University of Massachusetts Medical School

eScholarship@UMMS

GSBS Dissertations and Theses

Graduate School of Biomedical Sciences

2013-04-29

Application of a Naïve Bayes Classifier to Assign Polyadenylation Sites from 3' End Deep Sequencing Data: A Dissertation

Sarah E. Sheppard
University of Massachusetts Medical School

Repository Citation

Sheppard SE. (2013). Application of a Naïve Bayes Classifier to Assign Polyadenylation Sites from 3' End Deep Sequencing Data: A Dissertation. GSBS Dissertations and Theses. https://doi.org/10.13028/M20K68. Retrieved from https://escholarship.umassmed.edu/gsbs_diss/653

APPLICATION OF A NAÏVE BAYES CLASSIFIER TO ASSIGN POLYADENYLATION SITES FROM 3' END DEEP SEQUENCING DATA

A Dissertation Presented

By

SARAH ELIZABETH SHEPPARD

Submitted to the Faculty of the
University of Massachusetts Graduate School of Biomedical Sciences,
Worcester
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

April 29, 2013

Department of Molecular Medicine

Program in Gene Function and Expression

APPLICATION OF A NAÏVE BAYES CLASSIFIER TO ASSIGN POLYADENYLATION SITES FROM 3' END DEEP SEQUENCING DATA

A Dissertation Presented
By

SARAH ELIZABETH SHEPPARD

The signatures of the Dissertation Committee signify completion and approval as to style and content of the Dissertation

Nathan Lawson, Ph.D., Thesis Advisor

Victor Ambros, Ph.D., Member of Committee

Jeffrey Bailey, M.D., Ph.D., Member of Committee

John Keaney, M.D., Member of Committee

Charles Sagerström, Ph.D., Member of Committee

Alexander Schier, Ph.D., Member of Committee

The signature of the Chair of the Committee signifies that the written dissertation meets the requirements of the Dissertation Committee

Marian Walhout, Ph.D., Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies that the student has met all graduation requirements of the school.

Anthony Carruthers, Ph.D.,
Dean of the Graduate School of Biomedical Sciences

MD PhD Program
April 29, 2013

Dedication

My thesis is dedicated to my parents, Christine Wang and Norman.

Acknowledgements

I would first like to acknowledge my advisor and mentor, Nathan Lawson. His excitement, dedication, and love for his research are inspiring. He has a never ending drive to keep pushing a bit harder, discover something new, and think outside the box. He has encouraged me to independently investigate problems and overcome obstacles that appeared insurmountable. This has given me the self-confidence to believe I too can be a good scientist and independent thinker. I finally understand what it is to breathe and live my work everyday. For all this, I am grateful.

Julie Zhu has also been an incredible mentor. This project would not have been possible without her assistance and knowledge of machine learning. She has been incredibly enthusiastic and encouraging, which has helped to motivate me and work harder than I thought I could.

Victor Ambros, John Keaney, Charles Sagerstrom, and Marian Walhout have been instrumental in my development as a scientist and have provided great suggestions to aid my project. I'd specifically like to thank Marian for her additional encouragement and mentoring. I'd also like to thank Jeffrey Bailey for providing additional critiques for my thesis work and sitting on my defense committee. Additionally, I'd like to thank Alexander Schier for being my external thesis reviewer.

I was fortunate to have incredible grade school teachers, Mary Martines and Michael Griffin, and undergraduate mentors, Neil Kumar, Doug Lauffenburger, Ley Richardson, Linda Griffith, Jeff Coderre, and Emma Teng without whom I would not have pursued this path.

Additionally, I would like to thank the Lawson Lab members for all their support. Arndt Siekmann and Stefania Nicoli were both great post-docs that helped me learn the ins and outs of the lab. Siobhan Craige has opened her home to me, and has spent countless hours reading my thesis and listening to my oral defense. Ann Grosse has been so supportive. Tom Smith, Jacques Villefranc, George Kourkoulis, Aaron Aday, John Moore, Steph Chu, and Elizabeth Nourse have been a great lab “family”. I’d also like to thank Fatma Kok and Aurelie Quillie for reading portions of my thesis. Julien Dubrulle and Andrea Pauli, from Alex Schier’s lab, taught me how to successfully transplant germ cells to make maternal zygotic mutants- an invaluable skill! I’d also like to thank Shanna Spencer for her administrative support. I’d also like to thank Susie and Nina for keeping the lab running and John and Pat for taking care of the fish.

I’d like to thank Andrea D’Ambrogio for his friendship and insightful conversations about our projects. The friendship and shared experiences of Jimmy Young, Tim Rooney, and Steve Weicksel have been helpful. I’m also so thankful to Anne Michelson, my MD PhD program “Mom” who watches over everything and always has an open door for me.

I’d like to thank my family and friends. My parents, Christine Wang and Norman, and my pseudo-parents, Uncle Paul and Auntie Liz, are incredible scientists and role models. They all have always encouraged me to explore my intellectual curiosity and seek out unconventional pursuits. Even now, my mother continues to nurture and care for me (in addition to reviewing my thesis!). My brother, Peter, has also been incredibly supportive. My extended family, Auntie Andrea, Uncle Jay, Abby, Faith, and Grandma, have sent so many good thoughts, treats, and even simple groceries, to help me focus on my thesis. Although my Papa is not here with us today, I also want to thank you Papa. You encouraged me to study and work hard, and you always were so proud of me.

Ryan Ballentine has been like a brother to me. With no knowledge of biology he has spent hours with discussing my project and helping me clarify important points in my work and given me a better understanding of machine learning.

Finally, my fiancée, Evan, has shown such patience and nurturing throughout this process. He has taken care of everything and allowed me to focus solely on science. I couldn’t have made it through this without him or Ninja-pup.

I’d like to thank the American Heart Association for financial support through a pre-doctoral fellowship.

Abstract

Cleavage and polyadenylation of a precursor mRNA is important for transcription termination, mRNA stability, and regulation of gene expression. This process is directed by a multitude of protein factors and *cis* elements in the pre-mRNA sequence surrounding the cleavage and polyadenylation site. Importantly, the location of the cleavage and polyadenylation site helps define the 3' untranslated region of a transcript, which is important for regulation by microRNAs and RNA binding proteins. Additionally, these sites have generally been poorly annotated. To identify 3' ends, many techniques utilize an oligo-dT primer to construct deep sequencing libraries. However, this approach can lead to identification of artifactual polyadenylation sites due to internal priming in homopolymeric stretches of adenines. Previously, simple heuristic filters relying on the number of adenines in the genomic sequence downstream of a putative polyadenylation site have been used to remove these sites of internal priming. However, these simple filters may not remove all sites of internal priming and may also exclude true polyadenylation sites. Therefore, I developed a naïve Bayes classifier to identify putative sites from oligo-dT primed 3' end deep sequencing as true or false/internally primed. Notably, this algorithm uses a combination of sequence elements to distinguish between true and false sites. Finally, the resulting algorithm is highly accurate in multiple model systems and facilitates identification of novel polyadenylation sites.

Table of Contents

List of Tables

Chapter I

Chapter II

Chapter III

Appendices

List of Figures

Chapter I

Chapter II

Chapter III

Chapter IV

Appendices

List of Third Party Copyrighted Material

Figure Number	Publisher	License Number
Figure 1.2	Nature Publishing Group	3124151312904
Figure 1.6	Nature Publishing Group	3124160430455

List of Abbreviations

3’UTR: 3’ untranslated region

αSyn: α-synuclein

APA: alternative polyadenylation

CFI: Cleavage Factor I

CLIP: cross-linking (usually UV) followed by immunoprecipitation

CLIP-seq: cross-linking (usually UV) followed by immunoprecipitation and then deep sequencing

cMO: caged Morpholino

CPSF: Cleavage and Polyadenylation Specificity Factor

CPEB1: cytoplasmic polyadenylation element binding protein 1

CstF: Cleavage Specificity Factor

CYC1: iso-1-cytochrome c

DSE: downstream sequence element

EST: expressed sequence tag

GTP: guanine triphosphate

kDa: kilodalton

miRNA: micro RNA

mRNA: messenger RNA

MO: morpholino

nt: nucleotide

oligo: oligonucleotide

oligo-dT: Oligonucleotide of deoxythymines

PCR: polymerase chain reaction

pA site: cleavage and polyadenylation site

PAP: poly(A) polymerase

PAS: polyadenylation signal

poly(A): poly-adenosine

pre-mRNA: precursor-messenger RNA

RNAP II: RNA polymerase II

RRM: RNA recognition motif

SELEX: systematic evolution of ligands by exponential enrichment

SNP: single nucleotide polymorphism

SV40: simian virus 40

USE: upstream sequence element

UV: ultra violet

wt: wild type

Preface

The work presented in Chapter 2 was submitted to *Developmental Biology*. This project was in collaboration with John Moore, and the contributions are outlined in the chapter.

The work presented in Chapter 3 was submitted to *Bioinformatics*. The naïve Bayes classifier was implemented by Julie Zhu.

CHAPTER I

INTRODUCTION

The central dogma of molecular biology was first proposed by Francis Crick in the 1950s to explain the flow of genetic information. In a very simplified explanation, the central dogma states that DNA is transcribed into RNA, which is translated into protein. During transcription, RNA polymerase II (RNAP II) is recruited to a gene promoter to initiate transcription. The direct RNA transcript is called precursor-messenger RNA (pre-mRNA) and must be processed into messenger RNA (mRNA) before it can be used as a template for translation.

Key steps in RNA processing are the addition of a methylated guanine (GTP) cap; splicing of introns and exons; cleavage of the 3' end; and successive addition of a poly-adenosine (poly(A)) tail. The 5' end of the pre-mRNA is modified by the addition of a GTP cap when it is about 20 nucleotides (nt) long [1]. Splicing of the pre-mRNA to remove introns, which do not code for protein, is coupled with transcription elongation [2]. The 3' processing complex recognizes specific sequence elements (*cis* regulatory elements) located on the 3' end of the transcript leading to cleavage of the pre-mRNA by Cleavage and Polyadenylation Specificity Factor [3]. Finally, poly(A) polymerase (PAP) adds 200-300 adenosines to the 3' end (in humans), partially regulated by Poly(A) Binding protein, which binds after 11-14 adenosines are added [4]. Importantly, the C terminal domain of RNAP II increases cleavage efficiency *in vitro* [5]. It is thought that cleavage and polyadenylation are post-transcriptional modifications that occur while the C terminal domain of RNAP II is bound to cleavage and polyadenylation factors (Figure 1.1) [2]. 3' end processing has also been linked to proper transcription termination [6-8].

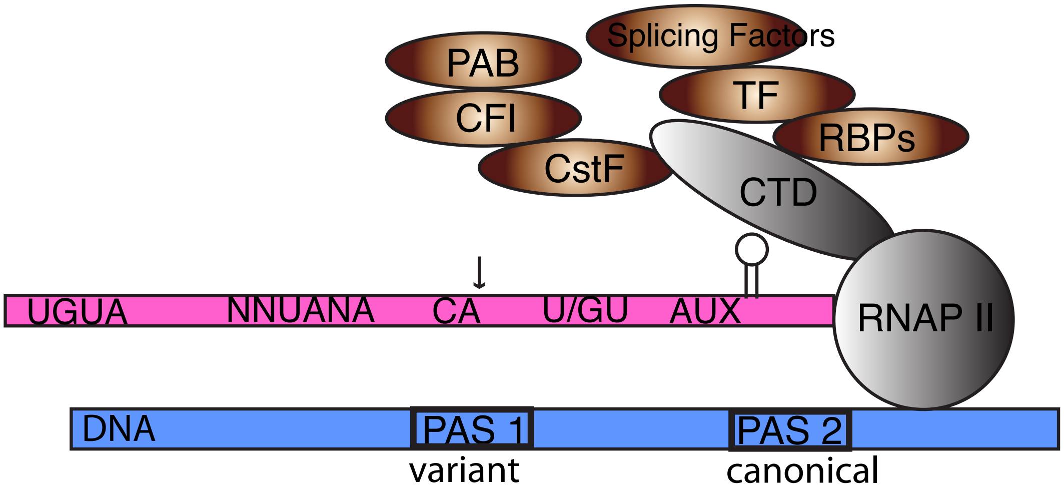
Cleavage and polyadenylation of a pre-mRNA is important for many reasons. First, the cleavage site determines the 3' end of the transcript. This defines an important regulatory region downstream of the stop codon known as the 3' untranslated region (3'UTR). The 3'UTR can be bound by microRNAs, short non-coding regulatory RNAs, or RNA binding proteins, both of which can affect the stability or the translatability of the transcript and thus gene expression. Cleavage and polyadenylation are also important for export of the mRNA into the cytoplasm and transcript stability. The poly(A) tail is bound by poly(A) binding protein, which in combination with the 5' cap, helps regulate translation [9].

Due to the importance of the 3'UTR and generally poor annotation of sites of cleavage and polyadenylation (pA) sites, techniques have been developed to identify pA sites on a genome-wide scale (reviewed in [10]). However, a major drawback is that the majority of these methods rely on priming by an oligonucleotide (oligo) of deoxythymines (oligo-dT). Oligo-dTs may bind to internal homopolymeric stretches of adenines as well as to the poly(A) poly(A) tail. Internal priming events are generally defined by

Figure 1.1: Cleavage and Polyadenylation is regulated by a multitude of factors. Purification of the human 3’ processing complex identified ~85 proteins involved in cleavage and polyadenylation [11]. These included known polyadenylation factors such as the CPSF complex, CstF complex, CF I complex, CF II, polyA polymerase, and multiple polyA binding proteins. Additionally, RNAP II, transcription factors such as TF I and TF II, and splicing factors were discovered.

Cis sequence elements may direct 3’ end processing. Cleavage and polyadenylation at a proximal or intronic site is associated with variant polyadenylation signals upstream and U rich sequences downstream [12-24]. Usage of distal polyadenylation sites is associated with canonical polyadenylation signals (AAUAAA) and GU rich sequences downstream. Far upstream U rich elements, such as UGUA, as well as auxilliary downstream G- rich elements or secondary structure may aid in cleavage and polyadenylation [25].

Additionally, changes in protein factors will also affect cleavage and polyadenylation. Increasing amounts of transcription factors, such as E2F and Mef2, are associated with 3’UTR shortening [26, 27]. Additionally increasing amounts of CstF are associated with proximal polyadenylation site usage [23, 28], while increasing amounts of CFI are associated with distal polyadenylation site usage [20]. Decreasing or blocking splicing factors is associated with polyadenylation at proximal sites [23, 29]. Additionally, RNAP II pausing, allowing polyadenylation factors to interact with a weaker polyadenylation signal, is associated with polyadenylation at a proximal site [27, 30-34]. Finally, RNA binding proteins or polyA binding proteins may facilitate interaction of polyadenylation factors with a specific polyadenylation site or block the interaction of a polyadenylation factors with a specific site [35-38]. Adapted from [39].


Splicing Factors
PAB
TF
CFI
RBPs
CstF
CTD
UGUA
NNUANA
CA
U/GU
AUX
RNAP II
DNA
PAS 1
variant
PAS 2
canonical

a proportion of adenines downstream of a putative site and removed during computational analysis. Simple filtering does not remove all instances of internal priming, and may remove true 3' ends [40, 41]. Thus, I used machine learning to reliably distinguish true 3' ends from internal priming events. This innovative approach is highly accurate in multiple organisms and facilitates identification of novel 3'UTRs.

In this introduction, I will review cleavage and polyadenylation of mRNAs, including a discussion of alternative polyadenylation. Next, I will describe technical approaches that have been used to identify 3' ends. Finally, I will introduce machine learning and its application in biology and medicine.

Cleavage and Polyadenylation of pre-mRNAs

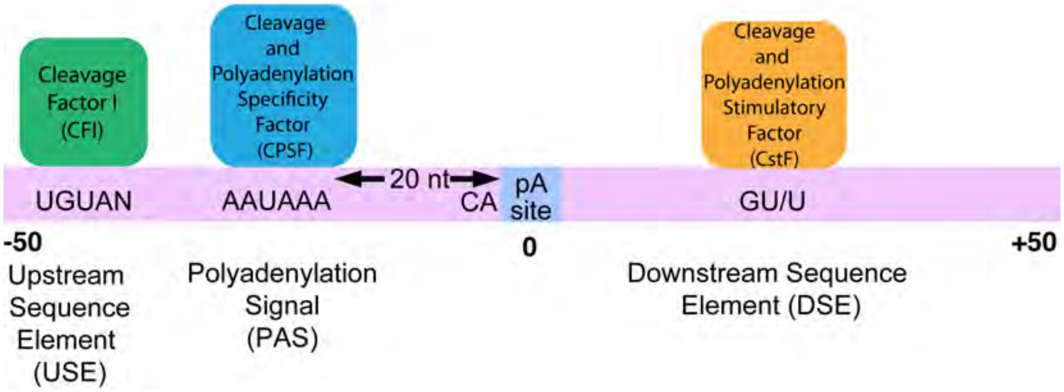
The 3' processing complex is composed of 85 proteins [11], including multi-subunit protein complexes, such as Cleavage and Polyadenylation Specificity Factor (CPSF), Cleavage Specificity Factor (CstF), Cleavage Factor I (CFI), that bind *cis* regulatory elements, known as the polyadenylation signal (PAS), downstream sequence element (DSE), and upstream sequence element (USE), respectively, to modulate cleavage and polyadenylation (Figure 1.1). In one of the earliest studies, Proudfoot and Brownlee aligned the sequences adjacent to the 3’ end of six different messenger RNAs (mRNA) from mouse, chicken, rabbit and human [42]. They discovered the sequence AAUAAA was present approximately 20 nt from the 3’ end in all six mRNAs (Figure 1.2). AAUAAA is now known as the canonical PAS, though single nt variants are also functional [43]. The PAS is known to bind CPSF (Figure 1.3) [44]. Proudfoot also noted in five out of the six mRNA ends surveyed that a cytosine was the last nucleotide before the poly(A) tail (Figure 1.2) [42]. Furthermore the dinucleotide CA is present most frequently 5’ of the pA site (Figure 1.1, 1.2, 1.3) [20, 43, 45]. Additionally, downstream of the pA site guanine/uracil or uracil rich sequences, known as DSE, bind CstF [46]. Uracil rich sequences upstream of the PAS, known as USE may recruit CFI [47]. CPSF, CstF, and CFI may be present along the entire transcriptional unit [31], thus the USE, PAS, and DSE signal where the transcript should be cleaved and polyadenylated.

Diverse combinations of the *cis* elements described above may define the site of cleavage and polyadenylation of a pre-mRNA. For example, only an adenine-rich sequence, rather than a canonical PAS, and strong uracil-rich DSE are needed for 3’ end processing of *JUNB* [48]. Stronger PASs contain uracil-rich elements 5 nt downstream and a slightly greater enrichment 25 nt downstream, whereas sites with no PAS have a

Figure 1.2. Alignment of the 3’ end sequence in six mammalian mRNAs. A. Rabbit a-globin. B. Rabbit b-globin. C. Human a-globin. D. Human b-globin. E. Mouse immunoglobulin light chain. F. Chicken ovalbumin mRNAs.

```
    30              20                  10
a   U-G-G-U-C-U-U-U-G-A-A-U-A-A-A-G-U-C-U-G-A-G-U-G-A-G-U-G-G-C-poly(A)

    30              20                  10
b   U-G-G-C-U-A-A-U-A-A-A-G-G-A-A-A-U-U-U-A-U-U-U-U-C-A-U-U-G-C-poly(A)

    30              20                  10
c   U-G-G-U-C-U-U-U-G-A-A-U-A-A-A-G-U-C-U-G-A-G-U-G-G-G-C-G-G-C-poly(A)

    30              20                  10
d   U-G-C-C-U-A-A-U-A-A-A-A-A-A-C-A-U-U-U-A-U-U-U-U-C-A-U-U-G-C-poly(A)

    30              20                  10
e   A-A-U-A-U-U-C-A-A-U-A-A-A-G-U-G-A-G-U-C-U-U-U-G-C-A-C-U-U-G-poly(A)

    30              20                  10
f   C-C-U-U-U-A-A-U-C-A-U-A-A-U-A-A-A-A-A-C-A-U-G-U-U-U-A-A-G-C-poly(A)
```

Figure 1.3 *Cis* regulatory elements bind proteins involved in cleavage and polyadenylation to define the cleavage and polyadenylation site. Uracil rich upstream sequence elements, such as UGUAN, bind Cleavage Factor I. The polyadenylation signal, AAUAAA or a single nucleotide variant, is located approximately 20 nucleotides upstream of the cleavage and polyadenylation site and is bound by Cleavage and Polyadenylation Specificity Factor. Cleavage usually occurs 3’ of a CA dinucleotide. Downstream sequence elements are bound by Cleavage and Polyadenylation Stimulatory Factor and are generally guanine/uracil or uracil rich.


Cleavage Factor I (CFI)
Cleavage and Polyadenylation Specificity Factor (CPSF)
Cleavage and Polyadenylation Stimulatory Factor (CstF)
UGUAN
AAUAAA
20 nt
CA
pA site
GU/U
-50
0
+50
Upstream Sequence Element (USE)
Polyadenylation Signal (PAS)
Downstream Sequence Element (DSE)

single strong peak of uracil-richness 20 nt downstream [24]. Furthermore, 3’ end sequencing in parallel with cross-linking immunoprecipitation of CstF-64 followed by deep sequencing (CLIP-seq) revealed putative pA sites that may contain CstF-64 bound downstream had uracil-rich motifs in the 20-40 nt downstream, whereas sites only identified by 3’ end sequencing had mostly guanine richness in the same region [23]. Additionally, sites identified by CLIP-seq of CstF-64 containing AAUAAA upstream showed UG-rich motifs, while those without AAUAAA showed U-rich motifs. Taken together, these results suggest that a combination of sequence elements is important for 3’ end processing.

CPSF and the Polyadenylation Signal

Though the canonical PAS is known as AAUAAA, different kingdoms tend to use different motifs, though all are located ~20 nt upstream of the pA site. In *Entamoeba histolytica*, a human parasite, AAWUDA motif is associated with polyadenylation [49]. Yeast tend to use AAAATA, with the motif AATAAA enriched slightly less [50, 51], though more recently AAWAAA was also identified [52]. AAUAAA is still the dominant PAS in plants, but is only present upstream of 7% of pA sites [40, 53]. Usage of canonical AAUAAA in 10 metazoans (*H. sapiens, C. familiaris, M. musculus, R. Norvegicus, G. galus, D. rerio, T. rubripes*, *D. melanogaster, A. gambiae*, and *C. elegans*) ranged from ~50-70% of pA sites, with the higher order organisms using the canonical PAS more frequently [12, 13, 17, 18, 26, 54-57]. Other sites may use a variant PAS or no PAS at all. The consensus PAS identified in humans is NNUANA, suggesting that positions 1,2,5 are highly variable while positions 3,4,6 are highly conserved [17]. Approximately 10-15% of *C. elegans*, mouse, and human pA sites use no PAS [12, 13, 17, 18, 24, 55, 57], although, the proportion is slightly higher in zebrafish [56] and *Drosophila* [55]. Finally, comparison of pA sites in human testis, liver, kidney, muscle, and brain showed these tissue use similar proportions of canonical and variant PASs [45].

To determine the functionality variant PASs, single nt mutations of AAUAAA in the simian virus 40 late (SV40) poly(A) signal were generated and examined in *in vitro* polyadenylation or cleavage reactions using HeLa cell extracts [43]. All single nt variants of AAUAAA decrease *in vitro* polyadenylation. Cleavage efficiency, tested in a subset of the variants used for the polyadenylation reaction, was decreased similarly. Interestingly, Graber et al. demonstrated the distribution of AAUAAA and variants in *Drosophila*, mouse, and human correlated with these *in vitro* activities [43, 58]. Additionally 269 vertebrate mRNAs were cleaved most frequently adjacent to the dinucleotide CA [43]. Single or double mutations of this dinucleotide shifted the position of cleavage site but did not affect cleavage efficiency.

CPSF is necessary for *in vitro* polyadenylation and preferentially binds RNAs containing AAUAAA. Only CPSF and PAP were needed for *in vitro* polyadenylation to occur [44]. Polyadenylation was inhibited by the addition of recombinant CPSF, which acts a dominant negative [59]. Furthermore, in a reaction with only PAP, immunodepletion of CPSF decreases polyadenylation efficiency, which can be rescued by addition of exogenous CPSF [59, 60]. Polyadenylation of RNAs containing AAUAAA was more efficient than RNAs containing a mutant PAS [43, 44]. CPSF preferentially bound RNAs containing AAUAAA better than mutants [44, 59, 60].

CstF and the Downstream Sequence Element

DSE tends to be uracil rich in all organisms and is bound by CstF [23, 26, 50, 51, 54, 61, 62]. Initially, a reporter assay suggested the sequence YGTGTTYY downstream of the pA site was important for cleavage and polyadenylation [6]. Moreover, inspection of 100 mammalian and viral genes showed that 67% contain the consensus YGTGTTYY 24-30 nt downstream of AATAAA [6]. RNA footprinting mapped CstF to uracil rich sequences 14-30 nt downstream of the pA site [63]. In agreement with the previous observation, CstF was not able to bind the pre-mRNA if the DSE was replaced with unrelated sequence [63]. Notably, this also inhibited cleavage, which could be rescued by the insertion of five uracils at the same position of the original DSE [63]. Systematic evolution of ligands by exponential enrichment (SELEX) experiments [46], nuclear magnetic resonance [30], and cross-linking immunoprecipitation of CstF-64 followed by deep sequencing (CLIP-seq) [20] confirmed that recombinant human CstF-64 binds to uracil rich sequences, interspersed with guanines, similar to the consensus YGTGTTYY identified in [6]. Correspondingly, SELEX experiments using the yeast homolog of CstF-64, which has ~50% sequence homology, identified a single consensus that is uracil/guanine rich, but also contains cytosines [46]. Interestingly, helix C of CstF-64 is strongly conserved in metazoans (human, mouse, xenopus, drosophila, worm) but not in plants or yeast [30]. As helix C is located perpendicular to the RNA recognition motif, this may explain the preference for cytosines in addition to uracil and guanine seen in yeast CstF-64 [46].

The location of the PAS in relation to the DSE may determine the cleavage efficiency and cleavage location. Shifting the PAS closer to the DSE inhibited cleavage [64]. Correspondingly, moving the PAS further upstream, shifting the PAS distally, or shifting both distally decreased cleavage efficiency at the original site and resulted in additional smaller or larger products respectively [64]. A more comprehensive analysis of 3' ends revealed the PAS was positioned 11-23 nt upstream of the pA site in nearly all genes examined [64]. The majority also contained a uracil rich DSE 10-30 nt downstream of the pA site. Taken together, these results demonstrate the restricted positional arrangement of the PAS and DSE.

CFI and the Upstream Sequence Element

Uracil-rich elements dominate the USE [50-53, 58, 61, 62] and may enhance cleavage and polyadenylation efficiency by binding CFI. SELEX [47] and CLIP-seq [20] discovered CFI bound UGUAN sequences, though other uracil-rich USEs are functional. Addition of short antisense oligonucleotide to bind USEs [7, 47] or mutation of USEs [29, 31] resulted in decreased cleavage and polyadenylation of multiple pre-mRNAs. Similarly, substitution of USE in *COL1A2* decreases polyadenylation, while addition of USE to the upstream region in the adenovirus IVA2, normally lacking USE, increases polyadenylation [7]. Accordingly, shifting USE closer to PAS or adding a second USE increased the cleavage activity of prothrombin pre-mRNA [29]. It is interesting to note that splicing factors, such as U2AF65 and hnRNP1 may facilitate cleavage in RNAs containing USE [29].

UGUAA elements are located 55 nt upstream, 22 nt upstream, and 7 nt downstream of the pA site in human 68 kiloDalton subunit of CFI

(CFI_M-68) pre-mRNA [47]. Low doses (0.02 or 0.2 pmol) of CFI_M led to increased cleavage of CFI_M-68 pre-mRNA but high doses (2 or 4 pmol) led to decreased cleavage. Mutation of all three UGUAA elements or just USE located 22 nt upstream of the pA site led to decreased cleavage compared to the wild type (wt), although addition of CFI_M led to a dose-dependent increase in cleavage efficiency. Mutation of the UGUAA 7 nt downstream led to a decrease in cleavage efficiency, while mutation of the USE 55 nt upstream did not change the cleavage efficiency. However, addition of CFI did not change the cleavage efficiency in either of these mutations. Conversely, increasing amounts of CFI_M to wt showed dose dependent increased polyadenylation, which was decreased by mutation of either of the upstream USEs. Taken together, these results suggest that CFI may regulate its own expression through three UGUAA elements.

Alternative Polyadenylation

Alternative polyadenylation (APA), or the choice of different pA sites within a transcript, is common among many organisms (Table 1.1).

Table 1.1: Prevalence of alternative polyadenylation. The percentage of genes identified with alternatively polyadenylated isoforms.

Sample	**% APA**	**Reference**
C. elegans	43%	[12]
Drosophila	54.3%	[15]
Danio rerio (zebrafish)	55%	[56]
Arabidopsis	59%	[65]
	70%	[66]
rice	51%	[53]
	82%	[65]
mouse	32%	[18]
mouse cells undergoing neuronal differentiation	52%	[16]
human	54%	[18]
normal and tumor matched human breast, colon, kidney, liver and lung	30%	[19]

Moreover, the relative location of a pA site within a gene may be associated with distinct polyadenylation signals. APA may alter both the coding sequence and non-coding 3’UTR. Longer 3’UTRs may contain additional binding sites for microRNAs or RNA binding proteins compared to their shorter counter parts, resulting in increased positive or negative regulation. However, changes in 3’UTR length among different organisms may not correlate with changes in microRNA regulation, as the average shorter 3’UTR in worm had increased predicted miRNA binding site density compared to fly or human [13]. In addition, a pA site located within an intron can alter the protein structure of a gene. APA may be stage- or tissue-specific and thus may contribute to the proper development and differentiation of an organism. Correspondingly, improper cleavage and polyadenylation may result in APA associated with disease.

The location or utilization of a pA site within a gene may correlate with the *cis* elements controlling cleavage and polyadenylation. Genes with a single 3’UTR tend to use canonical PASs compared to alternatively polyadenylated genes [18, 49]. Proximal or middle pA sites were more likely to use variant PASs, while the most 3’ distal pA sites were enriched for the canonical AAUAAA [12-22] and more uracil/guanine rich elements in the downstream region [15, 16]. Similarly, most frequently used pA sites tended to use the canonical AAUAAA [40, 41], as well as UGUA upstream and uracil or uracil/guanine rich elements downstream [41]. The distal pA sites also tended to be more frequently used, suggesting more efficient processing perhaps due to the usage of AAUAAA [17, 18, 20].

Intronic Polyadenylation Sites

20% of pA sites in 16,610 human genes are located in introns [33]. Cleavage and polyadenylation within an intron may result in an alternative last exon, thereby modifying coding sequence and protein structure. For instance, primary B cells express the membrane-bound form of immunoglobulin M heavy chain associated with a distal pA site, while differentiated B cells express soluble immunoglobulin M heavy chain associated with an intronic pA site. CstF-64 preferentially binds the distal site [28]. When stimulated to differentiate, B cells upregulated expression of CstF-64, likely allowing it to bind at the intronic pA site and producing transcript for soluble immunoglobulin M heavy chain [28].

Overall, introns containing pA sites are larger than introns without pA sites [33]. As the 3’ processing complex assembles slower on weak pA sites [32], pausing of splicing machinery in longer introns may allow for these alternative pA sites to be used. mRNA-seq from ten human tissues and five breast cancer cell lines showed alternative 3’UTRs were highly correlated with skipped exons and over-represented motifs were shared between introns and distal 3’UTR extensions [34], suggesting splicing may be involved in intronic polyadenylation. CstF bound additional intronic locations not identified by 3’ end sequencing, suggesting a mechanism exists to block cleavage and polyadenylation at these CstF binding sites [23]. Preventing U1 snRNP, a spliceosomal protein, from binding led to polyadenylation at these intronic sites, suggesting that splicing and 3’ end processing are interconnected.

3’UTR Usage in Proliferation and Transformation

Regulation of 3’UTR usage may be one way to control the proliferative capacity of the cell. Highly proliferative cells tend to show increased proximal pA site usage resulting in shortening of 3’UTRs. 3’ end profiling in multiple cancer tissues [19, 67] and proliferating cells [26] demonstrated increased 3’ end processing at proximal sites compared to controls. In opposition to the general trend, highly invasive breast cancer cells, that do not express the estrogen receptor, showed 3’UTR lengthening compared to a normal breast tissue cell line [67]. Nevertheless, stimulation of resting mouse T cells, human T cells, B cells, or monocytes decreased extended 3’UTR usage 48 hours after stimulation, demonstrated by a negative correlation (R = -0.81) between proliferation and 3’UTR length [68]. Similarly, *Cyclin D1, IMP-1, DICER1, Cyclin D2, RAB10, FGF2* had increased usage of shorter 3’UTRs in multiple cancer cell lines [21]. This 3’UTR shortening was associated with increased protein production that could be attributed to loss of microRNA regulation [21, 68]. However, in MCF7 breast cancer cells, which also showed increased usage of proximal pA site, there was no negative correlation between 3'UTR length and gene expression [67].

Multiple mechanisms may contribute to enhanced processing of proximal pA sites during proliferation. Transcription factors E2F1 and E2F2 were increased in proliferating cells [26]. Knockdown of E2F1/2 decreased expression of 3' end processing proteins and increased distal/proximal 3'UTR ratios, suggesting increased E2F1/2 during proliferation may up-regulate cleavage and polyadenylation proteins to facilitate usage of proximal pA sites [26]. Stimulation of neuronal cells was associated with increased MEF2, a transcription factor, and shortened 3'UTRs [27]. Moreover, RNAP II was enriched at the shortened end of mRNAs upon stimulation, but not the longer ends [27], indicating MEF2 stimulation may increase RNAP II pausing to allow for cleavage and polyadenylation at a proximal site.

3'UTR Usage During Development and Differentiation

Changes in 3'UTR length during development and differentiation may help regulate gene expression. Alternative pA site usage was significantly increased in *Arabidopsis* seedlings [65, 66]. In zebrafish, an initial shortening of 3'UTRs from 0-6 hours post fertilization (hpf), perhaps due to the maternal zygotic transition, then lengthening of 3’UTRs from 6 hpf to 120 hpf was seen [14], though overall there was a general trend toward lengthening into adulthood [56]. Mouse 3’UTRs also lengthen over the course of development [41, 69], then slightly shorten after transition to adulthood [69]. In contrast, the average 3’UTR length and the number of 3’UTR isoforms per gene decreased throughout *C. elegans* development [12]. Furthermore, increased distal pA site usage was seen in myogenesis [69], and the differentiation of mouse ES cells into neurons [16]. Together, these meta-analyses suggest alternative 3’UTRs may contribute to proper differentiation. Indeed, I discovered that alternative polyadenylation may help regulate endothelial cell specification (see Chapter 2).

