Bioinformatics, 34(17), 2018, 2889–2898
doi: 10.1093/bioinformatics/bty211
Advance Access Publication Date: 10 April 2018
Original Paper

Genome analysis

Inference of the human polyadenylation code

Michael K. K. Leung[1,2], Andrew Delong[1,2] and Brendan J. Frey[1,2,3,*]

[1]Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3S4, Canada, [2]Deep Genomics, MaRS Centre, Toronto, ON M5G 1L7, Canada and [3]Banting and Best Department of Medical Research, University of Toronto, Toronto, ON M5S 3E1, Canada

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on June 27, 2017; revised on October 15, 2017; editorial decision on March 31, 2018; accepted on April 9, 2018

Abstract

Motivation: Processing of transcripts at the 3′-end involves cleavage at a polyadenylation site followed by the addition of a poly(A)-tail. By selecting which site is cleaved, the process of alternative polyadenylation enables genes to produce transcript isoforms with different 3′-ends. To facilitate the identification and treatment of disease-causing mutations that affect polyadenylation and to understand the sequence determinants underlying this regulatory process, a computational model that can accurately predict polyadenylation patterns from genomic features is desirable.

Results: Previous works have focused on identifying candidate polyadenylation sites and classifying tissue-specific sites. By training on how multiple sites in genes are competitively selected for polyadenylation from 3′-end sequencing data, we developed a deep learning model that can predict the tissue-specific strength of a polyadenylation site in the 3′ untranslated region of the human genome given only its genomic sequence. We demonstrate the model's broad utility on multiple tasks, without any application-specific training. The model can be used to predict which polyadenylation site is more likely to be selected in genes with multiple sites. It can be used to scan the 3′ untranslated region to find candidate polyadenylation sites. It can be used to classify the pathogenicity of variants near annotated polyadenylation sites in ClinVar. It can also be used to anticipate the effect of antisense oligonucleotide experiments to redirect polyadenylation. We provide analysis on how different features affect the model's predictive performance and a method to identify sensitive regions of the genome at the single-based resolution that can affect polyadenylation regulation.

Contact: frey@psi.toronto.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Polyadenylation is a pervasive mechanism responsible for regulating mRNA function, stability, localization and translation efficiency. As much as 70% of human genes are subject to alternative polyadenylation (APA) and wide-spread mechanisms have been found which influence its regulation (Elkon *et al.*, 2013). By selecting which polyadenylation site (PAS) is cleaved, different transcript isoforms that vary either in their coding sequences or in their 3′ untranslated region (3′-UTR) can be produced. Transcripts differentially cleaved can influence how they are regulated. For example, longer variants can harbor additional destabilization elements that alter a transcript's stability (Shaw and Kamen, 1986), and shortened variants can escape regulation from microRNAs, which have been observed in various cancers (Lin *et al.*, 2012; Di Giammartino *et al.*, 2011). Furthermore, APA can be tissue-dependent, so a single gene can generate different transcripts, for instance, based on the tissue in which it is expressed (Tian and Manley, 2017). One mechanism of APA regulation occurs at the level of the sequences of the transcript. The presence or absence of certain regulatory elements can influence which PAS is selected. PAS selection is also influenced by a site's position relative to other sites. A computational model that can accurately predict how polyadenylation is affected by genomic features

© The Author(s) 2018. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

as well as cellular context is highly desirable to understand this widespread phenomenon. Moreover, several inherited diseases have been linked to errors in 3′-end processing (Danckwardt *et al.*, 2008). Such model would enable the exploration of the effects of genetic variations on polyadenylation and their implications for disease.

Here, we present the polyadenylation code, a computational model that can predict alternative polyadenylation patterns from transcript sequences. While there have been previous works in classifying whether a stretch of sequence contains a PAS (Akhtar *et al.*, 2010; Chang *et al.*, 2011; Cheng *et al.*, 2006; Ho *et al.*, 2013; Kalkatawi *et al.*, 2012; Xie *et al.*, 2013), or characterizing whether a PAS is tissue-specific (Hafez *et al.*, 2013; Weng *et al.*, 2016), many of them are aimed at improving gene annotations and understanding which features are involved in APA regulation, and do not address the question of predicting how APA sites are variably selected. Here, we tackle this question by developing a model that can predict a score, which we refer to as PAS strength (Shi, 2012), that describes the efficiency in which a PAS is recognized by 3′-end processing machinery for cleavage and polyadenylation. The ability to predict PAS strength enables this model to generalize to multiple prediction tasks, even though it is not explicitly trained for them. For example, the model can be applied to a gene with multiple PAS to determine the relative transcript isoforms that would be produced, in a tissue-specific manner. The model can predict the consequence of nucleotide substitutions on PAS strength, which can be used to prioritize genetic variants that affect polyadenylation. It can be used to assess the effects of anti-sense oligonucleotides to alter transcript abundance. It can also scan the 3′-UTR of the human genome to find potential PAS. We demonstrate examples of these applications in this work, and provide analysis on how different features affect the predictions of the model.

2 Materials and methods

2.1 Inferring the strength of a polyadenylation site

The goal of this work is to infer a score that describes the strength of a PAS, or the efficiency in which it is recognized by the 3′-end processing machinery. The problem would be straightforward if this target variable is directly measurable. However, current sequencing protocols only provide a measurement of the relative transcript abundance from APA. Various approaches exist in the literature which attempt to quantify the strength of a PAS. For example, normalized read counts are often used, but quantification can be affected by factors such as sequencing biases, transcript length and RNA decay (Gallego Romero *et al.*, 2014; Oshlack and Wakefield, 2009). Some studies classify PAS strength based on whether a canonical polyadenylation signal or other known sequence elements are present near the PAS (Akhtar *et al.*, 2010). We believe a more principled approach to predict a quantitative description of the strength of a PAS is to model it as a hidden variable, and infer it from data. Moreover, the position of a PAS relative to neighboring sites affects its selection. Some biological processes and tissues tend to favor PAS at the distal end, whereas cells under disease states tend to utilize PAS that are more proximal (Elkon *et al.*, 2013). Therefore, the model should include a variable that accounts for the distance between neighboring sites during training. Even though the position of a PAS is modeled, a desirable characteristic of the predictor is that during inference, positional information should be optional. This can be useful in regions of the genome where there are insufficient annotation sources to ascertain the distance to a nearby PAS. This would also enable one to apply this model to any DNA sequence associated with a site, optionally modify the bases within, and see the predicted effect on polyadenylation regulation. To determine which PAS in a gene with multiple sites is more likely to be selected, the model can be applied to each PAS separately to compare their relative strengths. Optionally, their positions can be factored in to the model's prediction if annotation sources are available in order to get a better estimate.

2.2 The polyadenylation code

The polyadenylation code is a model that can infer tissue-specific PAS strength scores from sequence, and optionally account for the influence of position if it is provided. It takes as input a sequence of length 200 bases centered on a PAS. We benchmark two models which operate on the sequence differently.

