Sequence analysis

Accurate identification of polyadenylation sites from 3' end deep sequencing using a naïve Bayes classifier

Sarah Sheppard¹, Nathan D. Lawson¹*, and Lihua Julie Zhu¹,²*

¹Program in Gene Function and Expression, University of Massachusetts Medical School, 364 Plantation St, Worcester, MA 01605
²Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, 364 Plantation St, Worcester, MA 01605
* co-corresponding authors

Associate Editor: Prof. Alfonso Valencia
ABSTRACT

Motivation: 3’ end processing is important for transcription termination, mRNA stability, and regulation of gene expression. To identify 3’ ends, most techniques utilize an oligo-dT primer to construct deep sequencing libraries. However, this approach can lead to identification of artificial polyadenylation sites due to internal priming in homopolymeric stretches of adenines. While heuristic filters have been applied in these cases, they typically result in a high proportion of both false positive and negative classifications. Therefore, there is a need to develop improved algorithms to better identify mis-priming events in oligo-dT primed sequences.

Results: By analyzing sequence features flanking 3’ ends derived from oligo-dT based sequencing, we developed a naïve Bayes classifier to classify them as true or false/interally primed. The resulting algorithm is highly accurate, outperforms previous heuristic filters, and facilitates identification of novel polyadenylation sites.

1 INTRODUCTION

3’ end processing of pre-mRNAs influences transcription termination, mRNA stability and localization, and dynamic regulation of translation. In plants, yeast, and metazoa, sequence elements in the 3’ untranslated region (3′UTR) direct cleavage and polyadenylation (reviewed in Millevoi and Vagner, 2010; Proudfoot, 2011). Among these elements is the polyadenylation signal (PAS), a defined hexameric sequence located 10-50 nucleotides (nt) upstream of the cleavage and polyadenylation site (pA site), which binds Cleavage and Polyadenylation Specificity Factor complex (Millevoi and Vagner, 2010; Proudfoot, 2011). The PAS predominantly comprises the sequence AAUAAA (Proudfoot and Brownlee, 1976), although single nucleotide variants are also functional (Beaudoing, et al., 2000; Sheets, et al., 1990). In addition to the PAS, a guanine/uracil- or uracil-rich downstream sequence element can be found 20-40 nt downstream of the pA site that is recognized by Cleavage Stimulatory Factor (Millevoi and Vagner, 2010; Proudfoot, 2011). In some instances, a uracil-rich sequence element is present upstream of the PAS, which may also act to enhance usage of a specific PAS by recruiting Cleavage Factor I (Millevoi and Vagner, 2010; Proudfoot, 2011). In combination, these sequence elements help define the site of cleavage and polyadenylation at the 3’ end of a pre-mRNA.

Most efforts to identify 3’ ends of mRNAs have relied on priming with an oligonucleotide of deoxothymines (oligo-dT). These efforts include early studies relying on expressed sequence tags (ESTs; Beaudoing, et al., 2000; Tian, et al., 2005; Zhang, et al., 2005), as well as more recent work utilizing deep sequencing (reviewed in Mueller, et al., 2013). These include Poly(A) Site sequencing (PAS-Seq; Shepard, et al., 2011) and PolyA-Seq (Derti, et al., 2012), which rely on oligo-dT containing primers for first strand cDNA synthesis. While these approaches are technically straightforward, oligo-dT binding can occur in internal homopolymeric stretches of adenines (Nam, et al., 2002) leading to identification of false positive pA sites. A more selective method is poly(A)-position profiling by sequencing (referred to as 3pseq), where a splint RNA:DNA oligonucleotide with overhanging thymines is hybridized and ligated to the polyadenylated tail of mRNAs to prevent internal priming (Jan, et al., 2011). However, 3pseq is technically demanding and most laboratories are more likely to utilize oligo-dT-primed approaches. In these latter cases, internal priming events are generally filtered from datasets based on the number of adenines in the genomic sequence downstream of the cleavage site (Beaudoing, et al., 2000; Brockman, et al., 2005; Fu, et al., 2011; Haenni, et al., 2012; Liu, et al., 2007; Shen, et al., 2011; Shepard, et al., 2011; Smibert, et al., 2012; Tian, et al., 2005; Wilkening, et al., 2013; Wu, et al., 2011; Zhang, et al., 2005). However, the strict definition of these heuristic filters inevitably misses some internal priming events (false positives) and also excludes true 3’ ends (false negatives; Sherstnev, et al., 2012). Thus, additional methods are needed to easily analyze oligo-dT primed deep sequencing data to identify true pA sites.