Tissue Specific 3’UTR Usage

A genome-wide study of ten human tissues and five breast cancer cell lines showed 80% of alternative 3’UTRs identified were tissue specific, implying tissue specific 3’UTRs may aid in differentiation [34]. For the most part, germ cells exhibit increased proximal pA site usage. In *Dropsophila,* testis displayed the shortest 3’UTRs while ovaries used intermediate length 3’UTRs [15]. However, in zebrafish ovaries expressed the shortest 3’UTRs and testis had the greatest number of tissue-specific pA sites [56]. Increased expression of *Cst1,2,3* in the ovary may be responsible for the increased usage of proximal pA sites [56], as knockdown of CstF resulted in increased processing at distal pA sites [23, 28]. In addition, germ cells undergoing spermatogenesis displayed shortening of 3'UTRs [70]. In comparison, brain contains the longest 3'UTRs in both zebrafish [56] and *Drosophila* [15]. A subset of *Drosophila* genes showed neural-specific lengthening of 3'UTRs [71], perhaps due to neural specific expression of ELAV, an RNA binding protein [38]. Neural-specific lengthening of 3'UTRs was also seen in mouse and human [72]. Importantly, longer neural expressed 3'UTRs are enriched for regulatory microRNA and RNA binding protein sites [15] and also show increased conservation for predicted miRNA target sites in their extensions [72].

Alternative 3'UTR Usage in Disease

Inappropriate expression of alternative 3'UTRs may lead to differential regulation by microRNAs or RNA binding proteins and cause disease. Up-regulation of a long 3'UTR isoform of α-synuclein (α-syn), which is stabilized by miR-34b, may account for the increased aggregation of α-syn in Lewy bodies, which is associated with Parkinson's disease [73, 74]. The distal portion of the Copine III 3'UTR was decreased in both Brodman Area 10 and the caudate in 20 post-mortem schizophrenic brains with normal matched controls, though a specific mechanism of how this is involved with schizophrenia has not been identified [75].

Mutations in the PAS of a gene may alter processing at the original pA site or lead to usage of an alternative PAS. For example, IPEX (immune dysfunction, polyendocrinopathy, enteropathy, X-linked) has been associated with an A to G (AAUAAA to AAUGAA) mutation in the polyadenylation signal of *FOXP3* [76]. Similarly, different PAS mutations in both α–thalessemia and β–thalessemia resulted in transcripts with longer 3'UTRs [77-80]. PAS mutations have also been associated with cancer. Heterozygotes with a single nucleotide polymorphism (SNP) in the PAS of *TP53* (AATAAA → AATACA) had decreased *TP53*, perhaps due to lengthening of the 3'UTR [81]. This SNP was significantly correlated with prostate cancer, colorectal adenoma, and glioma in a genome-wide association study of 457 Icelanders [81]. Deletions or mutations, in the genomic sequence of the cyclinD1 (*CCND1*) 3'UTR, that introduce a canonical PAS were found in the majority of patients with *CCND1* positive mantle cell lymphoma tumors over-expressing the short 3'UTR of *CCND1* [82]. The most highly proliferative tumors from patients with *CCND1* positive mantle cell lymphoma preferentially expressed shorter *CCND1* transcripts, which were found to be more stable. In agreement with this observation, the Kaplan-Meir curves showed the patients with tumors expressing the full length 3'UTR had longer median survival (3.28 years) compared to those expressing the short 3'UTR (1.38 years).

Methods for Identifying Polyadenylation Sites

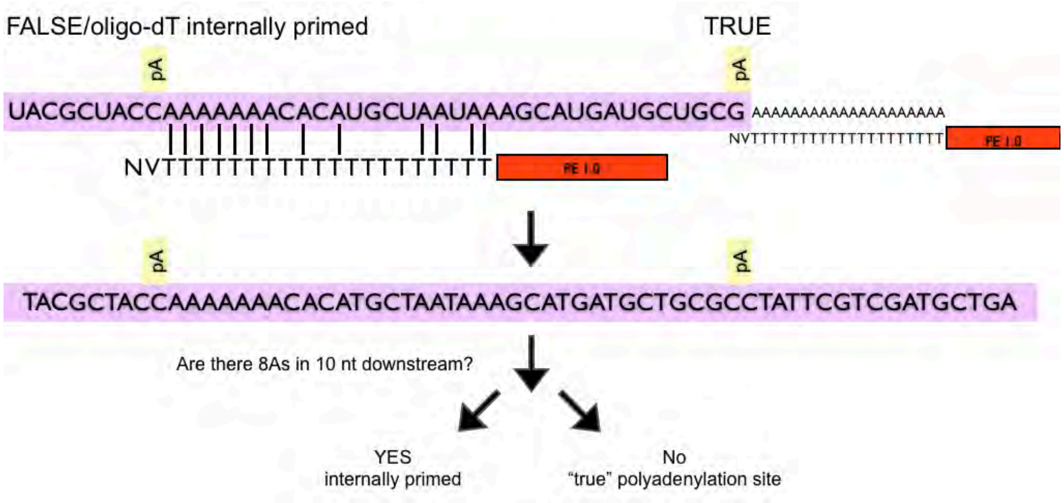
Defining the 3’ ends of transcripts is important for examining 3’UTR regulation in development and disease. The 3’ ends of genomes have generally been poorly annotated thus many tools have been developed to identify pA sites. For example, a recent study of *Arabidopsis* seedlings allowed for re-annotation of 10,215 3’UTRs and the first annotation of 165 3’UTRs [40]. Initial studies used publicly available expressed sequence tags (ESTs) to assemble putative pA sites and compile them into databases [17, 18, 33, 50, 54, 55, 58, 61, 62, 70, 83-90]. However, because the relatively low number of ESTs did not allow for thorough annotation of pA sites, models have been developed to predict putative pA sites [53, 55, 61, 89, 91-98]. More recently, deep sequencing has been used to identify 3’ ends of transcripts genome-wide [12, 15, 16, 19, 20, 35, 40, 41, 45, 49, 52, 56, 57, 65, 67, 99-102].

The majority of these methods rely on an oligo-dT. Oligo-dTs were used to prime reverse transcription from the poly(A) tail for the majority of ESTs (Figure 1.4 TRUE). However, oligo-dTs may also bind to homopolymeric stretches of adenines internal to 3’ ends, referred to as internal priming (Figure 1.4 False/oligo-dT internally primed). To thoroughly investigate the prevalence of internal priming in oligo-dT identified ESTs, Nam et al. performed reverse transcription using an oligo-dT (16) followed by polymerase chain reaction (PCR) for either a control transcript with 20 internal adenines or endogenous transcripts containing eight or more internal adenines [103]. Sequence verification of these PCRs showed internal priming products were present in a ratio of three to one with true 3' ends. Addition of sequence (A/G/CA/CG/CC) to the 3' end of the oligo-dT (referred to as an anchored oligo-dT) to anchor it to the junction of the 3'UTR end and the poly(A) tail decreased the proportion of internal priming products, present only in a one to one ratio with true 3' ends. Only half the number of adenines as the total length of the oligo-dT primer was needed for internal priming. Internal priming events were originally estimated to represent only 2-3% of cDNA libraries, suggesting this may not be of concern [104], though later analyses suggested 12-14% of human ESTs were due to internal priming [83, 103]. A motif of six consecutive adenines was identified ~15% of mouse and human annotated 3'UTR ends (Ensembl v65), suggesting internal priming events still plague genome annotations [72].

Polyadenylation Site Databases

Initial studies delineated pA sites from sequence tags with 3' terminal adenines not templated in the genome. ESTs with 5' proximal thymines or 3' terminal adenines were collected [17, 18, 33, 50, 54, 55, 58, 61, 62, 70, 83-90]. If the adenine tail did not align to the genome, the

Figure 1.4: Oligo-dT mispriming creates false positives*.* Oligo-dT may bind to internal homopolymeric stretches of adenines as will to the poly(A) tail. For the most part, internal priming events are defined by the number of adenines in the genomic sequence downstream of a putative pA site. For example, if there are eight adenines in the 10 nt downstream it will be defined as internally primed.


FALSE/oligo-dT internally primed
TRUE
pA
pA
UACGCUACCAAAAAAACACAUGCUAAUAAAGCAUGAUGCUGCG AAAAAAAAAAAAAAAAAAAA
NVTTTTTTTTTTTTTTTTTTTTT
PE 1.0
NVTTTTTTTTTTTTTTTTTTTTT
PE 1.0
pA
pA
TACGCTACCAAAAAAACACATGCTAATAAAGCATGATGCTGCGCCTATTCGTCGATGCTGA
Are there 8As in 10 nt downstream?
YES
internally primed
No
"true" polyadenylation site

sequences were clustered into putative pA sites. Finally, a simple filter based on the number of adenines downstream of the putative pA site was used to remove internal priming events (Figure 1.4).

Some of these EST-established pA sites were compiled into databases. These are summarized in Table 2. To aid in biological studies investigating the regulatory capacity of a specific 3'UTR, the majority of these databases contained additional features such as the supporting ESTs, tissue specific expression, repetitive elements, adenine-rich elements, and *cis* regulatory elements associated with cleavage and polyadenylation. Some studies used simple filters as described above to remove internal priming events [85, 86, 105]. Alternatively, some databases assigned a confidence level to the pA site rather than removed putative sites of internal priming [87, 90]. Though these databases provided some insight into alternative 3'UTR usage, they were not comprehensive. For example, PolyA_DB2 annotated only ~4500 pA sites [86] for the 27,490 genes in zebrafish [106].

Polyadenylation Site Prediction Programs

As the initial databases and gene annotations were not complete, prediction algorithms were created using various methodologies to aid in the identification of novel pA sites in both human and other model

Table 1.2: Polyadenylation site databases

Database	Organisms	Additional Information	Source
UTRdb	human, rodent, “other mammals”, “other vertebrates”, invertebrates, plants, and fungi	repetitive and functional sequence elements *also contained 5’UTRs	Pesole, Liuni et al. 1999
BodyMap	51 different human organs and tissues	Tissue specificity	Kawamoto, Yoshii et al. 2000
(no name)	Human, mouse	Color-coded putative PASs and associated putative pA sites, repetitive elements, adenine-rich elements individual ESTs (color coded by organ system)	Beaudoing and Gautheret 2001
PolyA_DB	Human, mouse	gene structure from RefSeq, the supporting cDNA or EST evidence tags, the orthologs between human and mouse, the associated PAS, tissue expression pattern	Zhang, Hu et al. 2005
polyA_DB2	Human, mouse, rat, chicken, and zebrafish	Same as polyA_DB *cis* elements identified in [62] conservation analysis	Lee, Yeh et al. 2007
polyA Cleavage Site Database (PACdb)	human, mouse, rat, dog, chicken, zebrafish, fugu, fruitfly, mosquito, worm, rice, and baker’s yeast	confidence level of the pA site based on characteristics of supporting ESTs and surrounding genomic sequence	Brockman, Singh et al. 2005

organisms. These algorithms and their performance are summarized in Table 3. Sequence regions or specific elements surrounding EST-established pA site, such as the PAS, were used to model and predict pA sites. Early attempts to predict human pA sites used discriminant functions, or scoring systems, based on position weight matrices of specific sequence regions [61, 91, 92]. Though these initial programs performed relatively well, they only detected sites with a canonical (AATAAA) or most common variant (ATTAAA) PAS. The same methodology applied to human pA sites with variant or no PASs resulted in diminished performance [95]. Using a machine learning method called support vector machine improved performance predicting human pA sites [93, 94]. Similar approaches were used to predict pA sites in yeast [89], *C. elegans* [96], *Drosophila* [55], *Arabidopsis* [97, 98], and rice [53]. However, with the exception of one support vector machine algorithm [94], they are not integrated into the current genome annotations.

While these methods achieved relatively high specificity and sensitivity, there are major drawbacks. The training sets used to develop these algorithms were not optimal. EST-established pA sites were used as the “True Positives”. However, these may be contaminated with instances of internal priming, and even heuristic filters may not remove all false positives. Additionally, in some cases the “True Negatives” were taken

Table 1.3: Polyadenylation Site Prediction Programs.

Method	**Organism**	**Performance**	**Source**

Linear discriminant function	Human	Sensitivity = 85.5%
		Specificity = 50.7%
		CC = 0.62
			Salamov and Solovyev 1997

Two quadratic discriminant functions	Human	All gene length:
		Sensitivity = 61.5%
		Specificity = 48.5%
		CC = 0.413
		Last two exons:
		Sensitivity = 64.1%
		Specificity = 83.3%
		CC = 0.512
			Polyadq
			Tabaska and Zhang 1999

Position weight matrix	Human	EMBL Annotated pA sites
		Sensitivity = 55.91%
		Specificity = 85.38%
		CC = 0.494
		“weak” pA sites (< 30% of ESTs in a 3’UTR):
		Sensitivity =31.38%
		Specificity = 80.21%
		CC = 0.262
			ERPIN
			Legendre and Gautheret 2003

Support vector machine	Human	Sensitivity = 94.4%
		Specificity = 92.2%
			Liu, Han et al. 2003

Support vector machine	Human	Sensitivity = 84.3%
		Specificity = 84.8%
		CC = 0.693
			Polya_svm
			Cheng, Miura et al. 2006

Linear discriminant function	Human	Canonical PAS
		Sensitivity = 80.8%
		Specificity 66.4%
		Variant PAS
		Sensitivity = 25.2%
		Specificity = 28.4%
		No PAS
		Sensitivity = 13.5%
		Specificity = 14.7%
			POLYAR
			Akhtar, Bukhari et al. 2010

Method	**Organism**	**Performance**	**Source**

Hidden Markov model	Yeast	Sensitivity = 95%
		Specificity = 95%
			Graber, McAllister et al. 2002

Hidden Markov model	*C. elegans*	Sensitivity = 68.3%
		Specificity = 68.3%
			Hajarnavis, Korf et al. 2004

Hidden Markov model	Human
	Drosophila
		Sensitivity = 50%
		Specificity = 77%
			Retelska, Iseli et al. 2006

Generalized hidden Markov model	*Arabidopsis*	coding sequences, randomly generated sequences SN/SP= 97% 5'UTR SN/SP = 82% intron SN/SP = 72%	PolyA Site Sleuth Ji, Zheng et al. 2007
Generalized hidden Markov model	Rice	SN/SP = 90% (score threshold = 4)	PASS-Rice Shen, Ji et al. 2008
Bayesian networks	*Arabidopsis*	CC = 0.65	Ji, Wu et al. 2010

CC = Matthew's Correlation Coefficient

from other genomic regions thought not to contain pA sites, e.g. coding regions and introns, which could be true sites of cleavage and polyadenylation. Additionally the algorithms must be used with these training sites, and are not available for re-training. Although some of these algorithms can be accessed through a web page, users may only input a few sequences for prediction and not entire genomes.

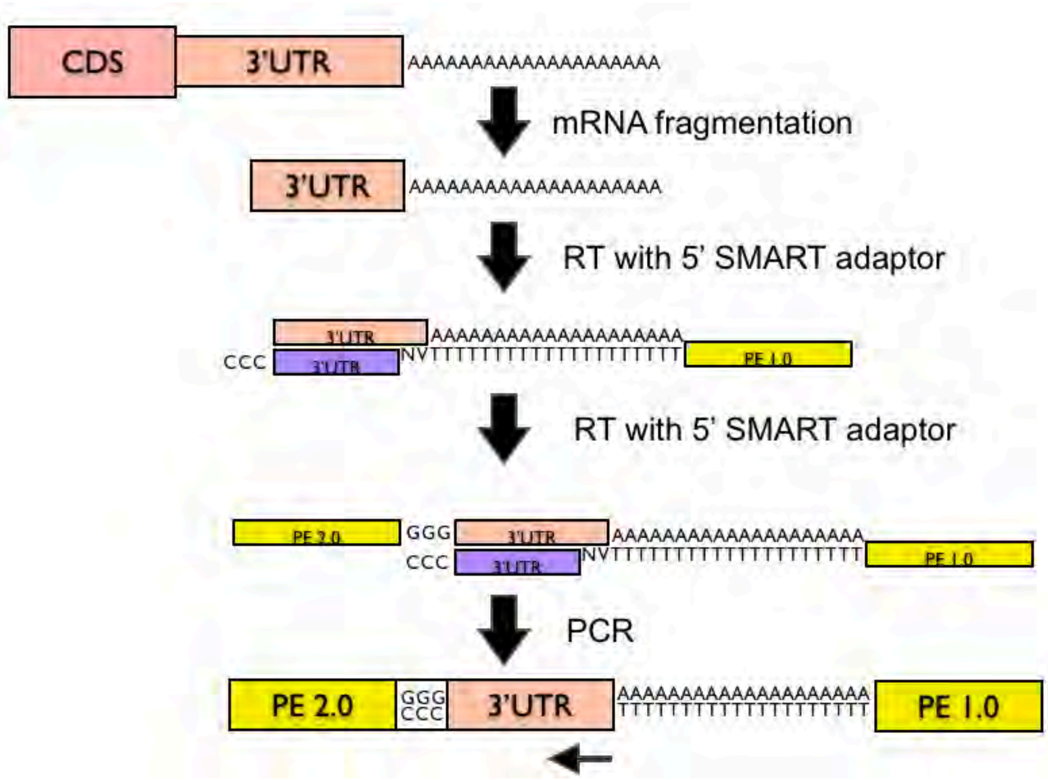
Genome Wide Methods for Identifying Polyadenylation Sites

More recently, genome-wide identification of 3’ ends was accomplished by deep sequencing. However, the majority of methods utilize oligo-dT priming to identify 3’ ends followed by heuristic filtering to remove internal priming events. The 3’UTR is adenine-uracil rich and the majority of heuristic filters use the number of downstream adenines in the genomic sequence downstream of a putative pA site to define internal priming, thus removing true 3’ ends while not removing all instances of internal priming. Both computational and technical methods have been developed to address this problem.

The majority of present-day deep sequencing techniques to identify polyadenylation sites use an anchored oligo-dT to amplify sequence from the poly(A) tail, therefore also identifying sites of internal oligo-dT priming (Figure 1.4). RNA-seq reads with 5’ proximal thymines or 3’ terminal adenines were used to identify pA sites [15, 49] in a similar manner as described above for ESTs. A biotinylated, anchored oligo-dT was used to prime reverse transcription and then cDNA ends were isolated using streptavidin magnetic beads [12, 57, 65, 99]. Others simply primed reverse transcription using an oligo-dT [16, 20, 35, 45, 67, 100]. In some protocols, the RNA was fragmented before reverse transcription, perhaps increasing the likelihood the oligo-dT may bind to adenine-rich sequence that is not a poly(A) tail [16, 20, 35, 67, 100]. Poly(A) Site Sequencing (PAS-Seq) utilized an anchored oligo-dT (20) and the template switching activity of a Maloney reverse transcriptase to build a deep sequencing library composed of 3' ends (Figure 1.5) [16]. PolyA-seq used an anchored oligo-dT (10) for reverse transcription followed by RNase H treatment to degrade the RNA template [45]. Klenow polymerase and a primer of random hexamers was used synthesize the cDNA. These methods are quick and relatively simple, but due to utilization of an oligo-dT there are also false positives due to internal priming.

Another oligo-dT based 3' end sequencing method called 3'READS (3' Region Extraction And Deep Sequencing) was created to decrease the number of internal priming events and remove partially degraded transcripts with short pA tails destined for exosome degradation was developed [41]. A chimeric oligo containing 45 thymines followed by 5 uracils (CU_5T_{45}) in combination with stringent washings enriched the ratio

Figure 1.5. Poly(A) Site Sequencing (PAS-Seq). Briefly, poly(A) RNA was chemically fragmented. Reverse transcription was primed with an anchored oligo-dT (20) containing Illumina adapter sequence (PE1.0). The template switching activity of a Maloney reverse transcriptase was used to incorporate another Illumina adapter (PE2.0) on the 5’ end. The library was PCR amplified. A custom sequencing primer with oligo-dT (20) is used to sequence at the junction of the cleavage site and the poly(A) tail on the Illumina platform. Adapted from [16].

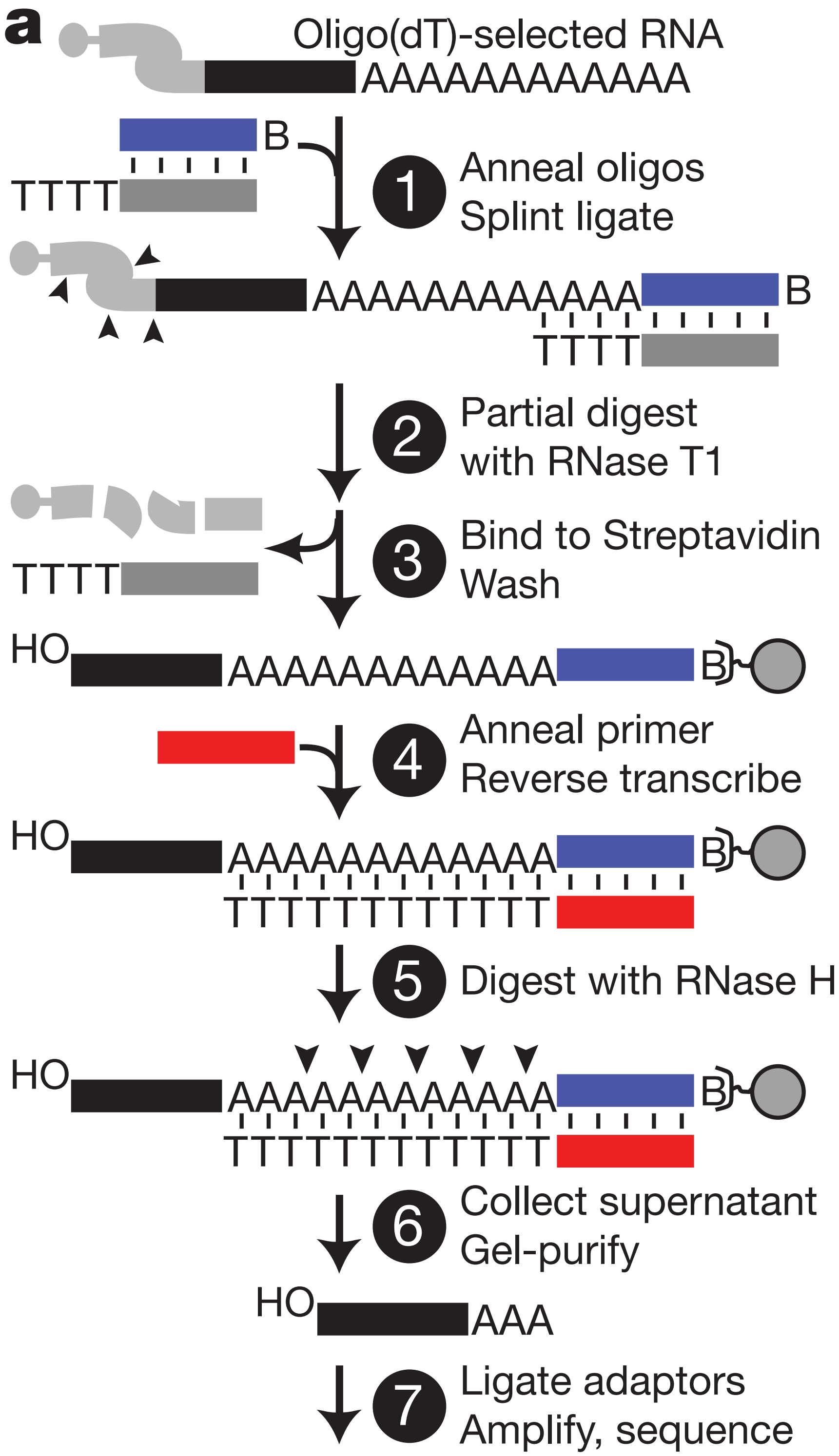

CDS
3'UTR
AAAAAAAAAAAAAAAAAAAA
mRNA fragmentation
3'UTR
AAAAAAAAAAAAAAAAAAAA
RT with 5' SMART adaptor
3'UTR
AAAAAAAAAAAAAAAAAAAA
CCC
3'UTR
NVTTTTTTTTTTTTTTTTTTTT
PE 1.0
RT with 5' SMART adaptor
PE 2.0
GGG
3'UTR
AAAAAAAAAAAAAAAAAAAA
CCC
3'UTR
NVTTTTTTTTTTTTTTTTTTTT
PE 1.0
PCR
PE 2.0
GGG
CCC
3'UTR
AAAAAAAAAAAAAAAAAAAA
TTTTTTTTTTTTTTTTTTTT
PE 1.0

ratio of transcripts isolated with long (60 nt) poly(A) tails to those with short (15 nt) poly(A) tails 12:1. However, the authors concede that about 6% of their mapped reads are due to internal priming.

Direct RNA sequencing (DRS) was developed to identify RNAs without the bias that may occur when converting RNA to cDNA, including internal priming, and to allow for identification of unstable transcripts using small amounts of RNA (e.g. 2 ng of polyA+) [101]. An oligo-dT (50) bound to the flow cell is used to hybridize polyadenylated RNAs. A “T” fill step is performed to fill in the rest of the poly(A) tail, and then a “lock” step is performed with A/C/G. Single nucleotide with fluorophores are incorporated, imaged, and cleaved to allow for single nucleotide resolution.

Poly(A) position profiling (3pseq) is an elegant procedure to only identify transcripts with polyadenylated ends (Figure 1.6) [13]. A biotinylated splint RNA:DNA hybrid oligo containing overhanging single-stranded thymines, which will force hybridization only at the 3’ end of the poly(A) tail, was ligated to the ends of polyadenylated transcripts. RNase T1, which cleaves 3’ to guanines, was used to partially digest the RNA. Subsequently, the 3’ ends were bound to streptavidin beads and washed, resulting in isolation of polyadenylated transcripts. Next, reverse transcription was performed using only thymines to fill in the poly(A) tail.

Figure 1.6. Poly(A) position profiling (3pseq). A biotinylated splint RNA:DNA hybrid oligo containing overhanging single-stranded thymines, which will force hybridization only at the 3’ end of the poly(A) tail, was ligated to the ends of polyadenylated transcripts. RNase T1, which cleaves 3’ to guanines, was used to partially digest the RNA. Subsequently, the 3’ ends were bound to streptavidin beads and washed, resulting in isolation of polyadenylated transcripts. Next, reverse transcription was performed using only thymines to fill in the polyA tail. RNAse H digested the DNA:RNA hybrid, the original poly(A) tail, leaving 2-4 3’ terminal adenines, releasing the 3’ ends from the biotinylated oligo. The supernatant was collected. Sequencing adapters were ligated to the 3’ ends, the library was PCR amplified, and sequenced. Republished from [102].


a
Oligo(dT)-selected RNA
AAAAAAAAAAAA
B
TTTT
1 Anneal oligos
Splint ligate
AAAAAAAAAAAA
B
TTTT
2 Partial digest
with RNase T1
TTTT
3 Bind to Streptavidin
Wash
HO
AAAAAAAAAAAA
B
4 Anneal primer
Reverse transcribe
HO
AAAAAAAAAAAA
B
TTTTTTTTTTTT
5 Digest with RNase H
HO
AAAAAAAAAAAA
B
TTTTTTTTTTTT
6 Collect supernatant
Gel-purify
HO
AAA
7 Ligate adaptors
Amplify, sequence

RNAse H digested the DNA:RNA hybrid, the original poly(A) tail, leaving 2-4 3' terminal adenines. Sequencing adapters were ligated to the 3' ends, the library was PCR amplified, and sequenced. 3pseq identifies only true 3' ends, however it is more technically challenging than the oligo-dT primed methods outlined above, and thus less likely to be selected to identify pA sites (compared with an oligo-dT primed method in Table 1.4).

Following oligo-dT primed 3' end sequencing (see above), internal priming events were removed, defined by a proportion of adenines in the sequence flanking a putative pA site. Comparison of *Arabidopsis* pA sites established by oligo-dT primed 3'end sequencing data or DRS demonstrated that these simple filters may not remove all internal priming events and can exclude true 3'ends, thereby increasing false positives and false negatives respectively [40]. Therefore, more computationally intensive filters were developed to remove sites of internal priming.

Derti et al. scored the 10 bp downstream of the pA site to distinguish true pA sites from oligo-dT primed artifactual pA sites [45]. Using universal human reference RNA, 3' end sequencing library was built with oligo-dT (10). Reads that mapped with at least three terminal adenines, excluding annotated sites or sites with a PAS, indicated internal priming events. Conversely, reads that did not map with at least three

Table 1.4: Comparison of PAS-Seq and 3pseq. PAS-Seq is an oligo-dT primed method for identifying 3’ ends of transcripts. 3pseq ligates a splint RNA:DNA hybrid oligo to the 3’ ends of transcripts. PAS-Seq will identify sites of internal oligo-dT priming in addition to true pA sites. 3pseq will only identify true 3’ ends.

	PAS-Seq [16]	**3pseq [102]**
method	Oligo-dT primed method utilizing template switching	Direct ligation to the 3' end of the transcript
polyadenylation sites identified	True pA site and also false positives due to oligo-dT mispriming	Only polyadenylated 3'UTR ends are sequenced
benefits & drawbacks of sample preparation	quick, easy with relatively little sample	long, complicated construction with larger amounts of sample

terminal adenines were deemed true pA sites. With these distinctions, they created positional discriminant function or a scoring system based on the 10 bp downstream of the cleavage site, similar to [54], to remove oligo-dT primed 3'ends. Though this method achieved relatively high sensitivity (85.6%), other groups have not applied this scoring system to filter 3' end data generated from oligo-dT priming.

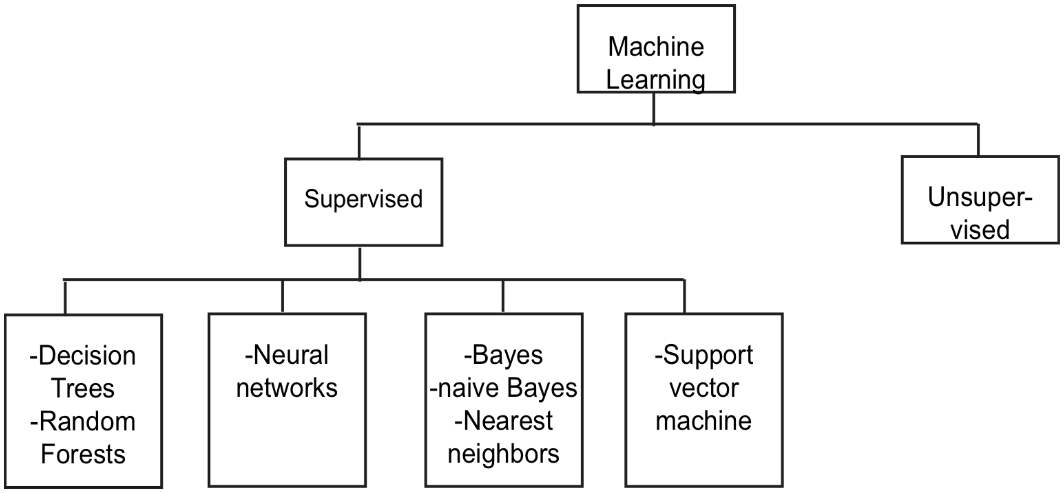
Categorization of putative pA sets before application of multiple heuristic filters was performed to remove sites of internal priming [107]. An oligo-dT primed 3' end sequencing technique was used to assess PAS usage in a human osteosarcoma-derived cell line. Putative pA sites were separated into four groups based on the presence of a canonical PAS (AAUAAA or AUUAAA) upstream of the cleavage site and adenine-richness downstream of the cleavage site. Each category was heuristically filtered by known annotations, presence of alternative PAS, or downstream uracil or guanine/uracil richness. This group reports a success rate of 88% by comparison of DRS data from human liver, which may not be accurate due to the comparison of different tissue types, as transcripts may be processed differently depending on the cell type or developmental stage. Using this method would require replicating the detailed analysis for each organism or cell type, therefore employing their lengthy filtering process is not desirable.

Thus, rapid and easy to construct 3' end sequencing libraries generated from oligo-dT priming would benefit from a user-friendly, but stringent bioinformatic analysis necessary to remove the associated internal oligo-dT priming events. pA site prediction algorithms could hypothetically be re-trained to remove instances of internal priming, however, these algorithms are not publicly available for retraining. So we examined machine learning methods that could be utilized to distinguish between sites of polyadenylation and internal oligo-dT priming.

Machine Learning

The term "machine learning" describes how computational systems can model patterns from data [108]. Machine learning is accurate, not biased by human interpretation, and fast. Within a data set used for learning ("training data"), certain "features" or characteristics are used to represent each sample. Machine learning can be divided into unsupervised and supervised learning methods (Figure 1.7). *Unsupervised learning* identifies patterns from a training set with unknown outcomes, for example by trying to cluster unknowns into groups. When the training data consists of samples labeled with known outcomes a

Figure 1.7: Machine Learning. Machine learning can be divided into supervised and unsupervised learning. Examples of logic-based supervised learning include decision trees and random forests. Neural networks is an example of a perceptron-based technique that classifies input into one of multiple outputs. Statistical based learning methods include Bayesian and naïve Bayesian classifiers, as well as nearest neighbor algorithms. Finally, support vector machine is one of the newest methods of supervised machine learning and uses a multi-dimensional plane to segregate the two different outcomes.


Machine Learning
Supervised
Unsuper-vised
-Decision Trees
-Random Forests
-Neural networks
-Bayes
-naive Bayes
-Nearest neighbors
-Support vector machine

supervised learning approach is used. The goal of the learning algorithms, which may also be called classifiers, is to predict the outcome of unknown instances with the same "feature" categories as the training set.

For example, a classifier could be designed to distinguish between true pA sites and internally primed sites. The training data could consist of sequence features and the outcomes could be labeled as true pA sites or internally primed sites. After "learning", the classifier could then be used to classify new sequences into the same outcomes.

Multiple machine learning methods that could be applied for this purpose (reviewed in [108]). *Simple decision trees* use an individual classification rule at each "branching" to model training data. For example, starting with a set of putative pA sites, the presence or absence of a canonical PAS upstream could be the first rule. Subsequently, additional rules help classify data. However, generally a large "tree" is needed for modeling which can lead to the creation of a model that is too complicated (referred to as overfitting). The *random forests* method combines multiple decision trees and allows for multiple classification rules at each split. Multiple nodes connected together to classify data are called a *neural network*. Neural networks are considered slower than other classification methods. *Support vector machines* create a multi-dimensional plane to segregate the two different outcomes of training data, for example true pA sites and internally primed sites. Additionally, both neural networks and support vector machines usually require large sets of training data.

