem as a discrimination task between the constitutive CPA sites and the cryptic sites, instead of discrimination between constitutive CPA sites and randomly selected pre-mRNA sequences as previously, but otherwise applied a similar framework as above.

We first examined the ability of U1 recognition elements to discriminate globally between the constitutive and cryptic CPA sites, because U1 can suppress CPA (20–23). We considered two representations of U1 binding: RNAhybrid (41), to calculate binding affinity of the classical U1 7-mer (44) to any given 7 base RNA sequence, and 5′MaxEntScan, which calculates the likelihood that a given 9-mer is a 5′ splice site (of which U1 recognition is a major component) (40). Both models displayed some bias in constitutive vs. cryptic CPA sites (lines show medians at each base relative to the predicted CPA site in Figure 5A and B). We observed a slight overall decrease in high-affinity U1 RNAhybrid scores for constitutive CPA sites, as expected if U1 suppresses CPA. MaxEntScan displayed more striking biases, including a depletion following constitutive CPA sites. But, both U1 measures also displayed high standard deviation (shading in Figure 5A and B), and we note that their overall scores would be biased by deviations in base content caused by the core CPA sequences.

We then asked how well the U1 scores serve in a classification framework. Here, we used a sequence window of –70 to +70, with 20 nt windows tiled every 10 nt, because initial

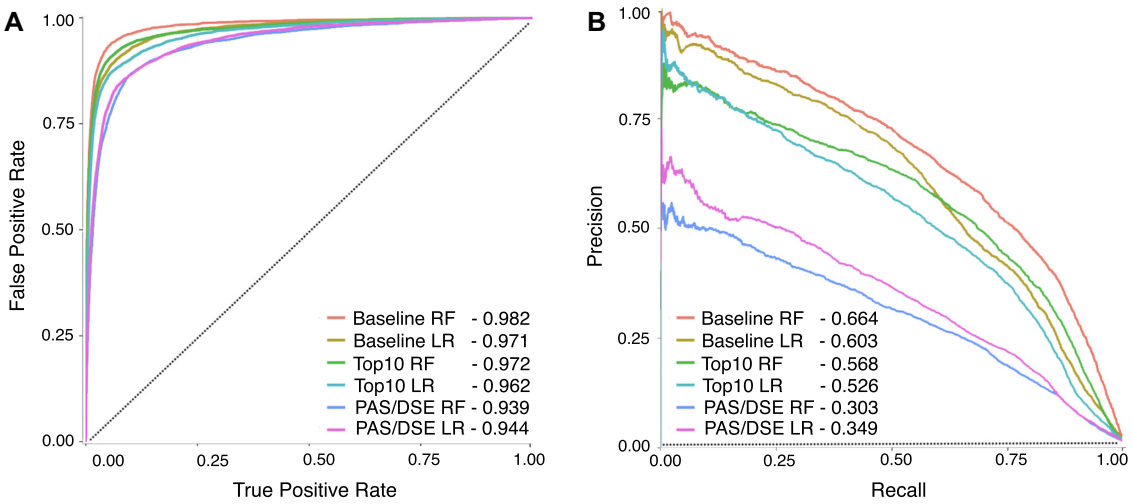

Figure 2. Performance of the baseline models (i.e. those built using elements recognized by known CPA machinery). (**A**) Receiver operating curve (ROC); (**B**) Precision recall curve (PRC), with a 50-fold excess of negatives. Models shown on the plots include baseline RF and LR models with all 705 features, Top10 RF and LR models with only Top10 features based on LR weights, and PAS/DSE RF and LR models that are built using only two features (i.e. PAS and DSEs), as described in the text. The legends show area under the curve values for each model. Dotted lines show the performance of a random guessing classifier.