A naïve Bayes classifier, based on Bayes theorem, is a supervised learning algorithm in which the features used to predict the class are considered conditionally independent (Mitchell, 2010). Naïve Bayes classifiers are computationally efficient, require relatively small training datasets, handle both continuous and discrete features, and ignore non-relevant features (Kotsiantis, et al., 2006; Mitchell, 2010). Here, we demonstrate the effectiveness of a naïve Bayes classifier to identify internal priming events in oligo-dT primed sequencing data. We find that our trained algorithm outperforms heuristic filters and enriches for 3’ ends in oligo-dT sequencing data that bear canonical motifs important for cleavage and polyadenylation. Biological validation shows that our method is highly accurate, facilitating identification of novel 3′UTRs and 3’ ends in multiple animal species.

2 METHODS

Zebrafish care and staging

Zebrafish were maintained as described in (Westerfield, 1993) and staged as described in (Kimmel, et al., 1995). Studies were performed under the approval of the University of Massachusetts Medical School Institutional Animal Care and Usage Committee.

RNA purification

Total RNA was purified from either 6 hpf or 24 hpf wild type CF zebrafish and treated with DNase I (Qiagen RNasey Midi Kit, Qiagen RNase-Free DNase Set). Polyadenylated RNA was selected using magnetic oligo-dT beads (Invitrogen mRNA Direct Kit).

RNA-seq libraries and data analysis

The 24 hpf zebrafish RNA-seq library was built using an Illumina mRNA-seq protocol (Part # 1004898 Rev. D) and paired-end sequenced on an Illumina Genome Analyzer II (76 nt reads) and an Illumina Hi-Seq (101 nt reads). Sanger 6 hpf RNA-seq data was downloaded from the European Bioinformatics Institute (run ERR022485). RNA-seq reads from both developmental stages starting with at least five thymines (the reverse complement of a polyadenylated mRNA) for potential internal oligo-dT priming and ended with at least five adenines were mapped to the zebrafish genome (Zv9) using Bowtie (Langmead, et al., 2009; Figure S1A). Those that mapped to the genome were taken as sites for potential internal oligo-dT priming and included in the True Negative training set (Figure S1A). The site of internal priming was assigned to the single nucleotide immediately upstream of the last mapped 3’ adenine in this set (referred to as RNA-seq internally primed sites). Sequence fragments that did not initially map were trimmed of terminal adenines (or thymines) and re-mapped (Figure S1A). Mapped reads (referred to as RNA-seq putative pA sites) were combined with the PAS-Seq data for establishment of the True Positive training sets (Figure S1A; see Training Sets).

3’ end deep sequencing datasets
We constructed PAS-Seq libraries as described in (Shepard, et al., 2011), using barcoded adapters, and paired-end sequenced on an Illumina Hi-Seq (101 nt reads) with a custom sequencing primer described in (Shepard, et al., 2011) designed to exclude the remainder of the poly(A) tail from sequencing. Libraries were de-convoluted using Perl scripts and mapped to the zebrafish genome (Zv9) using Tophat (Trapnell, et al., 2009). Zebrafish 6 hpf and 24 hpf 3pseq (Ulitsky, et al., 2011) and mammalian polyA-seq alignments (Derti, et al., 2012) were downloaded from the Gene Expression Omnibus (accession numbers GSE32880, GSE30198). cleanUpdTSeq (see below) was used to classify putative sites from unfiltered polyA-seq as true or false, using a probability assignment cutoff = 0.5. No additional filtering was performed on the 3pseq or the originally-filtered polyA-seq data sets.