Bayesian classifications, based on Bayes theorem of conditional probability, use the relative probability of modeling characteristics or features to generate the probability of an outcome [108].[1] Compared to decision trees and support vector machine, Bayesian classifiers have the additional benefit of giving a probability, rather than a discrete classification. However, this method assumes each characteristic is dependent on the others so a large number probabilities to be estimated for classification. For example, the presence of a PAS upstream, the presence of a downstream uracil rich motif and the number of As downstream could be used to model true pA sites and internally primed sites. Using a Bayesian classifier the following probabilities would need to be estimated:

1) the presence of a PAS upstream dependent on the presence of a downstream uracil motif

[1] Appendix I contains a comparison of conditional probability calculated by Bayes theorem and the probability assuming conditional independence (naïve Bayes).

2) the presence of a PAS upstream dependent on the number of As downstream
3) the presence of a downstream uracil motif dependent on the presence of a PAS upstream
4) the presence of a downstream uracil motif dependent on the number of As downstream
5) the number of As downstream dependent on the presence of a PAS upstream
6) the number of As downstream dependent on the presence of a downstream uracil motif

To simplify this problem, a “naïve” assumption can be made, that the features are conditionally independent of each other. In this case, each characteristic alone needs to be estimated, rather than each characteristic in relation to the others. Using a naïve Bayesian classifier now only three probabilities need to be estimated:

1) the presence of a PAS upstream
2) the presence of a downstream uracil motif
3) the number of As downstream

The simplicity of the naïve Bayes classifier results in many benefits. As evident by the simple example above, the number of parameters that need to be estimated is reduced from exponential, represented as O(2^N), to linear, represented as O(N) [109]. Consequently, a relatively small number of samples with known outcomes is needed for training and a large number of features can be used [108]. The training is quick and the required computational power is low. In real world instances, training data is often incomplete, and the information for each classification feature may not be available for all samples [108]. The naïve Bayes classifier will ignore missing information when computing probabilities, thus this is not an issue [108]. Finally, the naïve Bayes classifier performs well, even if the assumption of conditional independence is broken [110]. Tools to utilize naïve Bayes classification are available in several programming languages, including python, matlab, java, and R.

Naïve Bayes classification has been applied successfully to biological and medical problems where the training set might be small and the number of features is high. Ribosomal RNAs may be used to classify bacteria, however hundreds of thousands are maintained by the Ribosomal Database Project II and more are continually generated [111]. A naïve Bayes classifier bacterial ribosomal RNA sequences into different taxonomies with 88.7% accuracy [111]. Often millions of single nucleotide polymorphisms (SNP) are examined to determine if one may correlate with disease. For example, Naïve Bayes classifiers associated the same

SNP from patients with Late Onset Alzheimer's Disease or normal matched controls that had previously been discovered [112]. Finally, a naïve Bayes classifier designed to identify breast lesions from ultrasound images displayed 93.94% sensitivity [113].

Summary

Cleavage and polyadenylation of pre-mRNA transcripts is important for localization, stability, and gene regulation. Combinations of *cis* elements within the transcript direct cleavage and polyadenylation. Furthermore, alternative polyadenylation produces distinct 3'UTRs, which may be important for proper development and have been implicated in disease. Therefore, identification of polyadenylation sites genome-wide will aid studies investigating post-transcriptional regulation. Early databases and prediction models are incomplete. Multiple oligo-dT primed methods for identifying polyadenylation sites by deep sequencing are quick and efficient, but are inaccurate and lead to identification of false positives. The current simple filtering methods, defining internal priming by the number of adenine in the genomic sequence, do not remove all false positives and may create false negatives. More complicated computational analyses may be more accurate, but are not easily adaptable. The most accurate method, 3pseq, identifies only true polyadenylated 3’ ends, but is technically challenging and is not commonly used. Therefore, novel techniques for accurately removing false positives in combination with oligo-dT primed 3’ end sequencing will improve genome-wide identification of polyadenylation sites. Naïve Bayes classifiers require relatively small training data sets and require little computational power. Thus I developed a naïve Bayes classifier to accurately delineate true pA sites from internally primed sites, which is discussed in the body of this work.

CHAPTER II

POST-TRANSCRIPTIONAL MECHANISMS CONTRIBUTE TO DOWNREGULATION OF ETV2 DURING VASCULAR DEVELOPMENT

This manuscript was submitted to Developmental Biology: Moore, J.C., **Sheppard, S**., Shestopalov, I., Chen, J., and Lawson, N.

Contributions

Etv2 protein fragment cloning and isolation for antibody production.
Dr. Nathan Lawson, University of Massachusetts Medical School

Etv2 caged morpholino.
Dr. Ilya Shestopalov; Chen Lab, Stanford University

Etv2 expression analysis, Etv2 caged MO experiment and circulation phenotyping, Let7a overexpression analysis, Lin-28 overexpression analysis.
John Moore, Lawson Lab, University of Massachusetts Medical School

Overexpressed Etv2 mosaic transplant analysis. Etv2 3'UTR identification, cloning, 3'RACE, RT-PCR, *Let-7* binding site analysis and *let-7* binding site mutagenesis. Endothelial cell autonomous 3'UTR analysis, cloning, injection, imaging, quantification and statistical analysis. Overexpressed let-7 duplex mosaic transplant analysis. Maternal zygotic *dicer1* production.
Sarah Sheppard, Lawson Lab, University of Massachusetts Medical School

Etv2 3'UTR mRNA sensor experiments with *let-7* duplex and Western blotting.
Sarah Sheppard performed the preliminary experiments, but Figure A5.6 and A5.7 are the work of John Moore, Lawson Lab, University of Massachusetts Medical School

The manuscript was a collaborative effort between Nathan Lawson, John Moore, and Sarah Sheppard.

Introduction

The vertebrate circulatory system serves as an essential conduit for the systemic distribution of oxygenated blood, nutrients, hormones, immunological factors and the removal of metabolic waste. The formation of a patent and functional circulatory system begins before the onset of blood circulation with the specification of endothelial progenitors, or angioblasts, from the lateral mesoderm. As angioblasts differentiate and express an endothelial gene program, they migrate and coalesce to form vascular cords through a process called vasculogenesis [114]. This initial vascular plexus is subsequently remodeled and extended into a system of patent blood vessels through a process referred to as angiogenesis. While the morphological events that define vasculogenesis and angiogenesis are relatively well-defined, the transcriptional regulatory networks that control angioblast specification and subsequent endothelial differentiation are poorly understood.

Multiple transcription factor families have been implicated in the activation and maintenance of endothelial gene expression, including members of the Sox, Forkhead, GATA, and Kruppel-like families [115]. Among the most prevalent transcription factors involved in endothelial biology are members of the ETS family. ETS transcription factors are defined by the presence of a conserved, approximately 85 amino acid

DNA binding domain, referred to as the ETS domain, which consists of a winged helix-turn-helix motif that binds a core DNA sequence of 5’-GGA(A/T)-3’ [116]. There are approximately 19 and 12 ETS factors expressed in human and zebrafish endothelial cells, respectively, including ETS1, ETS2, ETV2 (etsrp/ER71), ETV6 (TEL), FLI1, ERG and ELK3 (NET/SAP2) [115, 117, 118]. Most characterized endothelial gene promoters or enhancers contain essential ETS binding sites [115, 118-120] and it has been proposed that nearly every endothelial gene may be regulated by ETS factors in some manner [118]. Indeed, the founding member of the ETS family, Ets1, which is highly expressed in endothelial cells in multiple species [121-123], is capable of directly binding to elements flanking genes encoding receptors important for vascular morphogenesis, including Vegf receptor-1 and -2 (Flt1 and Kdr, respectively) [124, 125], Tie-2 [126], and neuropilin-1 [127]. Targeted deletion or knockdown of individual ETS factors can cause specific developmental defects in embryonic vascular morphogenesis or function, although in many cases knockout mice are viable or display only mild phenotypes. For example, mouse embryos lacking *fli1* alone die at E12.5 due to poor blood vessel integrity and cranial hemorrhage [128]. By contrast, Ets1-deficient mice are viable with no overt vascular defects [129] and only mild defects have been noted following knockdown of *ets1* in zebrafish [122]. The highly conserved DNA binding domain shared between ETS factors and their overlapping expression in endothelial cells likely contributes to some degree of functional redundancy that reduces the severity of vascular defects in these cases.

Indeed, ETS factors share significant consensus DNA-binding specificity [130] and can bind to and transactivate the same consensus sequences in some promoters [119, 131]. Analysis of double knockout mice further supports at least partially overlapping functions among some ETS factors. For example, mouse embryos lacking either Ets1 or Ets2 alone display relatively normal vascular development. However, combined loss of both Ets1 and 2 leads to embryonic lethality between E11.5 and E15.5 due in part to defects in vessel remodeling and diminished angiogenic branching [132]. Similarly, combined reduction of related ETS factors in zebrafish results in a higher penetrance of defects and a block in angiogenesis [122].

In contrast to the mild vascular phenotypes associated with loss of most ETS factors, mouse or zebrafish embryos lacking Ets-variant protein 2 (Etv2; also known as Ets-related protein/Etsrp and ER71) show profound defects at the earliest stages of vascular development. Etv2-deficient mouse embryos fail to specify hematopoietic and endothelial cell lineages leading to embryonic lethality at E9.5 due to a failure to develop a functional circulatory system [133, 134]. Zebrafish *etv2* mutants and morphants display similar defects in vascular development and exhibit severe reduction in the expression of most endothelial genes, including, *kdrl*, *flt4*, *cdh5*, and *plxnd1*. In addition, *etv2*-deficient zebrafish embryos exhibit defects in the morphogenesis of the major trunk blood vessels [122, 135]. The severe early vascular defects and global effects on endothelial gene expression in both mouse and zebrafish embryos suggests that *etv2* plays an early role in specifying endothelial cell lineages. Consistent with this possibility, overexpression Etv2 in both zebrafish embryos and mouse embryoid bodies can expand endothelial cell lineages and induce concomitant expression of hundreds of vascular genes [135-138]. Furthermore, recent studies demonstrate that Etv2 is an essential component, along with Fli1 and Erg, during direct endothelial reprogramming of human amniotic cells [139]. Together, these studies suggest a central role for Etv2 in the initial specification of endothelial cells during the initial stages of vascular development.

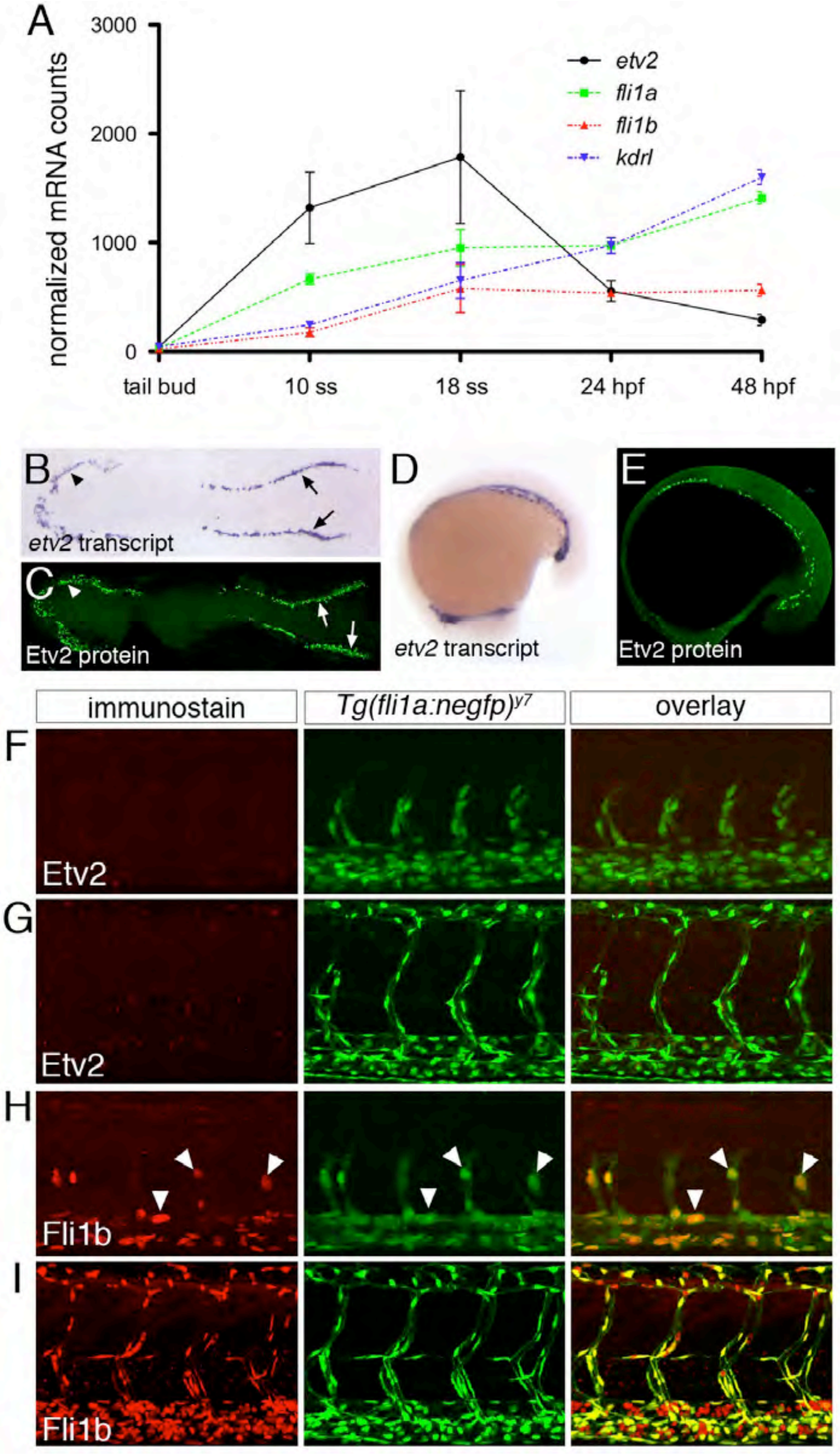
Despite the importance of Etv2 during early vascular development, its role during later stages is unclear. Evidence suggests that Etv2 may only be expressed in endothelial progenitors early during mouse development (E9.5), while expression in the zebrafish is evident in angioblasts but appears to be down-regulated by 36 hpf in endothelial cells of the axial vasculature [133-135]. Interestingly, mouse embryos are viable following conditional endothelial ablation of *etv2* using a Kdr:Cre driver [140], suggesting that its function is restricted to very early stages of vascular development prior to the onset of *kdr* expression. Although these studies suggest dynamic control and function of *etv2* expression during embryogenesis, carefully quantified and staged studies in this regard are still lacking. Furthermore, the mechanisms that exist to downregulate Etv2 during development have not been investigated.

In this work, we assessed the expression levels of *etv2* transcript and protein during early zebrafish vascular development. Both mRNA quantification and whole mount immunostaining revealed that Etv2 is expressed during early and mid-somitogenesis and subsequently downregulated as endothelial cells differentiate and form the major trunk blood vessels. Conditional knockdown of Etv2 using a caged morpholino demonstrated that it is required only during endothelial cell specification and appears to be largely dispensable for subsequent vascular morphogenesis and function. We further find that the *etv2* 3’ UTR is subjected to negative regulation in endothelial cells and that this effect can be mediated by members of the *let-7* microRNA family. Finally, we observe that Etv2 protein levels persist in endothelial cells of embryos lacking maternal and zygotic *dicer1*, which is required for microRNA maturation. Together, our results demonstrate that *etv2* is required during a defined developmental window for angioblast specification and is actively downregulated, in part, through microRNA-mediated post-transcriptional regulation.

Results

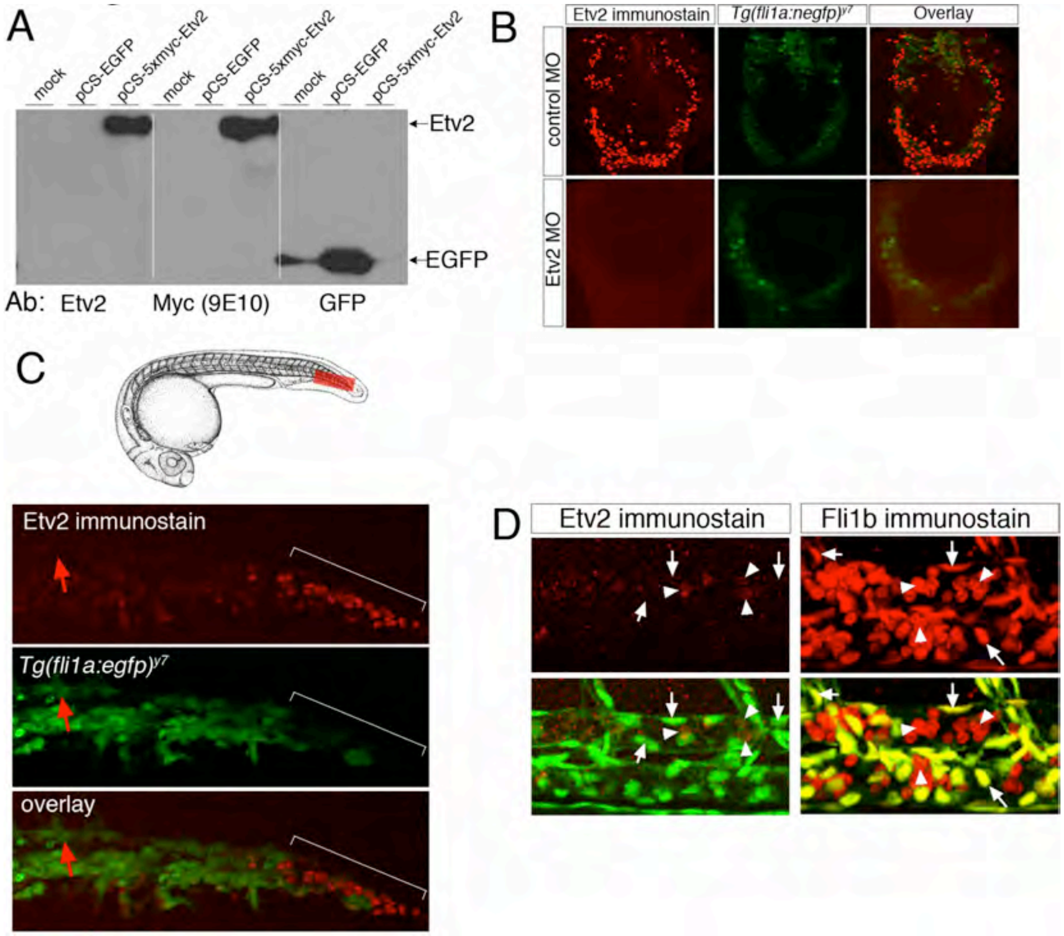
Based on previous studies that suggested *etv2* levels might be dynamically regulated during embryogenesis, we carefully investigated its expression during zebrafish vascular development. We first applied the NanoString nCounter gene expression assay to quantitatively measure *etv2* transcript levels at different stages of development. Using this approach, we observed that *etv2* transcript increases between tail bud and 10 somite stage (ss) and peaks at 18 ss, at which time it is expressed nearly 2 fold greater than endothelial transcripts encoding Fli1a and Fli1b, and the zebrafish Vegf receptor-2 ortholog, Kdrl (Figure 2.1A). Subsequently, *etv2* transcript decreases between 18 ss and 48 hours post fertilization (hpf), when it is expressed at levels five fold below that of *kdrl* (Figure 2.1A). By contrast, *fli1a*, *fli1b*, and *kdrl* transcripts continued to modestly increase from 10 ss until 48 hpf (Figure 2.1A). Thus, the *etv2* transcript displays an initial burst of expression during the time in which endothelial specification and vasculogenesis are taking place and is

Figure 2.1: Etv2 is down-regulated during vascular development. A. Graph of nCounter quantification for *etv2*, *fli1a*, *fli1b*, and *kdrl* at the indicated developmental stages. Values are normalized to *actb2* (*beta-actin*) and *eef1a1l1*(*ef1alpha*). B, D. Whole mount *in situ* hybridization using an antisense *etv2* riboprobe at 5ss and 18ss. C, E. Embryos at 5ss and 18ss immunostained with Etv2 antibody and anti-rabbit Alexa-488. B, C. Dorsal views of flat-mounted embryos, anterior to the left. D, E. Lateral views, anterior to the left. F-I. Two-photon micrographs of trunk vessels in fixed *Tg(fli1a:negfp)*$_{y7}$ embryos immunostained with antibodies against F, G Etv2 or (H, I) Fli1b. Left panels, immunostained protein detected with Alexa-568 secondary antibody. Middle panels, transgenic expression of nuclear localized EGFP. Right panels, overlay of Alexa-568 and EGFP signals. Embryos at F, H 25 hpf or G, I 48 hpf.


A
3000
2000
1000
0
normalized mRNA counts
etv2
fli1a
fli1b
kdrl
tail bud
10 ss
18 ss
24 hpf
48 hpf
B
etv2 transcript
C
Etv2 protein
D
etv2 transcript
E
Etv2 protein
immunostain
Tg(fli1a:negfp)y7
overlay
F
Etv2
G
Etv2
H
Fli1b
I
Fli1b

subsequently downregulated. We next raised an antibody that specifically recognized the divergent N-terminal domain of Etv2 (Figure 2.2A,B) and used this to perform whole mount immunostaining on zebrafish embryos. Similar to *etv2* transcript, we observed Etv2 protein in the anterior and posterior lateral mesoderm within nuclei of presumptive endothelial progenitors at 5 ss (Figure 2.1B,C) and during initial formation of the trunk blood vessels at 18 ss (Figure 2.1D,E). However, we did not observe vascular expression of Etv2 protein at 24 hpf or 48 hpf, while an endothelial-expressed nuclear localized EGFP (*Tg(fli1a:negfp)*^{y7}) was easily detectable at both stages in the same embryos (Figure 2.1F,G). By contrast, we observed robust expression of Fli1b protein in endothelial nuclei of *Tg(fli1a:negfp)*^{y7} embryos at the same time points (Figure 2.1H,I). Interestingly, *etv2* and *fli1b* transcript are expressed at similar levels at 24 hpf (Figure 2.1A). Despite its down-regulation in endothelial cells, Etv2 protein was still detected at 24 hpf in a subset of cells posterior to the caudal vein plexus, which may comprise hematopoietic precursors [141] (Figure 2.2C). We also detected a small population of weakly stained Etv2-positive cells in circulation at 48 hpf, while Fli1b was expressed in the majority of blood cells at this time point (Figure 2.2D). Taken together, these observations demonstrate that *etv2* transcript and protein are expressed in angioblasts during vasculogenesis, but are subsequently

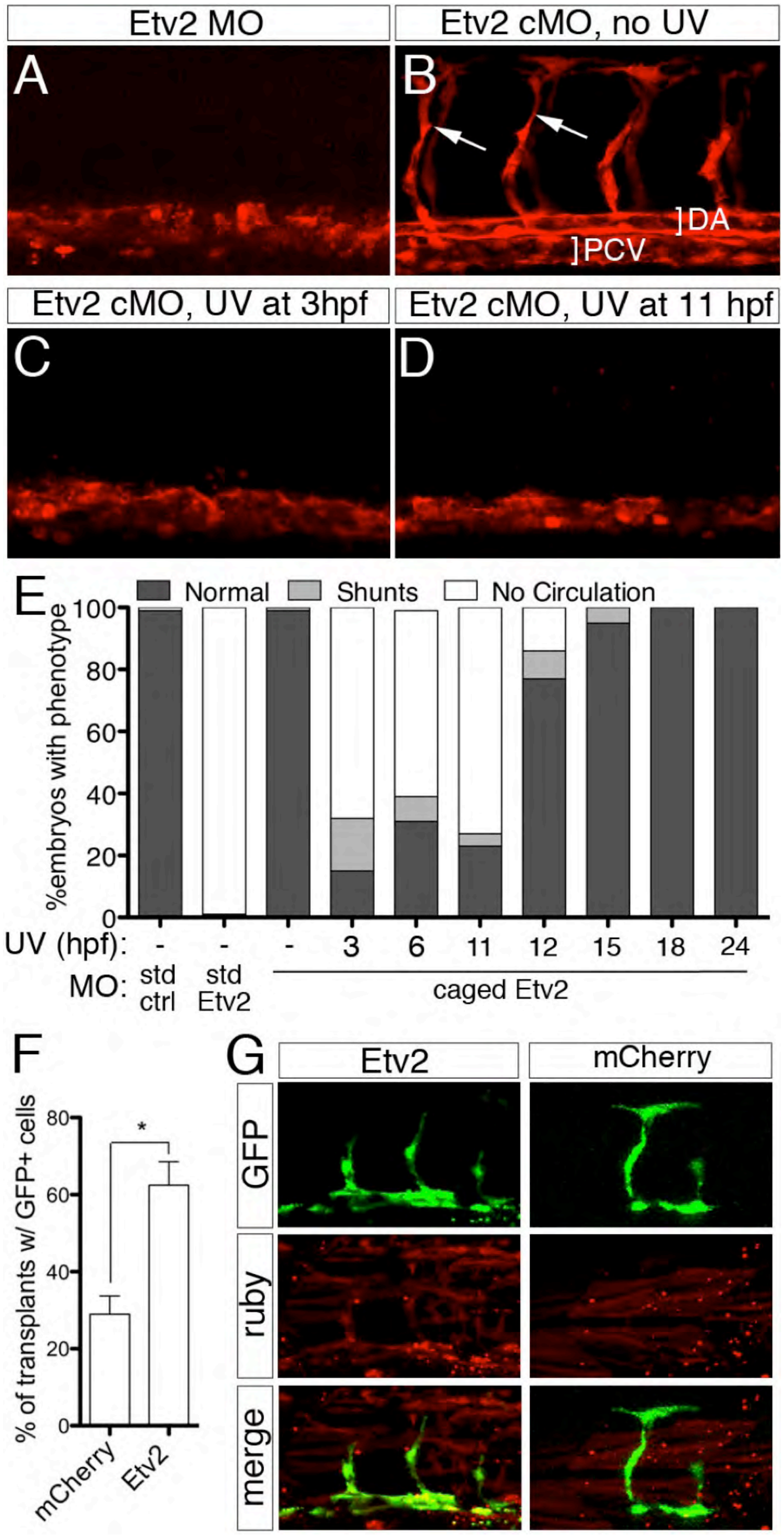
Figure 2.2: A polyclonal antibody against zebrafish Etv2. A. SDS-PAGE gel of HEK293T lysates transfected with mammalian expression vectors for EGFP (pCS- EGFP), myc-tagged zebrafish Etv2 (pCS-5xmycEtv2), or left untransfected (mock). Lysates from each sample were run on triplicate immunoblots, which were individually probed with Etv2 polyclonal antiserum, a monoclonal against the myc-epitope (9E10), or a polyclonal against GFP. The Etv2 polyclonal serum recognizes a single band that is the same size as that recognized following immunodetection for the myc epitope. B. *Tg(fli1a:negfp)y7* embryos at 18 hpf injected with 5 ng of control or Etv2 MO followed by immunostaining using Etv2 polyclonal serum and Alexa-568 secondary. Etv2 antibody staining is clearly visible in embryos injected with control MO, but absent in embryos injected with 5 ng of an Etv2 translation blocking morpholino. C. Top, camera lucida drawings of embryo at approximately 24 hpf. Bottom, immunostaining of an *Tg(fli1a:egfp)y1* embryo with Etv2 polyclonal serum and alexa-568 secondary at 24 hpf. Faint Etv2 expression can be observed in many EGFP-positive cells within the caudal vein plexus, while strong Etv2 expression is apparent in a separate EGFP-negative population of cells (indicated by a white bracket). Etv2 expression is not detectable in the dorsal aorta at this time point (red arrows). D. Two photon micrographs of *Tg(fli1a:negfp)y7* embryos immunostained with Etv2 (left panels) or Fli1b (right panels) polyclonal serum. Top panels are signal from Alexa-568 secondary antibody. Bottom panels are overlay of Alexa-568 and EGFP fluorescence. Images are higher magnification views of embryos shown in Figure 1G and I. *Left*, arrows indicate EGFP-positive endothelial nuclei that do not express Etv2; arrowheads indicate EGFP- negative/Etv2-positive cells within the dorsal aorta. *Right*, arrows indicate EGFP- positive endothelial nuclei that also express Fli1b; arrowheads denote EGFP- negative/Fli1b-positive blood cells circulating within the dorsal aorta.


A
mock
pCS-EGFP
pCS-5xmyc-Etv2
mock
pCS-EGFP
pCS-5xmyc-Etv2
mock
pCS-EGFP
pCS-5xmyc-Etv2
←Etv2
←EGFP
Ab: Etv2
Myc (9E10)
GFP
B
Etv2 immunostain
Tg(fli1a:negfp)y7
Overlay
control MO
Etv2 MO
C
Etv2 immunostain
Tg(fli1a:egfp)y7
overlay
D
Etv2 immunostain
Fli1b immunostain

downregulated in endothelial cells as vascular development proceeds.

The dynamic expression of *etv2* suggested that its function might only be required during early stages of vascular development. To investigate this possibility, we utilized a caged Morpholino (cMO) that is activated by exposure to UV light to conditionally block Etv2 translation at different developmental stages [142, 143]. We injected Etv2 cMO into 1-cell stage *Tg(fli1a.ep:DsRedex)^{um13}* zebrafish embryos, exposed them to UV light at distinct developmental stages, and subsequently assessed vascular morphology and function. As has been shown previously, embryos injected with a standard Morpholino targeting Etv2 exhibited loss of intersegmental vessels (ISV) and a poorly formed dorsal aorta (DA) at 30 hpf and did not display circulation at 48 hpf (Figure 2.3A,E). By contrast, *Tg(fli1a.ep:DsRedex)^{um13}* embryos injected with Etv2 cMO that were not exposed to UV light, or those that were uninjected and exposed to UV, were phenotypically normal (Figure 2.3B, E). Likewise, *Tg(fli1a.ep:DsRedex)^{um13}* embryos injected with scrambled control morpholino (MO) exhibited normal vascular morphology at 30 hpf and normal circulation at 48 hpf (Figure 2.3 and data not shown). However, most embryos injected with Etv2 cMO and exposed to UV light at 11 hpf or earlier exhibited defects in vascular morphology and loss of circulation (Figure 2.3C-E), similar to embryos injected with an uncaged Etv2 MO

Figure 2.3: Etv2 is required only during early stages of vascular development. A-D. Confocal images of trunk blood vessels in *Tg(fli1a.ep:DsRedex)*^{um13} embryos at 30 hpf. Lateral views, dorsal is up, anterior to the left. Embryos injected with (A) 5 ng standard Etv2 Morpholino (MO) or (B) 2 ng Etv2 caged MO (cMO), but not illuminated with UV light. ISVs (arrows), dorsal aorta (DA; bracket) and posterior cardinal vein (PCV; bracket) are indicated. C, D. Embryos injected with Etv2 cMO exposed to UV light at C 3 hpf or D 11 hpf. E. Penetrance of indicated circulatory defects in embryos at 48 hpf following injection with MO and UV exposure as indicated. F. Percentage of mosaic miniRuby-positive host embryos showing successful transplantation of *Tg(fli1:EGFP)*^{y1} donor cells. Donor embryos were injected with 100 pg of *mcherry* or *etv2* mRNA. *p < 0.05. G. Representative confocal images of wild type hosts with contribution to both vascular (green) and non-vascular (red) tissue.

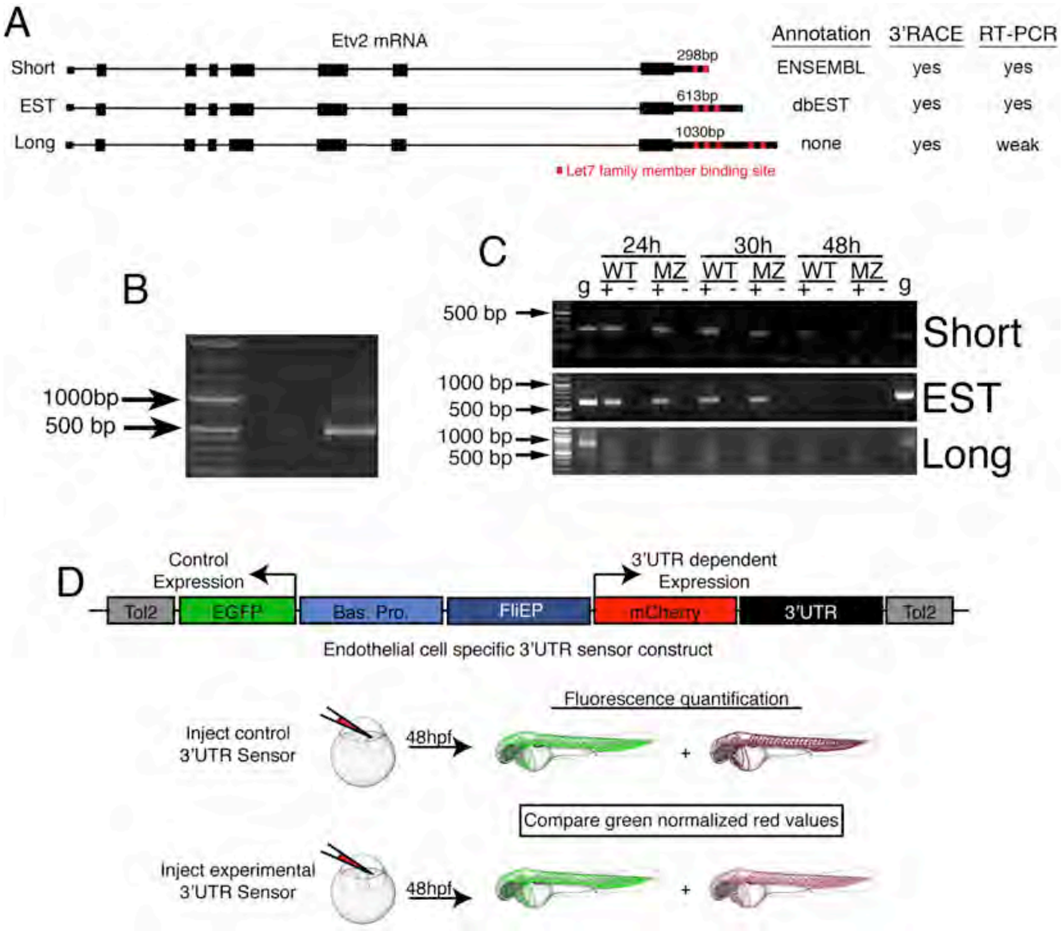

Etv2 MO
Etv2 cMO, no UV
A
B
]DA
]PCV
Etv2 cMO, UV at 3hpf
Etv2 cMO, UV at 11 hpf
C
D
E
Normal
Shunts
No Circulation
%embryos with phenotype
0
20
40
60
80
100
UV (hpf):
-
-
-
3
6
11
12
15
18
24
MO:
std ctrl
std Etv2
caged Etv2
F
% of transplants w/ GFP+ cells
0
20
40
60
80
*
mCherry
Etv2
G
Etv2
mCherry
GFP
ruby
merge

(Figure 2.3A,E). In all cases, we did not observe any overt effects on general morphology (data not shown). Many fewer Etv2 cMO-injected embryos exposed to UV light at 12 hpf displayed defects in circulation and UV illumination at later time points did not cause severe defects in trunk blood vessels or loss of circulation (Figure 2.3E). Thus, Etv2 appears to be required in a precisely defined early window during vascular development. Such an early requirement would be consistent with a role for Etv2 during specification of lateral mesodermal precursors to an endothelial cell fate. If this were the case we would expect exogenous Etv2 to increase the potential of cells to contribute to the vascular lineage. Indeed, mosaic analysis revealed that donor cells derived from *Tg(fli1a:egfp)*^{y1} embryos injected with *etv2* mRNA were much more likely to contribute to trunk blood vessels than donor embryos injected with mRNA encoding mCherry (Figure 2.3F, G). Together, these observations demonstrate that Etv2 plays an essential role during endothelial cell specification but is likely dispensable for later aspects of vascular development.