The first model is built on hand-crafted features. The genomic sequence is processed by a feature extraction pipeline, which divides the sequence into four regions relative to the PAS (Supplementary Material Section S1) (Hu, 2005). Some feature are limited to specific regions, namely the polyadenylation signals in the 5′–5′ and 5′–3′ regions, and hexamers defined in Hu (2005). Other features are computed in all regions, including counts of RNA-binding protein (RBP) motifs that may be involved in polyadenylation, all possible 1–4 n-mers counts and nucleosome positioning features from van der Heijden *et al.* (2012). The feature vector is mapped to a fully connected neural network. We will refer to this model as the Feature-Net.

The second model directly learns from the genomic sequence, using a convolutional neural network (Conv-Net) architecture (LeCun *et al.*, 1998), which can efficiently discover sequence patterns without prior knowledge even when the location of the patterns is unknown. The Conv-Net comprises of tunable motif filters which are free to adapt to the input sequence to optimize the predictive performance of the model. It also contains pooling operations that enable the model to focus on select locations in the input sequence whose composition maximally activates the motif filters. The use of convolutional neural networks to learn from raw genomic sequences have been successfully applied in other areas of biology (Alipanahi *et al.*, 2015; Angermueller *et al.*, 2017; Kelley *et al.*, 2016; Zhou and Troyanskaya, 2015).

To account for the positional preference of PAS, the log distance between sites is also an input feature for both models. Given two sites, the proximal (5′) site has a position feature of 0, whereas the distal (3′) site has a position feature that is equal to the logarithm of the distance between the distal and proximal site.

Figure 1 shows a schematic of both models. After the sequences are transformed by the Feature-Net and Conv-Net into a hidden representation, it is processed by separate fully connected hidden layers to make tissue-specific predictions. The architecture therefore factors predictions into two components: a score that describes the tissue-specific PAS strength, followed by predictions that represent the relative abundance of transcripts from RNA-Seq experiments between two competing PAS. The parameters of the fully connected layers model the cell state of tissues, which describes the steady-state environment of the cell, such as the protein concentrations in the cytosol, that can affect transcriptional modifications. We do not explicitly define what these cell state parameters consist of or how they factor in the predictions, but rather simply model them as hidden variables and learn them from data. A similar approach has been described in the splicing regulatory model by Xiong *et al.* (2014).

Seven distinct tissue types are available in the dataset used to train the models. Since there are two sets of sequencing reads for the naïve B-cells obtained from different donors (Lianoglou *et al.*, 2013), we treat them as separate tissues, and so our models have

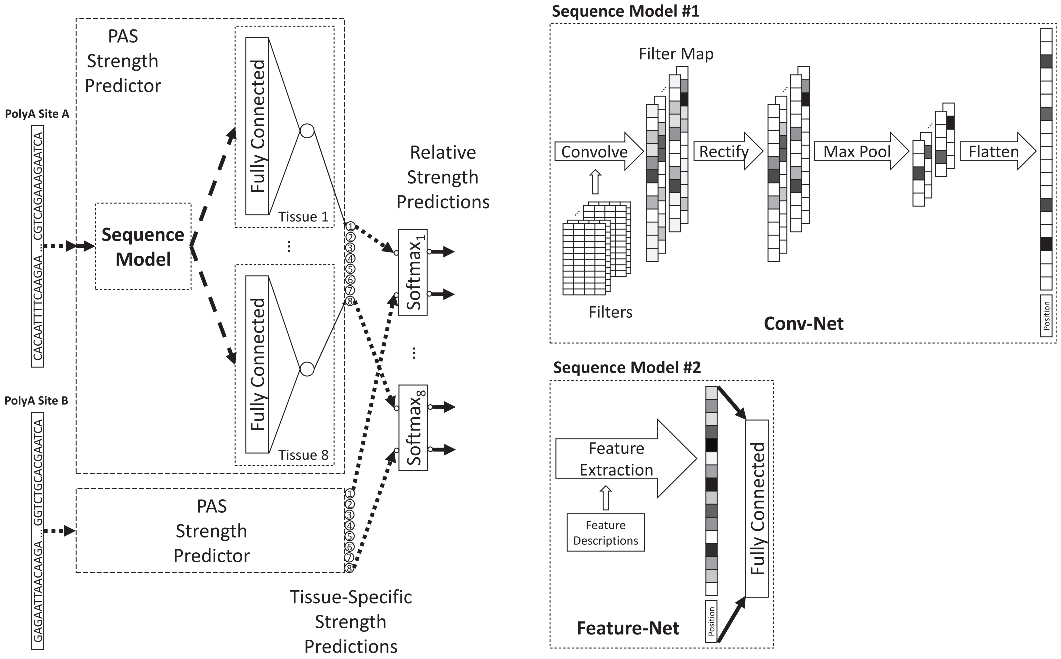

Fig. 1. (Left) A schematic of the components of the neural network that represent the polyadenylation model. The genomic sequence surrounding a polyadenylation site is an input to the strength predictor, which outputs eight tissue-specific scores describing the efficiency of the site for cleavage and polyadenylation. The model is trained from the relative strength between pairs of competing sites. (Right) Two architectures are compared for the sequence model, a convolutional neural network that operates directly on sequences and a fully connected neural network that takes in a feature vector processed by a feature extraction pipeline

eight polyadenylation strength prediction outputs. We choose not to rely on evolutionary conservation to force the models to learn patterns from the genome itself (Leung *et al.*, 2016). We also do not want to make use of additional data sources such as conservation tracks or expression data as input. For our model to be widely applicable to multiple tasks, it is beneficial for the input to be easily obtainable, such as sequences. Requiring anything beyond sequences makes a model more difficult to apply across diverse problem domains.

A training example consists of two PAS from the same gene, and requires the model to predict their relative strengths, which can be interpreted as the probability that each site would be selected for cleavage and polyadenylation. The relative strength is measured by the read counts from RNA-Seq that have been mapped to each site. As shown in Figure 1, a softmax function is used to squash the real-valued predictions from the PAS strength predictor into a normalized score that can be interpreted as the probability that one PAS is chosen over the other. The predictions are penalized against training targets of the relative abundances of transcripts for these PAS, which is measured from the sequencing experiment. Most of the results presented in this work are based on the predictions from the PAS strength predictor (i.e. the logits) instead of the relative strength predictions that follows the softmax.

In this work, we apply the predictive model to multiple tasks, even though it is trained only to the task of modeling competing site selection. All the predictions for these other tasks are evaluated without any additional task-specific training or data augmentation to demonstrate the general applicability of this model.