**RNA-Seq Transcriptome Analysis**

To assess the utility of our classifier on an annotated transcriptome, we used it on a previously published RNA-seq models for zebrafish embryos described in Pauli, et al. (2012). Unique 3' transcript ends were classified using cleanUpdTSeq as described below.

**pA site builds**

A custom Perl script clustered mapped sequencing reads into putative pA sites. Mapped reads were trimmed 3' terminal nt, which corresponds to the site of cleavage. Reads were clustered first for identically matching sites. An iterative process was used to cluster adjacent sites within +/- 5 nt, starting with the site with the highest number of reads. Within a cluster, the putative pA site was defined as the location with the most reads and the total reads were combined to give the height. Mann-Whitney test was performed to assess height differences between datasets (Hollander and Wolfe, 1999). Concordance between data sets was defined as being within +/- 10 nt using a Perl script. The distance from the putative PAS to the pA site was determined as the distance from the 3' end of the PAS to the pA site.

**Training Sets**

RNA-seq putative pA sites were combined with the PAS-Seq putative pA sites and clustered as described above (Figure S1A). Sites concordant between the PAS-Seq and the 3pseq data sets were assigned to the True Positive training set (Figure S1A). 3pseq coordinates were used if there was not an exact match. RNA-seq internally primed sites not concordant with 3pseq were assigned to the True Negative training set (Figure S1A). Only sites that were present in both the 6 hpf and 24 hpf data sets were used for training (Figure S1B). We did not take the number of sequencing reads that composed a putative pA site into account.

**cleanUpdTSeq**

The function buildFeatureVector in the cleanUpdTSeq package was used to build feature vectors for training dataset and test dataset. Features include: presence/absence of 4096 hexamers in the upstream of the pA sites; downstream mononucleotide count; downstream dinucleotide count; average distance of downstream adenines to the pA site (Figure S2B). The upstream features are modeled as binomial variables and the downstream features are modeled as normal variables. A naïve Bayes classifier was built using the training data and the function buildClassifier, which leverages the R package e1071 (Meyer, et al., 2012) with laplace set to 1. To classify the test dataset, the predictClass function was applied. These functions along with sequence fetching utilities and training data are available on our website (lawsonlab.umassmed.edu/cleanupdtseq.html). The package cleanUpdTSeq has been submitted to Bioconductor.

**Performance Metrics**

Precison, recall, true negative rate (TNR), false discovery rate (FDR), false positive rate (FPR), accuracy, F-score, and Matthew’s correlation coefficient (MCC) were calculated using the following equations. TP = true positive, TN = true negative, FP = false positive, FN = false negative.

\[
\text{Precision} = \frac{TP}{FP + TP}
\]

\[
\text{Recall} = \frac{TP}{FN + TP}
\]

\[
\text{TNR} = \frac{TN}{TN + FP}
\]

\[
\text{FDR} = \frac{FP}{FN + FP}
\]

\[
\text{FPR} = \frac{FP}{TN + FP}
\]

\[
\text{Accuracy} = \frac{TP + TN}{TN + FN + FP + TP}
\]

\[
\text{F - Score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

\[
\text{MisclassificationError} = \frac{FP + FN}{TP + TN + FP + FN}
\]

Pearson’s correlation coefficient was used to assess nucleotide profile correlation (Pietrokovski, 1996; R Core Team, 2013). To apply heuristic filters (8A and 8A plus top 10 PAS), Perl scripts were used to classify all putative pA sites from the training set (True Positives and True Negatives).

**Model Selection, Training Set Size, and Relative Feature Importance**

For performance evaluation, the training datasets were randomly split (70% used for training and 30% used for cross-validation) in 10 trials, each with a range of probability cutoffs from 0-1 at an interval of 0.1 for each combination of upstream (20-50 nt in increments of 10 nt) and downstream (30-50 nt in increments of 10 nt) sequence. We calculated average precision, recall, F-score, accuracy, true negative rate, false discovery rate, false positive rate and MCC from 10 cross-validations of each model, using a probability of true cutoff of 0.5 (Supplementary Table 1). We chose 40 nt of upstream and 30 nt of downstream sequence for subsequent PAS identification.