Recent studies in mouse demonstrate that persistent transgenic expression of Etv2 leads to defects in vascular morphology and causes a block in endothelial maturation [144]. Coupled with our observations that Etv2 is robustly downregulated at early stages of vascular development, these results suggest the existence of mechanisms that actively reduce Etv2 expression. Furthermore, the discrepancy between Etv2 and Fli1b protein expression compared to their relative levels of transcript at 24 hpf (compare Figure 2.1A,F,H) suggested that a post-transcriptional mechanism might contribute to reduction in Etv2 protein. To investigate this possibility, we tested the ability of the *etv2* 3'UTR to repress reporter gene expression. In the process of cloning the appropriate regulatory sequences for this assay, we observed evidence of alternative 3'UTRs encoded by the *etv2* locus[2]. In ENSEMBL (version 69, Zv9), the annotated *etv2* 3'UTR spans only 298 nucleotides, while two separate expressed sequence tags (ESTs) identified in the NCBI database extend past this sequence by an additional 315 nucleotides (Figure 2.4A). To further characterize expressed *etv2* 3' UTR sequences, we performed 3' rapid amplification of cDNA ends (3'RACE) from 24 hpf zebrafish embryos. Sequence analysis of cloned 3'RACE products and subsequent RT-PCR confirmed the expression of both the ENSEMBL- and EST-annotated 3'UTRs (hereafter referred to as Short and EST *etv2* 3' UTR, respectively) as well as a third isoform encoding a 3'UTR of approximately 1030 nucleotides (Long *etv2* 3' UTR; Figure 2.4A-C). To

[2] I discovered alternative *etv2* 3'UTRs by polyadenylation site sequencing, though not noted in this manuscript. See Appendix V.

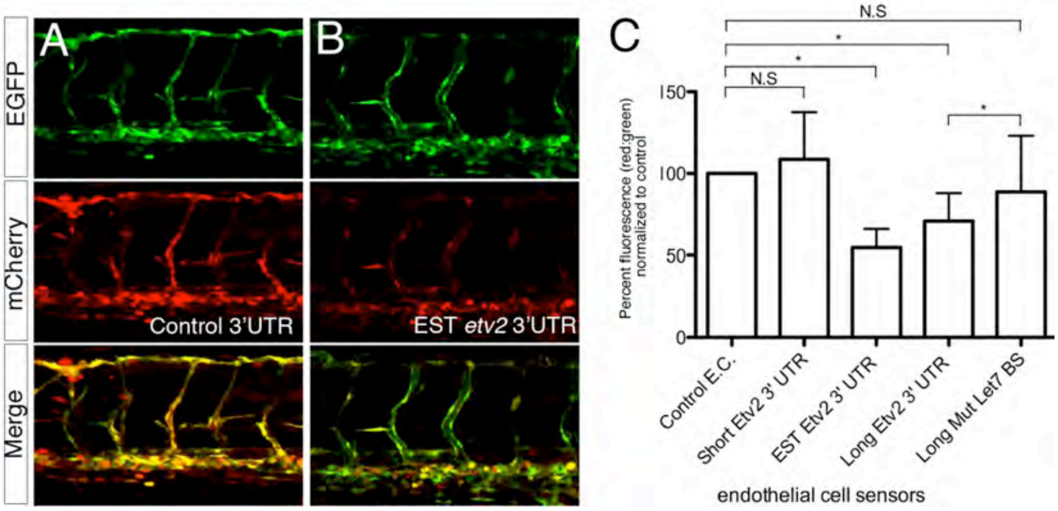
Figure 2.4: Evidence for alternative 3’UTRs encoded by the zebrafish etv2 locus. A. Schematic depicting etv2 intron/exon structure and alternative 3’UTR lengths. Evidence for the existence of each isoform derived from annotation, 3’RACE, and RT-PCR is indicated. B. 3’RACE products amplified from 24 hpf embryos. C. RT- PCR from 24, 30, or 48 hpf wild type or MZDicer embryos was performed using primers specific to the short, EST, or long etv2 3’UTR. Genomic (g) DNA was used as a positive control. + denotes reverse transcribed cDNA template; - indicates template without reverse transcription to rule out genomic DNA contamination. D. Diagram of endothelial cell autonomous 3’ UTR sensor construct and experimental procedure for measuring post-transcriptional regulation of 3’UTRs.


A
Etv2 mRNA
Short
EST
Long
298bp
613bp
1030bp
Let7 family member binding site
Annotation
3'RACE
RT-PCR
ENSEMBL
yes
yes
dbEST
yes
yes
none
yes
weak
B
1000bp
500 bp
C
24h
30h
48h
WT
MZ
WT
MZ
WT
MZ
g
+
-
+
-
+
-
+
-
+
-
+
-
g
500 bp
Short
1000 bp
500 bp
EST
1000 bp
500 bp
Long
D
Control
Expression
3'UTR dependent
Expression
Tol2
EGFP
Bas. Pro.
FliEP
mCherry
3'UTR
Tol2
Endothelial cell specific 3'UTR sensor construct
Fluorescence quantification
Inject control
3'UTR Sensor
48hpf
+
Compare green normalized red values
Inject experimental
3'UTR Sensor
48hpf
+

determine the regulatory potential of these 3'UTRs, we employed an endothelial cell autonomous transient transgenic reporter assay in which a 3'UTR of interest is placed downstream of a red fluorescent protein (mCherry; Figure 2.4D) [145]. This reporter construct also contains a separately expressed enhanced green fluorescent protein (EGFP) reporter fused to a control 3'UTR as an internal reference (Figure 2.4D). We cloned each *etv2* 3'UTR downstream of mCherry and assessed their effect on reporter expression in endothelial cells *in vivo* compared to the internal EGFP cassette. In embryos injected with a transgenic construct encoding mCherry fused with a control 3'UTR, we observed robust co-expression of both mCherry and EGFP in endothelial cells within trunk blood vessels (Figure 2.5A). By contrast, both the EST and Long *etv2* 3'UTRs caused significant reduction of mCherry expression when compared to the co-expressed EGFP control (Figure 2.5B, C), while the Short *etv2* 3'UTR did not appear to alter expression reporter expression (Figure 2.5C and data not shown). These results suggest that post-transcriptional regulation of alternative *etv2* 3'UTRs may contribute to its regulation during vascular development.

microRNAs (miRNA) are short non-coding RNAs that can repress gene expression through interaction with target sequences usually located in the 3'UTR of transcripts [146]. Since the *etv2* EST and Long 3'UTRs

Figure 2.5. The etv2 3’UTR represses a heterologous reporter in endothelial cells. A, B. Representative confocal micrographs of 48 hpf wild type embryos co-injected with 25 pg of a Tol2 bis-cistronic endothelial cell autonomous sensor construct encoding mCherry fused to a A control 3’UTR or the B EST etv2 3’UTR sensor and 25 pg of transposase mRNA. Top, endothelial expression of the control EGFP transgene. Middle, endothelial expression of the mCherry sensor transgene. Bottom, merge of green and red channels. Lateral views, dorsal is up, anterior to the left. C. Quantification of relative mCherry fluorescence levels compared to EGFP following 3’ UTR sensor injection. *p< 0.05, N. S. = Not significant.


A
B
EGFP
mCherry
Merge
Control 3'UTR
EST *etv2* 3'UTR
C
N.S
*
*
N.S
*
150
100
50
0
Percent fluorescence (red:green) normalized to control
Control E.C.
Short Etv2 3' UTR
EST Etv2 3' UTR
Long Etv2 3' UTR
Long Mut Let7 BS
endothelial cell sensors

caused repression of reporter gene expression, we investigated whether this may be due to miRNA regulation. Analysis of the *etv2* 3'UTR sequence revealed 5 putative binding sites for members of the *let-7* family of miRNAs in the longest defined *etv2* 3'UTR, with the short and EST *etv2* 3'UTRs having two and three binding sites, respectively (Figure 2.4A). Additionally, *let-7* binding sites were also evident in the mouse and human *etv2* transcripts (data not shown), suggesting a conserved role for *let-7* in regulating Etv2 levels. To test the possibility that *let-7* regulates the *etv2* 3'UTR we co-injected zebrafish embryos with mRNA encoding EGFP fused to the EST or short 3'UTR with either *let-7a* or a mis-match (mm) control duplex RNA, as well as *mCherry* mRNA with a control 3'UTR and subsequently compared levels of green and red fluorescence. Coinjection of the *etv2* EST 3'UTR sensor with *let-7a* duplex decreased EGFP expression compared to control duplex without an effect on mCherry levels (Figure 2.6A, B). Quantification of EGFP and mCherry in the same embryos by Western analysis demonstrated that *let-7a* led to significant and potent repression of the *etv2* EST 3'UTR (Figure 2.6E, F). We also observed modest but significant repression of the Short *etv2* 3' UTR by *let-7a* (Figure 2.6C-F). Consistent with the highly conserved sequence within *let-7* microRNAs, we observed similar repression of the EST etv2 3'UTR by other *let-7* family members (Figure 2.7), which have been

Figure 2.6: let-7a negatively regulates the etv2 3'UTR. A-D. Transmitted light (left column), green fluorescence (middle column) and red fluorescence (right column) images of embryos injected with sensor mRNAs. A, B. Embryos co-injected with 25 pg gfp-est- etv2-3' UTR and 25 pg mcherry mRNAs and A control or B let-7a duplex. C, D. Embryos co-injected with 25 pg gfp-short-etv2-3' UTR and 25 pg mcherry mRNAs and C control or D let-7a duplex. E. Western analysis for GFP and mCherry protein on embryo lysates at 24 hpf following injection with est- or short-etv2 3' UTR sensor mRNA, mcherry mRNA, and indicated duplex. F. Quantification of Western analysis from three independent experiments. Bars represent the average ratio of GFP band intensities from embryos injected with control duplex compared to let-7a duplex from either the EST or Short GFP-etv2 3' UTR sensor. Significance was calculated using the student t-test. *$p < 0.05$

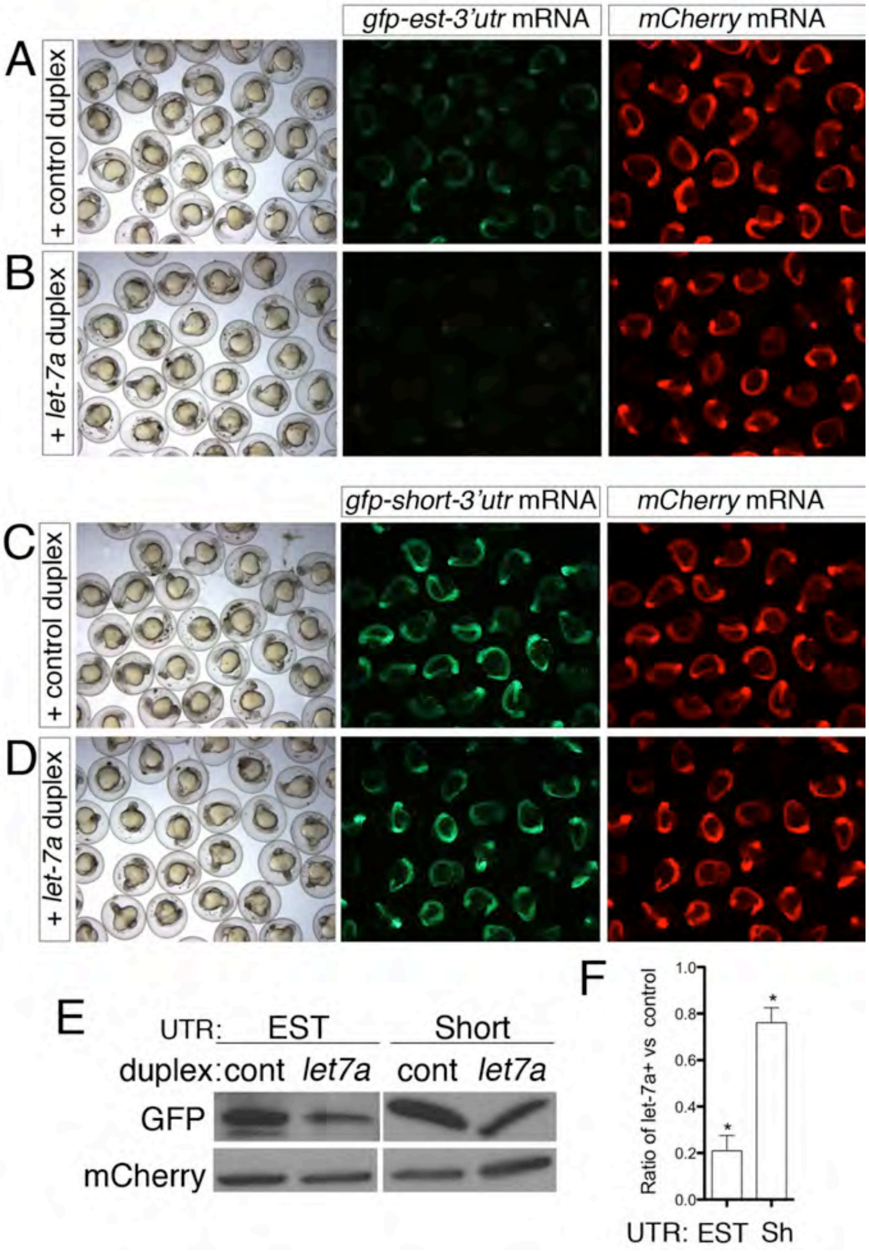
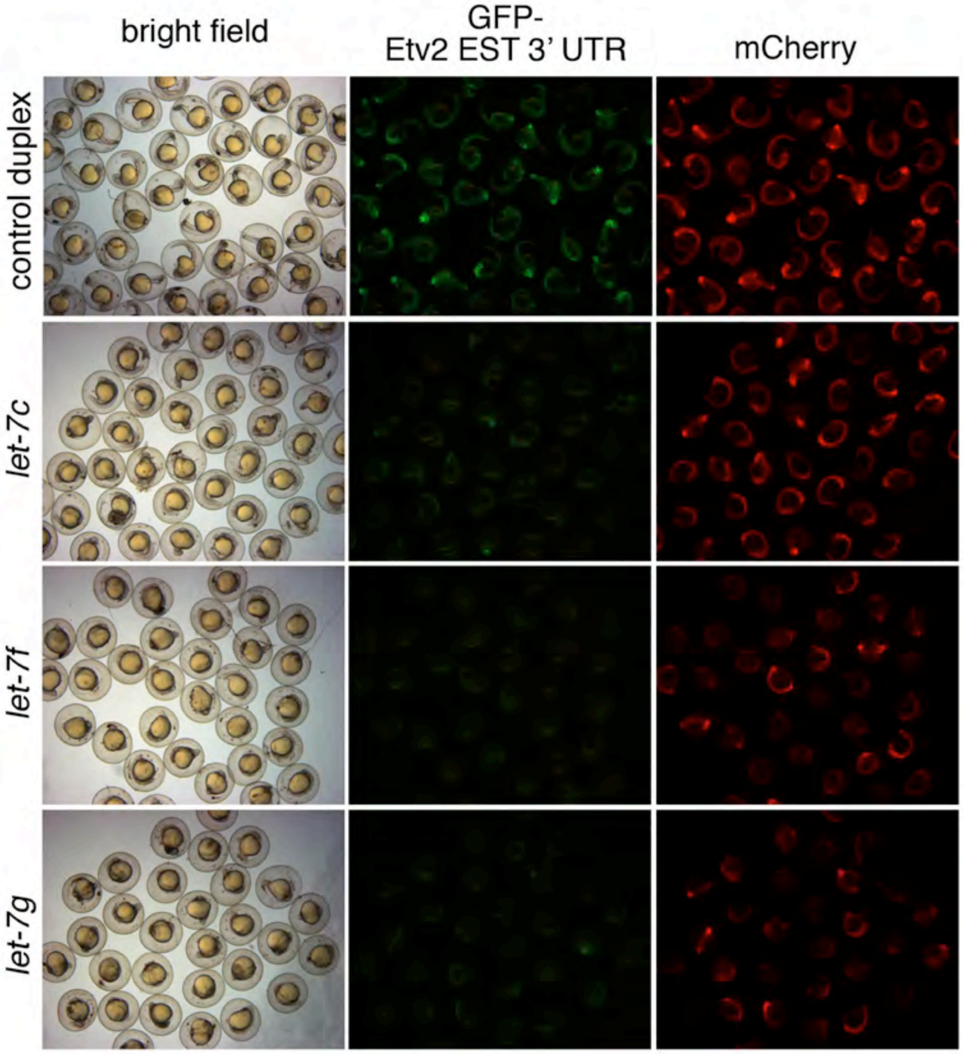

gfp-est-3'utr mRNA
mCherry mRNA
A
+ control duplex
B
+ let-7a duplex
gfp-short-3'utr mRNA
mCherry mRNA
C
+ control duplex
D
+ let-7a duplex
E
UTR: EST Short
duplex: cont let7a cont let7a
GFP
mCherry
F
Ratio of let-7a+ vs control
1.0
0.8
0.6
0.4
0.2
0.0
*
*
UTR: EST Sh

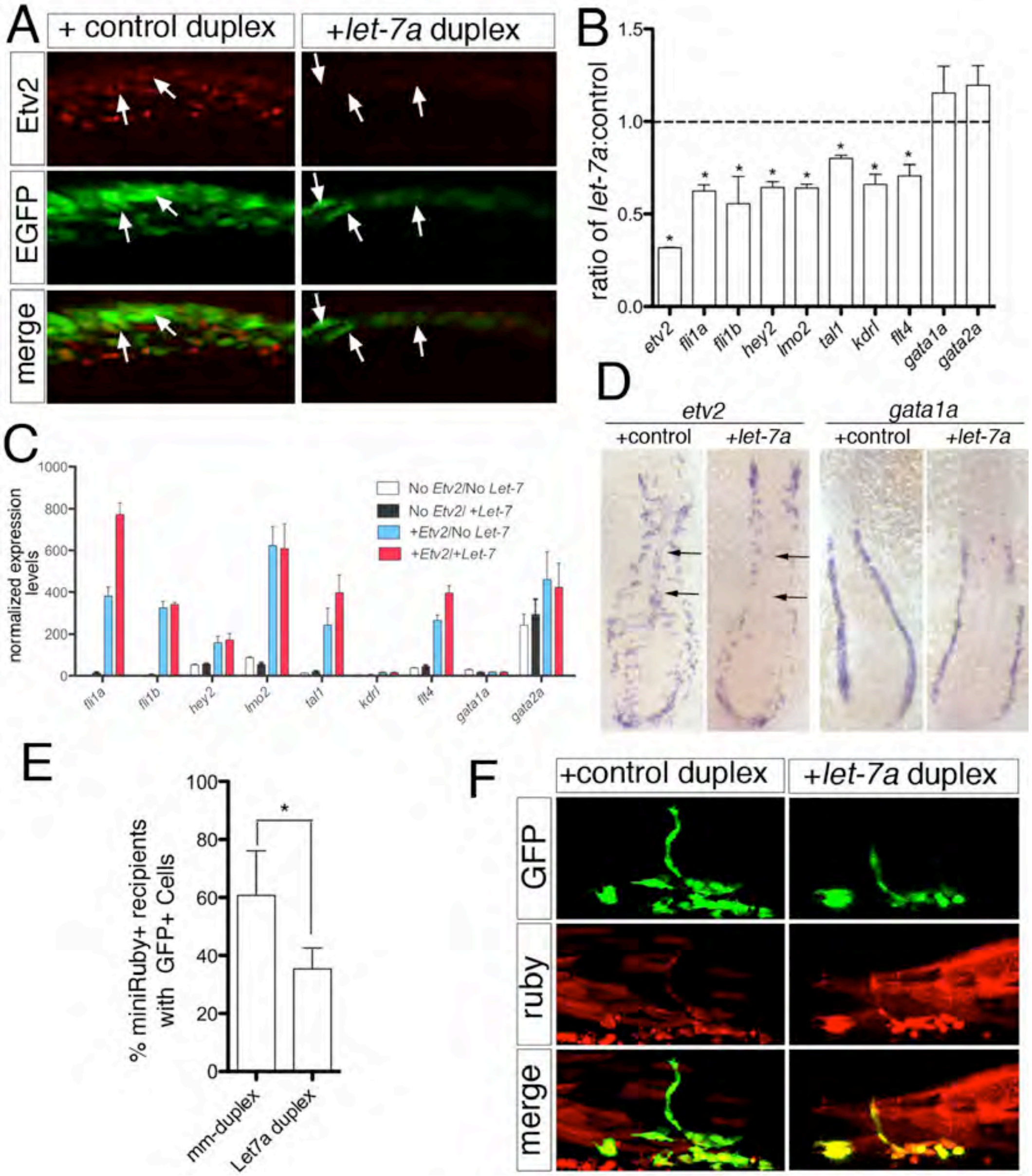
Figure 2.7: Multiple *let-7* family members can repress the *etv2* 3'UTR. Embryos were co-injected with gfp-est-*etv2* 3' UTR sensor (25 pg) and mcherry mRNAs (25 pg), along with indicated RNA duplexes. Bright field (left column), green fluorescent (middle column) and red fluorescent (right column) images of injected embryos were captured at 24 hpf.


bright field
GFP-
Etv2 EST 3' UTR
mCherry
control duplex
let-7c
let-7f
let-7g

reported to be expressed in zebrafish endothelial cells at 24 hpf [147]. Furthermore, deletion of the five putative *let-7* binding sites in the *etv2* Long 3' UTR resulted in a significant increase in mCherry reporter expression in the endothelial autonomous sensor assay compared to the wild type Long 3'UTR (Figure 2.5C). These observations suggest that *let-7* microRNAs can contribute to the negative regulation of Etv2 in endothelial cells.

To determine if *let-7* could repress endogenous *etv2*, we injected *let-7a* or control duplex into zebrafish embryos and assessed Etv2 protein and transcript levels. While Etv2 protein was apparent at low levels in *Tg(fli1a:egfp)*^{y1} -positive cells lining the nascent dorsal aorta at 15 ss following injection with control duplex, we failed to detect it in embryos injected with *let-7a* duplex (Figure 2.8A). Furthermore, we also noted reduced expression of the *fli1a:egfp* transgene in *let-7a* duplex injected embryos (Figure A5.8A), which is likely due to endothelial differentiation defects as a result of reduced Etv2 function [135]. We also observed significant effects on *etv2* transcript as a result of ectopic *let-7a* expression. We found that endogenous *etv2* transcript was significantly down regulated at 15 ss following injection of the *let-7a* duplex compared to embryos injected with control mismatch duplex (Figure 2.8B). Furthermore, we noted concomitant reduction in *fli1a, fli1b, hey2, lmo2,*

Figure 2.8: Endogenous Etv2 is repressed by let-7a. A. Two photon images of Tg(fli1a:egfp)y1 embryos injected control or let-7a duplex and immunostained with Etv2 polyclonal serum and Alexa-568 secondary antibody. Lateral view, dorsal is up, anterior to the left. Arrows denote Etv2/GFP-positive cells (left panels) or Etv2-negative/GFP- positive cells in the forming dorsal aorta (right panels). B. Histogram showing fold change in expression of indicated genes at 15 ss in embryos injected with let-7a compared to those injected with control duplex measured by the nCounter system. Genes normalized to actb2 (beta-actin) and eef1a1l1 (ef1alpha). *p<0.05. C. Histogram of relative nCounter expression counts normalized as in B for indicated genes following injection with mRNA encoding Etv2 (+Etv2) or Etv2 lacking the DNA binding domain (no Etv2) and mismatch (no let-7) or let-7a duplex (+ let-7). D. Whole mount in situ hybridization using riboprobes against etv2 (left) or gata1a (right) at 15 ss in embryos injected control or let-7a duplex RNA. Angioblasts that have migrated to the midline, or lack thereof, are indicated by arrows. Dorsal view of flat mounted embryo, anterior is up. E. Histogram showing percentage of successfully transplanted wild type host embryos (miniRuby-positive) that display contribution to vascular tissue, as indicated by presence of Tg(fli1a:egfp)y1-positive cells. Donors were injected with control or let-7a duplex as above. Data are from three independent experiments and significance was calculated using the Fisher’s exact test; *p < 0.05. F. Confocal micrographs showing contribution of Tg(fli1a:egfp)y1 positive cells (green channel) from donors that were injected with control or let-7 RNA duplex. miniRuby-positive cells (red channel) indicate overall contribution of donor cells in the trunk.

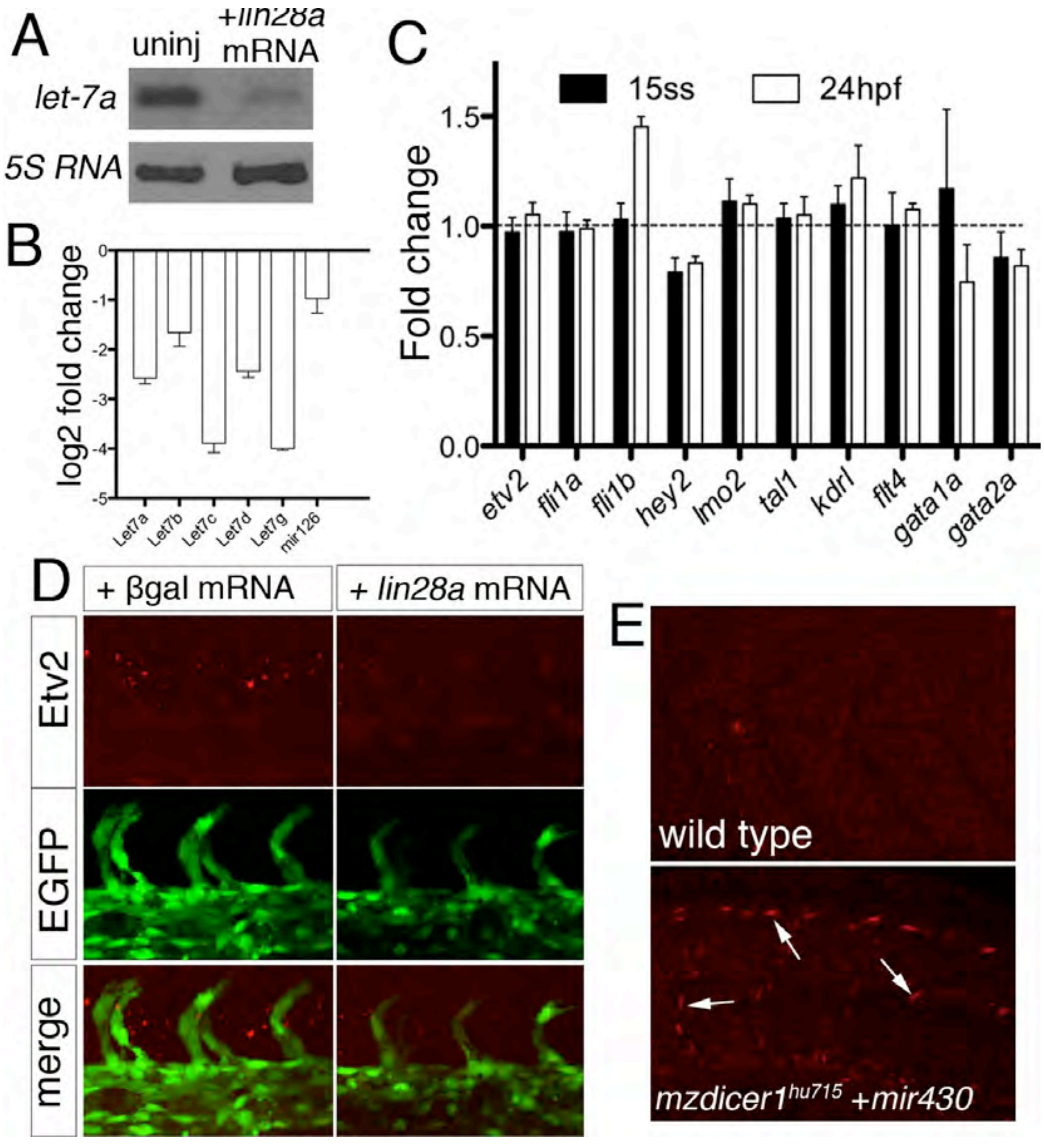

A
+ control duplex
+let-7a duplex
Etv2
EGFP
merge
B
ratio of let-7a:control
1.5
1.0
0.5
0.0
*
etv2
fli1a
fli1b
hey2
lmo2
tal1
kdrl
flt4
gata1a
gata2a
C
normalized expression levels
1000
800
600
400
200
0
No Etv2/No Let-7
No Etv2/ +Let-7
+Etv2/No Let-7
+Etv2/+Let-7
fli1a
fli1b
hey2
lmo2
tal1
kdrl
flt4
gata1a
gata2a
D
etv2
gata1a
+control
+let-7a
+control
+let-7a
E
% miniRuby+ recipients with GFP+ Cells
100
80
60
40
20
0
*
mm-duplex
Let7a duplex
F
+control duplex
+let-7a duplex
GFP
ruby
merge

tal1, kdrl, and *flt4* in *let-7a* duplex-injected embryos (Figure 2.8B), consistent with the observation that Etv2 can induce expression of these endothelial genes [136, 138, 148]. Co-injection of *etv2* mRNA containing a heterologous 3'UTR along with *let-7a* duplex rescued the expression of these Etv2-responsive genes, suggesting that they are not directly targeted by *let-7a* (Figure 2.8C). While we observed repression of several endothelial genes, there was no change in early hematopoietic markers such as *gata1a* and *gata2a* following injection of *let-7a* duplex (Figure 2.8B). Analysis of *etv2* expression by whole mount in situ hybridization in *let-7a* injected embryos at 15 ss revealed a slight down-regulation in *etv2* expression in the lateral mesoderm. More strikingly, we observed loss of midline-positioned *etv2*-positive cells, which normally comprise the forming aorta (Figure 2.8D). This defect is also known to be associated with Etv2 deficiency [135] and is consistent with reduced Etv2 function as a consequence of ectopic *let-7a* expression. We did not note any obvious changes in *gata1a* (Figure 2.8C), consistent with the gene expression data (Figure 2.8B). Together these results demonstrate that *let-7a* can potently repress expression of the Etv2 protein and lead to reduction in endothelial gene expression.

To investigate the effect of *let-7a* over-expression on endothelial cell commitment, we transplanted cells from *Tg(fli1a:egfp)*^{y1} embryos injected with *let-7a* into wild type embryos and assessed the frequency of successfully transplanted host embryos with EGFP-positive donor cells. Consistent with our observation that *let-7a* can repress endogenous *etv2,* significantly fewer host embryos transplanted with *let-7a* overexpressing donor cells displayed contribution to vascular tissue compared to control mis-match injected embryos (Figure 2.8E). Despite the negative effect of *let-7a* over-expression on endothelial cell contribution, *let-7a* duplex-injected donor cells were otherwise able to contribute to other cell types (Figure 2.8F). Taken together these data suggest that *let-7* family members can act to limit the ability of Etv2 to induce endothelial specification during development.

Determining the effect of *let-7* deficiency is a challenge due to the large number of related family members. Indeed, *let-7a* itself can be expressed from at least 6 distinct loci in the zebrafish genome and, including duplication events, there are a total of 18 different *let-7* family genes (ENSEMBL, Zv9). As mentioned above, several of these are expressed in endothelial cells and can repress the Etv2 3'UTR. Therefore, we over-expressed *lin28a*, which uridylates *let-7* miRNAs and causes their rapid degradation [149-151], to reduce *let-7* function. We find that injection of mRNA encoding Lin28a causes a significant decrease in *let-7a* expression in zebrafish embryos at 24 hpf (Figure 2.9A). This effect

Figure 2.9: Contribution of *let-7a* and other microRNAs to Etv2 repression. A. Northern analysis of RNA isolated from 24 hpf embryos left uninjected or injected with 1 ng *lin28a* mRNA. Blots were hybridized with using DIG labeled probes against *let-7a* and *5s* RNA. B. Histogram showing log2 fold change comparison of *let-7* family members at 15 ss assessed by miScript qPCR quantification between embryos injected with 1 ng *lin28a* mRNA and those left uninjected, quantification from triplicate experiments. C. Histogram showing fold change comparison of indicated genes assessed by nCounter quantification between embryos injected with 1 ng *lin28a* and 1 ng ®galactosidase mRNA. Genes normalized to *actb2* (*beta-actin*) and *eef1a1l1*(*ef1alpha*). D. Two-photon micrographs of trunk blood vessels in *Tg(fli1a:egfp)*$_{y1}$ embryos immunostained with Etv2 polyclonal antiserum and Alexa-568 secondary antibody at 24 hpf following injection with 1ng of *β-galactosidase* (left panels) or *lin28a* mRNA (right panels). E. Confocal micrographs of embryos immnuostained with Etv2 polyclonal antiserum and Alexa-568 secondary antibody at 48hpf. Top, wild type embryo. Bottom, *MZdicer1*$_{hu715}$ mutant embryos injected with *mir-430* duplex RNA (bottom). Etv2-positive nuclei in the endothelial cells of trunk blood vessels are denoted by arrows (bottom).