2.3 Assembling the polyadenylation atlas

Analysis of human polyadenylation events is confined to the 3′-UTR, where PAS are most frequently located. To identify the 3′-UTR regions of the human genome, 3′-UTR annotations from UCSC (Kent *et al.*, 2002), GENCODE (Harrow *et al.*, 2012), RefSeq (Pruitt, 2004) and Ensembl (Yates *et al.*, 2016) are combined, where overlapping regions are merged, and each 3′-UTR segment is further extended by 500 bases to capture potential uncharacterized regions. Then, to generate a comprehensive atlas of PAS, multiple polyadenylation annotations and reads from different 3′-end sequencing experiments are mapped to the 3′-UTR to generate an atlas of human PAS. The polyadenylation annotations used include PolyA_DB 2 (Lee *et al.*, 2007), GENCODE (Harrow *et al.*, 2012) and APADB (Müller *et al.*, 2014). Mapped reads that lie in the 3′-UTR from PolyA-Seq (Derti *et al.*, 2012) and 3′-Seq (Lianoglou *et al.*, 2013) are also used to expand the repertoire of PAS, where the genomic positions of reads from these sequencing experiments are used to mark the locations of PAS in the genome. PAS from different sources largely overlap, but some sites can be unique to one study due to the differences in cell lines or tissue types as well as sequencing protocol. Due to the inexact nature of 3′-end processing (Proudfoot, 2011), PAS that are within 50 bases of each other are clustered, and the resulting peak marked as the location of the PAS. The final PAS atlas contains 19 320 3′-UTR regions with two or more PAS from genes in the hg19 assembly for a total of 92 218 sites.

2.4 Quantifying relative polyadenylation site usage

The model is trained from the relative abundance of transcripts from a 3′-end sequencing experiment of seven distinct human tissues, including the brain, breast, embryonic stem (ES) cells, ovary, skeletal muscle, testis and two samples of naïve B cells (Lianoglou *et al.*, 2013). Other cell lines are also available in the dataset, but they are not used. The version of aligned reads which have been processed through the studies' computational pipeline is used, which include removal of internally primed and antisense reads, as well as application of minimum expression requirements to reduce sequencing noise. These reads are assigned to our PAS atlas, resulting in read counts associated with each PAS (Supplementary Data).

To quantify the relative PAS usage for each gene which acts as the target to train the model, we adopted the Beta model derived from Bayesian inference described in Xiong *et al.* (2016), treating the percent read counts of one site relative to another site as the parameter of a Bernoulli distribution. With this model, the relative PAS usage of one site relative to another, referred to as Φ, is $p(\Phi) = Beta(1 + N_{site1}, 1 + N_{site2})$, where N_{site1} and N_{site2} are the number of reads from two different sites. We use the mean of this distribution as the target to train the model, that is, the PAS usage of site 1 relative to site 2 is $(1 + N_{site1})/(2 + N_{site1} + N_{site2})$. For 3′-UTR regions with more than 2 PAS, different combinations of pairs of sites are generated as training targets and quantified as above. The assumption is that the relative strength of neighboring PAS can be described by the relative read counts at those sites, even if there are other sites present in the same gene. This assumption simplifies the architecture of the computational model and quantification of relative strength between sites.

2.5 Training the neural networks

The model is constructed and trained in Python using the TensorFlow library (Abadi *et al.*, 2015; Rampasek and Goldenberg, 2016). All hidden units of the neural network consists of rectified linear activation units (Glorot *et al.*, 2011). For the Feature-Net, the feature vectors are normalized with mean zero and standard deviation of one. For the Conv-Net, the input uses a one-hot encoding representation for each of the 4 nucleotides. For a sequence of length n, the dimension of the input would be $4 \times n$. Padding is inserted at both ends of the input so that the motif filters can be applied to each position of the sequence from beginning to end. For a motif filter of length m, the additional padding on each side of the sequence would be $4 \times (m - 1)$, where these additional padding would be filled with the value 0.25, equivalent to an N nucleotide in IUPAC notation. This is similar to what is done in Alipanahi *et al.* (2015).

Each training example consists of a pair of PAS from a gene, where the input is the two sites' genomic sequences, and the target is their relative read counts computed as described in Section 2.4. For genes with more than two PAS, different combinations of pairs of sites are generated as examples. Only examples with more than 10 reads are kept. This resulted in a dataset of 64 572 examples, which is split for training and testing.

The parameters of the neural network are initialized according to Glorot and Bengio (2010), and trained with stochastic gradient descent with momentum and dropout (Hinton *et al.*, 2012). Predictions from each softmax output are penalized by the cross-entropy function, and its sum across all tissue types is backpropagated to update the parameters of the neural network. Training and testing of the model are performed in a similar fashion as described in Leung *et al.* (2014). Briefly, data is split into approximately five equal folds at random for cross validation. Each fold contains a unique set of genes that are not found in any of the other folds. Three of the folds are used for training, one is used for validation, and one is held out for testing. By selecting which fold is held out for testing, five models are trained. The prediction of these five models on their corresponding test set is used for performance assessment, as well as to estimate variances, for all the tasks analyzed in this work.

The validation set is used for hyperparameters selection. The selected hyperparameters for our models can be found in (Supplementary Material Section S6). A graphics processing unit is used to accelerate training and hyperparameter selection by randomly sampling the hyperparameter space.

3 Results

3.1 Polyadenylation site selection

The performance of the model to predict the likelihood that a PAS is selected for cleavage and polyadenylation against a competing site in the same gene is shown in Table 1. These are the tissue-specific relative strength predictions for pairs of PAS that's shown in Figure 1. Performance is assessed using the area under the receiver-operator characteristic (ROC) curve (AUC) metric on held-out test data. To compare the models' performance against a baseline, we also trained a logistic regression (LR) classifier, which is essentially the Feature-Net with hidden layers removed. Predictions from the model based on the Conv-Net architecture are consistently the best performer. There is sizable performance gain from using the neural network models compared to the logistic regression classifier.

For the more general task of predicting which PAS would be selected in a gene with multiple sites, the model is applied to all PAS in the 3′-UTR of each gene. A score for each site is computed from the logits (the output of the PAS strength predictor shown in Fig. 1), where a larger value suggests that the site is more likely to be selected. The target is defined by the PAS in each gene which has the most measured reads in the 3′-Seq data. The metric we report here is the prediction accuracy, or the percentage of genes in which the model has correctly predicted the PAS that has the most reads. This is shown in Table 2 for genes with two to six sites, averaged across all tissues. The number of genes used in this evaluation is 2270, 2043, 1745, 1364 and 1163, respectively, where a gene is included only if at least one of its sites has more than 10 reads.