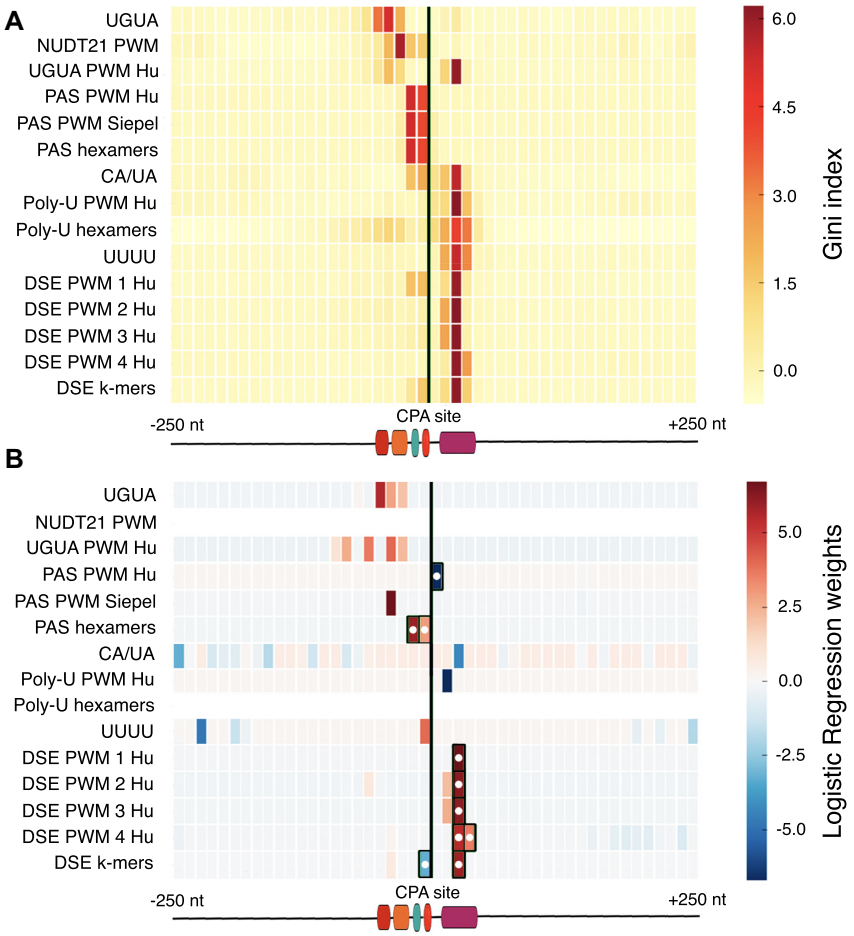

Figure 3. Heatmaps showing feature importance scores identified from (**A**) RF and (**B**) LR baseline models. Black vertical lines indicate the location of the CPA site relative to top-scoring features. In panel (B), the matrix rows are normalized prior to plotting, and black boxes and white dots indicate the Top10 features.