To evaluate training set size, we trained the classifier with 15995 (50%), 19194 (60%), 22393 (70%), 25592 (80%), or 28791 (90%) peaks and used the remainder for cross validation. Average precision, recall, F-score, accuracy, true negative rate, false discovery rate, false positive rate and MCC from 10 cross-validation trials were calculated as above.

To evaluate the relative importance of each feature, p-values using prop.test in R and odds ratio were calculated for the binary features (Mladenic and Grobelnik, 1999). For the continuous features, p-values were calculated using t-test. Features were ordered by p-values, with lowest values indicating greatest importance. Upstream features (binary) and downstream features (continuous) are listed separately in Supplementary Tables 2A and 2B, respectively. In addition, the top 200 positive upstream features are listed in Supplementary Table 2C, which only include top upstream features with odds ratio less than 0. The presence of positive features is associated with increased probability of being a true polyadenylation site.

**PAS flanking sequence characterization**
Commonly used zebrafish PASs used in zebrafish were identified by applying Multiple Em for Motif Elicitation (MEME) (Bailey and Elkan, 1994) to 50 nt upstream of 3' ends annotated in Ensembl (v61). These were used to build a Perl script to search for a canonical or variant PASs. For mammals, a Perl script was used to search for a canonical or variant PAS in order of decreasing usage in polyA-seq data (Derti, et al., 2012). For the True Positive and True Negative training sets, 50 nt upstream for all of the sites was examined using MEME (Bailey and Elkan, 1994) with the following settings: -minw 5 -maxw 50 –oops. 50 nt downstream of all of the sites was examined using the options: -minw 5 -maxw 50 –oops. For the other data sets, 40 nt upstream of the pA site and 30 nt downstream of 10,000 randomly chosen sites within the data set were used for analysis (upstream: -minw 5 -maxw 10 –oops; downstream: -minw 5 -maxw 30 –oops).

**Poly(A) tail length assays**

Total RNA was purified from 24 hpf wild type CF zebrafish (Qiagen RNeasy Mini Kit). For the G-tail assay, we used the Affymetrix Poly(A) Tail-Length Assay Kit to add guanosines and inosines to the 3' end of the polyadenylated mRNAs (Martin and Keller, 1998). Subsequently, reverse transcription was performed with a poly-cytosine anchored primer. Alternatively, we used an oligo-dT(10) primer to make cDNA (Murray and Schoenberg, 2008). In both cases cDNAs were used as a template in a 20 cycle primary PCR with Hot Master Taq DNA polymerase (5Prime) to amplify the 3' end with poly(A) tail with a forward primer and assay-specific reverse primer (G-Tail: Affymetrix Poly(A) Tail-Length Assay Kit Universal Primer, oligo-dT: GGGGATCCGCGGTTTTTTTTTTT; Murray and Schoenberg, 2008). Nested PCR was performed for 20 to 35 cycles using 1 ul of a 1:50 dilution of the primary PCR as template, a nested forward primer and the assay-specific reverse primer. PCR products were run on a 2% agarose gel. Gene specific oligonucleotides were also used to help estimate the size of the 3'UTR without any poly(A) tail. The lower part of the smear or single band were excised from the gel, column purified (Qiagen MinElute Gel Extraction Kit), shotgun cloned (Promega pGEM-T Easy Vector System I), and sequence verified.

### 3 RESULTS AND DISCUSSION

To distinguish between true and false pA sites in oligo-dT primed deep sequencing data, we trained a naïve Bayes classifier using defined True Positive and True Negative sites. Given the demonstrated technical rigor of 3pseq, True Positives were defined as the intersection of 3' ends identified by both 3pseq and PAS-Seq datasets from the same stage of zebrafish embryos (Figure S1A). True Negatives were derived from oligo-dT primed RNA-Seq reads with at least five genomically templated terminal adenines or proximal thymines and were not present in 3pseq (Figure S1A). The training set consists of 22770 True Positives and 9219 True Negatives, representing the majority of zebrafish coding genes (data not shown). As expected of pA sites (Graber, et al., 1999), True Positives exhibited a prevalence of adenines and thymines upstream of the cleavage site (Figure 1A) and a canonical PAS clustered near the 3' end (Figure 1B, C). Downstream of the pA site we noted thymine and a slight enrichment of guanine (Figure 1A). By contrast, True Negative sites failed to exhibit characteristics associated with pA sites (Figure 1D, E) and the small fraction of PASs in these sequences did not cluster near the 3' end (Figure 1F, G).