A
uninj
+lin28a mRNA
let-7a
5S RNA
B
log2 fold change
0
-1
-2
-3
-4
-5
Let7a
Let7b
Let7c
Let7d
Let7g
mir126
C
15ss
24hpf
Fold change
1.5
1.0
0.5
0.0
etv2
fli1a
fli1b
hey2
lmo2
tal1
kdrl
flt4
gata1a
gata2a
D
+ βgal mRNA
+ lin28a mRNA
Etv2
EGFP
merge
E
wild type
mzdicer1^hu715 +mir430

was similar for other *let-7* family members (Figure 2.9B), while *mir-126,* an unrelated miRNA expressed specifically in endothelial cells, was not significantly reduced. Despite reduction in let-7 miRNAs, we did not observe any significant changes in endothelial gene expression at 15 ss or 24 hpf (Figure 2.9C). Furthermore, neither *etv2* transcript or protein levels appeared to change significantly in Lin28a over-expressing embryos and vascular development proceeded normally (Figure 2.9C, D). These results raise the possibility that other miRNAs may contribute to post-transcriptional repression of Etv2 in the absence of *let-7*. Alternatively, sufficient levels of *let-7* remain to repress Etv2, even in the presence of increased *lin28a*. Therefore, we further investigated whether miRNAs contribute to Etv2 down-regulation by determining its expression in embryos lacking maternal and zygotic (MZ) *dicer* function. *dicer1* encodes an essential nuclease required for miRNA maturation and *MZdicer* mutant embryos are devoid of mature miRNAs [152]. Wild type embryos at 48 hpf did not exhibit Etv2 expression in endothelial cells (Figure 2.9E). By contrast, Etv2 protein expression was apparent at this stage in embryos lacking both maternal and zygotic *dicer1* (*MZdicer1*) that had been injected with *miR-430* to rescue early developmental defects associated with a lack of *dicer1* (Figure 2.9E) [152]. Despite the persistence of Etv2 protein, down-regulation of *etv2* transcript seemed to occur normally in wild type and MZdicer1 mutant embryos (Figure 2.4C). These results suggest that microRNAs may play a role in post-transcriptionally regulating levels of Etv2 protein during vascular development.

Discussion

The ETS transcription factor Etv2 is essential for vascular development, but less is known about its dynamic regulation or functional requirements during different stages of vascular development. Using the zebrafish as a model system, we find that both *etv2* transcript and protein are expressed during angioblast specification and vasculogenesis, but are subsequently downregulated as development proceeds. This expression pattern is mirrored by its functional requirement, which we find is restricted to early stages that correspond to angioblast emergence from the lateral mesoderm. We further provide evidence that post-transcriptional control of Etv2 levels likely contribute to its down-regulation, and this regulation occurs in part, through *let-7* miRNAs.

The phenotypes of *etv2*-deficient zebrafish and mouse embryos suggest that it should be considered as a master regulator of endothelial cell fates. In both species, loss of *etv2* leads to profound defects in vascular morphogenesis and a global loss of endothelial gene expression [122, 134, 135]. Conversely, exogenous Etv2 expression can precociously and ectopically induce an endothelial gene program [138, 148]. Accordingly, we find that Etv2-overexpression can increase the commitment of cells to the endothelial lineage in mosaic embryos, similar to its effect in mouse embryoid bodies [137]. While our results indicate that Etv2 is essential for endothelial cell specification, it appears to be dispensable for later steps of vascular development. Conditional knockdown of Etv2 at early, but not later stages, severely perturbed vascular morphogenesis, demonstrating that its function is only required during an early window of development in which the first endothelial progenitors are known to emerge. This early functional restriction is likely due the highly dynamic expression of *etv2*, which peaks during somitogenesis but is barely detectable by 24 hpf in the zebrafish embryo. Our results are consistent with recent studies in mouse embryonic stem cells where Etv2 expression can be detected in Brachyury-positive mesodermal cells that have not yet initiated expression of endothelial markers, such as *vegf receptor-2* (*vegfr2*) [140]. Furthermore, *etv2* expression also appears to be reduced in mouse embryos as development proceeds [133]. Finally, conditional ablation of *etv2* in mouse embryos using a Flk1:Cre driver does not effect endothelial differentiation or vascular morphogenesis [140], suggesting that *etv2* is also dispensable for later steps in vascular development in mammals as well. Taken together with our studies, these results suggest that Etv2 plays an essential and conserved role to commit early lateral mesoderm progenitors to an endothelial cell lineage, yet is not required during subsequent endothelial differentiation and morphogenesis.

In many cases, ETS transcription factors are essential throughout the ontogeny of a particular cell lineage, often acting reiteratively during commitment and differentiation. For example, *Spi1* (also known as *pu.1*) is essential for development of the myeloid lineage [153, 154], where it is required for early expression of receptors for macrophage and granulocyte specific growth factors on progenitor cells [155]. Subsequently, *spi1* plays an essential role in the terminal differentiation and function of both macrophages and neutrophils, where it continues to be expressed [155, 156]. By contrast, Etv2 function is restricted to early endothelial specification and its persistent expression has deleterious effects on differentiation and vascular morphogenesis. Continued endothelial expression of a Cre-activated conditional ROSA26:Etv2 transgene leads to abnormal yolk sack vascular morphology and a failure to induce expression of genes involved in vascular maturation [144]. A similar effect is observed during direct endothelial reprogramming of human amniotic cells, which requires ETV2, along with ERG and FLI1 [139]. While ETV2 is essential for direct reprogramming in this context, it must be subsequently down-regulated for normal endothelial differentiation to occur [139]. Together with our findings, these studies underscore the need to actively repress *etv2* expression to allow normal endothelial differentiation.

Our results suggest that miRNA-mediated post-transcriptional repression plays an important role in reducing the levels of Etv2 to allow normal differentiation. The *etv2* 3’UTR is capable of mediating repression in endothelial cells and Etv2 protein persists in MZ*dicer1* mutant embryos, which lack miRNAs. We further find that post-transcriptional repression of *etv2* is mediated, in part, by members of the *let-7* family of microRNAs. Our results are consistent with the role of *let-7* microRNAs in other animal species, where they are known to promote cellular differentiation or block transformation by negatively regulating genes associated with growth and proliferation, such as RAS and MYC [157-160]. We also find evidence for multiple *etv2* transcript isoforms with different length 3’UTRs, which varied in their potential to repress reporter expression and their responsiveness to *let-7* over-expression. While only the shortest 3’UTR did not appreciably repress reporter gene expression in our endothelial autonomous sensor assay, it retained two *let-7* binding sites and could still be partially repressed by *let-7a*. Importantly, *let-7* potently repressed endogenous Etv2 protein levels, suggesting an important role for this miRNA in regulating *etv2 in vivo*. However, the precise role of alternative 3'UTR usage in *etv2* regulation is not clear at this time. Interestingly, previous studies demonstrate that shortened 3'UTRs, which can presumably evade microRNA regulation, are prevalent in proliferating or transformed cells [21, 68]. Furthermore, 3'UTRs generally lengthen during embryonic zebrafish development as cells differentiate [14, 56]. These observations suggest that there may be a transition in 3'UTR usage as angioblasts differentiate into endothelial cells allowing for *let-7* to contribute to Etv2 down-regulation during differentiation.

Despite the effect of *let-7a* on endogenous Etv2 expression, we did not note any vascular defects caused by over-expressing *lin28a*, which drastically reduced *let-7a* levels. In addition, *lin28a* over-expression did not appear to cause Etv2 to persist at later stages. There could be several reasons for these results. First, the degree of *let-7a* knockdown caused by exogenous *lin28a* expression, while significant based on our quantification, may not be sufficient to mimic a true genetic null given the number and diversity of *let-7* microRNAs expressed in endothelial cells at this time point [147]. Thus, sufficient levels of *let-7* microRNAs may remain to repress Etv2 even in the presence of high levels of Lin28a. Future application of *lin28a* transgenes that would allow persistent high level expression throughout development are likely required to fully address this issue. Second, other miRNAs may contribute to Etv2 repression in the absence of *let-7* family members. While we are currently not able to rule out either of these possibilities, the persistent expression of Etv2 protein in *MZdicer1* mutant embryos underscores the importance of miRNAs in mediating its repression during development. Finally, there are likely to be additional regulatory mechanisms that contribute to Etv2 down-regulation and may be sufficient to allow normal vascular development in the absence of *let-7*. Indeed, while Etv2 protein persisted in *MZdicer1* mutant embryos, *etv2* mRNA was reduced as development proceeded similar to wild type siblings. Although exogenous *let-7a* appeared to reduce *etv2* transcript following whole embryo analysis, comparison of *in situ* results and immunostaining suggested a much more potent effect on Etv2 protein rather than mRNA levels. Furthermore, the observed decrease in *etv2* mRNA was likely due to embryos displaying fewer angioblasts, a phenotype associated with Etv2 deficiency [122, 135] and consistent with the observed reduction in Etv2 protein following *let-7a* over-expression. Together, these observations suggest that down-regulation of etv2 mRNA may occur through transcriptional mechanisms. Consistent with this possibility, a recent study has identified an enhancer element in the zebrafish *etv2* locus that drives expression in endothelial cells only during early stages of embryonic development but not later

[161]. The activity of this enhancer mirrors that which we observe for the endogenous transcript, suggesting that transcriptional mechanisms can contribute to the dynamic expression of *etv2*. Thus, while our work suggests that miRNA-mediated post-transcriptional repression contributes to Etv2 down-regulation, it is likely that other mechanisms play an important role in this process.

Whether *let-7* or other miRNAs contribute to the repression of Etv2 in mammals is unknown. We did observe that human and mouse ETV2 3' UTRs also contain *let-7* binding sites (data not shown). Moreover, zebrafish and human endothelial cells highly express several *let-7* family members[162], but do not express appreciable amounts of *ETV2* [163]. These observations suggest that *let-7*-mediated repression of *etv2* may be a conserved aspect of its regulation. Based on its powerful ability to induce endothelial gene programs and block differentiation, along with the evidence that it is normally actively repressed, we would further speculate that Etv2 is likely to play a role in the pathogenesis of syndromes associated with dysregulated endothelial growth. In this regard, it will be of interest to determine if *etv2* expression is persistent in cases of infantile hemangioma or angiosarcoma. At the same time, further investigation into the mechanisms that contribute to *etv2* regulation will likely provide novel insights into how this master regulator contributes to endothelial lineage specification.

Material and Methods

Zebrafish Handling and Maintenance

Zebrafish and their embryos were handled according to standard protocols [164] and in accordance with the University of Massachusetts Medical School IACUC guidelines. *Tg(fli1a:egfp)^{y1}*, *Tg(fli1a:negfp)^{y7}* and *Tg(fli1a.ep:DsRedex)^{um13}* lines have been described [165-167]. Maternal zygotic (MZ) *dicer1* embryos were made using the germline replacement technique as previously described [152, 168, 169].

Etv2 Caged Morpholino Injections

The Etv2 caged morpholino (cMO) used in this study has been previously reported [143]. 230 fmol (2 ng) of Etv2 cMO was injected into *Tg(fli1a.ep:DsRedEx)^{um13}* embryos at 1-cell stage. Embryos were subjected to UV illumination for 10 seconds at indicated stages using a Zeiss Axioskop2 Plus compound microscope with a DAPI filter and an Achroplan (Zeiss) 20x water immersion objective. Following photoactivation, embryos were grown in egg water at 28.5°C. Control embryos were left in the dark. 5 ng of scrambled control or 5 ng Etv2 MO [135] were injected as negative and positive controls, respectively.

Vascular morphology was assessed at 30 hpf. Embryos were imaged using an MZFLIII fluorescent dissection microscope or using a using a Leica DMIRE2 confocal microscope (Objective: HC PL APO 20x/0.70CS). Circulatory defects were observed using a MZ12 stereomicroscope (Leica) and captured with a DMK21F04 camera (Imagesource) using Quicktime Pro or iMovie.

Plasmid Construction

The *etv2* open reading frame was amplified from 24 hpf whole embryo cDNA and used in a BP recombination reaction with plasmid pDONR221 (Invitrogen) to make pME-*etv2*. The zebrafish *lin28a* open reading frame was amplified from a full-length Zebrafish Gene Collection (ZGC) clone (Clone ID: 2643384; Thermo Scientific; see Table 2.1 for primers), Then subjected to BP recombination with plasmid pDONR221 to generate pME-*lin28a*. pME-*etv2*, pME-*lin28a*, or pME-*mcherry* [170] were used in LR reactions with pCSDest or pCSMTDest [171] to generate pCS-*etv2*, pCSMT-*etv2*, pCS-*lin28a*, and pCS-*mCherry*. Alternative *etv2* 3’ UTRs were cloned through PCR amplification using attB2 and attB3 primers (see Table 2.1) followed by BP recombination into pDONRP2r-P3 (Invitrogen) to give p3E-EST *etv2* 3’UTR, p3E-short *etv2* 3’UTR and p3E-long *etv2* 3’UTR. *let-7* binding sites were identified by miRANDA, RNAhybrid, and a perl script. Bases 1, 3, 4, 5, 6 were mutated to adenines within 5 identified

Table 2.1: Primers used in Chapter 2.

Primer #	Primer Name	Primer Sequence (5'-3')
906	for GST-Etv2 F	gatcggatccGAAATGTACCAATCTGGATT
906	for GST-Etv2 R	gatcctcGAGCGCTGCGTCTTTTGACCA
964	attB1 etv2 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTtaaccatggaaatgtaccaatctg
824	attB2 etv2 R	GGGGACCACTTTGTACAAGAAAGCTGGGTctaatgtgtccaggactctgt
4173	etv2 3'RACE F	CATCATTCACAAAACGGCGGGAAAGCGCTACG
4174	etv2 3'RACE nested F	CCGCTTTGTCTGTGACGTGCAGGGCATGCTTG
4053	attB1 lin28 F	GGGGACAAGTTTGTACAAAAAAGCAGGCTGCgccaccatgcccccggcaaatccgc
4054	attB2 lin28 R	GGGGACCACTTTGTACAAGAAAGCTGGGTcctaatcagtgctctctggc
1751	etsrp 3'UTR short F attB2	GGGGACAGCTTTCTTGTACAAAGTGGCCTGGACACATTAGAGGAGGA
1752	etsrp 3'UTR short R attB3	GGGGACAACTTTGTATAATAAAGTTGtgtaatcgtccgtcttcaaca
1753	etsrp 3'UTR long F attB2	GGGGACAGCTTTCTTGTACAAAGTGGTGTTGAAGACGGACGATTACA
1754	etsrp 3'UTR long R attB3	GGGGACAACTTTGTATAATAAAGTTGtctgttgaagcttttggagag
4219	attB2-etv2 3'utrF	GGGGACAGCTTTCTTGTACAAAGTGGAGGAGGAATTCTCGAAGGAT
4278	attB3 etv2 peak3 3'utr R	GGG AC AAC TTT GTA TAA TAA AGT TG ATGCCACAACAACAGTTTTATTGTAAATAA
1793	F attB2 miR sensor control	GGGG ACA GCT TTC TTG TAC AAA GTG G GGCGCGCCTACGTAACTAGT
1794	R attB3 miR sensor control	GGGG AC AAC TTT GTA TAA TAA AGT TG CTCGAGACTAGTTACGTAGG
587	ets1 b1 fwd	GGGGACAAGTTTGTACAAAAAAGCAGGCTgc gtg acc atg acggcagct
953	ets1a attB2 R	GGGGACCACTTTGTACAAGAAAGCTGGGTcagactttactcgtccgtgtc
3332	Mm etsrp attB1	GGGGACAAGTTTGTACAAAAAAGCAGGCTtaaccatggacctgtggaactgggatgagg
3333	Mm etsrp attB2	GGGGACCACTTTGTACAAGAAAGCTGGGTcttattggccttctgcacctggcagatgcc

let-7 binding site seed sequences identified by all three methods, as described in [172]. The mutant *let-7* etv2 3' UTR fragment was synthesized by Genewiz (pUC57-kan-etv2_3putr_mut_let7) followed by subcloning into p3E-mcs1 with AscI and XhoI to give p3E-mut*let-7 etv2* 3' UTR. To generate mRNA sensors constructs, p3E-EST*etv2* 3'UTR or p3E-short*etv2* 3'UTR were recombined with pCSDEST2 and pENTR-EGFP2 [171] to yield pCS2-egfp-EST*etv2* 3'UTR and pCS2-egfp-short*etv2* 3'UTR. Endothelial 3' UTR sensor constructs were generated by performing an LR Gateway recombination reaction between pTolBasPegfpfliEPmcherryR2-R3 and one of the following 3' entry clones: p3E-mcs1, p3E-shortEtv2-3'UTR, p3E-ESTEtv2-3'UTR, p3E-longEtv2-3'UTR, p3E-mut-let7-Etv2-3'UTR.

mRNA Synthesis and Injections

Capped mRNA was synthesized from pCS plasmids that had been linearized with NotI using the SP6 mMessage mMachine kit (Ambion). mRNAs were injected into 1-cell stage embryos according to standard protocols [164].

3'UTR Sensor Assays

For whole embryo sensor assay, 50 pg of *mCherry* mRNA and 50 pg of indicated *gfp etv2 3'UTR* mRNA was co-injected along with 50 µM of indicated miRNA duplexes into 1-cell stage zebrafish embryos. Embryos were visualized at 24 hpf using an MZFLIII dissection microscope equipped with epifluorescence and digital images were captured using an AxioCam mRC (Zeiss). Alternatively, equal numbers of dechorinated embryos were lysed by boiling in 2x Laemmli buffer. Lysates were run on an SDS-PAGE gel and transferred to Western blots, which were probed with antibodies against EGFP (Invitrogen, A11122) and mCherry (Clontech, 632496). Blots were stripped in between each antibody detection. Expression levels were quantified by measuring the optical density of bands using ImageJ following incubation with a horseradish peroxidase conjugated secondary antibody and chemiluminscence detection. For endothelial autonomous sensor assays, 25 pg of indicated pTol sensor construct was co-injected with 25pg *transposase* mRNA into one-cell stage wild type embryos. Individual 3'UTR constructs were always injected with control sensor in parallel. At 24 hpf, embryos were transferred to egg water containing 0.2mM 1-phenyl-2-thiourea (PTU) to inhibit the pigment formation. At 48-50hpf, approximately five embryos from each group per experiment displaying robust transgenesis were imaged by confocal microscopy. Gain settings were set using embryos injected with the control sensor and remained constant throughout the experiment. Quantification of fluorescence levels was performed using Imaris by creating a surface based on GFP fluorescence and examining the average values intensity sum of green and red channels. The red/green ratio of an experimental embryo was normalized against the red/green ratio of a control embryo imaged on the same day. All sensor experiments were done and quantified in quadruplicate, except the EST-3'UTR which was done in triplicate. Significance was calculated by a Welsh test and significance determined by a p value < 0.03.

Antibody Production

A fragment encoding the N-terminal 218 amino acids of zebrafish Etv2 was amplified from 24hpf zebrafish cDNA (see Table 2.1 for primers), cloned into pCR2.1 by TOPO cloning (Invitrogen), and sequence verified. The *etv2* fragment was subcloned into pGEX-6P-1 using BamHI and XhoI sites. pGEX-Etv2 was transformed into BL21(DE3) and glutathione S-transferase (GST) fusion protein expression was induced with IPTG. Expressing bacteria were lysed using Bug Buster (Novagen), and proteins were purified using Glutathione Sepharose 4B(GE Healthcare), followed by release of the Etv2 fragment and removal of the GST using PreScission Protease (GE Healthcare). Purified Etv2 protein was used for rabbit polyclonal antibody production (Caprologics, Gilbertville, MA). Etv2 antiserum was validated using Western analysis of lysates from HEK293T over-expressing myc-tagged zebrafish Etv2. The myc epitope was detected using a 1:10,000 dilution of 9E10 (Sigma) and Etv2 protein was detected using a 1:5,000 dilution of anti-Etv2 polyclonal antibody serum.

Whole Mount Immunostaining

Staged zebrafish embryos were fixed overnight at 4°C in 2% paraformaldehyde (w/v) dissolved in phosphate buffered saline containing 0.1% Tween-20 (PBSTw). Embryos were washed 4 times for 5 minutes at room temperature in PBSTw and in PBS containing 0.5% TritonX-100 (PBSTr) at room temperature for 30 minutes. Embryos were blocked for a minimum of 2 hrs in blocking solution (PBSTw, 0.1% TritonX-100, 10% normal goat serum, 1% BSA, 0.01% sodium azide) at room temperature. Fli1b and Etv2 rabbit polyclonal serum was diluted 1:1000 and 1:500, respectively, in blocking solution and embryos incubated over night at 4°C. Embryos were washed 6 times in PBSTw for 4 hrs minimum at room temperature and then incubated in Alexa Fluor 488 or Alexa Fluor 568 (Invitrogen) anti-rabbit secondary antibody diluted 1/1000 as indicated in blocking solution overnight. Immunostained embryos were imaged on a LSM7 MP microscope (Zeiss; Objective: 20x/1.0 DIC(UV) VIS-IR 421452-9800) equipped with a Chameleon Ti:Sapphire pulsed laser (Coherent, Inc.) Alexa Flour 488 was excited using 904 nm light and Alexa Flour 568 was excited using 1057 nm light.

miRNA Duplexes

RNA oligonucleotides (Integrated DNA technologies) corresponding to the mature and star sequences of zebrafish *let-7a*, *let-7c*, *let-7f*, and, *let-7g* (see Table 2.2) were diluted to 250 mM in nuclease-free water. Equal volumes of mature and start oligonucleotides were combined, heated to 95°C and annealed at 37°C for 30 minutes. miRNA duplexes were aliquotted and stored at -80°C. 2 nl of miRNA duplexes were injected into embryos at a concentration of 50 µM . A mis-match duplex in which 4 out of 8 bases in the seed sequenced were changed (Table 2.2) was used as a negative control (referred to as “control duplex”).

Quantification of Endothelial Gene Expression

mRNA was quantified using the NanoString nCounter gene expression system (Nanostring Technologies, Seattle, WA)[173]. Total RNA was isolated from embryos using a Qiagen RNAeasy kit. For embryos injected with 50µM *let-7a* or *mm-let7a* duplex, RNA was isolated at 15 ss. To assess over-expression of Etv2 and *let-7a*, embryos were co-injected with *let-7a* duplex as above along with 50 pg of mRNA encoding Etv2 or Etv2 minus its DNA binding domain [Etv2(-DBD)] and RNA was isolated at shield stage. For each experiment, 100 ng of total RNA was hybridized for 12 to 20 hrs with the Nanostring probeset (Table 2.3) at 65^0C in a thermocycler. Samples were then loaded into the nCounter prep station and fluorescence signal was quantified using the nCounter Digital

Table 2.2: let7 duplexes and probes used in Chapter 2.

dre-Let-7a mature	rUrGrArGrGrUrArGrUrArGrGrUrUrGrUgArUrArGrUrU
dre-Let-7a anti-sense	rArArCrUrArUrArCrArArCrCrUrArCrCrUrCrA
dre-Let-7c mature	rUrGrArGrGrUrArGrUrArGrGrUrUrGrUrArUrGrGrUrU
dre-Let-7c anti-sense	rArArCrCrArUrArCrArArCrCrUrArCrUrArCrA
dre-Let-7f mature	rUrGrArGrGrUrArGrUrArGrArUrUrGrUgArUrArGrUrU
dre-Let-7f anti-sense	rArArCrUrArUrArCrArArUrCrUrArCrUrArCrCrUrCrA
dre-Let-7g mature	rUrGrArGrGrUrArGrUrArGrUrUrUrGrUrArUrArGrUrU
dre-Let-7g anti-sense	rArArCrUrArUrArCrArArArCrUrArCrUrArCrCrUrCrA
mutant-Let7-sense	rUrCrArCrCrUrUrGrUrArGrGrArUrGrUrArUrArGrUrU
mutant-Let-7 anti-sense	rArArCrUrArUrArCrArUrCrCrUrArCrArArGrGrUrGrA
let-7a LNA	5’ –dig AACTATACAACCTACTACCTCA-dig- 3’
5S DIG-oligo probe	5’ –N(dig)ATCGGACGAGATCGGGCGTA - 3’

Table 2.3: Nanostring probes used in Chapter 2.

Gene	Accession	Targeted Region	Tm CP	Tm RP	Gene	PN(CP;RP)
etv2	NM_001037375.1	790-890	79	82	etv2	340352;240352
kdrl	NM_131472.1	455-555	82	81	kdrl	340356;240356
flt4	NM_130945.1	620-720	82	82	flt4	340353;240353
fli1a	NM_131348.2	620-720	81	81	fli1a	340355;240355
fli1b	NM_001008780.1	1365-1465	82	82	fli1b	340350;240350
hey2	NM_131622.2	990-1090	83	82	hey2	346488;246488
actb2	NM_181601.3	1647-1747	84	82	bactin2	328374;228374
eef1a1l1	NM_131263.1	1455-1555	78	82	ef1a	328447;228447
gata2a	NM_131233.1	2030-2130	81	82	gata2a	328441;228441
tal1	NM_213237.1	635-735	82	80	tal1	340349;240349
lmo2	NM_131111.1	215-315	79	82	lmo2	340354;240354
gata1a	NM_131234.1	175-275	81	83	gata1	328442;228442

Gene	Accession	Target Sequence
etv2	NM_001037375.1	CTTTGGCAGTTTCTGCTAGAACTCCTGCTGGATTCTGCTTGCCACACTTTTATAAGTTGGACTGGTGATGGCTGGGAGTTTAAAATGTCAGATCCCGCTG
kdrl	NM_131472.1	AACATACCCAAACCAAAACGTTATCCTTGAGACGCAGATGAATCCTATGGCAGATGATGTTAAAAGAGGGGTACAGTGGGATCCAAAAAAAGGTTTCACG
flt4	NM_130945.1	TCCTGACCTAAAAGTCACTCTCTTCTCGTTAGTGCCGTATCCAGAGCCTGTGGATGGCAGTGTGGTCACCTGGAATAATAAAAAGGGTTGGTCGATTCCC
fli1a	NM_131348.2	ACTTCCTGAGACTCACCAGCGTTTATAACACCGAGGTCCTTCTCTCACATCTCAATTACCTCAGGGAAAGTAGCTCATCGATATCATACAACACGCCATC
fli1b	NM_001008780.1	GTAATTTCTTCACGCCTCAATCCACCTACTGGAACTCCGCAACCAGTGTGGTTTATCCCAGTTCACCGATGCCACGACATCCCAGCACTCACACTCACTT
hey2	NM_131622.2	CGCTGGATTCCCACTCTTCAGCCCCAGCGTTACAGCATCTTCAGTGGCTTCTTCCACCGTGAGCTCTTCCGTTTCCACATCCACCACATCCCAACAGAGC
actb2	NM_181601.3	CCTGGGCATATTGTAAAAGCTGTGTGGAACGTGGCGGTGCCAGACATTTGGTGGGGCCAACCTGTACACTGACTAATTCAATTCCAATAAAAGTGCACAT
eef1a1l1	NM_131263.1	CCAAGTGAATTTCCCTCAATCACACCGTTCCAAAGGTTGCGGCGTGTTCTTCCCAACCTCTTGGAATTTCTCTAAACCTGGGCACTCTACTTAAGGACTG
gata2a	NM_131233.1	ATTTACTGAGTCACTTTGGTACTGAAAGAGCGGACGCAGAATCACTGTGTGGTAGTCAAAACGGCCACCTCAAAACTCTCATAAAGGACTCGCTTTGAGC
tal1	NM_213237.1	TAGCAATCGAGTCAAGCGCAGACCTGCACCTTATGAGGTTGAAATCAACGATGGTTCGCAGCCCAAAATTGTGCGACGGATTTTCACGAACAGTCGCGAG
lmo2	NM_131111.1	GCGTACACAATGTGTGCTGGATGTTTCTGACCTTTGATACACTTGCTAAGACAGCAGAACAGGTGCATCTCTGAAGCGTTTTGTGCGGCAGATGGTCTTT
gata1a	NM_131234.1	ACAGACTCTGGTTTACTGCCACCCGTTGATGTAGATGAACCTTTCTACTCAAGCTCTGAGACTGACCTACTGCCATCGTATTATTCCACCAGCGTCCAGA

Analyzer. Gene normalization and fold change calculations were done using Nsolver Analysis Software (Nanostring Technologies). In all cases, biological triplicates were performed and gene counts were normalized to *eukaryotic translation elongation factor 1 alpha 1 like 1* (*eef1a1l1*) and *actin, beta 2* (*actb2*). Either the average normalized gene count or the average fold change was plotted and error bars represent the Standard Error of the Mean (SEM).

***In Situ* Hybridization**

An antisense DIG-labeled *etv2* riboprobe was synthesized by linearizing pCS2-*etv2* with EcoRI followed by in vitro transcription using T7 polymerase. A *gata1a* riboprobe was synthesized as described elsewhere [174]. Whole mount in situ hybridization was performed according to standard protocols [175].

Mosaic Analysis

Tg(fli1a:egfp)^{y1} embryos were used as donors in all cases and 0.35% miniRuby (dextran, tetramethylrhodamine and biotin 10,000MW) (Invitrogen D-3312) was co-injected as a lineage tracer. To assess the effect of Etv2 overexpression, we injected 100pg of *myc-etv2* or *mCherry* mRNA into 1-cell stage donor embryo. For *let-7a* overexpression we injected 2 nl of either 50µm control or *let-7a* duplex. At sphere stage, approximately 20 cells were transplanted from the ventral blastoderm margin of donors into wild type hosts, which were subsequently screened at 30 hpf for the appearance of red and green fluorescence. Embryos were imaged using an MZFLIII fluorescent dissection microscope or using a using a Leica DMIRE2 confocal microscope (Objective: HC PL APO 20x/0.70CS). The proportion of successfully transplanted embryos (i.e. exhibiting miniRuby-positive cells in any trunk tissue) with contribution to blood vessels was determined in three separate experiments and significance was calculated by Fisher's Exact test. $p < 0.05$ was deemed significant.

Northern Blotting

Northern blot analysis for microRNA expression was performed as previously described [176]. Zebrafish RNA was isolated using a miRNeasy Micro kit (Qiagen) and 5µg of total RNA was loaded per lane. Blots were hybridized with a DIG labeled *let-7a* locked-nucleic acid probe (Exiqon), stripped using boiling water, and hybridized with a DIG-labeled 5s rRNA DNA probe (see Table 2.2). Chemilumenscence detection was performed following incubation with a horseradish peroxidase-conjugated antibody against DIG. Northern blots were performed using RNAs from three separate experiments and quantified by measuring the optical density of bands using ImageJ to compare levels in uninjected versus *let-7a* injected embryos. Average fold difference from three independent experiments was plotted and error bars represent SEM. Significance was measured using a student t-test.

3’ RACE and Etv2 3’ UTR Cloning

3’ RACE was performed using the SMART RACE kit (Clontech). *etv2*-specific primers for primary and nested PCR are listed in Table 2.1. Amplified fragments were gel purified, cloned into pGEM-t (Promega) and sequence verified.

Quantitative PCR of miRNAs

RNA was purified from uninjected zebrafish embryos injected at 24 hpf or those injected with 1 ng *lin28a* mRNA using a miRNeasy micro kit (Qiagen). qRT-PCR to detect mature miRNAs was performed using the miScript System (Qiagen). Two µg of whole RNA was used to synthesize cDNA. qPCR was performed from 100 ng of cDNA template with a commercially available primers for indicated miRNA (Qiagen) and the miScript universal primer using the miScript SYBR green PCR Kit (Qiagen). *snord61.2* expression was assessed in parallel and used to normalize microRNA expression levels. PCR quantification was performed on a StepOnePlus real time PCR system (Applied Biosystems). Each reaction was run in triplicate and performed on at least two experimental replicates and 2-log fold change calculated by comparing uninjected to Lin28a injected.

CHAPTER III

APPLICATION OF A NAÏVE BAYES CLASSIFIER TO ASSIGN POLYADENYLATION SITES FROM 3' END SEQUENCING DATA

This manuscript was submitted to Bioinformatics: **Sheppard, S**, Lawson, N, and Zhu, L.J.

Introduction

3’ end processing of pre-mRNAs in the nucleus influences transcription termination, mRNA stability and localization, and dynamic regulation of translation. In plants, yeast, and metazoans, specific sequence elements in the 3’ untranslated region (3’UTR) direct cleavage and polyadenylation (reviewed in [1, 2]). Among these elements is the polyadenylation consensus signal (PAS) [3], a defined hexameric sequence located 10-30 nucleotides (nt) upstream of the cleavage and polyadenylation site (pA site), which binds the protein complex Cleavage and Polyadenylation Specificity Factor [4-6]. While the PAS is predominantly known to comprise the sequence AAUAAA [3], numerous single nucleotide variants are also functional [7, 8]. 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 [9-11] that is recognized by Cleavage Stimulatory Factor [12, 13]. In some instances, a uracil-rich sequence element is present upstream of the PAS [10, 14], which may also act to enhance usage of a specific PAS [15, 16] by recruiting Cleavage Factor I [17]. In combination, these sequence elements help define the site of cleavage and polyadenylation at the 3’ end of a pre-mRNA.

In many instances, differential PAS usage results in alternative polyadenylation (APA) and the formation of distinct transcript isoforms. In some cases, APA can affect protein structure. For example, use of an alternative exon containing a common variant PAS (ATTAAA) in FLT1, an angiogenic growth factor receptor, truncates the full length sequence resulting in a soluble form of the receptor [18]. The soluble FLT1 receptor acts as competitive inhibitor of vascular endothelial growth factor stimulated angiogenesis [19] and has been associated with pathological conditions, such as preeclampsia [20]. APA may also modify the regulatory potential of the 3'UTR. Longer 3'UTRs may contain more cis elements for binding of microRNAs or RNA binding proteins when compared to shorter 3'UTRs, subjecting 3'UTR isoforms to distinct regulation. Interestingly, shorter 3'UTRs have been associated with uncontrolled growth and cancer [21-24] and highly expressed housekeeping genes in humans [25]. Conversely, the use of longer 3'UTRs, in some cases, correlates with differentiation [26-36]. Alternative 3'UTR usage can also be stage- or tissue-specific. For example, miR-206 is expressed at the same levels in diaphragmatic and limb satellite muscle cells, but the transcript encoding one of its targets, Pax3, possesses distinct 3'UTRs in these tissues [37]. A longer 3'UTR used in limb satellite cells leads to repression of Pax3, while Pax3 expression in diaphragmatic satellite muscle cells persists due to evasion of miR-206 regulation through usage of a shorter 3'UTR [37]. In addition to normal biological processes, alternative PAS usage is also associated with Parkinson's disease [38], schizophrenia [39], thalessemias [40-42], and cancer [43, 44], suggesting a significant role for aberrant 3' end processing in a variety of disease settings.