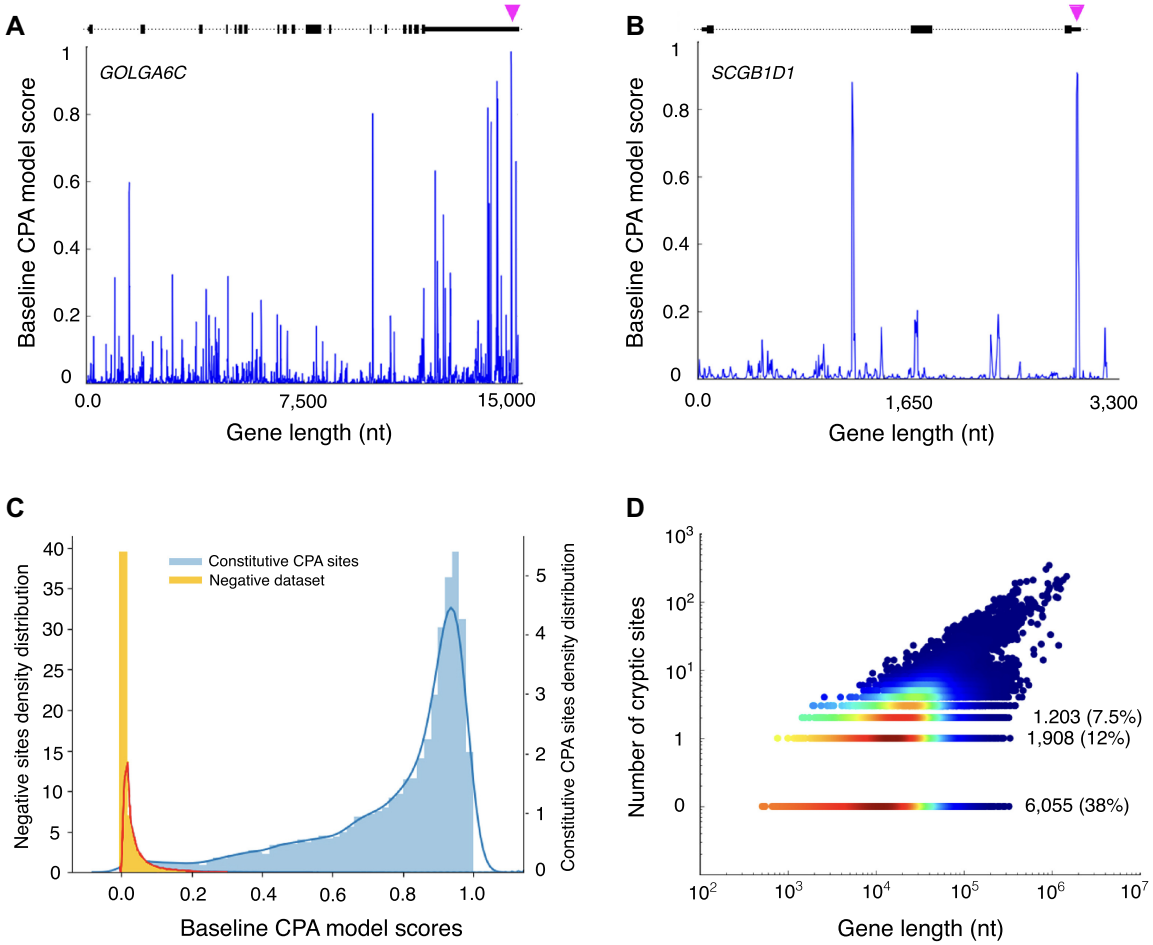

Figure 4. Overview of cryptic CPA sites. (**A, B**) Baseline model CPA probability scores for the genes GOLGA6C (a) and SCGB1D1 (b) at base-level resolution. (**C**) Distribution of CPA site probabilities the model assigns to constitutive CPA sites and all other sites. (**D**) Scatter plot showing the number of cryptic sites per gene (i.e. those with D>0.8, grouping adjacent bases exceeding this metric) vs. the length of the corresponding gene. The number of genes having 0, 1 and 2 cryptic sites and their proportion are indicated.

trials with different window sizes indicated that most of the predictive signal is near the centre (i.e. the CPA site), and also in order to accommodate a larger number of features (see below). We used an LR model as it allowed us to perform L1 regularization with the goal of reducing potential redundancy in training data, and generating a subset of top scoring features. Together, the two U1 representations do provide some classification ability (AUROC of 0.58) (Figure 5C and D), with depletion just upstream of CPA having greatest impact (Figure 5E). By the conventional interpretation of AUROC, addition of these U1 measures represents a ∼16% performance increase over random guessing.

We then extended the constitutive versus cryptic CPA model to include all known RNA binding motifs for human RBPs. Because the number of CPA sites is large, there is sufficient statistical power to simultaneously consider all 324 different PWMs for human RBPs (obtained from CisBP-RNA (38) and from (39)) as well as U1. The LR model resulting from a standard procedure of feature reduction (see Materials and Methods) is roughly four times better than the model utilizing only U1 (0.86 versus 0.58 ROC, i.e. 72% versus 16% performance increase over random guessing). This model retains many features (71 out of 325, including U1) with further reduction resulting in rapid loss of performance. A heatmap of feature importance scores is shown in Figure 6C. A clear outcome of this analysis is that there are potentially a large number of sequence elements and corresponding trans-acting factors that modulate CPA usage. RNA binding proteins shown in Figure 6C include many known or suspected CPA regulators (names colored in blue). U1 recognition sites are retained, but do not have a major influence on the model performance.

Given that we are including 324 different PWMs, it is possible that the predictive power of the model comes from capturing base content in the region and is not necessarily related to structure of the corresponding motifs. To assess this possibility, we trained an additional model, where nucleotide positions of each PWM were permutated. The resulting model had a dramatic decrease in performance (see Figure 6A and B), indicating that the precise composition of motifs is important.

Finally, we asked whether combining the two models (the 'baseline', and the 'constitutive CPA vs cryptic' models), by simply multiplying their probabilities, would increase the