#### 3.1 Algorithm Training and Performance

To delineate the True Positives from the True Negatives, we chose algorithm features to represent sequence elements known to direct cleavage and polyadenylation (Figure S2A). In the upstream sequence region, this included all hexamer permutations to allow for self-discovery of potential canonical and variant PASs and uracil-rich elements (Figure S2B). Downstream of a putative cleavage site, guanine/uracil- or uracil-rich elements may signify a true pA site (Proudfoot, 2011), while adenine richness may indicate internal oligo-dT priming. Therefore, we also included mono- and dinucleotide counts, as well as the average distance of the adenines to the pA site, as features (Figure S2B). The upstream features were modeled as a binomial distribution and the downstream features were modeled as a normal distribution. The relative importance of each feature is listed in Supplementary Table 2A. As expected, canonical PASs, AATAAA and ATTAAA, are among the top four most important binary features (Supplemental Table 2A), while downstream T/GT rich elements, adenine richness and proximity to the pA site are among the most important continuous features (Supplemental Table 2B). Variant PASs also aid in identification of true pA sites (Supplemental Table 2C: top 200 positive hexamers upstream). Variation of up- (20-50 nt) and downstream (30-50 nt) sequence lengths for these features (Figure S2C) demonstrated low variability between the different models (Supplementary Table 1). We chose 40 nt of upstream sequence, as not to miss any possible PASs in the upstream region due to variations in cleavage site usage, and 30 nt downstream for subsequent training (Pauws, et al., 2001).
To develop and test the naïve Bayes classifier, we randomly sampled 70% of the training set to build the classifier (training) and the remaining 30% to evaluate performance (cross-validation) and averaged the results of 10 trials. Following training, we found that the naïve Bayes classifier recalled 92.2% of True Negatives (true negative rate, Table 1, Figure 2A) and 93.8% of True Positives (recall, Table 1, Figure 2B), while it incorrectly categorized only 3.2% of predicted positives (false discovery rate, Table 1, Figure 2C). By contrast, a heuristic filter defined as 8 or more adenines in 10 nt downstream of the pA site (referred to hereafter as 8A) generally performed worse at identifying True Negatives and False positives, although recall with this filter was quite good (Figure 2, Table 1). While removing all sites without a putative PAS in combination with the 8A filter (PAS+8A) improved performance (Figure 2A-C), the naïve Bayes classifier generally outperformed both of these heuristic filters, as determined by Matthew’s correlation coefficient (MCC), a balanced measure of true positives, false positives, true negatives and false negatives (Matthew’s correlation coefficient, Table 1, Figure 2D; Matthews, 1975). While the size of the training set may lead to over-fitting due to biased sequence composition, algorithm performance was similar using 50%, 60%, 70%, 80%, or 90% of the True Positives and True Negatives for training (Figure S3), demonstrating that our initial training set was of sufficient size. Taken together, the naïve Bayes classifier outperforms the heuristic filters on these initial training and cross-validation sets. Furthermore, the increased specificity appears to come with little cost to sensitivity.