Given the relevance of APA, reliable identification of cleavage and polyadenylation sites is necessary. Early studies relied on expressed sequence tags (ESTs), constructed from cDNA libraries primed with an oligonucleotide of deoxythymines (oligo-dT), to broadly identify the 3' ends of mRNAs [45-49]. More recently, deep sequencing has been applied for this purpose (reviewed in [50]). Poly(A) Site sequencing (PAS-Seq) [29] and Sequencing of APA Sites (SAPAS) [24] utilize an anchored oligo-dT (20) primer and template switching to construct 3' end libraries. PolyA-seq is similar to these methods, but uses a shorter anchored oligo-dT (10) and random hexamers during second strand cDNA synthesis [51]. While these approaches are technically straightforward, a major drawback is that the oligo-dT primer can bind to internal homopolymeric stretches of adenines, as well as to the poly-adenosine (poly(A)) tail [52]. Additionally, fragmentation of RNA occurs before cDNA synthesis in a majority of these protocols, potentially allowing the oligo-dT access to mis-prime internally adenine-rich regions. Together, these technical issues lead to identification of false positive pA sites. To reduce biases from internal oligo-dT priming and conversion of RNA to cDNA, Direct RNA Sequencing (DRS), using an oligo-dT (50) bound directly to a flow cell, was developed [10, 23, 53, 54]. A more selective method is poly(A)-position profiling by sequencing (referred to as 3pseq), in which a splint RNA:DNA oligonucleotide containing 3' overhanging thymines is ligated to the polyadenylated tail of mRNAs leading to identification of only true 3' ends of transcripts [36]. However, because this elegant method is technically daunting, most laboratories are likely to perform a simpler, oligo-dT primed approach, coupled with simple computational filters to remove sequences associated with mis-priming.

Efforts to identify internally primed (referred to as false) sites from oligo-dT primed deep sequencing have mostly employed heuristic filters comprising a defined number of adenines in the sequence downstream of the cleavage site [7, 23, 24, 29, 32, 34, 35, 47-49, 55-65]. Although filtering may not be needed for more technically stringent techniques, such as DRS [53, 54], heuristic filtering does remove internally primed sites in these cases as well [23]. However, given the strict definition of a heuristic filter, these approaches will inevitably miss some internal priming events (false positives) and will also exclude true 3' ends (false negatives) [54].

Alternative methods combine computational and technical methods to establish scoring systems that allow identification of oligo-dT primed 3’ends [51]. Though these can achieve relatively high sensitivity, they require additional technical steps (e.g. control library construction), adding to cost. In addition, subsequent analysis and filtering may not be straightforward or easily applicable [31, 51, 66]. Therefore, additional methods are needed to easily analyze oligo-dT primed deep sequencing data to identify true pA sites.

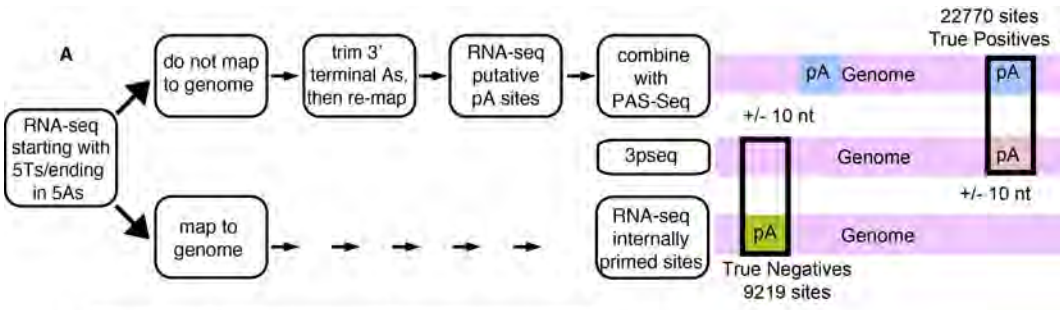
A naïve Bayes classifier is a supervised learning algorithm in which the features used to predict the class are considered conditionally independent given the class [67, 68]. Despite its simplicity, naïve Bayes classifiers have successfully addressed biological and medical problems, especially when many features are used for modeling [111-113]. In this study, we demonstrate its effectiveness in identifying internal priming events from oligo-dT primed 3’end sequencing data. We used PAS-Seq, 3pseq [69], and RNA-seq data to build training data sets containing true and false (internally primed) pA sites expressed in zebrafish. We developed a naïve Bayes classifier to assign the probability of being true or false to a putative pA site based on features from the surrounding sequence. Our algorithm outperforms several previously published heuristic filters and enriches for canonical motifs important for cleavage and polyadenylation. Furthermore, the nucleotide profiles and PAS distribution from algorithm-filtered oligo-dT primed 3’ end sequencing data mirror those from 3pseq data. Biological validation shows that our method is highly accurate, facilitating identification of novel 3’UTRs. Finally, we demonstrate the utility of the naïve Bayes classifier in other model organisms.

Results

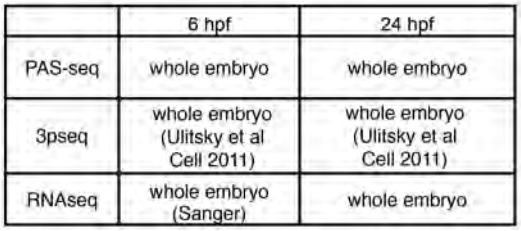
Establishment of True Positives and True Negatives

To develop an algorithm that could reliably distinguish true and false pA sites in oligo-dT primed deep sequencing data, we trained a naïve Bayes classifier. For this purpose, we defined training sets consisting of True Positive pA sites and True Negative sites from a combination of data sources. Given the demonstrated technical rigor of 3pseq, we utilized published datasets generated using this technique to build a True positive training set. An additional criterion of this set was presence of a pA site in both 3p-seq and PAS-Seq datasets from the same stage of zebrafish embryos (Figure 3.1A). We defined True Negatives as sequence fragments derived from oligo-dT primed RNA-Seq data with five terminal adenines or five proximal thymines that mapped to genomic sequence and were not present in 3pseq (Figure 3.1A). In both

Figure 3.1: Development of True Positive and True Negative training sets. A. RNA-seq reads starting with five thymines or five adenines were mapped to the genome. Reads that mapped were assigned to be putative sites of internal priming for the development of the True Negative training set. Reads that did not map had the terminal adenines or proximal thymines trimmed and were remapped. These reads were assigned as RNA-seq putative pA sites and were combined with PAS-Seq data for the development of the True Positive training set. The 22770 True Positives contain putative pA sites present in both 3pseq and RNA-seq putative pA sites/PAS-Seq (within +/- 10 nt). The 9219 True Negatives contain RNA-seq internally primed sites that are not present in 3pseq (within +/- 10 nt). B. True Positives and True Negatives had to be present at both 6 hpf and 24 hpf.



B

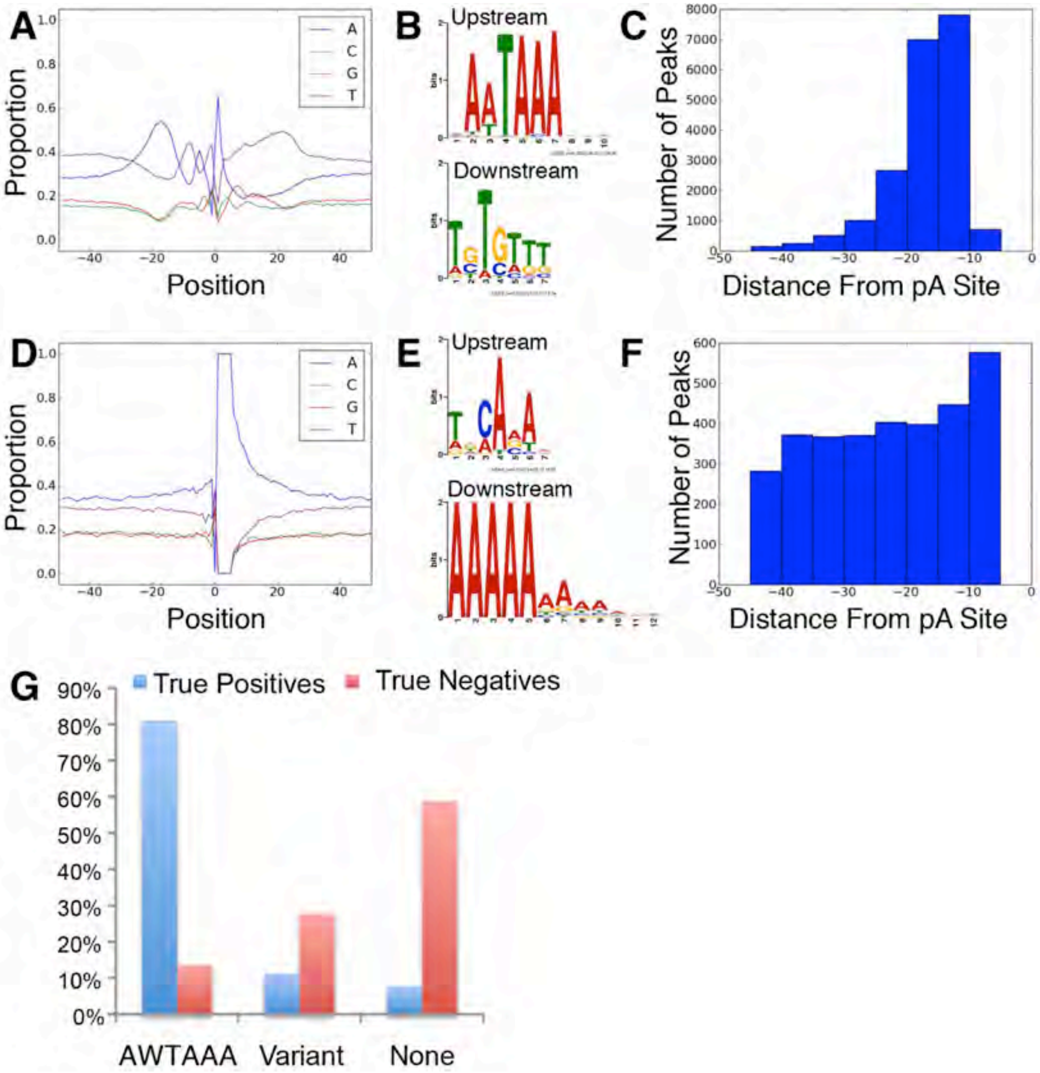


	6 hpf	24 hpf
PAS-seq	whole embryo	whole embryo
3pseq	whole embryo (Ulitsky et al Cell 2011)	whole embryo (Ulitsky et al Cell 2011)
RNAseq	whole embryo (Sanger)	whole embryo

cases, True Positives and True Negatives had to be present in both 6 hours post fertilization (hpf) and 24 hpf data sets (Figure 3.1B). In zebrafish, about 55% of the ~26,000 genes are thought to be alternatively polyadenylated [30]. Together, the training set consists of 22770 True Positives, representing 21902 genes, and 9219 True Negatives, representing 5391 genes, demonstrating sites from a majority of zebrafish coding genes are represented in our training data.

Prior to using these data for training, we examined the characteristics of flanking sequences up- and downstream of predicted cleavage sites in the True Positive and True Negative training sets. Analysis of nucleotide frequencies in genomic sequence flanking the True Positives demonstrated a prevalence of adenines and thymines upstream of the cleavage site (Figure 3.2A). Downstream of the pA site, we noted enrichment of thymines and a slight enrichment of guanines (Figure 3.2A), consistent with observations in yeast, plants, and metazoans [11, 70]. A search for over-represented motifs in True Positive sequences revealed a canonical PAS upstream of the cleavage sites and a consensus in the downstream region similar to that identified in [11] (Figure 3.2B). Furthermore, the cleavage distance (Figure 3.2C), measured from the 3' end of the PAS, clustered between 10 and 25 nt in the majority (87.5%) of True Positives with a canonical or variant PAS (Figure 3.2C), consistent

Figure 3.2: Training sets display characteristics of true pA sites and internally oligo-dT primed sites. A. Nucleotide composition of pA site-flanking sequences in the True Positive training set. B. Sequence logo of over-represented motifs identified in 50 nt upstream and 50 nt downstream of the True Positives. The canonical PAS (AATAAA) is identified upstream. C. Distribution of cleavage distance for consensus PASs of True Positive training set, measured from the 3' end of the PAS. D. Nucleotide composition surrounding True Negative training set. E. Sequence logo of over-represented motifs 50 nt upstream and 50 nt downstream of the True Negatives. F. Distribution of cleavage distance for canonical or variant PASs of True Negative training set, measured from the 3' end of the PAS. G. PAS distribution in 50 nt upstream sequence of True Positive and True Negative training sets.

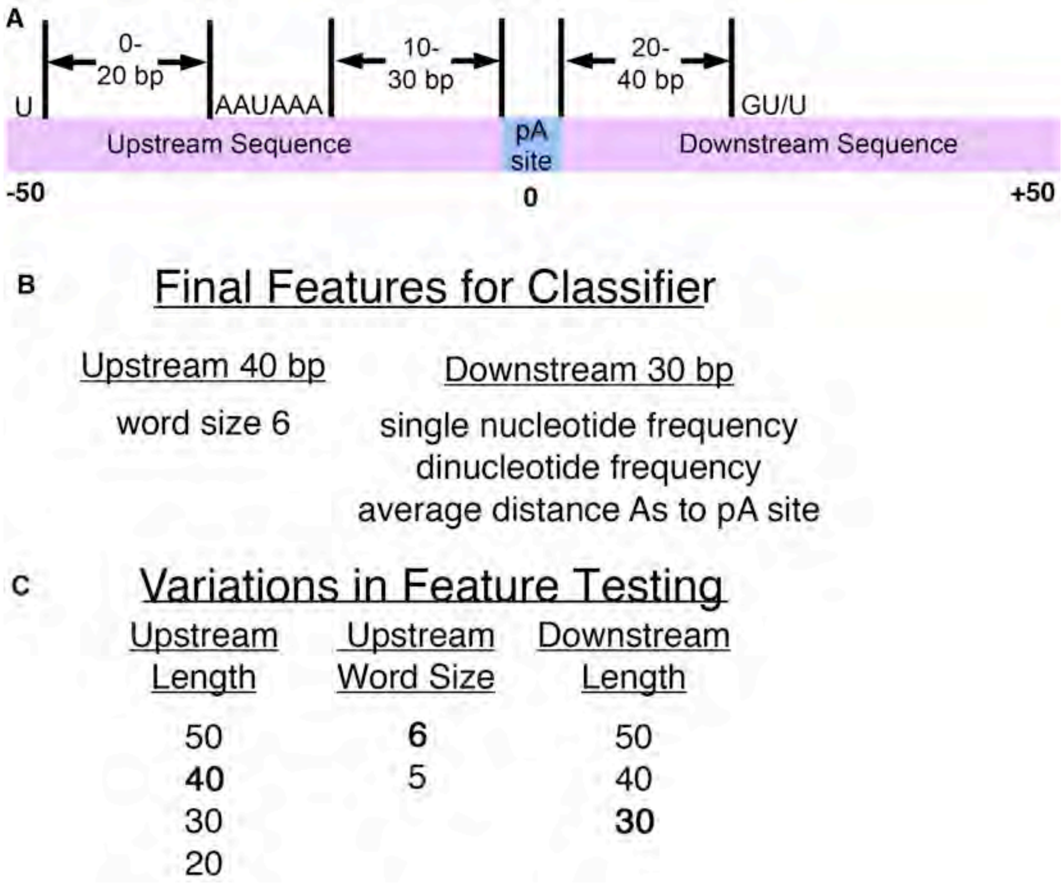

A
Proportion
Position
A
C
G
T
1.0
0.8
0.6
0.4
0.2
0.0
−40
−20
0
20
40
B
Upstream
Downstream
bits
C
Number of Peaks
Distance From pA Site
8000
7000
6000
5000
4000
3000
2000
1000
0
−50
−40
−30
−20
−10
0
D
Proportion
Position
A
C
G
T
E
Upstream
Downstream
bits
F
Number of Peaks
Distance From pA Site
600
500
400
300
200
100
0
−50
−40
−30
−20
−10
0
G
True Positives
True Negatives
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
AWTAAA
Variant
None

with previously published results [64]. In contrast to the True Positive set, sequences flanking True Negative sites showed a relatively constant nucleotide frequency upstream of the pA site, with an enrichment for adenines in the first five nt downstream (Figure 3.2D), as expected based on the criteria used to build this dataset (see Materials and Methods). Searches for over-represented motifs failed to reveal elements resembling a canonical PAS upstream of the oligo-dT priming site in the True Negative set, while downstream the stretch of adenines used to define these sequences are clearly evident (Figure 3.2E). Accordingly, only a small fraction of sequences from the True Negative training set contain a canonical or variant PAS (41.2%) and those that do fail to cluster at a defined distance proximal to the oligo-dT priming site (Figure 3.2F, G). Thus, our True Positive and True Negative training data represent two distinct populations that should clearly model the difference between true pA sites and internally primed sites.

Algorithm and Parameter Tuning

We incorporated sequence elements known to flank pA sites as features for building a naïve Bayes classifier (Figure 3.3A). Canonical and variant hexameric PASs can be identified 10-30 nt upstream of the pA site [177], while uracil-rich elements can be found 0-20 nt upstream of the PAS [177].

Figure 3.3: pA site features used for algorithm training. A. Schematic of region surrounding the pA site B. Features used in naïve Bayes classifier. C. Variations in upstream and downstream sequence length for finding features, and upstream word size. Boldface denotes features used for all subsequent training.



Final Features for Classifier

Upstream 40 bp	Downstream 30 bp
word size 6	single nucleotide frequency dinucleotide frequency average distance As to pA site

Variations in Feature Testing

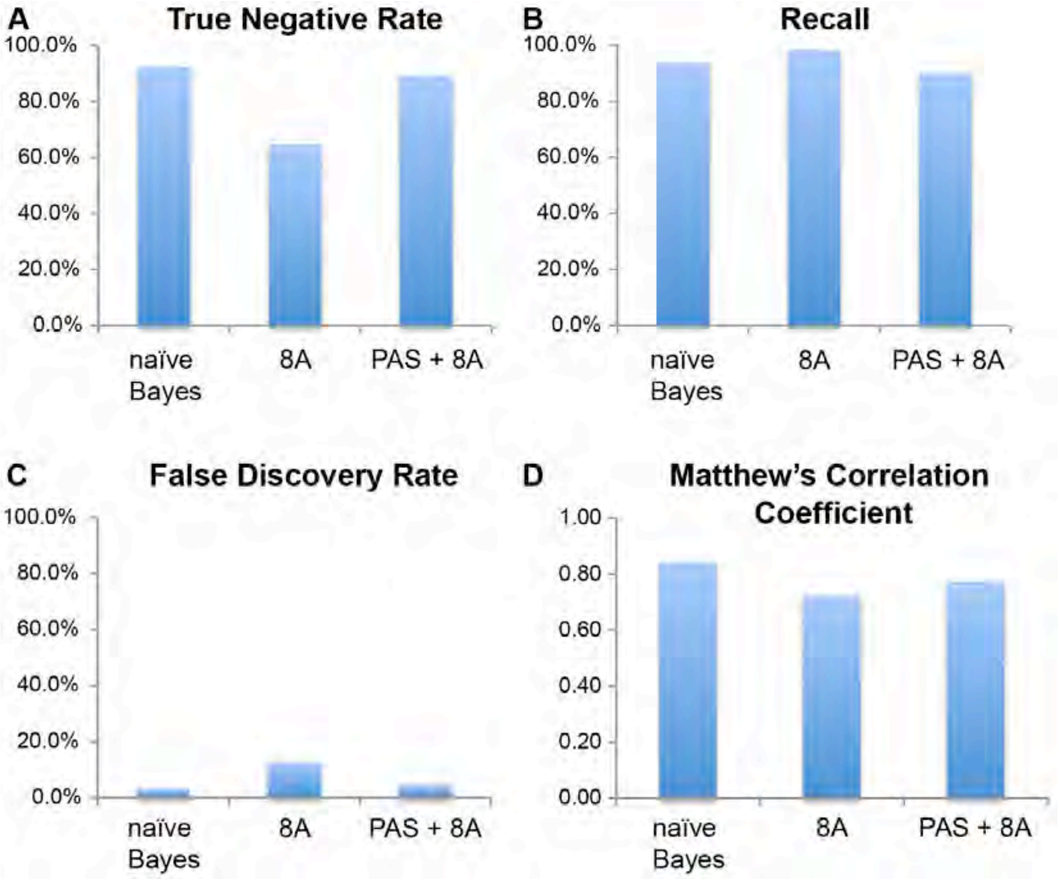
Upstream Length	Upstream Word Size	Downstream Length
50	**6**	50
40	5	40
30		**30**
20		

For that reason, we used all combinations of 6 nt (word size 6) as features in the upstream sequence region to allow for self-discovery of potential PAS and uracil-rich elements by the algorithm (Figure 2.S2B). Guanine/uracil- or uracil-rich elements 20-40 nt downstream of a pA site help direct cleavage and polyadenylation [177], while the presence and location of adenine richness downstream may indicate internal oligo-dT priming. Therefore, features in the downstream sequence region included mono- and di-nucleotide frequencies as well as the average distance of the adenines to the pA site (Figure 2.S2B). After varying the upstream (20-50 nt) and downstream (30-50 nt) sequence length for the features described above (Figure 2.S2C), using 30 or 40 nt of upstream sequence and 30 nt of downstream sequence consistently outperformed other models, though the variability between the different models was low (Appendix II). We chose to use 40 nt of upstream sequence, as not to miss any possible PASs in the upstream region due to slight variations in cleavage site usage [178]. To develop and test the functionality of the naïve Bayes classifier to identify true and false pA sites, we randomly sampled 70% of the training set to build the classifier (training) and the remaining 30% to evaluate the performance (cross-validation) and averaged the results of 10 trials. Following training, we found that the naïve Bayes classifier was capable of recalling 92.2% of True Negatives

Table 3.1. Performance measurement from naïve Bayes classifier and indicated heuristic filters.

	naïve Bayes	8A only	PAS + 8A
True Negative Rate	0.922	0.645	0.891
Recall	0.938	0.984	0.899
False Discovery Rate	0.032	0.127	0.047
Matthew's Correlation Coefficient	0.843	0.722	0.773
Precision	0.968	0.873	0.953
F-score	0.953	0.925	0.926
Accuracy	0.934	0.886	0.897
False Positive Rate	0.078	0.355	0.109

Figure 3.4: The trained algorithm outperforms previously published heuristic filters. Performance metrics for naïve Bayes classification compared to previously published heuristic filters, 8A or PAS+8A (see text for description of filters). A. True Negative Rate. B. Recall. C. False Discovery Rate. D. Matthew's Correlation Coefficient.


A
True Negative Rate
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
naïve Bayes
8A
PAS + 8A
B
Recall
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
naïve Bayes
8A
PAS + 8A
C
False Discovery Rate
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
naïve Bayes
8A
PAS + 8A
D
Matthew's Correlation Coefficient
1.00
0.80
0.60
0.40
0.20
0.00
naïve Bayes
8A
PAS + 8A

(True Negative Rate, Table 3.1, Figure 3.4A) and 93.8% of True Positives (Recall, Table 3.1, Figure 3.4B). Furthermore, the naïve Bayes classifier incorrectly categorized only 3.2% of predicted positives (False Discovery Rate, Table 3.1, Figure 3.4C). (Matthew's correlation coefficient (MCC), a balanced measure of true positives, false positives, true negatives and false negatives [179, 180], was calculated to be 0.84 (Matthew's Correlation Coefficient, Table 3.1, Figure 3.4D).

To compare the effectiveness of the naïve Bayes classifier with previously published methods, we categorized the training set with heuristic filters. For this purpose, we applied a heuristic filter (referred to hereafter as 8A) in which putative pA sites with 8 or more adenines in the 10 nt downstream of the pA site were assigned as false. We also combined this heuristic approach with a requirement for a canonical or variant PAS in the 40 nt upstream sequence (referred to as PAS+8A), an updated version of the criteria used in [181]. The 8A filter recalled only 64% of True Negatives while PAS+8A recalled 89% (True Negative Rate, Table 3.1, Figure 3.4A). Although the 8A recalled True Positives at a higher rate than the classifier (Recall, Table 3.1, Figure 3.4B), 8A and PAS+8A incorrectly called more predicted positives (False Discovery Rate, Table 3.1, Figure 3.4C). Comparison of MCC values revealed that the naïve Bayes classifier performs significantly better than either heuristic filter (Matthew's Correlation Coefficient, Table 3.1, Figure 3.2D).

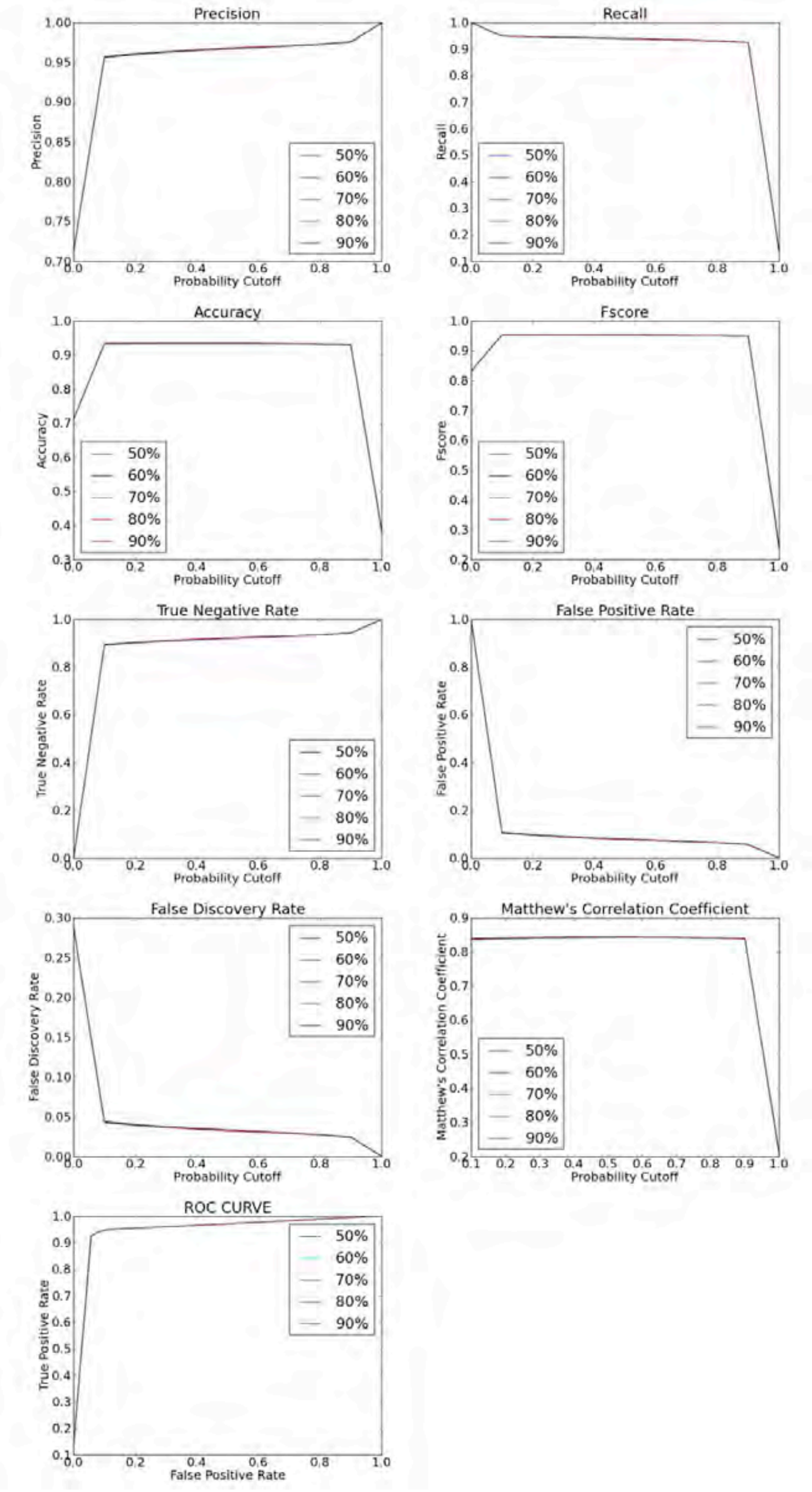
It is possible that the size of the training set may lead to over-fitting due to biased sequence composition. Therefore, we examined whether training set size influenced the predictive performance of the algorithm by varying the fraction of the training set used for algorithm training and subsequent cross validation. Algorithm performance for a variety of metrics was similar using 50%, 60%, 70%, 80%, or 90% of the True Positives and True Negatives for training (Figure 3.5), which indicates that our initial training set was of sufficient size.

Taken together, the naive 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.

Application to PAS-Seq data

To test the performance of the naïve Bayes classifier on a new data set, we used all True Positives and True Negatives for training the naïve Bayes classifier. We subsequently categorized unfiltered oligo-dT primed 3' end deep sequencing (PAS-Seq) data from 24 hpf zebrafish embryos as true or false (internally primed). Following classification, we compared the output to the same dataset filtered with the 8A heuristic filter (see

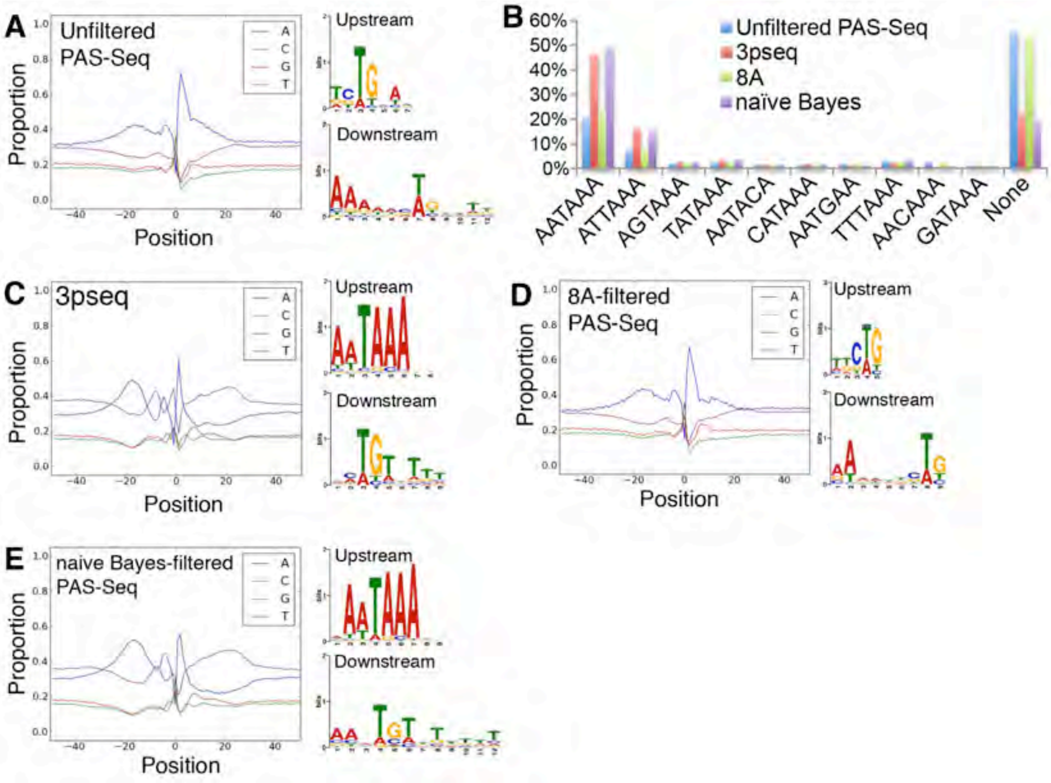
Figure 3.5: Algorithm for training set size variation. Precision, Recall, Accuracy, F-score, True Negative Rate, False Positive Rate, False Discovery Rate, Matthew's Correlation Coefficient, and Receiver Operating Curve (ROC) displayed for different sizes of training sets.


Precision
Precision
Probability Cutoff
Recall
Recall
Probability Cutoff
Accuracy
Accuracy
Probability Cutoff
Fscore
Fscore
Probability Cutoff
True Negative Rate
True Negative Rate
Probability Cutoff
False Positive Rate
False Positive Rate
Probability Cutoff
False Discovery Rate
False Discovery Rate
Probability Cutoff
Matthew's Correlation Coefficient
Matthew's Correlation Coefficient
Probability Cutoff
ROC CURVE
True Positive Rate
False Positive Rate
50%
60%
70%
80%
90%

above) or to unfiltered 3pseq data from zebrafish embryos at the same stage [182], which should contain only bona fide pA sites.

Analysis of the genomic sequence composition flanking 3' ends from unfiltered PAS-Seq data shows enrichment for adenines upstream and downstream of the pA site (Figure 3.6A), similar to our True Negative training set (see Figure 2.2D, E) ($r_A = 0.89$, $r_C = 0.78$, $r_G = 0.79$, $r_T = 0.76$). Furthermore, we failed to identify a canonical PAS as an over-represented motif upstream of the putative pA site in unfiltered PAS-Seq data (Figure 2.6A). Indeed, only 20.6% of the putative pA sites contain AATAAA upstream of the putative cleavage site and 55.4% have no identifiable PAS (Figure 3.6B). By contrast, the nucleotide profile of the 3pseq data is highly correlated with our original True Positive set ($r_A = 0.99$, $r_C = 0.99$, $r_G = 0.99$, $r_T = 0.99$) and shows enrichment of adenines ~20 and 5 nt upstream, enrichment of thymines ~10 nt upstream, and thymine richness downstream of the cleavage site (Figure 3.6C). Accordingly, a canonical PAS is easily recognizable in sequence upstream of the putative cleavage site while the downstream motif displays preference for thymines and guanines (Figure 3.6C), characteristics of known pA sites [6, 177]. In contrast to unfiltered PAS-Seq data, 46.5% of the 3pseq data set has AATAAA in the 40 nt upstream and only 22.2% have no identifiable PAS (Figure 3.6B). The majority (86.2%) of PASs is located 10 to 25 nt

Figure 3.6: Algorithm-filtered PAS-Seq 3’ ends resemble those identified by 3pseq. Sequence characteristics displayed by nucleotide profile, 40 nt upstream sequence logo, and 30 nt downstream sequence logo in A. 24 hpf unfiltered PAS-Seq, C. 24 hpf 3pseq, D. 8A filtered 24 hpf PAS-Seq, and E. 24 hpf PAS-Seq filtered by naïve Bayes classifier. B. PAS distribution for unfiltered 24 hpf PAS-Seq, 24 hpf 3pseq, 8A filtered 24 hpf PAS-Seq, naïve Bayes classified 24 hpf PAS-Seq.