3.2 Pathogenicity prediction of polyadenylation variants

An advantage of our model is that the PAS strength predictor can be used to characterize individual sites based only on the input

Table 1. PAS selection performance between competing sites in different tissues

Tissue Type	AUC		
	LR	Feature-Net	Conv-Net
Brain	0.826 ± 0.010	0.869 ± 0.007	0.895 ± 0.005
Breast	0.825 ± 0.006	0.862 ± 0.003	0.886 ± 0.004
ES cells	0.849 ± 0.006	0.898 ± 0.002	0.911 ± 0.006
Ovary	0.830 ± 0.009	0.873 ± 0.006	0.895 ± 0.003
Skel. Muscle	0.828 ± 0.006	0.872 ± 0.005	0.893 ± 0.004
Testis	0.787 ± 0.007	0.828 ± 0.005	0.856 ± 0.007
B cells 1	0.838 ± 0.005	0.880 ± 0.005	0.896 ± 0.004
B cells 2	0.832 ± 0.004	0.880 ± 0.008	0.893 ± 0.007
All	0.824 ± 0.005	0.866 ± 0.004	0.889 ± 0.003

Table 2. PAS selection performance in genes with 2–6 sites

Number of sites	Accuracy (%)		
	LR	Feature-Net	Conv-Net
2	79.6	82.5	83.5
3	68.3	73.0	75.5
4	58.9	64.4	69.8
5	55.6	62.8	64.0
6	48.5	56.4	59.7

sequence. We evaluate whether this model can be used for pathogenicity predictions. The basic approach involves applying the model to the 200 nucleotides sequence associated with a PAS from the reference genome to first generate a prediction of its strength, and then performing another prediction when one or more nucleotides in the sequence is altered. A difference is then computed between the reference and variant predictions. Since there are eight predictions, one for each tissue, we take the largest difference as the score to assess pathogenicity. A similar approach has been applied to splicing variants (Xiong *et al.*, 2014). The postulate is that if a variant causes a large change to the strength of a PAS, this can change the relative abundance of differentially 3′-UTR terminated transcripts that deviates from normal, potentially indicating disease associations.

To evaluate the efficacy of this approach, we extracted variants that overlap with our PAS atlas (within 100 bases on either side of an annotated PAS) from the ClinVar database (Landrum *et al.*, 2014). Some of these variants overlap with the terminal exon (e.g. missense mutations) and are removed. There are 12 variants that are labeled as pathogenic (CLNSIG = 5) and 48 that are labeled as benign (CLNSIG = 2) (Supplementary Material Section S2). Figure 2 shows the ROC curve for this classification task. The model can predict pathogenic variants from benign ones with an AUC of 0.98 ± 0.02 and 0.97 ± 0.02, for the Conv-Net and Feature-Net respectively, both with a *P*-value of $<1 \times 10^{-8}$. Even though the AUC's are essentially identical for both models, there is clear advantage in the performance characteristic of the Conv-Net: it outperforms in the low false positive rate region where variant classification matters. For these predictions, we used an input of zero for the position feature of the strength model, since each variant is not analyzed with respect to neighboring sites. However, in general, it may be advantageous to incorporate this information. For example, a variant may cause a large change to a nearby PAS, but if there is a much stronger neighboring PAS in the same gene, the effects of the variant may be dwarfed by this neighbor, and therefore not have any significant mechanistic effects.

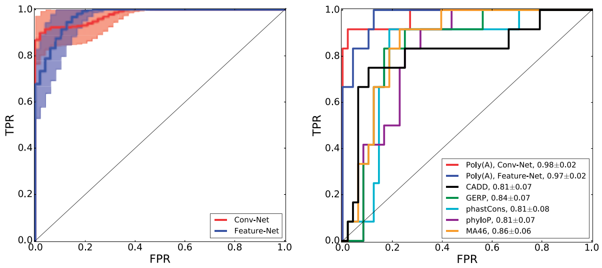

Fig. 2. Classification performance of ClinVar variants near polyadenylation sites. (Left) ROC curves comparing the variant classification performance of the Conv-Net and the Feature-Net. The shaded region shows the one standard deviation zone computed by bootstrapping. (Right) ROC curves comparing our model's performance against other predictors. AUC values are shown in the figure legend

We further evaluate how the model compares with four phylogenetic conservation scoring methods: Genomic Evolutionary Rate Profiling (GERP) (Cooper *et al.*, 2005), phastCons (Siepel *et al.*, 2005), phyloP (Pollard *et al.*, 2010) and the 46 species multiple alignment track from the UCSC genome browser (Blanchette, 2004). We also compare the predictions with Combined Annotation-Dependent Depletion (CADD), a tool which scores the deleteriousness of variants (Kircher *et al.*, 2014). Overall, as shown in Figure 2, the pathogenicity score from our model compares favorably, even though it has not been explicitly trained for this task. It is worth noting that although the model performed well for this ClinVar dataset, in general, a large difference in PAS strength does not necessarily imply pathogenicity, which is a phenotype that can be many steps downstream of 3′-end processing (Manning and Cooper, 2016).

The model can also be used to search for potential variants that would affect the regulation of polyadenylation. To visualize this approach, we applied the model and generated a mutation map (Alipanahi *et al.*, 2015) to a 100 nucleotide sequence in the human genome, where a ClinVar mutation that affect the polyadenylation signal is associated with *β*-thalassemia (Rund *et al.*, 1992). As shown in Figure 3, the polyadenylation signal is identified as an important region relative to other bases in the sequence.

3.3 Polyadenylation site discovery

The model is trained by centering the input sequence around a PAS at the cleavage site. If a PAS is off-center of the 200 nucleotides input sequence, or when no PAS is present, it stands to reason that the predicted PAS strength of the sequence would be small, due to the lack of sequence elements necessary for cleavage and polyadenylation. Alternatively, if the output of the PAS strength predictor is large, it would suggest that a PAS is present and is positioned near the center of the input sequence. Naturally, we ask whether the model can be translated across the genome to find potential PAS. While there have been previous works on this task (Akhtar *et al.*, 2010; Chang *et al.*, 2011; Cheng *et al.*, 2006), our model is not explicitly trained for this.

To illustrate an example of a predicted PAS track, we selected a section of the human genome and applied the Conv-Net strength model to it in a base-by-base manner (Supplementary Material Section S3). The average strength prediction from all eight tissues, without application of any filtering or thresholding, is shown. For this example, we chose a region of the genome with multiple PAS, and where there are differences between annotation sources.