### 3.2 Application to PAS-Seq data

To more generally test the performance of the naïve Bayes classifier, we used all True Positives and True Negatives to build the classifier and used it to categorize unfiltered oligo-dT primed 3’ end deep sequencing (PAS-Seq) data from 24 hpf zebrafish embryos. Genomic sequence flanking 3’ ends from unfiltered PAS-Seq data shows enrichment for adenines upstream and downstream of the pA site (Figure 3A), similar to our True Negative training set (see Figure 1D, E; \( r_A = 0.89, r_T = 0.78, r_G = 0.79, r_C = 0.76 \)). Furthermore, we failed to identify a canonical PAS as an over-represented motif upstream of putative pA sites in unfiltered PAS-Seq data (Figure 3A). Indeed, only 20.6% of the putative pA sites contain an upstream AATAAA and 55.4% have no identifiable PAS (Figure 3B). Thus, this dataset likely contains a high proportion of sequences derived from internal oligo-dT priming. Filtering these data with the trained naïve Bayes classifier calls 65.4% of putative pA sites from PAS-Seq as false and the nucleotide profile of pA sites called True resembled that of pA sites identified at the same developmental stage by the more technically rigorous 3pseq approach (compare Figure 3C, D; \( r_A = 0.83, r_T = 0.68, r_G = 0.85, r_C = 0.86 \)). Similar upstream (canonical PAS) and downstream sequence (guanine/thymine-rich) elements were also easily identified in both 3pseq and filtered PAS-Seq datasets (Figure 3C, D). 49% of the true pA sites called by the classifier contain AATAAA and only 19.2% have no PAS, in agreement with genome-wide PAS distributions (Li, et al., 2012; Ulitsky, et al., 2012) and similar to 3pseq data (Figure 3B). In contrast to the naïve Bayes classifier, categorizing PAS-Seq using the 8A heuristic filter classified only 18.1% of putative pA sites as false, leading to only slightly better correlation of sequence composition of the remaining sites with those from 3pseq data (\( r_A = 0.45, r_T = 0.12, r_G = 0.65, r_C = 0.61 \)) compared with the unfiltered PAS-Seq. However, the adenine
richness in the downstream sequence region and lack of identifiable consensus PAS suggest a large number of internally oligo-dT primed sites are called as positives after applying the 8A filter (Figure 3B, E). The PAS+8A heuristic filter performs better than the 8A filter, likely by excluding more false positives ($r_A = 0.76$, $r_C = 0.57$, $r_G = 0.83$, $r_T = 0.82$; Figure 3F). However, based on its strict definition, the PAS+8A filter eliminates all true pA sites that do not contain a consensus PAS (Figure 3B). Together, these results demonstrate that our naïve Bayes classifier performs better than heuristic filters, resulting in a set of putative pA sites that closely resembles 3pseq. Importantly, our classifier is also able to identify 3’ ends that do not bear a consensus PAS. Further examination of this subset reveals 53% contain the top 50 positive hexamers upstream, and 74% contain the top 100 positive hexamers upstream (Supplemental Table 2C).

Comparison of the proportion of pA sites common to PAS-Seq and 3pseq data in 24 hpf zebrafish embryos revealed an increase from 13.0% to 35.7% after filtering PAS-Seq data with the naïve Bayes classifier (Figure 4A). Interestingly, both common and unique pA sites exhibit characteristics typical of true PAS (Figure 4B-D), although those unique to only 3pseq or PAS-Seq datasets show a higher proportion of variant PAS usage (Figure 4E). The occurrence of these sites in only 3pseq or PAS-Seq datasets may be due to low levels of expression that are not consistently detected at this sequencing depth. Accordingly, pA sites common to PAS-Seq and 3pseq comprise significantly ($p < 2.2e-16$) more sequencing reads than those unique to either PAS-Seq (mean of 288.69 vs. 8.85) or 3pseq (mean of 204.94 vs. 21.4; Figure 4F). Other technical issues may also contribute to a pA site being uniquely found in either dataset. For example, internal oligo-dT binding may block extension from an oligo-dT bound to the poly(A) tail (Nam, et al., 2002), thus inhibiting identification of true 3’ ends in adenine-rich genomic loci in PAS-Seq. Variable PAS usage due to polymorphisms between the zebrafish strains used to generate the PAS-Seq and 3pseq datasets may also be possible (Howe, et al., 2013).