A
Unfiltered
PAS-Seq
Proportion
Position
A
C
G
T
Upstream
Downstream
B
60%
50%
40%
30%
20%
10%
0%
Unfiltered PAS-Seq
3pseq
8A
naïve Bayes
AATAAA
ATTAAA
AGTAAA
TATAAA
AATACA
CATAAA
AATGAA
TTTAAA
AACAAA
GATAAA
None
C
3pseq
Proportion
Position
Upstream
Downstream
D
8A-filtered
PAS-Seq
Proportion
Position
Upstream
Downstream
E
naive Bayes-filtered
PAS-Seq
Proportion
Position
Upstream
Downstream
1.0
0.8
0.6
0.4
0.2
0.0
−40
−20
0
20
40

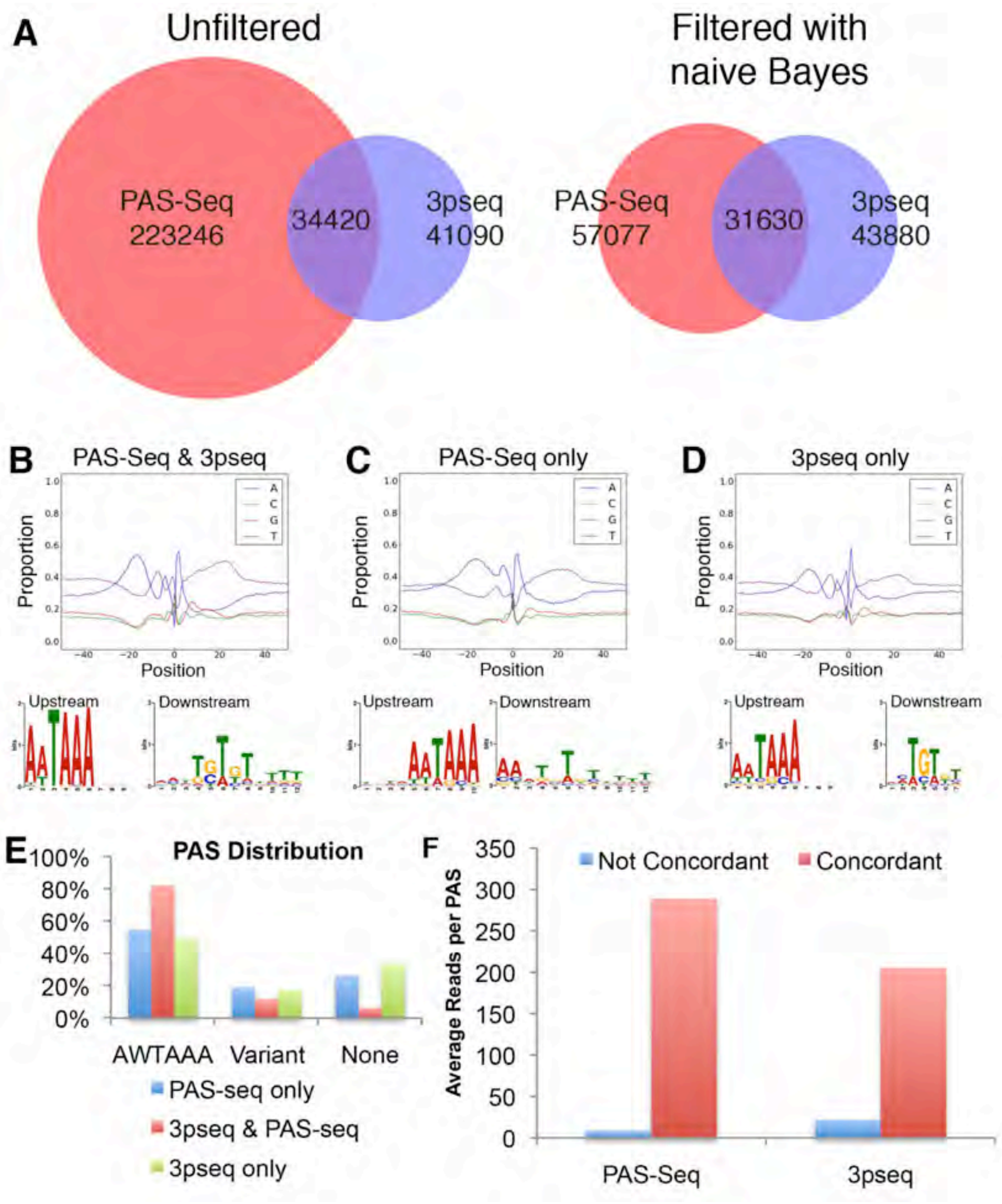
upstream of the putative pA site (data not shown). Consistent with these differences, the 3pseq and PAS-Seq nucleotide profiles are poorly correlated (r_A = 0.28, r_C = 0.06, r_G = 0.48, r_T = 0.45). Together, these results indicate that the PAS-Seq dataset likely contains a high proportion of sequences derived from internal oligo-dT priming.

Categorizing PAS-Seq using the 8A heuristic filter or the naïve Bayes classifier generated drastically different results. The 8A filter classified 18.1% of putative pA sites as false, leading to slightly better correlation of sequence composition of the remaining sites with those from 3pseq data (r_A = 0.45, r_C = 0.12, r_G = 0.65, r_T = 0.61). However, the adenine richness in the downstream sequence region and lack of enrichment of consensus PAS suggest internally oligo-dT primed sites remain called as positives by the 8A filter (Figure 3.6D). By contrast, the naïve Bayes classifier calls 65.4% of putative pA sites from PAS-Seq as false and the nucleotide profile of the putative pA sites predicted to be true resembled that of 3pseq (Figure 3.6E; r_A = 0.83, r_C = 0.68, r_G = 0.85, r_T = 0.86). Furthermore, a canonical PAS, AATAAA, was easily detected as over-represented upstream of the putative cleavage site, and thymines were enriched in the downstream flanking sequence (Figure 3.6E). In agreement with PAS distributions in genome-wide studies of pA sites [14, 55, 56], 49% of the putative pA sites contain AATAAA and only 19.2%

have no PAS (Figure 3.6B). Therefore, our naïve Bayes classifier successfully removes more contaminating sites of internal priming than the 8A filter, resulting in a set of putative pA sites that closely resembles 3pseq both qualitatively and quantitatively.

Comparison of the proportion of pA sites common to both PAS-Seq and 3pseq datasets in 24 hpf zebrafish embryos revealed a sizeable increase from 13.0% to 35.7% after filtering the PAS-Seq data with the naïve Bayes classifier (Figure 3.7A). Not surprisingly, pA sites common to both sets exhibit characteristics similar to our True Positive training set (Figure 3.7B). In addition, we noted that sequences flanking pA sites present in only 3pseq or PAS-Seq datasets also display characteristics observed in the True Positive training set (Figure 3.7C, D), although they show a slightly higher proportion of variant PAS usage (Figure 3.7E). The unique appearance of these sites in only 3pseq or PAS-Seq datasets may be due to tissue-specific 3'UTR isoforms that are expressed at low levels in the whole embryo and thus may not be consistently detected at this sequencing depth. Indeed, common pA sites present in both PAS-Seq and 3pseq have significantly ($p < 2.2e-16$) more sequencing reads contributing to them compared to either PAS-Seq (mean of 288.69 vs. 8.85) or 3pseq (mean of 204.94 vs. 21.4) alone (Figure 3.7F). Other technical issues may also contribute to a pA site being uniquely found in either dataset. For

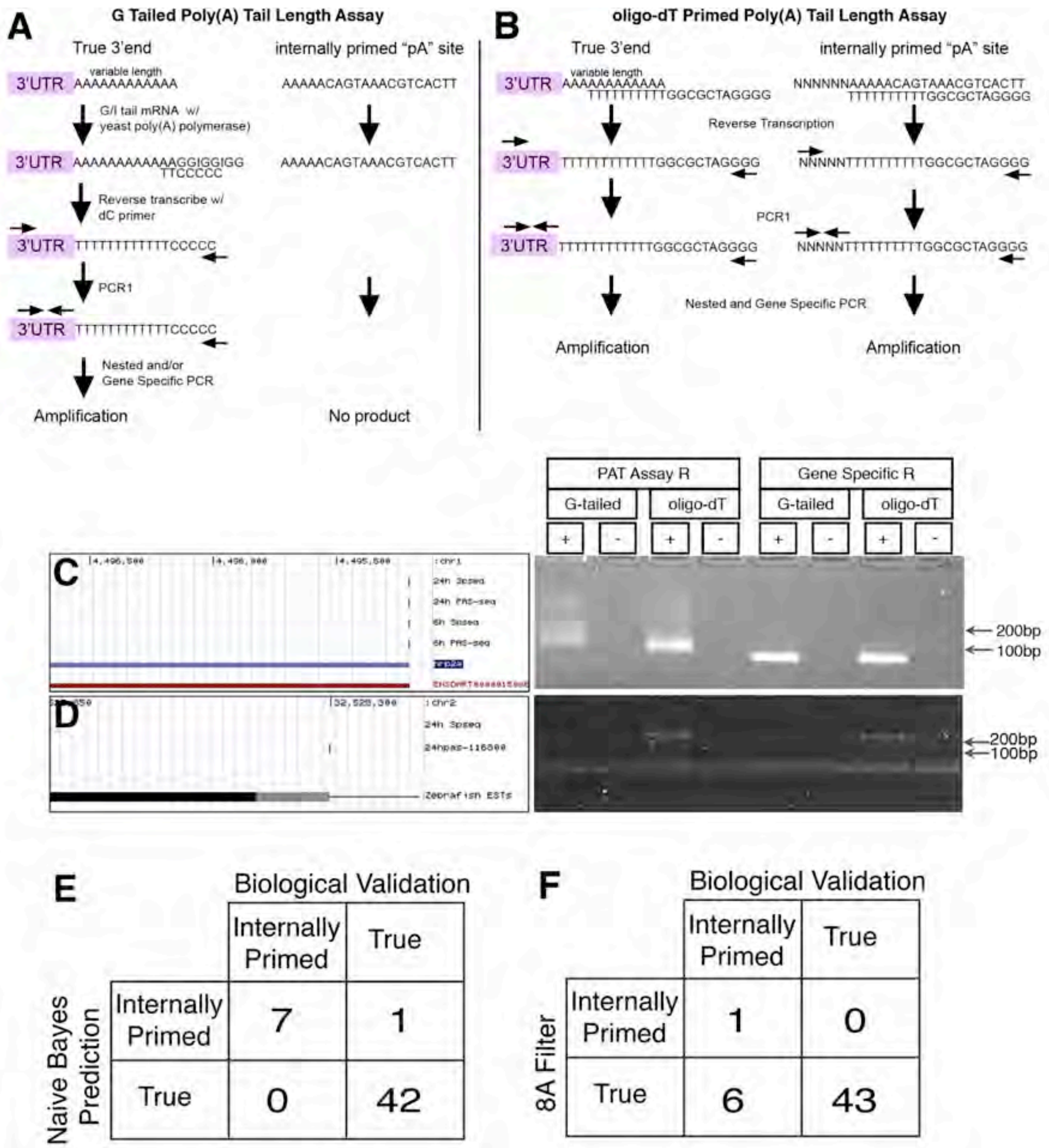
Figure 3.7: Comparison of raw and filtered PAS-Seq 3' ends with those from 3pseq. A. Overlap of 24 hpf zebrafish putative pA sites from PAS-Seq and 3pseq before and after filtering of PAS-Seq by the naïve Bayes classifier. B-D. Nucleotide composition graphs, and sequence logos for over-represented motifs 40 nt upstream and 30 nt downstream of pA sites B. common to PAS-Seq and 3p seq, or uniquely found in C. PAS-Seq or D. 3pseq datasets only. E. PAS distribution F. Mean number of sequencing reads contributing to a putative pA site.


A
Unfiltered
PAS-Seq
223246
34420
3pseq
41090
Filtered with naive Bayes
PAS-Seq
57077
31630
3pseq
43880
B
PAS-Seq & 3pseq
C
PAS-Seq only
D
3pseq only
Proportion
Position
A
C
G
T
1.0
0.8
0.6
0.4
0.2
0.0
-40
-20
0
20
40
Upstream
Downstream
E
PAS Distribution
100%
80%
60%
40%
20%
0%
AWTAAA
Variant
None
PAS-seq only
3pseq & PAS-seq
3pseq only
F
Average Reads per PAS
350
300
250
200
150
100
50
0
Not Concordant
Concordant
PAS-Seq
3pseq

example, internal oligo-dT binding may block extension from an oligo-dT bound to the poly(A) tail [103], thus inhibiting identification of true 3' ends in adenine-rich genomic loci in PAS-Seq. Finally, there may be differences in PAS usage due to polymorphisms between the different zebrafish strains used to generate the PAS-Seq and 3pseq datasets [49].

To biologically cross-validate our *in silico* predictions, we conducted two separate poly(A) tail length (PAT) assays to verify a putative pA site as true (polyadenylated) or false (internally primed). 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, which is then reverse transcribed with an anchored poly-cytosine primer (Figure 3.8A). Alternatively, an oligo-dT containing primer was used for reverse transcription (dtPAT; Figure 3.8B) [183]. 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 using gene-specific forward and reverse primers to amplify fragments without the poly(A) tail (Figure 3.8A,B). Due to different poly(A) tail lengths or variable oligo-dT binding along the poly(A) tail, validation of a true 3' end will result in a smear on a 2% agarose gel in both assays (Figure 3.8A-C). Conversely, an oligo-dT internally primed site will result in no product in the GPAT assay, but will yield a single product in the

Figure 3.8: Biological validation of filtered 3’ ends. A, B. Schematics depicting A. G Tailed poly(A) Tail Length Assay and B. oligo-dT Primed poly(A) Tail Length Assay. C. *Left*, UCSC genome browser screenshot of 3’end of *nrp2a* annotated by Ensembl (v68) and RefSeq in 6 hpf and 24 hpf PAS-Seq and 3pseq datasets. Right, GPAT and dtPAT assays for 3’ end of *nrp2a*. D. *Left,* UCSC genome browser (reversed to show negative strand in same orientation as C) shows a putative false pA site in an EST expressed in 24 hpf PAS-Seq but not 24 hpf 3pseq. Right, GPAT and dtPAT assay for 3’ indicated at left. C, D. “+”: reaction included reverse transcriptase; “-” : no reverse transcriptase. “PAT Assay R” denotes use of assay specific reverse primer. “Gene Specific R” denotes use of gene specific reverse primer. “G-tailed” or “oligo-dT” indicate the method by which the initial cDNA template was made, and which assay-specific reverse primer was used for the lanes labeled “PAT Assay R”. Total RNA from 24 hpf whole embryos was used for biological validation. Confusion matrices for biologically validated sites compared with E. naïve Bayes classifier or F. 8A filter



E

		Biological Validation	
		Internally Primed	True
Naive Bayes Prediction	Internally Primed	7	1
	True	0	42

F

		Biological Validation	
		Internally Primed	True
8A Filter	Internally Primed	1	0
	True	6	43

dtPAT assay (Figure 3.8A, B, D). For biological validation, we applied the GPAT and dtPAT assays to 50 putative poly(A) sites in the zebrafish genome defined as True or False by our classifier (Appendix III). 42 of these sites were called True by the classifier and 22 of these correspond to annotated 3'UTR ends (Zv9, ENSEMBL v68), while 20 represented possible novel 3' ends. Of the True sites, all were successfully amplified using the GPAT assay, indicating these are true polyadenylated 3' ends (Figure 3.8C, E; Appendix III). Notably, 9 of the novel 3' ends biologically validated as True were identified by 24 hpf PAS-Seq but not 3pseq. 13 validated True sites contained PASs other than the canonical AAUAAA within 40 nt upstream of the cleavage site, while 1 lacked any consensus motif, suggesting that our classifier can identify true 3' ends that lack a consensus PAS (Appendix III).

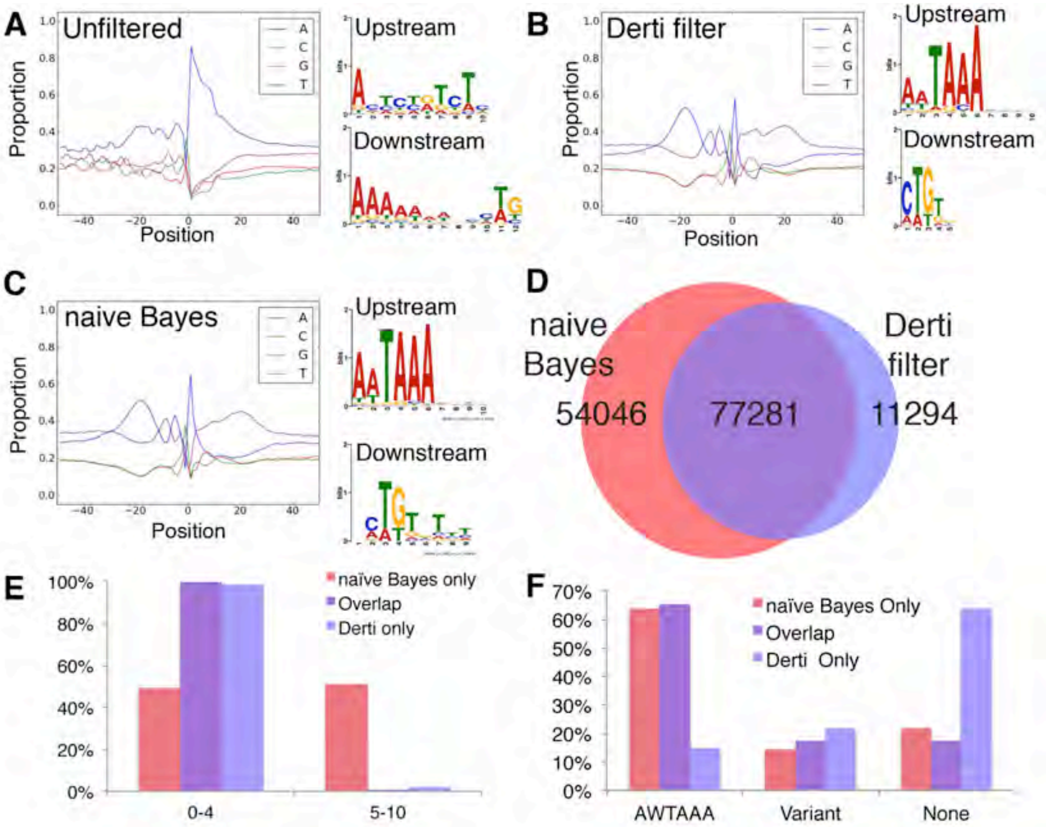
Along with the putative True set, we assayed eight sites that were classified as False. In this case, only one site was annotated as a 3' end in ENSEMBL (Appendix III). Half of these sites displayed a variant PAS in the vicinity of the putative 3' end, while the remaining sites contained no PAS 40 nt 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 (Appendix III; Figure 3.8D,F). Furthermore, six of these sites contained fewer than eight adenines in the downstream region and were called true by the 8A heuristic filter (Figure 3.8F; Appendix III). One False site, which did not possess a consensus PAS and contained only three downstream adenines, was amplified by the GPAT assay (Figure 3.8E). 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.

Naïve Bayes classifier displays utility in other species

The *in silico* and biological validation described above clearly demonstrates the utility of our naïve Bayes classifier to identify true pA sites from zebrafish PAS-Seq datasets. To determine if our algorithm, which was trained using zebrafish datasets, could be applied to similar data from other species, we used it to filter 3' end sequences generated by polyA-seq with an anchored oligo-dT(10) primer on RNA from human, rhesus, dog, rat and mouse [45]. In this particular study, Derti et al developed an empirical model to identify likely internally primed sequence fragments by constructing 3' end libraries using an unanchored oligo-dT (10) primer [45]. In this case, reads that mapped to genomic sequence with at least three terminal adenines indicated internal priming while those that did not map with at least three terminal adenines were deemed true pA sites [45]. The distribution of nucleotide frequencies downstream of the putative cleavage site was then compared between true pA sites and likely internal sites to derive a positional discriminant filter (referred to hereafter as Derti filter). This filter was then applied to the sequence 10 nt downstream of the cleavage site to distinguish true pA sites from oligo-dT primed artifactual pA sites in polyA-seq [45]. On the defined training set, this filter demonstrated 85% sensitivity and 97.5% specificity. Therefore, we filtered the same dataset with our naïve Bayes classifier to compare its utility to the Derti filter.

Analysis of unfiltered polyA-seq data from human kidney, failed to reveal a PAS over-represented upstream of putative pA sites, and identified strong adenine enrichment downstream of the pA site, consistent with a significant contribution of mapped reads from internal oligo-dT priming (Figure 3.9A). From approximately half million putative 3' ends in the unfiltered data, we found that the Derti filter calls 94,945 of these as true pA sites, approximately half the number of transcript ends annotated in human Ensembl v70. Accordingly, a canonical PAS is enriched upstream and guanines and thymines are enriched downstream of these filtered sites (Figure 3.9B), as expected of true pA sites. Using the naïve Bayes classifier identified more than 130,000 pA sites from the

Figure 3.9: 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. called true by Derti et al. positional discriminant function [45], or C. assigned as true by naïve Bayes classifier. D. Overlap of putative pA sites called true by naïve Bayes classifier and Derti et al. positional discriminant function [45]. 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.


A
Unfiltered
Proportion
Position
A
C
G
T
Upstream
Downstream
B
Derti filter
Proportion
Position
A
C
G
T
Upstream
Downstream
C
naive Bayes
Proportion
Position
A
C
G
T
Upstream
Downstream
D
naive Bayes
54046
77281
Derti filter
11294
E
naïve Bayes only
Overlap
Derti only
100%
80%
60%
40%
20%
0%
0-4
5-10
F
naïve Bayes Only
Overlap
Derti Only
70%
60%
50%
40%
30%
20%
10%
0%
AWTAAA
Variant
None

same dataset and these sites also exhibited expected characteristics for true polyadenylated 3' ends (Figure 3.9C). Between the two filters, we identified approximately 77,000 commonly assigned polyA sites (Figure 3.9D). To determine the discrepancies in uniquely called pA sites, we more carefully analyzed the differences in these sequences. 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 3.9E), consistent with the focus of this filter on nucleotide frequencies in the downstream region. 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 [45]. Importantly, the majority of sites uniquely identified by our classifier possess a canonical PAS, suggesting that they are true 3' ends (Figure 3.9F). By contrast, the majority of true sites uniquely identified by the Derti filter did not display a PAS in the upstream region (Figure 3.9F), 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. We would point out that our naïve Bayes classifier successfully identifies pA sites lacking a PAS at a rate similar to unfiltered 3pseq, at least on zebrafish data (see Figure 3.7). 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 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 10 base downstream mononucleotide frequencies.

The application of a trained naïve Bayes classifier is clearly beneficial to identify true pA sites from oligo-dT primed 3’ end sequencing data from other animals. Its usage could also be extended to RefSeq and other established 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 [103] and our own analysis demonstrated that the 3’ end of *vegfc,* as annotated by ENSEMBL, is due to mis-priming (Appendix III). 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. Though sequence elements important for polyadenylation are conserved, the PAS appears to have less importance in yeast and plants. As oligo-dT primed 3’ end sequencing and DRS have been performed on both yeast and *Arabidopsis*, similar True Positive and True Negative training sets to the ones we described above could be created.

Conclusions

Oligo-dT primed 3’ end library construction methods identify true pA sites as well as instances of internal priming. Attempts to polish this data by heuristic filtering produce a high rate of false positive and false negative pA sites. Using PAS-Seq, 3pseq [182], and RNA-seq data, we trained a naïve Bayes classifier to distinguish between true pA sites and oligo-dT internally primed sites based on sequence features flanking pA sites. Our algorithm outperforms other heuristic approaches and classifies pA sites in zebrafish, mouse, rat, dog, rhesus, and human with high accuracy. In summary, our method for separating internally oligo-dT primed pA sites from true 3’ ends will be of use in further genome-wide studies to identify novel cleavage and polyadenylation sites and examine alternative polyadenylation.

Materials and Methods

Zebrafish Care and Staging

Zebrafish were maintained as described in [164] and staged as described in [184]. 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 RNeasy 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) or ending with at least five adenines were mapped to the zebrafish genome (Zv9) using Bowtie [185] (Figure 3.1A). 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 3.1A). 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 map were trimmed of terminal adenines (or thymines) and re-mapped (Figure 3.1A). 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 2.1A) (see Training Sets).

3’ End Deep Sequencing Datasets

We constructed PAS-Seq libraries as described in [16], using barcoded adapters, and paired-end sequenced on an Illumina Hi-Seq (101 nt reads) with a custom sequencing primer described in [16] 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 [186]. Zebrafish 6 hpf and 24 hpf 3pseq [182] and mammalian polyA-seq alignments [45] were downloaded from the Gene Expression Omnibus (accession numbers GSE32880, GSE30198). cleanUpdTSeq 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.

Clustering of Deep Sequencing Reads into Putative pA Sites

A custom perl script clustered mapped sequencing reads into putative pA sites. Mapped sequencing reads were trimmed to the 3' most nt, as this would likely correspond to the site of cleavage and polyadenylation. Reads were clustered first for identically matching sites. A reiterative 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 whether the height is different between PAS-Seq and 3pseq concordant peaks and those present in one dataset alone [187]. Concordance or overlap between two 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 3.1A). Sites concordant between the PAS-Seq and the 3pseq data sets were assigned to the True Positive training set (Figure 3.1A). 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 3.1A). Only sites that were present in both the 6 hpf and 24 hpf data sets were used for training (Figure 3.1B). 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 frequency; downstream dinucleotide frequency; average distance of downstream adenines to the pA site (Figure 3.3B). 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 [188] 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.

Performance Metrics

Precision, 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.

$$\Pr ecision = \frac{TP}{FP + TP}$$

$$\mathrm{Re}call = \frac{TP}{FN + TP}$$

$$TNR = \frac{TN}{TN + FP}$$

$$FDR = \frac{FP}{FP + TP}$$

$$FPR = \frac{FP}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TN + FN + FP + TP}$$

$$F - score = \frac{2 \times precision \times recall}{precision + recall}$$

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

$$MisclassificationError = \frac{FP + FN}{TP + TN + FP + FN}$$

We calculated Pearson’s correlation coefficient, using R, to assess how correlated the nucleotide profiles were between the predicted PAS-Seq and the training sets [189, 190].

Model Selection and Training Set Size

To evaluate the performance of the naïve Bayes classifier, 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 the average precision, recall, F-score, accuracy, true negative rate, false discovery rate, false positive rate and MCC from the 10 cross-validations of each model, using a probability of true cutoff of 0.5 (Appendix II). The top two models, using 30 or 40 nt of upstream sequence and 30 nt of downstream sequence, yielded very similar performance (MCC of 0.845 vs. 0.843). We chose 40 nt of upstream sequence and 30 nt of downstream sequence as to not miss any potential PASs due to cleavage site microheterogeneity [178]. Additionally, using a word size of 5 in the upstream region led to slightly decreased performance.

To evaluate training set size, we trained the classifier with different percentages of the training data and used the remainder for cross validation. The percentages of training data used and the number of total peaks for each are summarized below. We calculated the average precision, recall, F-score, accuracy, true negative rate, false discovery rate, false positive rate and MCC from 10 cross-validation trials and graphed these for the different probability cutoffs.

% of training data set used for training	# of peaks used for training
50%	15995
60%	19194
70%	22393
80%	25592
90%	28791

Performance of the Algorithm Compared to Heuristic Filters

To examine the performance of the algorithm, we used the average of 10 cross validation trials using a probability cutoff of 0.5 as described above. To compare the algorithm to previous heuristic filters, we used a perl script to separate all putative pA sites from the training set (True Positives and True Negatives) with at least 8 adenines in the 10 nt downstream (False) from those that did not (True). We used a perl script to add the additional requirement of a canonical or variant PAS in the 40 nt upstream to be called True.

PAS Distribution

To identify canonical and variant PASs used in zebrafish, we examined the 50 nt upstream of 3'ends annotated in Ensembl (v61) for overrepresented motifs using Multiple Em for Motif Elicitation (MEME) [191]. In subsequent zebrafish analyses, a perl script was used to search for a canonical or variant PAS in order of decreasing importance. For mammals, a perl script was used to search for a canonical or variant PAS in order of decreasing importance as denoted in polyA-seq data [45].

Motif Finding

Multiple Em for Motif Elicitation (MEME) was used to search for motifs enriched in the sequence upstream and downstream of the putative pA sites [191]. For the True Positive and True Negative training sets, 50 nt upstream for all of the sites was examined using MEME [191] with the following settings: -minw 5 -maxw 10 –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. The upstream sequence was searched using options (-minw 5 -maxw 10 -oops), and the downstream sequence was searched using options (-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 (Figure 3.8A) [192]. Subsequently, reverse transcription was performed with a poly-cytosine anchored primer (Figure 3.8A). Alternatively, we used an oligo-dT(10) primer to make cDNA (Figure 3.8B) [183]. 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: GGGGATCCGCGGTTTTTTTTTT [183]) (Figure 3.8A, B). Nested PCR was performed for 20, 25, 30, 35 cycles, so not to over amplify products, using 1 ul of a 1:50 dilution of the primary PCR as template, a nested forward primer and the assay-specific reverse primer (Figure 3.8A, B). 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.

CHAPTER IV

CONCLUSIONS AND FUTURE DIRECTIONS

Contained in this dissertation is a novel method to accurately classify true pA sites, with both canonical and variant polyadenylation signals, and internal priming events from simple, oligo-dT primed 3’ end sequencing. By reliably defining 3’ ends of transcripts, our method will facilitate examination of post-transcriptional gene regulation important in both development and disease processes. The current standard removes internal priming events by simple, heuristic filtering of sites with a high proportion of adenines in the genomic sequence flanking the putative pA site. Not only does the naïve Bayes classifier outperform these simple filters [15, 17, 53, 65, 99, 181, 193], it also shows utility in zebrafish, mouse, rat, dog, rhesus, and human. Importantly, the functions, training data, and documentation are available on the Lawson Lab website[3]. Each function has its own documentation that explains the usage and arguments. In addition, a detailed step-by-step user’s guide is available (Appendix IV). Thus, the naïve Bayes classifier developed in these studies is an easily applicable, highly accurate solution to remove false positives from oligo-dT primed 3’end sequencing data and will be of use in future genome-wide studies identifying pA sites.

Comparison of the Naïve Bayes Classifier and Derti Filter

[3] http://lawsonlab.umassmed.edu/cleanupdtseq.html

Comparison of putative pA sites identified by the naïve Bayes classifier or the Derti filter [45] showed that the majority of true pA sites were identified by both. However there were also discrepancies between the two methods. The naïve Bayes classifier predicted additional sites as true 3' ends. The Derti filter likely incorrectly identifies these sites as false because the Derti filter biases against sites with 5 or more adenines in the 10 nt downstream of a putative pA site. In this regard, the Derti filter may be no better than a simple, heuristic filter. In fact, almost 20% of 3pseq *C. elegans* pA sites, which should all be true 3' ends, were called false by the Derti filter, demonstrating it does create a large number of false negatives [45]. Conversely, the Derti filter identified a smaller proportion of sites as true, which were called false by the naïve Bayes classifier. The majority of these sites use no PAS, suggesting a large proportion of them may actually be false.

To further examine the validity of the naïve Bayes classifier in comparison to the Derti filter [45], at least two experiments could be conducted. Biological validation of ambiguous putative pA sites as conducted for PAS-Seq could definitively identify whether the naïve Bayes classifier or the Derti filter was more accurate. To more comprehensively examine the problem, after normalizing for sequencing depth, pA sites in human brain identified by polyA-seq could be compared with Direct RNA

Sequencing [52]. The naïve Bayes classifier or the Derti filter could be applied to the normalized pA site set and the performance of either method could easily be validated. These experiments would likely confirm the superior performance of the naïve Bayes classifier compared to the Derti filter.

Improvements to the Naïve Bayes Classifier

The naïve Bayes classifier performed significantly better than simple filters, however its performance may still be improved. Clarifications to the training data may lead to improvement, as would integration of additional training data. The high performance of the naïve Bayes classifier is likely due to the multiple features used for categorization. The consideration of additional features known to be important in cleavage and polyadenylation may indeed increase performance. Finally, a different supervised learning method may result in improved classification.

Additional training data could be generated with the same methodology as originally performed. 3pseq data is available for multiple zebrafish stages and specific tissues [56], only two of which were used. PAS-Seq could be performed to complement these. Additionally, 3pseq data from mouse embryonic stem cells was recently submitted to the

Gene Expression Omnibus (GSM1089084). PAS-Seq data has also been performed on mouse embryonic stem cells [16]. In combination with RNA-seq data, additional True Positives and True Negatives could be established. This may result in a more complete training set. Moreover, a sizeable increase in the number of training sites may also allow for an increased number of features that can be used for training.

Supplementing the original training data with biologically validated sites that may represent outliers would likely improve algorithm performance. The misclassification rate of the naïve Bayes classifier, or the fraction of incorrect predictions, is 6.63% (Appendix II). The misclassification rate suggests samples within the training data are mislabeled, which may be distorting the performance of the classifier. Though I established the True Positives and True Negatives with high confidence from deep sequencing data, I did not biologically validate them. Therefore it is possible a True Negative could be a true 3' end or vice versa. Thus identification of the misclassified sites followed by biological validation is imperative. The single misclassified biologically validated putative pA site (Appendix III) contained no PAS upstream. Additional biological validation of putative pA sites containing no PAS will be a gainful addition to the training data. Interestingly, all putative pA sites containing more than five adenines downstream were validated as internally primed. However, previous studies demonstrate the existence of true pA sites with a large proportion of adenines downstream [41]. Further biological validation of putative sites with a large proportion of adenines in the downstream region may also be a worthy addition. Integration of the biological validation of these three subsets into the original training data will likely improve performance of the naïve Bayes classifier.