The set of predicted peaks labeled region A are present in all annotation sources. It is not a single sharp peak, indicating that various PAS are possible in that region. This agrees with the GENCODE Poly(A) track, which indicates that there are two peaks in this region, as well as 3′-Seq, which shows that there are RNA-Seq reads that map across a broad region for various tissues. As mentioned earlier, the location for cleavage and polyadenylation is not exact. Region B is less well-defined, is weaker, and approximately aligns

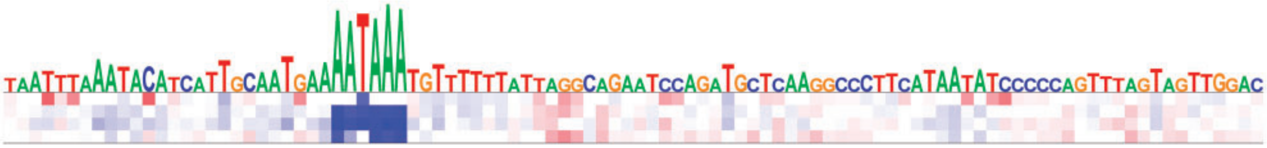

Fig. 3. A mutation map of the genomic region chr11: 5,246,678–5,246,777. Each square represents a change in the model's score if the original base is substituted. The substituted base is represented in each row in the order 'ACGT'. Red/blue denote a mutation that would increase/decrease the likelihood of the PAS for cleavage and polyadenylation

with the predicted positions from another PAS predictor (Cheng *et al*., 2006), as well as the muscle track from PolyA-Seq (in light gray). Finally, a small peak is observed in Region C, predicted to be a very weak PAS, which is present in PolyA-Seq. Note that the model is trained only from 3′-Seq reads and has no knowledge of RNA-Seq information from other datasets or other annotation sources.

To assess the model's ability in discovering PAS, we created a dataset with positive and negative examples to assess its classification performance. There is no general consensus from previous works on what constitutes a proper criteria to construct negative sequences or a standardized dataset for this task (Ji *et al*., 2015). We therefore defined the evaluation dataset based on our annotations and reads from 3′-Seq. Positive targets consist of annotated PAS in the 3′-UTR that has 10 or more reads. Since it is generally not appropriate to simply use random genomic sequences or locations for the negative set, we extracted the two immediately adjacent genomic regions near a PAS to ensure that both the negative and positive sequences have similar compositions (Supplementary Material Section S4). Each sequence is fed as input into the strength predictor, and the output from all tissues is averaged into a single value which is used for classification. The positional information of the sequence is not used (i.e. it has a position feature of zero). The AUC to classify sequences with PAS from negative sequences for the LR, Feature-Net and the Conv-Net are respectively 0.887 ± 0.003, 0.895 ± 0.004 and 0.907 ± 0.004. It is worth mentioning that of the negative sequences, 19% contain one of the two canonical polyadenylation signals (AAUAAA and AUUAAA), and 74% contain at least one of the known polyadenylation signals (Supplementary Material Section S1), meaning the model can distinguish real PAS from background. It does not simply look for the presence of polyadenylation signals to detect PAS in the genome.

It is interesting to observe that there is a relatively smaller difference in the AUC's for all models, especially between the Conv-Net and the logistic regression model, compared to previous tasks, which differed more drastically in performance. Identification of PAS from the genome is a simpler problem, characterized by the presence of features that are generally well-documented in the literature (Tian and Manley, 2017). For this, a logistic regression classifier may be sufficient. On the other hand, predicting the strength of a PAS given its sequence is arguably more complex. Instead of a binary classification problem, a strength predictor must quantify a PAS by integrating its genomic signature, and predict how it compares with another site, which may also contain all the core polyadenylation signatures, but differ in other ways with respect to its sequence. This observation is supported by the larger differences in the models' performance to the PAS selection problems in Tables 1 and 2, which require strength quantification.

3.4 Predicting the effect on oligonucleotide treatment

Anti-sense oligonucleotides therapies involve targeting RNAs via complementary base pairing, and can modulate RNA function by blocking the access of cellular machinery to the RNA (Kole *et al*., 2012). Application of this approach was demonstrated by Vickers *et al.* in the 3′-UTR, where oligonucleotides targeting polyadenylation signals and sites modulated the abundance of an mRNA (Vickers *et al*., 2001). Based on this, we show the utility of our model to provide an in-silico evaluation of oligonucleotides targeting regions near the PAS.

Three distinct forms of the transcript, Type 1, 2 and 3 are described in the study. A schematic of the E-selectin mRNA and the position of the polyadenylation signal, along with the targeted region of the oligonucleotides used is shown in Figure 4. All three forms harbor the canonical polyadenylation signal AAUAAA. A non-canonical polyadenylation signal AGUAAA is also present between the Type 1 and Type 2 cleavage site, which is selected when the corresponding signals from Type 1 and Type 2 are blocked. Here, it is referred to as the Type 4 form of the transcript.

According to the study, Type 3 is by far the dominant form of the transcript, followed by Type 1 and Type 2 (no differentiation is reported between them). Type 4 is the least common. Using the model, the predicted strengths for the corresponding PAS for Type 1 to 4 transcripts are respectively: -0.242, -0.420, 0.020, -0.765. These values do not account for the position of the PAS. If the relative positions of the 4 PAS are provided to the model, then the strengths become: -0.242, -0.170, 0.606, -0.584 (where Type 1 is assumed to be in position zero). These predictions match the observed abundances of the mRNA from the study.

The Vickers study performed a non-quantitative RT-PCR to assess the abundance of isoforms by administering different combinations of oligonucleotides targeting select regions of the transcript. To simulate this, we blocked the same regions of the input sequence complementary to the oligonucleotides by replacing the nucleotides with an N base, and predicted the resulting strengths of each PAS. The results are depicted in Figure 4, where the predicted values are arranged in an image to match the gel from the original paper. Each column is scaled such that the sum of the intensities of each column is constant, but otherwise, no additional processing is performed. The original paper does not provide values from RT-PCR that

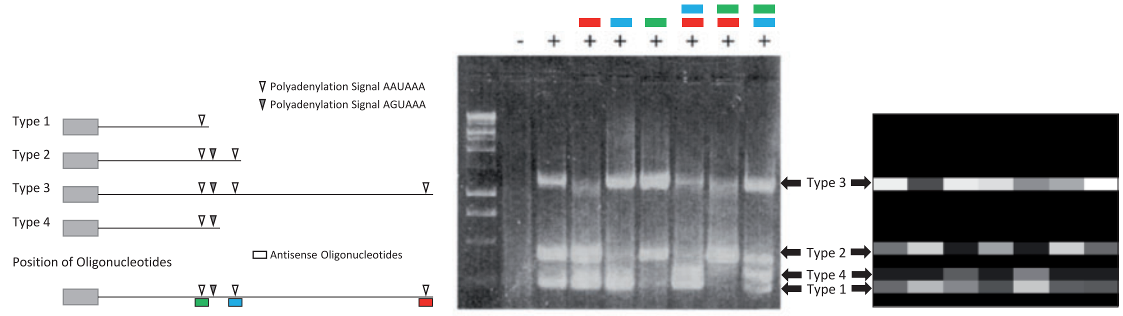
Polyadenylation Signal AAUAAA
Polyadenylation Signal AGUAAA
Type 1
Type 2
Type 3
Type 4
Position of Oligonucleotides
Antisense Oligonucleotides
- + + + + + + +
Type 3
Type 2
Type 4
Type 1

Fig. 4. Predicting the effect of an antisense oligonucleotide experiment. (Left) Schematic of human E-selectin 3′-UTR and the possible transcripts from polyadenylation site selection, reproduced from Vickers *et al.* (2001). The regions targeted by the oligonucleotides are shown. (Right) Predicted PAS strength, simulating the effects of blocked nucleotides due to oligonucleotide treatment. (Center) The figure from the original paper is reproduced here for ease of comparison. The oligonucleotides applied are shown on top of each column

would permit quantitatively comparison with the output of our model, but qualitatively, patterns of polyadenylation are generally captured. Note that the original paper mentions that Type 1 and 2 transcripts are shorter and therefore more efficiently amplified by PCR, and thus appear brighter than expected compared to Type 3. This experimental bias does not affect our simulated RT-PCR results in Figure 4.