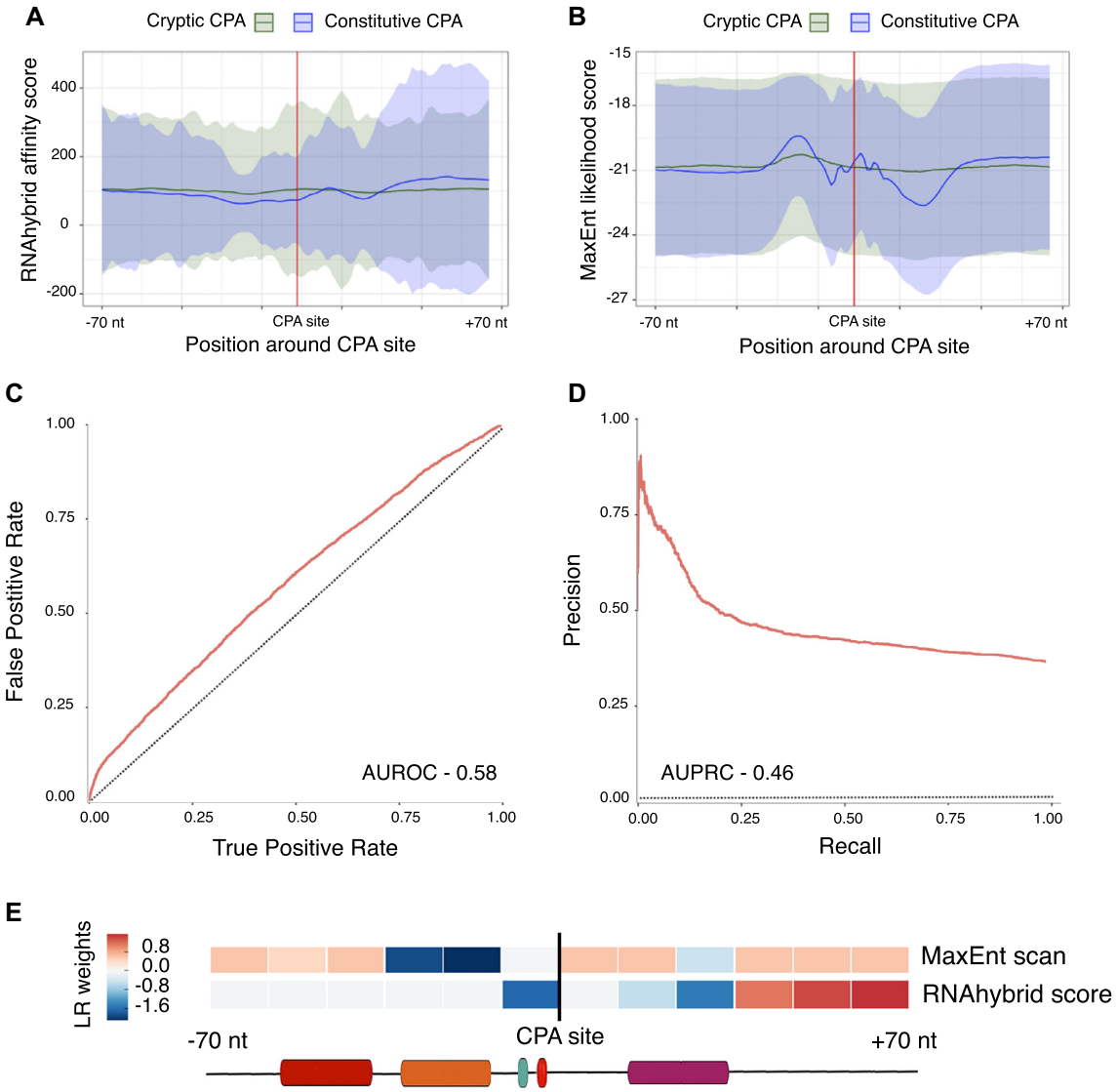

Figure 5. Overview of U1 analyses. (**A**, **B**) Average scores for RNAhybrid (A) and MaxEntScan (b) in a 140 nt sequence window centered on CPA sites. The line shows median and shading shows the standard deviation at each position. (**C**, **D**) ROC and PRC of the cryptic CPA versus constitutive CPA model trained using U1 RNAhybrid and MaxEntScan scores. (**E**) LR weights for RNAhybrid and MaxEntScan scores at each position.

proportion of genes for which the constitutive CPA site is given the highest score. Indeed, this combined score yielded the highest value at the constitutive CPA site for 49% of all genes (relative to 41% for the baseline model alone). Thus, while a clear increase in specificity is achieved by the two-stage model, there is still a large proportion of gene ends that are not fully explained even by the combined model (discussed below).

DISCUSSION

A major outcome of this study is that a model utilizing representations of the sequence preferences of only four different components of the 'core' CPA machinery has good ability at the difficult task of pinpointing the ends of human genes. To our knowledge, this is the first such global demonstration encompassing the majority of human gene sequences. The model identifies 'cryptic' sites, but they are not nearly as prevalent as would be expected from considering only the PAS, which is often used to identify cryptic sites. We propose this outcome to signify that the use of PWMs, and models that can incorporate spatial preferences and (potentially) interactions among the features is beneficial, relative to simple calculations based on the appearance of k-mers and consensus sequences. As a corollary, this outcome also shows that components of the known CPA machinery are quite specific in combination: specificity of the models is cut in half when only simplified PAS and DSE are included.