To biologically cross-validate the in silico predictions, we conducted two different poly(A) tail length (PAT) assays to verify a putative pA site. In the G-tailed PAT assay (GPAT), yeast poly(A) polymerase is used to ligate guanosines and inosines to the 3’ end of polyadenylated RNA, followed by reverse transcription with an
anchored poly-cytosine primer (Figure 5A). Alternatively, an oligo-dT containing primer was used for reverse transcription (dtPAT; Figure 5B) (Murray and Schoenberg, 2008). In both assays, nested PCR was performed using a gene-specific forward primer and an assay-specific reverse primer to amplify the 3' end of the transcript including the poly(A) tail, as well as gene-specific forward and reverse primers to amplify fragments without the poly(A) tail (Figure 5A,B). Due to different poly(A) tail lengths or variable oligo-dT binding along the poly(A) tail, validation of a true 3' end resulted in a smear on an agarose gel in both assays (Figure 5A-C). Conversely, an internally primed site will result in no product in the GPAT assay and a single product in the dtPAT assay (Figure 5A, B, D). We applied GPAT and dtPAT assays to 50 putative pA sites in the zebrafish genome defined by our classifier (Supplementary Table 3). 42 of these sites were called True by the classifier, of which 22 corresponded to annotated 3'UTRs (Zv9, ENSEMBL v68) and 20 represented novel 3' ends. All of these True sites were amplified using the GPAT assay, indicating these are true polyadenylated 3' ends (Figure 5C, E; Supplementary Table 3), including 11 novel 3' ends that were identified by 24 hpf PAS-Seq but not 24 hpf 3pseq. 13 validated True sites contained PASs other than the canonical AAUAAA within 40 nt upstream of the cleavage site and 1 lacked any consensus motif, demonstrating that our classifier can identify 3' ends without a consensus PAS (Supplementary Table 3). Along with the putative True set, we assayed eight sites that were classified as False, only one of which was annotated as a 3' end in ENSEMBL (Supplementary Table 3). Half of these sites displayed a variant PAS near the putative 3' end, while the remaining sites contained no PAS 40 nt upstream. 7 out of the 8 sites classified as False failed to amplify in the GPAT assay, but were detected by dtPAT suggesting that they arise from internal oligo-dT priming (Supplementary Table 3; Figure 5D, E). Furthermore, six of these sites contained fewer than eight adenines in the downstream region and were called true by the 8A heuristic filter (Figure 5F; Supplementary Table 3). One False site, which did not possess a consensus PAS and contained only three downstream adenines, was amplified by the GPAT assay (Figure 5E). Together, our biological cross-validation of putative pA sites demonstrates the high accuracy of the naïve Bayes classifier. Importantly, our classifier facilitated the identification of novel pA sites from PAS-Seq allowing the discovery of new 3'UTRs in the zebrafish transcriptome.

3.3 Naïve Bayes classifier displays utility in other species

To determine if our algorithm, which was trained using zebrafish datasets, could be applied to other species, we used it to filter mammalian datasets generated using polyA-seq, an alternative oligo-dT primed 3' end sequence method (Derti, et al., 2012; see following and data not shown). In this previous study, data were filtered using an empirically derived threshold of a log ratio, calculated as a product of mono-nucleotide frequencies in the 10 nt downstream of putative 3' ends relative to those from internal priming sites (referred to hereafter as the “Derti filter”). As a metric to assess the efficacy of our classifier, we compared the results of filtering polyA-seq data with the naïve Bayes classifier to the output of the Derti filter. For example, unfiltered polyA-seq data from human kidney exhibited similarities to our negative training set, consistent with internal oligo-dT priming (Figure 6A). From approximately half million putative 3' ends in the unfiltered data, application of the naïve Bayes classifier identified more than 130,000 pA sites, which exhibited the expected characteristics for true polyadenylated 3' ends (Figure 6B). In comparison, the Derti filter (Derti, et al., 2012) classified 94,945 as true pA sites, which also exhibited the expected sequence patterns (Figure 6C). Of these, approximately 77,000 were commonly assigned by the two filtering methods (Figure 6D). Closer inspection revealed that nearly all pA sites identified uniquely by the Derti filter have fewer than 5 adenines in the 10 nt downstream (Figure 6E), consistent with the focus of this filter on nucleotide frequencies in the downstream region (Derti, et al., 2012). In comparison, our naïve Bayes classifier identifies pA sites with all proportions of adenines in the downstream region, including 54,046 true pA sites called false by the Derti filter (Derti, et al., 2012). Importantly, the majority of sites uniquely identified by our classifier possess a canonical PAS, suggesting that they are true 3' ends (Figure 6F). By contrast, the majority of true sites uniquely identified by the Derti filter did not display a PAS in the upstream region (Figure 6F), suggesting that many may be false positive calls. However, without biological cross-validation, it is difficult to assess the false-positive rate within this group. In any event, these observations suggest that our naïve Bayes classifier, trained on zebrafish 3' end sequencing data, performs well in the identification of pA sites from mammalian species. Furthermore, our classifier discovered many more likely