3.5 Effect of genomic features on the model's predictions

To understand how different features contribute to performance, we train models using only individual feature groups. Table 3 shows each model's classification performance. Even though the polyadenylation signals are generally considered to be a main signature of PAS, they only partially account for the predictive performance for PAS selection compared to the full feature set. Overall, n-mers features are major contributors to the Feature-Net's performance, which is sufficiently rich to capture many motif patterns. It should be noted that each feature group has a different number of features (Supplementary Material Section S1), and therefore individual features in the larger feature groups may contribute only weakly, but as a whole affect predictions considerably. Position alone have very poor predictive capability, even though it was suggested to be a key feature in determining whether a PAS is used for tissue-specific regulation (Weng *et al.*, 2016). We also conducted an investigation on the uniqueness of each feature group, by training models with all features minus each feature group from Table 3. Removing the polyadenylation signals from the feature set reduces the performance from 0.866 ± 0.004 to 0.840 ± 0.008. All other groups, when removed, do not significantly reduce the performance of the model compared to the full feature set. This suggests that many features are redundant, and if removed, can be compensated by features in another group.

To see the contributions of individual features, we computed the gradient of the output with respect to the input feature vector of the neural network. This is referred to as the feature saliency of a prediction, and the gradients of features with large magnitudes can be interpreted as those that need to change the least to affect the prediction the most (Simonyan *et al.*, 2014). For this, we computed the feature saliency of all sites in our test set, and selected the features that on average have the largest magnitude. Table 4 shows the top 15 features computed using this method and the direction in which the feature affects the strength of a PAS, where an up arrow indicates that the effect is positive.

The top three features are consistent for all tissue types. Other features vary slightly between tissues and are grouped together unordered. As expected, the two most common canonical polyadenylation signals are the top features which increase the strength of a PAS. The log distance between PAS is also deemed to be important. Some features in this list are consistent with mechanisms of core elements known to be involved in cleavage and polyadenylation, including the upstream UGUA motif which the cleavage factor Im complex binds to, and a GU-rich sequence near the polyadenylation site (Tian and Graber, 2012). The genomic context upstream of the PAS appears to be more important, as most of the top features are in either the 5′–5′ and 5′–3′ region. Interestingly, three of the features reduce the strength of a PAS. They are the frequencies of C and AG nucleotides in the upstream region and the CA nucleotides downstream of the cleavage site, the latter of which is in line with the knowledge that the C-terminal domain of RNA polymerase II interacts with CA-rich RNA sequences, and is known to play a role in inhibiting polyadenylation (Kaneko and Manley, 2005).

Table 3. Comparison of Feature-Net PAS selection performance between competing sites using feature subsets

Feature group	AUC
All	0.866 ± 0.004
Poly(A) signal	0.728 ± 0.004
Position	0.553 ± 0.004
Cis-elements	0.608 ± 0.009
RBP motifs	0.676 ± 0.009
Nucleosome occupancy	0.656 ± 0.006
1-mers	0.762 ± 0.004
2-mers	0.794 ± 0.002
3-mers	0.817 ± 0.004
4-mers	0.833 ± 0.005

Table 4. Top 15 features of the Feature-Net, and the direction in which each feature can increase (↑) or decrease (↓) the strength of a polyadenylation site

Rank	Region	Feature name	Direction
1	5′–3′	PolyA Signal, AAUAAA	↑
2	—	Log distance between PAS	↑
3	5′–3′	PolyA Signal, AUUAAA	↑
4–15	5′–3′	1-mer, C	↓
	5′–3′	1-mer, U	↑
	5′–3′	2-mer, AG	↓
	3′–5′	2-mer, CA	↓
	3′–5′	3-mer, AAA	↑
	5′–3′	3-mer, UGU	↑
	5′–5′	3-mer, UGU	↑
	3′–5′	4-mer, AAAA	↑
	5′–5′	Cleavage factor Im, UGUA	↑
	5′–3′	PolyA signal, CAAUAA	↑
	5′–3′	PolyA signal, AUAAAG	↑
	5′–5′	PolyA signal, AGUAAA	↑

3.6 Determining tissue-specific polyadenylation features

Given that APA is used to achieve tissue-specific gene expression, we investigate whether our model can provide insights to this phenomenon. Previous computational approaches to address this problem are present in the literature. In Hafez *et al.*, an A-rich motif was found to be enriched in brain-specific PAS (Hafez *et al.*, 2013). In Weng *et al.*, the position of a PAS relative to another PAS and its position in the gene was found to be the strongest indicator of whether it is tissue-specific (Weng *et al.*, 2016). The computational models for both these works were trained to directly classify whether a PAS is tissue-specific. To be consistent with the methodology presented in this work, we will analyze our models without re-training them.

We use the set of tissue-specific and constitutive PAS defined in Weng *et al.* (2016) and apply the Feature-Net to generate predictions. To determine which feature is associated with tissue-specific PAS, we use the same gradient-based method as described in Section 3.5 to examine the top 200 most confident predictions for tissue-specific PAS, where our model predicts that at least one of the tissue outputs is considerably different than the rest, and for constitutive PAS, where our model predicts that all tissue outputs do not differ significantly. The magnitude of the gradients is then analyzed to see which features have a statistically greater effect on tissue-specific PAS compared to constitutive PAS. Statistical significance was determined by a permutation test by shuffling the predictions indicating whether a PAS it tissue-specific or constitutive. Applying a conservative P-value of 0.05/1506 (# of features) = 3×10^{-5}, 15 features were found to be associated with the model's ability to predict tissue-specific PAS. This is shown in Table 5. In the column indicating direction, an up arrow means the presence of the feature makes the site more likely to be tissue-specific, and vice versa.