The models themselves have properties that may reveal intriguing biology. UGUA, for instance, is important to the models only between 30–50 nt upstream, although it is described in the literature as appearing <100 (13), and the LR model assigns weights up to 100 bp from the CPA site. We note that CFIm is a dimer, such that a second UGUA up-

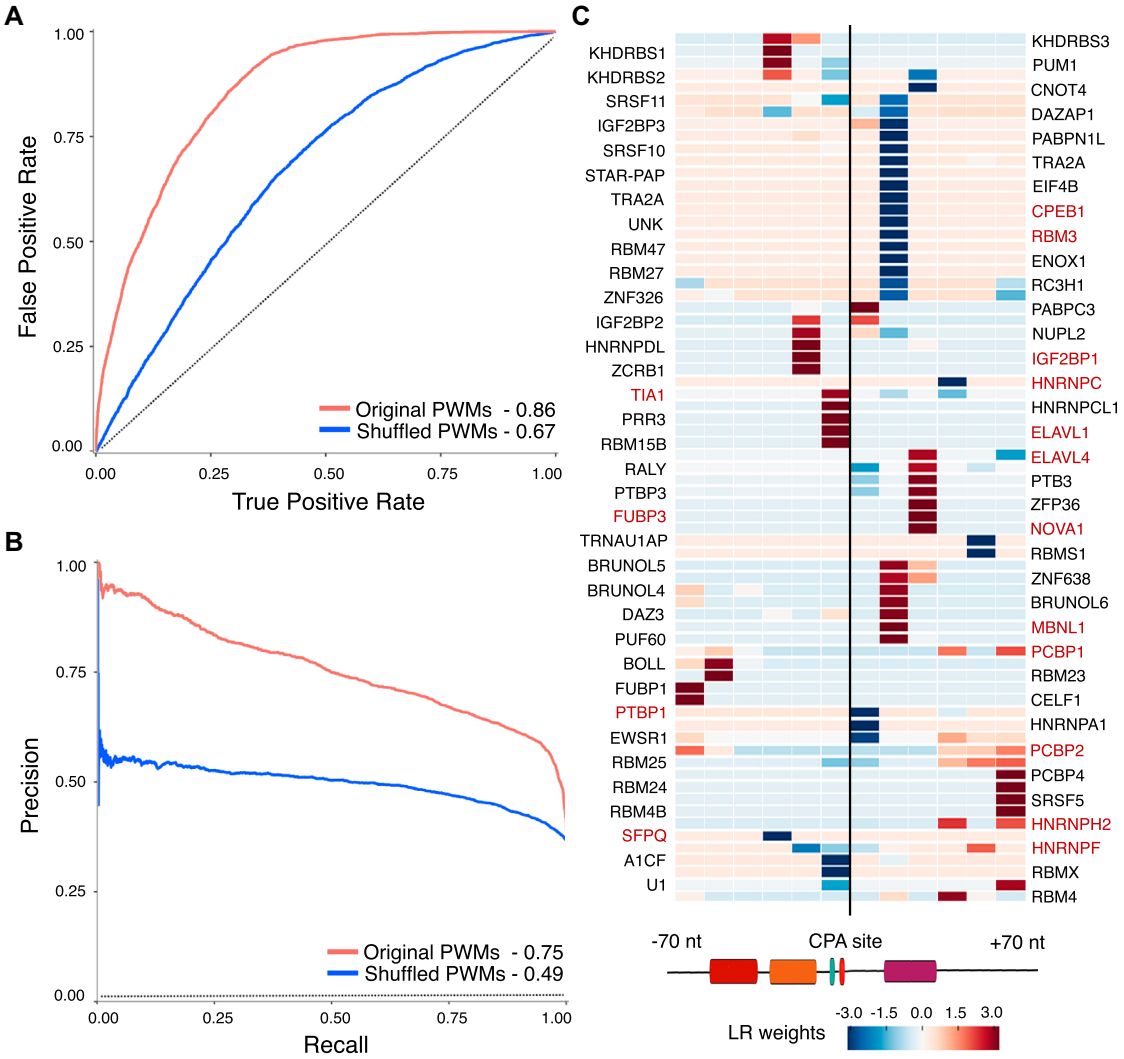

Figure 6. Overview of cryptic vs constitutive CPA model incorporating all known RBP motifs. (**A**, **B**) ROC and PRC of the model that uses both U1 representations, and 70 retained RBP PWMs, and ROC and PRC of the model trained using shuffled PWMs. (**C**) LR weights for the top-scoring RBP PWMs and U1 feature. RBP names highlighted in red have been previously reported in the literature to act as CPA regulators.

stream at a constrained distance may contribute to binding. But, if this were the case, the RF model should have detected it. If the second site did not have a constrained position, it would not be expected to greatly impact specificity. We also note that both the RF and LR models indicate that incorporation of multiple representations of the DSE is important. A trivial explanation would be that the current PWMs are inadequate, but it is also possible that a single PWM motif cannot fully capture the sequence specificity of the RNA binding activity, and/or that the DSE sequence impacts the spatial organization of the larger complexes. If so, this observation may provide an explanation why no simple regular expression has emerged that accurately identifies CPA sites.