Fig. 6. Naïve Bayes classifier shows utility in filtering human 3' end sequencing datasets. Nucleotide profiles and sequence logos of over-represented motifs 40 nt upstream and 30 nt downstream of putative pA sites that were A. Unfiltered, B. assigned as true by naïve Bayes classifier, or C. called true by Derti et al. positional discriminant function (Derti, et al., 2012). D. Overlap of putative pA sites called true by naïve Bayes classifier and Derti et al. positional discriminant function (Derti, et al., 2012). E. Number of downstream adenines or F. PAS distribution for putative pA sites called true by naïve Bayes, both naïve Bayes and Derti et al. positional discriminant function, or just Derti et al. positional discriminant function.
true positive pA sites from unfiltered data than the Derti filter. This is likely due to the interrogation and analysis of multiple sequence elements during the training of this classifier, while the Derti filter is restricted to consideration of only mononucleotide frequencies immediately downstream of the pA.

Based on our work, a trained naïve Bayes classifier is clearly beneficial to identify true pA sites from oligo-dT primed 3' end sequencing data from both zebrafish and other animal species. Additionally, this approach may also be helpful to assess and improve the quality of 3' ends of RNA-seq transcript models built from standard transcript annotation software, such as Cufflinks or Scripture (Guttman et al., 2010; Trapnell et al., 2010). Indeed, application of our classifier to previously published RNA-seq transcript models from zebrafish (Pauli et al., 2012) suggests a significant number of transcript ends may be due to internal oligo-dT priming, while filtering enriches for canonical and variant PAS hexamers (Supplementary Figure 4). The usage of our classifier could also be extended to RefSeq and other sequence databases, as these gene models have been largely built from oligo-dT primed cDNAs and likely contain a significant number of incorrect 3' end annotations. Indeed, an estimated 12% of ESTs labeled as 3' ends in dbEST human (release 10/04/2001) are due to internal oligo-dT priming (Nam et al., 2002) and our naïve Bayes classifier correctly predicted that the 3' end of vegfc, as annotated by ENSEMBL in the zebrafish genome (Zv9), is due to mis-priming (Supplementary Table 3). Thus, naïve Bayes filtering of annotated sequences in available databases, in addition to previously published genome-wide oligo-dT primed sequencing data, will likely lead to identification of new pA sites and eliminate false internally primed sites. Further studies are needed to assess the performance of the naïve Bayes classifier, trained on zebrafish data, in yeast and plants. In conclusion, the naïve Bayes classifier developed in these studies will facilitate the identification of novel pA sites in combination with simple oligo-dT primed 3' end sequencing.

ACKNOWLEDGEMENTS

We would like to thank Jianhong Ou for his help with the Bioconductor Package. We would like to thank George Kourkoulis and Ira Male for fish care and maintenance. We thank the Sanger center for use of unpublished data. We thank the UMass Medical School Deep Sequencing Core Facility, the Genomics Tufts University Core Facility, and the University of California Riverside HT Sequencing Core Facility. Additionally, we would like to thank Victor Ambros, Jeffrey Bailey, Andrea D’Ambrogio, John Keaney, Charles Sagerström, and Marian Walhout for helpful discussion about this project. S.S. was supported by a pre-doctoral fellowship from the American Heart Association (10PRE4170016). N.L. and L.J.Z. were supported by R01HL101374, from the National Heart, Lung, and Blood Institute for this work.

Funding: S.S. was supported by a pre-doctoral fellowship from the American Heart Association (10PRE4170016). N.L. and L.J.Z. were supported by R01HL101374, from the National Heart, Lung, and Blood Institute for this work.

REFERENCES