All but one of the entries in the table describe features that are in the 5′–5′ and 3′–3′ region, that is, most of them are located away from the cleavage site (Supplementary Material Section S1). Various G/U-rich features top the list, where its position upstream suggests the PAS is more likely to be constitutive but if downstream, tissue-specific. Polyadenylation signals are absent from the list. No hexamers other than UUUGUA was found, which was previously identified as a feature by statistical analysis from Hu (2005). However, we found no association of this hexamer with tissue-specific polyadenylation from the literature. Given that the model only sees sequences from ±100 bases from the cleavage site, it may be possible that other more distal tissue-specific signatures may be present. Alternatively, sequence signatures may not be fully predictive since tissue-specific proteins can act by modulating core polyadenylation proteins instead of directly binding to the transcript (MacDonald and McMahon, 2010).

3.7 A convolution neural network model of polyadenylation to predict the effect of genomic variations

We initially began this work with a feature-based model, and subsequently added a Conv-Net for comparison expecting it to approach the performance of the Feature-Net, not necessarily surpassing it. Given that the polyadenylation features were derived from many publications and multiple research groups, the prior work that went into obtaining the feature-based models, which include the logistic regression classifier used as a baseline in this work, should not be underestimated. The fact that the Conv-Net could learn a better model absent any insights or hypotheses about mechanism is an interesting result on its own. This is surprising at first, but perhaps not so if viewed in the context of other applications of machine learning like computer vision, where hand-crafted features have been largely superseded by models which learn directly from image pixels (LeCun *et al.*, 2015).

Table 5. Features associated with prediction of tissue-specific polyadenylation sites, and whether the presence of the feature makes a polyadenylation site more (↑) or less (↓) tissue-specific

Region	Feature name	P-value	Direction
5′–5′	4-mer, UUGU	8.0×10^{-11}	↓
3′–3′	3-mer, UUG	9.9×10^{-09}	↑
3′–3′	4-mer, CCCC	5.7×10^{-08}	↓
5′–5′	3-mer, UGU	6.8×10^{-08}	↓
3′–3′	4-mer, UCCC	1.1×10^{-07}	↓
5′–3′	4-mer, CGGC	1.0×10^{-06}	↓
5′–5′	Cis-element, UUUGUA	1.7×10^{-06}	↓
5′–5′	Cleavage Factor Im, UGUA	2.2×10^{-06}	↓
5′–5′	3-mer, UUG	3.4×10^{-06}	↓
5′–5′	3-mer, AUC	7.4×10^{-06}	↑
3′–3′	3-mer, UCC	1.2×10^{-05}	↓
5′–5′	2-mer, UC	1.7×10^{-05}	↑
5′–5′	4-mer, AUCC	1.9×10^{-05}	↑
5′–5′	2-mer, UU	2.0×10^{-05}	↓
3′–3′	3-mer, CCU	2.1×10^{-05}	↓

On top of this, the Conv-Net has additional advantages that are not available in feature-based models. For instance, it is completely free to discover novel sequence elements that may be relevant for polyadenylation regulation from data. An example set of filters from the Conv-Net model is shown in (Supplementary Material Section S5). It also has the potential to be more computationally efficient. Feature extraction from sequences can be the most computational intensive aspect of a model during inference. This is not required for models that directly operate on sequences. There are additional operations that are required in the Conv-Net, but these computations can be significantly sped up by graphics processing units, which can be important for application of the model to entire genomes.

Since the Conv-Net operates directly on the genomic sequence, it also enables one to perform analysis at the single-base resolution more naturally. By analyzing the flow of gradients, the Conv-Net can determine how each base in the input sequence changes the output of the model. If a model requires feature extraction, such as the Feature-Net, the output must be analyzed relative to each feature. Furthermore, in the Feature-Net, many features are derived in discrete sections of the genome (four in this case, see Supplementary Material Section S1) to reduce the dimensionality of the input. The Conv-Net on the other hand, is more efficient at sharing model parameters, thereby enabling the motif filters to be applied at much finer spatial steps across a genomic sequence (a stride of 1 is used, see "Materials and methods" section), while still make overfitting manageable during training. By computing the gradients (Simonyan *et al.*, 2014), analysis regarding the magnitude and direction of the effect of each base on the model's output can be performed. This has the potential to offer a prescription to the design of oligonucleotides for anti-sense therapies. Figure 5 shows the saliency map of a region of the oligo-targeted mRNA examined in the Section 3.4, which spans the first three polyadenylation signals. This is different than the previous mutation map approach, which visualizes the change in the model's predictions between the reference genome and mutation at each base for the alternate nucleotides. Here, the gradient of each base relative to the model's prediction is shown, which includes the reference genomic sequence. It is also computed differently, involving a single backpropagation step in the Conv-Net. This operation is not readily available in the Feature-Net, where the genomic sequence is separated from the model by a feature extraction pipeline, and therefore dependent on the complexity and choices in the pipeline. This saliency map can be generated for large stretches of the genome to look for potential sensitive regions to alter polyadenylation for therapeutic purposes.

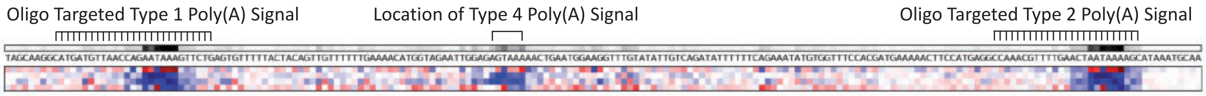

Fig. 5. Saliency map from the Conv-Net of a section of the oligo-targeted mRNA from Vickers *et al.* (2001). The base is represented in each row in the order 'ACGT'. Red means the base increases the likelihood of the sequence for cleavage and polyadenylation. Blue is the reverse. The sum of the magnitude of the gradient is shown above the saliency map to suggest how sensitive the nucleotide is to the strength of the polyadenylation site. The position of the oligonucleotide used in the study is shown at the top. The Type 4 Poly(A) signal is labeled also, but was not targeted in the original study

4 Conclusion

Regulation of polyadenylation is a crucial step in gene expression, and mutations in DNA elements that control polyadenylation can lead to diseases. Accurate, predictive models of polyadenylation will enable a deeper understanding of the sequence determinants of gene regulation and provide an important new approach to detecting and treating damaging genetic variations. We have presented here the polyadenylation code, a versatile model that can predict alternative polyadenylation patterns from transcript sequences and can generalize to multiple tasks that it was not trained on. Beyond its original trained usage to predict PAS selection from competing sites, it can classify variants near PAS and can be used for PAS discovery. We provided analysis of what sequences increase and decrease the strength of a PAS, and identified features that are associated with tissue-specific and constitutive PAS. We also illustrate the potential of our model to infer, and design for, the effects of antisense oligonucleotide treatment in the 3′-UTR.

Acknowledgements

The authors would like to thank Babak Alipanahi for helpful discussions and comments. They would also like to thank members of the Frey Lab especially Alice Gao, Leo Lee and Hui Xiong for discussions and technical assistance with this work.

Funding

This work was supported in parts by Natural Science and Engineering Research Council of Canada and Deep Genomics.

Conflict of Interest: none declared.