The outstanding question remains as to how bona fide CPA sites are distinguished from cryptic sites. Models that incorporate U1 sites and all known RBP motifs provide some benefit, but not a satisfying explanation. The RBPs identified in our cryptic vs. constitutive model (Figure 6C) do encompass several that were previously associated with alternative polyadenylation, supporting their relevance in this framework, and also showing that their importance can be captured in the models. Since this analysis included all known RBP RNA binding motifs, it would seem that there may be missing information—either a known RBP has an unknown or erroneous motif, or there are as-yet unknown RBPs (or other factors such as miRNAs or specific secondary structures)—otherwise, the model would have been more effective.

It is likely that other aspects of mRNA transcription and processing play a role in definition of gene ends, e.g. by licensing CPA activity. Indeed, the specificity of CPA for Pol II transcripts is believed to be controlled largely by the physical association of CPA-related proteins with the carboxyl-terminal domain (CTD) of the RNA polymerase II (Pol II) large subunit (45), which is in turn associated with the phosphorylation of Ser2 residues of the CTD heptad repeats (46,47), observed mainly near gene ends (48). Nuclear export factors also accumulate on the terminal exon, and

can influence CPA (49). Perturbation of other aspects of transcription and RNA processing (e.g. mRNA capping) can have an effect on polyadenylation (50), suggesting that longer-distance interactions along transcripts can occur.

How all of these events are dictated by the DNA and RNA sequence is unclear. The terminal exon structure itself is an obvious candidate: terminal exons are distinguished by large size, and by lacking a 5′ splice site. Surprisingly, however, 48% of terminal exons in our data set do contain at least one sequence scoring above 7.5 in MaxEntScan (the median score of bona fide 5′ sites we examined in known internal exons); thus, absence of a 5′ splice site sequence appears unlikely to be a discriminating factor. It has long been observed that the 3′-terminal intron is important for efficient RNA 3′-end formation (51), and *in vitro*, the terminal 3′ splice cite and the CPA site are coupled and mutually reinforcing (52). We speculate that sequence features analogous to splicing enhancers may exist. Identifying such elements is complicated due to the large size of terminal exons and the presence of multiple constraints: both coding sequence and 3′ UTRs are often highly conserved, presumably due to mechanism other than CPA specification. Nonetheless, dissecting how other sequence properties of gene ends interact with the core CPA sequence elements, potentially over long distances, may be the key to a complete understanding of how gene ends are recognized. The models we present provide a strong framing for the problem and will also be instrumental in this endeavor.

DATA AVAILABILITY

We employed public data sources as described above. The constitutive CPA sites, negative samples, matrixes for producing heatmaps used in this paper, and all PWMs are posted at https://hugheslab.ccbr.utoronto.ca/supplementary-data/HumanGeneEnds/. Executable code and descriptions thereof are found at https://github.com/AlekseiShkurin/HumanGeneEnds

ACKNOWLEDGEMENTS

T.R.H. is the Billes Chair of Medical Research at the University of Toronto and holds a Canada Research Chair. We are grateful to Daniel Dominguez, Chris Burge, Amit Blumberg, Andre Martins, and Adam Siepel for providing PWMs. We thank Aiden Sabibi, Sam Lambert, Mihai Albu, Quaid Morris, and Ben Blencowe for helpful conversations, and Nick Stepankiw, Debashish Ray, Kaitlin Laverty and Sara Pour for critical evaluation of this manuscript.

FUNDING

Canadian Institutes of Health Research [FDN-148403]. Funding for open access charge: Canadian Institutes of Health Research.
Conflict of interest statement. None declared.

REFERENCES

1. Neve,J., Patel,R., Wang,Z., Louey,A. and Furger,A.M. (2017) Cleavage and polyadenylation: ending the message expands gene regulation. *RNA Biol.*, **14**, 865–890.
2. Xiang,K., Tong,L. and Manley,J.L. (2014) Delineating the structural blueprint of the pre-mRNA 3′-end processing machinery. *Mol. Cell. Biol.*, **34**, 1894–1910.
3. Tian,B. and Manley,J.L. (2013) Alternative cleavage and polyadenylation: the long and short of it. *Trends Biochem. Sci.*, **38**, 312–320.
4. Brown,K.M. and Gilmartin,G.M. (2003) A mechanism for the regulation of pre-mRNA 3′ processing by human cleavage factor Im. *Mol. Cell*, **12**, 1467–1476.
5. Ruegsegger,U., Beyer,K. and Keller,W. (1996) Purification and characterization of human cleavage factor Im involved in the 3′ end processing of messenger RNA precursors. *J. Biol. Chem.*, **271**, 6107–6113.
6. Proudfoot,N.J. and Brownlee,G.G. (1976) 3′ non-coding region sequences in eukaryotic messenger RNA. *Nature*, **263**, 211–214.
7. Tian,B. and Graber,J.H. (2012) Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip. Rev. RNA*, **3**, 385–396.
8. Beaudoing,E., Freier,S., Wyatt,J.R., Claverie,J.M. and Gautheret,D. (2000) Patterns of variant polyadenylation signal usage in human genes. *Genome Res.*, **10**, 1001–1010.
9. 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.
10. Chan,S.L., Huppertz,I., Yao,C., Weng,L., Moresco,J.J., Yates,J.R., Ule,J., Manley,J.L. and Shi,Y. (2014) CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3′ processing. *Genes Dev.*, **28**, 2370–2380.
11. Schonemann,L., Kuhn,U., Martin,G., Schafer,P., Gruber,A.R., Keller,W., Zavolan,M. and Wahle,E. (2014) Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. *Genes Dev.*, **28**, 2381–2393.
12. Barabino,S.M., Hubner,W., Jenny,A., Minvielle-Sebastia,L. and Keller,W. (1997) The 30-kD subunit of mammalian cleavage and polyadenylation specificity factor and its yeast homolog are RNA-binding zinc finger proteins. *Genes Dev.*, **11**, 1703–1716.
13. Hu,J., Lutz,C.S., Wilusz,J. and Tian,B. (2005) Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA*, **11**, 1485–1493.
14. Danckwardt,S., Kaufmann,I., Gentzel,M., Foerstner,K.U., Gantzert,A.S., Gehring,N.H., Neu-Yilik,G., Bork,P., Keller,W., Wilm,M. *et al.* (2007) Splicing factors stimulate polyadenylation via USEs at non-canonical 3′ end formation signals. *EMBO J.*, **26**, 2658–2669.
15. McDevitt,M.A., Hart,R.P., Wong,W.W. and Nevins,J.R. (1986) Sequences capable of restoring poly(A) site function define two distinct downstream elements. *EMBO J.*, **5**, 2907–2913.
16. Takagaki,Y. and Manley,J.L. (2000) Complex protein interactions within the human polyadenylation machinery identify a novel component. *Mol. Cell. Biol.*, **20**, 1515–1525.
17. Bacchetta,R., Barzaghi,F. and Roncarolo,M.G. (2018) From IPEX syndrome to FOXP3 mutation: a lesson on immune dysregulation. *Ann. N. Y. Acad. Sci.*, **1417**, 5–22.
18. Stacey,S.N., Sulem,P., Jonasdottir,A., Masson,G., Gudmundsson,J., Gudbjartsson,D.F., Magnusson,O.T., Gudjonsson,S.A., Sigurgeirsson,B., Thorisdottir,K. *et al.* (2011) A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nat. Genet.*, **43**, 1098–1103.
19. Danckwardt,S., Gantzert,A.S., Macher-Goeppinger,S., Probst,H.C., Gentzel,M., Wilm,M., Grone,H.J., Schirmacher,P., Hentze,M.W. and Kulozik,A.E. (2011) p38 MAPK controls prothrombin expression by regulated RNA 3′ end processing. *Mol. Cell*, **41**, 298–310.
20. Oh,J.M., Di,C., Venters,C.C., Guo,J., Arai,C., So,B.R., Pinto,A.M., Zhang,Z., Wan,L., Younis,I. *et al.* (2017) U1 snRNP telescripting regulates a size-function-stratified human genome. *Nat. Struct. Mol. Biol.*, **24**, 993–999.
21. Lou,H., Neugebauer,K.M., Gagel,R.F. and Berget,S.M. (1998) Regulation of alternative polyadenylation by U1 snRNPs and SRp20. *Mol. Cell. Biol.*, **18**, 4977–4985.
22. Almada,A.E., Wu,X., Kriz,A.J., Burge,C.B. and Sharp,P.A. (2013) Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature*, **499**, 360–363.
23. Kaida,D., Berg,M.G., Younis,I., Kasim,M., Singh,L.N., Wan,L. and Dreyfuss,G. (2010) U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*, **468**, 664–668.