References

Abadi,M. *et al.* (2015) Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv: 1603.04467*.

Akhtar,M.N. *et al.* (2010) POLYAR, a new computer program for prediction of poly(A) sites in human sequences. *BMC Genomics*, **11**, 646.

Alipanahi,B. *et al.* (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.

Angermueller,C. *et al.* (2017) DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.*, **18**, 67.

Blanchette,M. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.

Chang,T.H. *et al.* (2011) Characterization and prediction of mRNA polyadenylation sites in human genes. *Med. Biol. Eng. Comput.*, **49**, 463–472.

Cheng,Y. *et al.* (2006) Prediction of mRNA polyadenylation sites by support vector machine. *Bioinformatics*, **22**, 2320–2325.

Cooper,G.M. *et al.* (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.

Danckwardt,S. *et al.* (2008) 3′ end mRNA processing: molecular mechanisms and implications for health and disease. *Embo J.*, **27**, 482–498.

Derti,A. *et al.* (2012) A quantitative atlas of polyadenylation in five mammals. *Genome Res.*, **22**, 1173–1183.

Di Giammartino,D.C. *et al.* (2011) Mechanisms and consequences of alternative polyadenylation. *Mol. Cell.*, **43**, 853–866.

Elkon,R. *et al.* (2013) Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.*, **14**, 496–506.

Gallego Romero,I. *et al.* (2014) RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biol.*, **12**, 42.

Glorot,X. *et al.* (2011) Deep sparse rectifier neural networks. *Proc. 14th Int. Conf. Artif. Intell. Stat.*, 315–320.

Glorot,X. and Bengio,Y. (2010) Understanding the difficulty of training deep feedforward neural networks. *Proc. 13th Int. Conf. Artif. Intell. Stat.*, **9**, 249–256.

Hafez,D. *et al.* (2013) Genome-wide identification and predictive modeling of tissue-specific alternative polyadenylation. *Bioinformatics*, **29**, i108–i116.

Harrow,J. *et al.* (2012) GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.*, **22**, 1760–1774.

Hinton,G.E. *et al.* (2012) Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580*.

Ho,E.S. *et al.* (2013) A multispecies polyadenylation site model. *BMC Bioinformatics*, **14**, S9.

Hu,J. (2005) Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA*, **11**, 1485–1493.

Ji,G. *et al.* (2015) Genome-wide identification and predictive modeling of polyadenylation sites in eukaryotes. *Brief. Bioinf.*, **16**, 304–313.

Kalkatawi,M. *et al.* (2012) Dragon PolyA Spotter: predictor of poly(A) motifs within human genomic DNA sequences. *Bioinformatics*, **28**, 127–129.

Kaneko,S. and Manley,J.L. (2005) The mammalian RNA polymerase II C-terminal domain interacts with RNA to suppress transcription-coupled 3′ end formation. *Mol. Cell.*, **20**, 91–103.

Kelley,D.R. *et al.* (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, **26**, 990–999.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Kircher,M. *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.

Kole,R. *et al.* (2012) RNA therapeutics: beyond RNA interference and antisense oligonucleotides. *Nat. Rev. Drug Discov.*, **11**, 125–140.

Landrum,M.J. *et al.* (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.

LeCun,Y. *et al.* (1998) Gradient-based learning applied to document recognition. *Proc. IEEE*, **86**, 2278–2323.

LeCun,Y. *et al.* (2015) Deep learning. *Nature*, **521**, 436–444.

Lee,J.Y. *et al.* (2007) PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res.*, **35**, D165–D168.

Leung,M.K.K. *et al.* (2014) Deep learning of the tissue-regulated splicing code. *Bioinformatics*, **30**, i121–i129.

Leung,M.K.K. *et al.* (2016) Machine learning in genomic medicine: a review of computational problems and data sets. *Proc. IEEE*, **104**, 176–197.

Lianoglou,S. *et al.* (2013) Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.*, **27**, 2380–2396.

Lin,Y. *et al.* (2012) An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Res.*, **40**, 8460–8471.

MacDonald,C.C. and McMahon,K.W. (2010) Tissue-specific mechanisms of alternative polyadenylation: testis, brain, and beyond. *Wiley Interdiscip. Rev. RNA*, **1**, 494–501.

Manning,K.S. and Cooper,T.A. (2016) The roles of RNA processing in translating genotype to phenotype. *Nat. Rev. Mol. Cell Biol.*, **18**, 102–114.

Müller,S. *et al.* (2014) APADB: a database for alternative polyadenylation and microRNA regulation events. *Database (Oxford)*, doi/10.1093/database/bau076/2634812.

Oshlack,A. and Wakefield,M.J. (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct.*, **4**, 14.

Pollard,K.S. *et al.* (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.

Proudfoot,N.J. (2011) Ending the message: poly(A) signals then and now. *Genes Dev.*, **25**, 1770–1782.

Pruitt,K.D. (2004) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.

Rampasek,L. and Goldenberg,A. (2016) TensorFlow: biology's gateway to deep learning? *Cell Syst.*, **2**, 12–14.

Rund,D. *et al.* (1992) Two mutations in the beta-globin polyadenylylation signal reveal extended transcripts and new RNA polyadenylylation sites. *Proc. Natl. Acad. Sci. USA*, **89**, 4324–4328.

Shaw,G. and Kamen,R. (1986) A conserved AU sequence from the 3′ untranslated region of GM-CSF mRNA mediates selective mRNA degradation. *Cell*, **46**, 659–667.

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

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

Simonyan,K. *et al.* (2014) Deep inside convolutional networks: visualising image classification models and saliency maps. In: Proc. of the Int. Conf. on Learn. Representations.

Tian,B. and Graber,J.H. (2012) Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip. Rev. RNA*, **3**, 385–396.

Tian,B. and Manley,J.L. (2017) Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.*, **18**, 18–30.

van der Heijden,T. *et al.* (2012) Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy. *Proc. Natl. Acad. Sci.* **109**, E2514–E2522.

Vickers,T. *et al.* (2001) Fully modified 2′ MOE oligonucleotides redirect polyadenylation. *Nucleic Acids Res.*, **29**, 1293–1299.

Weng,L. *et al.* (2016) Poly(A) code analyses reveal key determinants for tissue-specific mRNA alternative polyadenylation. *RNA*, **22**, 813–821.

Xie,B. *et al.* (2013) Poly(A) motif prediction using spectral latent features from human DNA sequences. *Bioinformatics*, **29**, i316–i325.

Xiong,H.Y. *et al.* (2014) The human splicing code reveals new insights into the genetic determinants of disease. *Science*, **347**, 1254806.

Xiong,H.Y. *et al.* (2016) Probabilistic estimation of short sequence expression using RNA-Seq data and the positional bootstrap. *bioRxiv: 046474*.

Yates,A. *et al.* (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.

Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods*, **12**, 931.