24. Derti,A., Garrett-Engele,P., Macisaac,K.D., Stevens,R.C., Sriram,S., Chen,R., Rohl,C.A., Johnson,J.M. and Babak,T. (2012) A quantitative atlas of polyadenylation in five mammals. *Genome Res.*, **22**, 1173–1183.
25. Hoque,M., Ji,Z., Zheng,D., Luo,W., Li,W., You,B., Park,J.Y., Yehia,G. and Tian,B. (2013) Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nat. Methods*, **10**, 133–139.
26. Beaudoing,E. and Gautheret,D. (2001) Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res.*, **11**, 1520–1526.
27. Lianoglou,S., Garg,V., Yang,J.L., Leslie,C.S. and Mayr,C. (2013) Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.*, **27**, 2380–2396.
28. Zhang,H., Lee,J.Y. and Tian,B. (2005) Biased alternative polyadenylation in human tissues. *Genome Biol.*, **6**, R100.
29. Licatalosi,D.D., Mele,A., Fak,J.J., Ule,J., Kayikci,M., Chi,S.W., Clark,T.A., Schweitzer,A.C., Blume,J.E., Wang,X. *et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464–469.
30. Zhu,H., Zhou,H.L., Hasman,R.A. and Lou,H. (2007) Hu proteins regulate polyadenylation by blocking sites containing U-rich sequences. *J. Biol. Chem.*, **282**, 2203–2210.
31. Legendre,M. and Gautheret,D. (2003) Sequence determinants in human polyadenylation site selection. *BMC Genomics*, **4**, 7.
32. Tabaska,J.E. and Zhang,M.Q. (1999) Detection of polyadenylation signals in human DNA sequences. *Gene*, **231**, 77–86.
33. Xie,B., Jankovic,B.R., Bajic,V.B., Song,L. and Gao,X. (2013) Poly(A) motif prediction using spectral latent features from human DNA sequences. *Bioinformatics*, **29**, i316–325.
34. Hafez,D., Ni,T., Mukherjee,S., Zhu,J. and Ohler,U. (2013) Genome-wide identification and predictive modeling of tissue-specific alternative polyadenylation. *Bioinformatics*, **29**, i108–116.
35. Leung,M.K.K., Delong,A. and Frey,B.J. (2018) Inference of the human polyadenylation code. *Bioinformatics*, **34**, 2889–2898.
36. Perez Canadillas,J.M. and Varani,G. (2003) Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. *EMBO J.*, **22**, 2821–2830.
37. Wang,R., Nambiar,R., Zheng,D. and Tian,B. (2018) PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic. Acids. Res.*, **46**, D315–D319.
38. Ray,D., Kazan,H., Cook,K.B., Weirauch,M.T., Najafabadi,H.S., 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.
39. Dominguez,D., Freese,P., Alexis,M.S., Su,A., Hochman,M., Palden,T., Bazile,C., Lambert,N.J., Van Nostrand,E.L., Pratt,G.A. *et al.* (2018) Sequence, structure, and context preferences of human RNA binding proteins. *Mol. Cell*, **70**, 854–867.
40. Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.
41. Rehmsmeier,M., Steffen,P., Hochsmann,M. and Giegerich,R. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.
42. Core,L.J., Martins,A.L., Danko,C.G., Waters,C.T., Siepel,A. and Lis,J.T. (2014) Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.*, **46**, 1311–1320.
43. Ray,D., Kazan,H., Chan,E.T., Castillo,L.P., Chaudhry,S., Talukder,S., Blencowe,B.J., Morris,Q. and Hughes,T.R. (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.*, **27**, 667–670.
44. Pomeranz Krummel,D.A., Oubridge,C., Leung,A.K., Li,J. and Nagai,K. (2009) Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution. *Nature*, **458**, 475–480.
45. 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.
46. Fusby,B., Kim,S., Erickson,B., Kim,H., Peterson,M.L. and Bentley,D.L. (2016) Coordination of RNA polymerase II pausing and 3′ end processing factor recruitment with alternative polyadenylation. *Mol. Cell. Biol.*, **36**, 295–303.
47. Davidson,L., Muniz,L. and West,S. (2014) 3′ end formation of pre-mRNA and phosphorylation of Ser2 on the RNA polymerase II CTD are reciprocally coupled in human cells. *Genes Dev.*, **28**, 342–356.
48. Schlackow,M., Nojima,T., Gomes,T., Dhir,A., Carmo-Fonseca,M. and Proudfoot,N.J. (2017) Distinctive patterns of transcription and RNA processing for human lincRNAs. *Mol. Cell*, **65**, 25–38.
49. Viphakone,N., Sudbery,I., Griffith,L., Heath,C.G., Sims,D. and Wilson,S.A. (2019) Co-transcriptional loading of RNA export factors shapes the human transcriptome. *Mol. Cell*, **75**, 310–323.
50. Marini,F., Scherzinger,D. and Danckwardt,S. (2021) TREND-DB-a transcriptome-wide atlas of the dynamic landscape of alternative polyadenylation. *Nucleic. Acids. Res.*, **49**, D243–D253.
51. Nesic,D. and Maquat,L.E. (1994) Upstream introns influence the efficiency of final intron removal and RNA 3′-end formation. *Genes Dev.*, **8**, 363–375.
52. Rigo,F. and Martinson,H.G. (2008) Functional coupling of last-intron splicing and 3′-end processing to transcription in vitro: the poly(A) signal couples to splicing before committing to cleavage. *Mol. Cell. Biol.*, **28**, 849–862.