Additional features may improve the performance of the naïve Bayes classifier, at the expense of increasing the complexity of training. I used the presence/absence of all hexamer permutations in the 40 nt upstream, in addition to the single and dinucleotide frequencies, and the average distance of adenines to the pA site in the 30 nt downstream. However, over-represented motifs have also been identified in the 40-100 nt upstream, 0-40 nt downstream, and 40-100 nt downstream [62]. Therefore, adding the presence or absence of hexamers or pentamers in these specific regions may provide additional beneficial information for classification. On the other hand, the sequence of the RNA recognition motif of CstF-64 diverges and this correlates with slightly different motifs identified in the region downstream of the pA site for different organisms. These are all uracil or uracil/guanine rich, suggesting the original features of single and dinucleotide frequency may model pA sites for a variety of organisms better. Tethering the distance to the cleavage site with each word found upstream could improve performance, as the PAS is generally located 20 nt upstream of the cleavage and polyadenylation site [54, 55, 58], and moving the position of the PAS has been shown to affect cleavage efficiency [64]. Open structure surrounding the PAS is important for the accessibility of CPSF to the PAS [194], while RNA pseudoknots or binding sites for *trans* acting factors in the auxiliary downstream region may help to stabilize or localize the 3' processing complex [25]. Therefore, the secondary structure of the pre-mRNA should be considered as an additional feature. More frequently used sites tend to use stronger *cis* elements, while less frequently used sites may use variant signals (Figure 3.4) [17, 18, 20, 41]. Thus, utilizing the peak height in relation to the total number of reads in a data set may allow for separate modeling of strong and weak pA sites.

Clearly, many features have been proposed that could alter the performance of the naïve Bayes classifier. Too many features may lead to overfitting, or creating a model that is too complicated for the training data. Therefore, to determine if the proposed additional features would result in a significant increase in performance, feature selection should be conducted. Feature selection eliminates features that may not have any additional predictive power from training. Additionally, different models may be tested to examine different combination of features. Indeed, model averaging (compared to the original model selection I performed), may also aid in successful integration of additional features. Instead of choosing the best performing model, the average of all models can be used if the initial training and cross validation is thought not to correctly model performance when the algorithm is applied to a set of unknowns. In fact, previous studies have shown increased performance with feature selection or model averaging [112], thus, these steps may result in increased performance.

Comparison of supervised learning algorithms showed other methods, such as random forests, outperformed naïve Bayes on average in 11 test cases [195]. Decision trees are a discrete classification method, in which a single feature is used as a branch point [108]. Due to this simple tree structure, these models are transparent and generally quick, though they require some "pruning" e.g. feature selection so the decision tree does not over-fit. Random forests is an adaptation of simple decision trees, in which many trees are grown and multiple, randomly selected features can be used at branch points [196]. The average of all the trees in the "forest" is the model and increasing the number of randomly generated trees increases the accuracy of the model. No pruning of trees is needed and overfitting is not an issue. Using a random forest, which contained 200 trees with 64 variables at each split, to model the True Positives and True Negatives yielded an out of bag error rate (equivalent to the misclassification rate of the naïve Bayes classifier) of 4.19% (Julie Zhu, unpublished data). Thus random forests may yield minimal performance increases over the naïve Bayes classifier.

Applications for the Naïve Bayes Classifier

Many biologists (including myself before this project) know nothing about scripting and bioinformatic analysis. Galaxy is a web server with a user-friendly graphical user interface for deep sequencing data analysis [197-199]. “Workflows” or protocols can be publicly shared among users. I could create a workflow on Galaxy combining the mapping capabilities of Galaxy, the original scripts I wrote for analysis of 3’ end sequencing data, the naïve Bayes functions, and the training data. Thus, a user could upload the raw deep sequencing data, utilize the workflow, and obtain a highly accurate set of putative pA sites. This would be the ultimate user-friendly solution.

We did not test the performance of the naïve Bayes classifier in yeast or plants. The divergence of polyadenylation signals in yeast, plants, and metazoans [40, 50], suggests the naïve Bayes classifier trained on zebrafish sites may not perform well. However, because DRS

[40, 52], which should contain no false positives, as well as oligo-dT primed 3'end deep sequencing [65, 66, 99] has been performed for both yeast and *Arabidopsis*, similar training sets to the ones described in the previous chapter could be generated. The naïve Bayes classifier could then easily be retrained.

The naïve Bayes classifier could be used to filter previously published 3' end data sets that originally used simple filtering based on the number of adenines in the downstream region. Applying the classifier to these studies will likely remove all false positives and result in identification of novel polyadenylation sites, as demonstrated in the previous chapter.

Current gene annotations are built, in part, from Expressed Sequence Tags (ESTs) generated using oligo-dT priming. Previous analysis estimated 12% of 3' labeled ESTs (dbEST human release 10/04/2001) are artifacts of internal oligo-dT priming [103]. A recent study identified adenine-rich motifs at the 3' ends of 15% of transcripts annotated in Ensembl v65 for mouse and human, suggesting internal priming does plague the current annotations [72]. Therefore the current 3' end genome annotations may be contaminated with sites of internal priming. Indeed, I biologically validated the annotated 3' end of *vegfc* in zebrafish as internally primed. Thus the naïve Bayes classifier will likely identify more internally primed sites from RefSeq and Ensembl transcript models. These sites should be biological validated, and the changes should be noted in a new set of transcript models.

Interestingly, similar proportions of True Negatives and True Positives are miscalled, suggesting the 3pseq data may have rare artifacts of internal priming. In fact, I did biologically validate one site identified by 3pseq as internally primed, which was correctly predicted by the naïve Bayes classifier (Appendix III). Thus, this method may even have utility in combination with 3pseq.

Tools for Studying Cleavage and Polyadenylation in Zebrafish

I could develop an *in vivo* method to examine cleavage and polyadenylation in zebrafish, as zebrafish have genetic conservation with mouse and human. Previously, the Lawson lab utilized the *Tol2* transposon system to develop a Tol2 plasmid that uses a bi-directional enhancer to express GFP as a control and mcherry fused to a 3'UTR of interest as a sensor [145]. The mcherry 3'UTR sensor uses the SV40 late poly(A) signal. Using a control 3'UTR downstream of the mcherry, the SV40 poly(A) signal could be substituted with other poly(A) signals to test whether they are sufficient for cleavage and polyadenylation. The GFP expression would still act as a control to screen for proper injection and efficient transgenesis. The mcherry with SV40 poly(A) signal would act as the control for cleavage and polyadenylation. Though mcherry expression may be able to be used as a read out of proper 3’ end processing, the GPAT assay should be performed on RNA from these embryos to confirm cleavage and polyadenylation. To start, 100 nt up- and down-stream of a putative pA site could be tested. If that is sufficient, smaller fragments of upstream or downstream sequence could be tested or mutations could be made in the sequence containing the putative polyadenylation elements. Additionally, short antisense oligonucleotides, known as morpholinos (MOs) [200], could be designed against putative polyadenylation elements in a specific pre-mRNA. It would be important to also include some 3’UTR specific sequence, as to only affect a specific pA site. The MOs could be injected in parallel with a control scrambled MO. The GPAT assay could be used to confirm if this method blocks 3’ end processing of a putative pA site. This assay should be optimized first on known signals, such as AAUAAA.

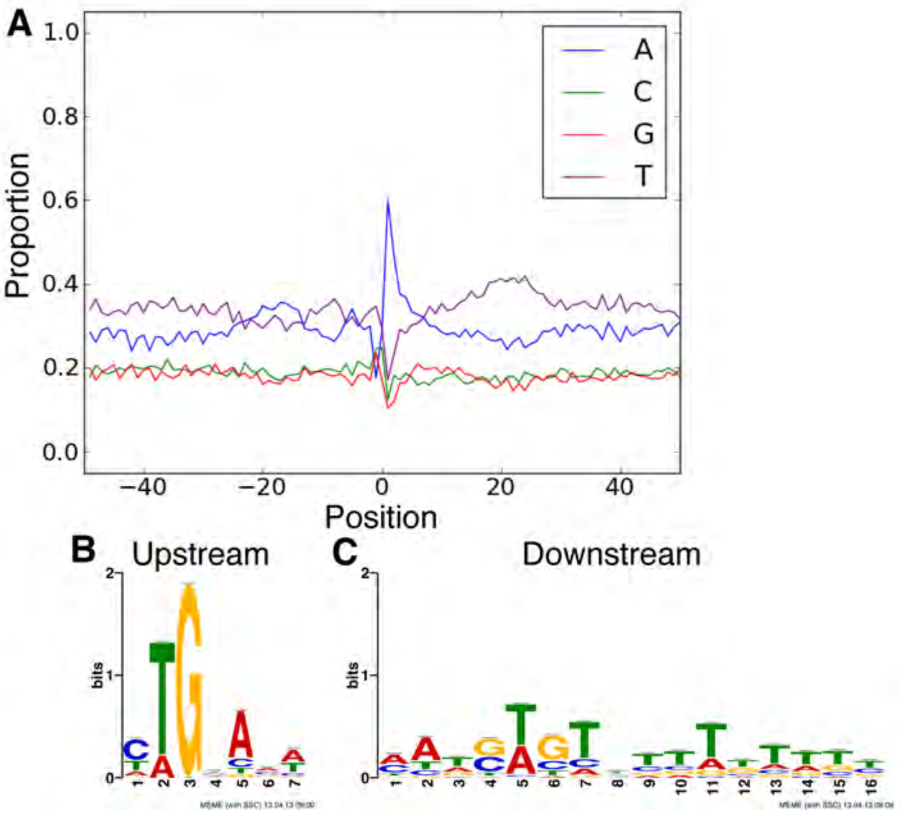
Non-canonical Polyadenylation

Approximately 20% of pA sites do not use a PAS, implying this may be a regulated mechanism rather than a stochastic event. In some genes, such as *JUNB*, only an adenine-rich sequence, rather than a canonical

PAS, and strong uracil-rich DSE are needed for 3’ end processing [48]. Examination of over 10,000 human pA sites, represented by ESTs containing at least 30 adenines in the poly(A) tail, demonstrated 5-10% of pA sites contain a hexamer with at least 5 adenines (not including AAUAAA and not overlapping with AWUAAA) in the 40 nt upstream [48]. pA sites with adenine-rich sequences upstream were significantly enriched for uracil or uracil/guanine elements. In the same regard, sites identified by CstF-64 CLIP-seq containing AAUAAA upstream contained uracil/guanine rich motifs, while sites not using AAUAAA contained uracil rich motifs [23]. These results suggest strong downstream sequence elements may be needed for regulation of non-canonical pA sites. Moreover, analysis of tissue-specific RNA-seq data demonstrated pA sites with adenine rich sequences were more likely to be expressed tissue-specifically compared to pA sites using AWUAAA [48], indicating these may play a role in differentiation. Also, UGUAN elements upstream of the pA site in *PAPOLG* were important for cleavage and polyadenylation, likely by recruiting CFI [31].

There are 1754 pA sites with no PAS in the 50 nt upstream in the True Positives training set which may allow us to investigate this mechanism in zebrafish. In these sites, there is a smaller enrichment of adenines 20 nt upstream and thymines downstream (Figure 4.1),

Figure 4.1: Sequence composition of True Positives with no PAS in the 50 nt upstream. A. Nucleotide composition pA site-flanking sequences. Sequence logo of over-represented motifs identified in B. 50 nt upstream and C. 50 nt downstream.


A
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
−40
−20
0
20
40
Position
A
C
G
T
B
Upstream
C
Downstream
bits
2
1
0
1
2
3
4
5
6
7
MEME (with SSC) 13.04.13 09:00
8
9
10
11
12
13
14
15
16
MEME (with SSC) 13.04.13 09:04

compared to all True Positives (Figure 3.1). Adenines compose a larger fraction of the downstream sequence region in the True Positives with no PAS (Figure 4.1) compared to all True Positives (Figure 3.1). The region upstream of the True Positives with no PAS contained a similar motif as the unfiltered PAS-Seq data (Figure 4.1, Figure 3.3), while downstream was enriched for thymines (Figure 4.1).

The thymine richness downstream previously identified in non-canonical pA sites [48] and identified in the True Positives containing no PAS upstream suggests CstF may play a role, as this protein is known to bind uracil rich sequences [20, 30, 46, 63]. Interestingly, knockdown of CstF increased distal pA site usage [23, 201]. Zebrafish have *cstf1, cstf2, cstf3,* and *cenpi*, an orthologue of mouse *Cstf2*. *In situ hybridization* and whole mount immunostaining should be performed to examine the expression of these proteins in multiple stages of development, as tissues may have differential expression of CstF [202]. CLIP-seq for either or all of these proteins, in combination with PAS-Seq followed by naïve Bayes classification of pA sites, could be performed to investigate if these proteins are enriched at pA sites with no PAS. If CstF regulates 3’ end processing at sites with no PAS, loss of CstF may result in a decreased proportion of pA sites with no PAS. To test this hypothesis, transcription activator-like effector nucleases (TALENs) could be used to knock-out other *cstf* genes in mutants for *cstf3* or *cenpi* [203]. PAS-Seq followed by naïve Bayes classification would be performed to define pA sites. Then, the PAS distribution could be compared between *cstf* mutants and controls.

To identify other proteins that may be involved in regulating cleavage and polyadenylation of sites with no PAS, the sequences surrounding True Positives with no PAS should be further examined. It may be important to look at multiple sequence regions, as performed for human in [62]. If additional motifs are identified, they could be searched against databases of known motifs to perhaps identify novel proteins that are involved in cleavage and polyadenylation.

These studies may reveal mechanisms regulating the processing of pA sites with no PAS. While the algorithm is able to correctly distinguish putative pA sites with variant PASs, it also has the most trouble correctly predicting sites with no PAS and few numbers of adenines downstream. Importantly, any insight into how cleavage and polyadenylation is regulated at sites with no PAS may allow for better modeling of these sites and improved algorithm performance.

APA in Vascular Development

Alternative polyadenylated transcripts have shown to be important in development and disease processes. I identified alternative 3'UTRs for *etv2*, a transcription factor important for endothelial cell specification, which may act to help down-regulate its expression as development proceeds (Moore et al, submitted, see Chapter 2). APA may play a role in other vascular processes such as defining artery and vein identity.

Perhaps distinct 3'UTRs could play a role in artery vein identity. Some transcripts that are over-expressed in arteries compared to vein have decreased H3K27 acetylation and P300 in artery compared to vein (Samir Sissoui, unpublished data), indicating more active transcription may be occurring in vein. This result suggests that this subset of transcripts may have increased transcript stability in artery compared to vein. A miRNA or RNA binding protein may be differentially expressed in artery and vein that may result in different regulation.

Another possibility is that these transcripts may express distinct 3'UTRs in each tissue, allowing for different regulation by a miRNA present in both cell types. A similar relationship has been shown in muscle cells, where the transcript *pax3* is expressed with unique 3'UTRs in two different tissue types, leading to differential miRNA regulation [204]. To further investigate this hypothesis, 3' end sequencing could be performed in artery and vein cells isolated from the zebrafish embryo. Importantly, a control RNA should be added in a known concentration to the original pool of mRNAs to allow for accurate quantification. 3' end sequencing should also be performed on a population of endothelial cells before artery and vein differentiation occurs so that any changes in alternative 3'UTR usage during differentiation could be identified. Putative pA sites would be accurately identified using the naïve Bayes classifier. pA sites unique to artery or vein or differential usage of the same pA site in artery versus vein may indicate potential candidates of post-transcriptional regulation. The Lawson lab has profiled miRNAs in zebrafish endothelial cells [147]. Thus, with the potential 3'UTRs and potential miRNAs, target prediction algorithms could be used to further pinpoint candidates of miRNA regulation.

After identifying putative candidates for miRNA regulation, validation of the 3'UTRs should be confirmed. Polyadenylated mRNAs could be G/I-tailed using yeast poly(A) polymerase, as in the GPAT assay (Figure 3.8A), followed by PCR with a gene specific primer and a cytosine-rich primer and shotgun cloning. This would confirm the presence and sequence of true polyadenylated 3'UTRs. Nanostring technology could be adapted to quantify usage of the 3'UTR common to both tissues and the extended 3'UTR, which may only be present in vein or artery [173]. *In situ* hybridization analysis could be performed for the common and extended portions of the 3’UTR to further confirm tissue specificity. Next these candidate 3’UTRs could be tested for post-transcriptional regulation using an endothelial cell-specific 3’UTR sensor in both wild type and *MZDicer* embryos [152], which contain no mature miRNAs. Subsequently, MOs to block miRNA binding to potential target sites within the 3’UTR or mutation of a putative binding site within the sensor construct would allow for identification of the specific regulatory miRNA. MOs designed to block the polyadenylation elements for a particular pA site would allow me to modulate alternative 3’UTR usage and examine the effect on artery vein identity. Further characterization may of these miRNA target interactions may elucidate important mechanisms defining artery-vein identity.

Conclusions

Use of the naïve Bayes classifier, described in this dissertation, in combination with oligo-dT primed 3’ end sequencing will accurately identify novel pA sites. The naïve Bayes classifier is easy to use and outperforms simple heuristic filters to remove internal priming events. Additional training data or features may improve the performance of the naïve Bayes classifier. The current implementation can also be used to remove internal priming events from genome annotations. It could also be retraining using data from yeast and plants. Tools to study cleavage and polyadenylation in zebrafish may allow for discovery of novel mechanisms regulating this process. Finally, the naïve Bayes classifier may help examine alternative 3’UTR usage important for artery and vein differentiation.

APPENDIX I

BAYESIAN PROBABILITY

A sample space contains all the possible events within an experiment, represented by the box in Figure A1.1A [109, 205][4]. Within the sample space represented in Figure A1.1A, are events A and B. This sample space holds events "A but not B", "A and B", "B but not A", and "not A and not B". These can be represented by the following probabilities [205]:

$$\Pr obabilityEventA = p(A)^{5}$$

$$\Pr obabilityEventB = p(B)$$

$$\Pr obabilityEventsA \,\&\, B = p(A \cap B)$$

$$\Pr obabilityNotEventA, B = 1 - p(A \cap B)$$

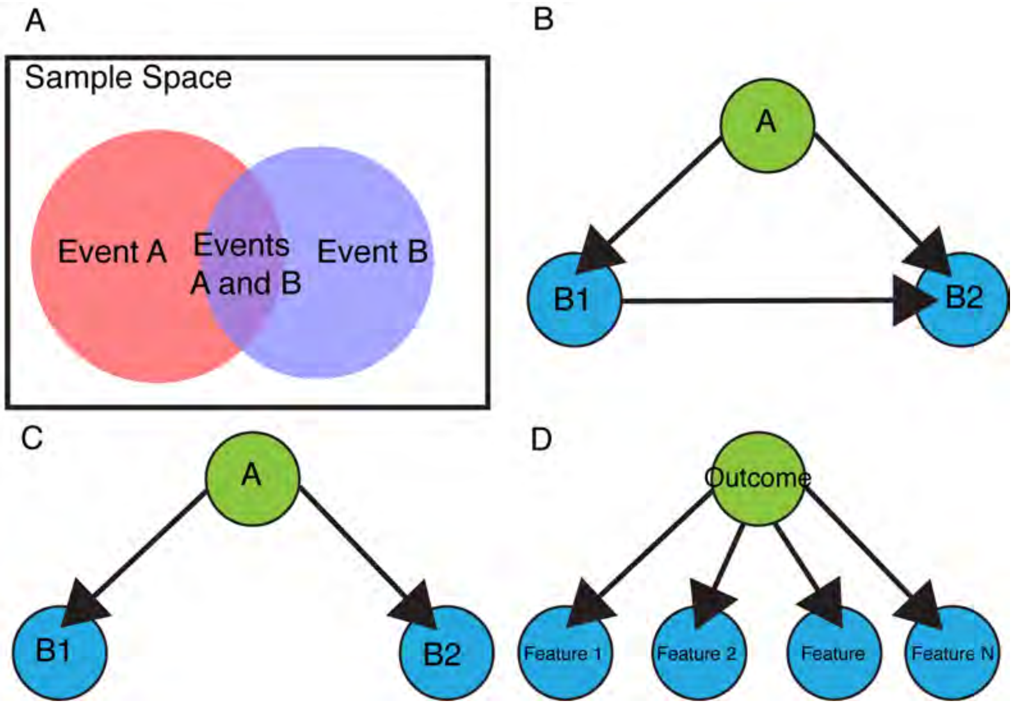
Sometimes it is helpful to know what the conditional probability of an event is. For example, what is the likelihood of event B, given A? First, we need to determine the probability of event A in the sample space. Then, within that subset, how many times does event B occur? Therefore, the probability of event B, given event A is equal to the probability of the events "A and B" divided by the probability of event A [205]. This is called the rule of conditional probability [205].

$$p(B \mid A) = \frac{p(A \cap B)}{p(A)} \quad (1)$$

[4] I performed the derivations in this section using the two references for guidance.

[5] The ∪ represents "union/or" and the ∩ represents "intersection/and".

Figure A1.1: Pictorial representations of probability. A. Events A, represented by the pink circle, and B, represented by the blue circle, lie within the sample space. B. Bayesian Networks can be represented by a directed acyclic graph. Notice the dependency of B1 and B2. C. Graphical representation of a naïve Bayesian Network with two features. Features are assumed to be conditionally independent. D. Graphical representation of a naïve Bayesian Network with N number of features. Features are assumed to be conditionally independent.


A
Sample Space
Event A
Events
A and B
Event B
B
A
B1
B2
C
A
B1
B2
D
Outcome
Feature 1
Feature 2
Feature
Feature N

Similarly, if we want to know the probability of event A, given event B [205]:

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)} \text{ (2)}$$

From equation (1) it follows:

$$p(A \cap B) = p(B \mid A)p(A) \text{ (3)}$$

With these equations (2) and (3) we can derive Bayes Theorem, an equation to relate conditional probabilities named for Thomas Bayes (1701-1761) [205]. Substituting equation (3) into equation (2):

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)} \text{ (4)}$$

What happens if we add a third variable? For example, what is the probability of event A, given events B_1 and B_2? This can be depicted in a directed acyclic graph (Figure A1.1B.). Bayes theorem defines the probability as such:

$$p(A \mid B_1 \cap B_2) = \frac{p(B_1 \cap B_2 \mid A)p(A)}{p(B_1 \cap B_2)} \text{ (5)}$$

The rule of conditional probability allows us to define the probability of events B_1 and B_2, given event A:

$$p(B_1 \cap B_2 \mid A) = \frac{p(B_1 \cap B_2 \cap A)}{p(A)} \text{(6)}$$

We can expand the numerator of equation (6):

$$p(B_1 \cap B_2 \cap A) = p(B_1 | B_2 \cap A)p(B_2 | A)p(A) \quad (7)$$

Thus substituting equation (7) into equation (6) results in:

$$p(B_1 \cap B_2 | A) = \frac{p(B_1 | B_2 \cap A)p(B_2 | A)p(A)}{p(A)} = p(B_1 | B_2 \cap A)p(B_2 | A) \quad (8)$$

If we use equation (8) in Bayes theorem (5):

$$p(A | B_1 \cap B_2) = \frac{p(B_1 | B_2 \cap A)p(B_2 | A)p(A)}{p(B_1 \cap B_2)} = p(B_1 | B_2 \cap A)p(B_2 | A) (9)$$

Naïve Bayesian probability assumes conditional independence of events [109]. In this example, B_1 and B_2 are conditionally independent given A (Figure A1.1C):

$$p(A | B_1 \cap B_2) = \frac{p(B_1 | A)p(B_2 | A)p(A)}{p(B_1 \cap B_2)} \quad (10)$$

For two events this simplification hardly seems warranted. However, using Bayes theorem for N events:

$$p(A | B_1 \cap ... \cap B_N) = \frac{p(B_1 \cap ... \cap B_N | A)p(A)}{p(B_1 \cap ... \cap B_N)} \quad (11)$$

We see that the number of parameters that need to be estimated grows exponentially, represented as $O(2^N)$. For example, if we wanted to use 20 features to classify an outcome over 2 million parameters would need to be estimated, which would require an exponentially large training dataset. However, If we assume conditional independence of events (Figure A1.1D):

$$p(A \mid B_1 \cap ... \cap B_N) = \frac{p(B_1 \mid A)...p(B_N \mid A)\,p(A)}{p(B_1 \cap ... \cap B_N)} \quad (12)$$

By assuming conditional independence of events, naïve Bayes classification significantly decreases the number of parameters that need to be estimated from training data to a linear proportion, represented O(N) [109].

APPENDIX II

PERFORMANCE OF DIFFERENT NAÏVE BAYES MODELS

Appendix II. Performance of different models investigated. The first number represents the length of upstream sequence for finding features and the second number represents the length of downstream sequence for finding features. 0.5 indicates the probability cutoff used to determine if a site was true or false.

Sample	Precision	Recall	F-score	Accuracy	True Negative Rate	False Discovery Rate	False Positive Rate	Matthew's Correlation Coefficient	Misclassification Error
20-30-0.5	0.959	0.932	0.945	0.923	0.901	0.041	0.099	0.817	0.077
20-40-0.5	0.951	0.930	0.940	0.916	0.882	0.049	0.118	0.800	0.084
20-50-0.5	0.956	0.916	0.935	0.910	0.895	0.044	0.105	0.789	0.090
30-30-0.5	0.968	0.939	0.953	0.935	0.924	0.032	0.076	0.845	0.065
30-40-0.5	0.959	0.935	0.947	0.925	0.901	0.041	0.099	0.822	0.075
30-50-0.5	0.956	0.933	0.944	0.922	0.894	0.044	0.106	0.813	0.078
40-30-0.5	0.968	0.938	0.953	0.934	0.922	0.032	0.078	0.843	0.066
40-40-0.5	0.962	0.935	0.948	0.927	0.908	0.038	0.092	0.827	0.073
40-50-0.5	0.955	0.932	0.944	0.921	0.892	0.045	0.108	0.811	0.079
50-30-0.5	0.964	0.937	0.950	0.930	0.912	0.036	0.088	0.833	0.070
50-40-0.5	0.958	0.935	0.946	0.925	0.900	0.042	0.100	0.820	0.075
50-50-0.5	0.953	0.932	0.942	0.919	0.885	0.047	0.115	0.805	0.081

APPENDIX III

BIOLOGICALLY VALIDATED PUTATIVE POLYADENYLATION SITES

Gene	Novel or Annotated 3'End	PAS 0-40 nt upstream	#A in 10 nt down	3pseq at 24h?	3pseq at 6h?	pas-seq at 24h?	pas-seq at 6h?	Bio. Valid.	naïve Bayes prob false	naïve Bayes prob true	naïve Bayes predict	8A filter
EST coverage	Novel	None	3	no	no	yes	no	TRUE	0.683651	0.316349	FALSE	TRUE
praf2	novel	AAUACA	4	no	no	yes	yes	FALSE	9.58E-01	0.041796	FALSE	TRUE
EST coverage	Novel	None	5	no	no	yes	yes	FALSE	9.44E-01	0.05568	FALSE	TRUE
cyyr1	Novel	UUUAAA	5	yes	no	yes	no	FALSE	8.12E-01	0.188092	FALSE	TRUE
none	Novel	AACAAA	7	no	no	yes	no	FALSE	0.999743	2.57E-04	FALSE	TRUE
vegfc annotated 3'utr	Annotated	None	7	no	no	yes	no	FALSE	1.00E+00	5.38E-11	FALSE	TRUE
EST coverage	Novel	None	7	no	no	yes	no	FALSE	0.95898	0.04102	FALSE	TRUE
none	Novel	AACAAA	10	no	no	yes	no	FALSE	1.00E+00	1.33E-11	FALSE	FALSE
slc25a3b 3'utr end	Annotated	AAUAAA	2	yes	yes	yes	no	TRUE	1.30E-09	1	TRUE	TRUE
tnnt3a 3'utr	Novel	AAUAAA	1	yes	no	no	yes	TRUE	5.56E-09	1	TRUE	TRUE
col1a1a 3'utr end (refseq)	Annotated	AAUAAA	1	yes	no	yes	yes	TRUE	1.48E-11	1	TRUE	TRUE
rpl26 3'utr end	Annotated	AAUAAA	0	yes	yes	yes	yes	TRUE	2.18E-20	1	TRUE	TRUE
ENSDART00000051763	Annotated	AAUAAA	1	yes	yes	yes	yes	TRUE	4.10E-15	1	TRUE	TRUE
ets2 longer 3'utr	Annotated	AAUAAA	1	yes	yes	yes	yes	TRUE	7.05E-04	0.999295	TRUE	TRUE
tcf7l1a 3'utr	Novel	AAUAAA	1	no	yes	yes	yes	TRUE	3.11E-15	1	TRUE	TRUE
ENSDART00000110621	Annotated	AAUAAA	1	yes	yes	yes	yes	TRUE	2.89E-20	1	TRUE	TRUE
flot2b 3'utr end	Annotated	AAUAAA	1	yes	yes	yes	yes	TRUE	2.01E-14	1	TRUE	TRUE
ap1g1 extended 3'utr	Novel	AAUAAA	1	yes	yes	yes	yes	TRUE	9.07E-10	1	TRUE	TRUE
hnrnpk 3'utr	Annotated	AAUAAA	1	yes	yes	yes	yes	TRUE	7.60E-23	1	TRUE	TRUE
ENSDART00000099208	Annotated	AAUAAA	2	yes	yes	yes	yes	TRUE	2.46E-09	1	TRUE	TRUE
ranbp1 3'utr (novel)	Novel	AAUAAA	2	yes	yes	yes	yes	TRUE	1.56E-08	1	TRUE	TRUE
rbb4	Annotated	AAUAAA	4	yes	yes	yes	yes	TRUE	5.96E-15	1	TRUE	TRUE
ENSDART00000084090	Annotated	AAUAAA	5	no	yes	yes	yes	TRUE	6.30E-04	0.99937	TRUE	TRUE
adam28 extended 3'utr	Novel	UAUAAA	0	no	yes	yes	yes	TRUE	4.81E-12	1	TRUE	TRUE
ENSDART00000150863	Annotated	AAUAAA	2	yes	no	yes	yes	TRUE	2.42E-16	1	TRUE	TRUE
sort1a (extended 3'utr)	Novel	AUUAAA	1	no	no	yes	yes	TRUE	1.79E-13	1	TRUE	TRUE
elovl6 3'utr	Novel	AUUAAA	3	no	no	yes	yes	TRUE	2.26E-01	0.774347	TRUE	TRUE
vegfc extended 3'utr	Novel	AAUAAA	3	no	yes	no	no	TRUE	1.16E-10	1	TRUE	TRUE
ENSDART00000033980	Novel	AAUACA	5	no	yes	no	no	TRUE	4.98E-05	0.99995	TRUE	TRUE

Gene	Novel or Annotated 3'End	PAS 0-40 nt upstream	#A in 10 nt down	3pseq at 24h?	3pseq at 6h?	pas-seq at 24h?	pas-seq at 6h?	Bio. Valid.	naïve Bayes prob false	naïve Bayes prob true	naïve Bayes predict	8A filter
wu:fa26c03	Anotated	UAUAAA	2	yes	yes	no	no	TRUE	1.24E-08	1	TRUE	TRUE
ENSEMBL00000102500	Annotated	AAUAAA	0	yes	no	yes	no	TRUE	1.81E-20	1.00E+00	TRUE	TRUE
gbp1	Annotated	AAUAAA	0	no	no	yes	no	TRUE	3.31E-19	1	TRUE	TRUE
pdcd4a 3'utr	Annotated	AAUAAA	1	yes	no	yes	no	TRUE	3.72E-15	1	TRUE	TRUE
vegfc extended 3'utr	Novel	AAUAAA	1	yes	no	yes	no	TRUE	1.53E-15	1	TRUE	TRUE
ctnna2 intron (EST coverage)	Novel	AAUAAA	1	no	no	yes	no	TRUE	0.003317	0.996683	TRUE	TRUE
tcf7l1a 3;utr annotated end	Annotated	AAUAAA	1	yes	no	yes	no	TRUE	1.19E-04	0.999881	TRUE	TRUE
EST coverage	Novel	AAUAAA	1	yes	no	yes	no	TRUE	1.67E-13	1	TRUE	TRUE
EST coverage	Novel	AAUAAA	1	no	no	yes	no	TRUE	1.02E-12	1	TRUE	TRUE
EST coverage	Novel	AAUAAA	1	no	no	yes	no	TRUE	3.61E-16	1	TRUE	TRUE
EST coverage	Novel	AAUAAA	2	no	no	yes	no	TRUE	1.24E-06	0.999999	TRUE	TRUE
ldb3a annotated 3'utr	Annotated	AAUAAA	2	yes	no	yes	no	TRUE	9.48E-10	1	TRUE	TRUE
ensdart77304	Annotated	AAUAAA	4	no	no	yes	no	TRUE	2.71E-12	1	TRUE	TRUE
cldn1 3'utr end (annotated)	Annotated	AUUAAA	0	yes	no	yes	no	TRUE	6.68E-12	1	TRUE	TRUE
praf2	Novel	AUUAAA	0	yes	no	no	no	TRUE	6.60E-23	1	TRUE	TRUE
stab1l 3'utr	Novel	AUUAAA	1	yes	no	yes	no	TRUE	2.06E-13	1	TRUE	TRUE
npb	Annotated	AUUAAA	1	no	no	yes	no	TRUE	6.14E-16	1	TRUE	TRUE
ipo7 3'utr end	Novel	AUUAAA	2	yes	no	yes	no	TRUE	4.08E-07	1	TRUE	TRUE
ENSDART00000054925	Annotated	AUUAAA	3	yes	no	yes	no	TRUE	5.15E-08	1	TRUE	TRUE
ranbp1 3'utr	Novel	None	2	yes	no	yes	no	TRUE	0.014708	0.985292	TRUE	TRUE
no gene	Novel	UUUAAA	3	no	no	yes	no	TRUE	6.28E-06	0.999994	TRUE	TRUE

APPENDIX IV

CLEANUPDTSEQ USER'S GUIDE

The cleanUpdTSeq User's Guide

Authors: Sarah Sheppard, Nathan Lawson, Lihua Julie Zhu

Introduction

3' ends of transcripts have generally been poorly annotated. With the advent of deep sequencing, many methods have been developed to identify 3' ends. The majority of these methods use an oligo-dT primer, which can bind to internal adenine-rich sequences, and lead to artifactual identification of polyadenylation sites. Heuristic filtering methods rely on a certain number of adenines in the genomic sequence downstream of a putative polyadenylation site to remove internal priming events. We introduce a package to provide a robust method to classify putative polyadenylation sites. cleanUpdTSeq uses a naïve Bayes classifier, implement through the R package e1071 [188], and sequence features surrounding the putative polyadenylation sites for classification.

The input for this package is a bed file of putative polyadenylation sites with or without sequence. First, the function BED2RangedDataSeq converts the bed information to RangedData. Next, buildFeatureVector builds a data frame containing the features for the naïve Bayes classifier.

An option is included to get the sequence surrounding the putative polyadenylation site from BSgenome [206].

Task 1: Use cleanUpdTSeq to classify a list of putative polyadenylation sites

First, read in the bed file and then use the function BED2RangedDataSeq to convert it to RangedData.

```
----------------------------------
library(cleanUpdTSeq)
testSet = read.table("test.bed", sep = "\t", header = TRUE)
peaks = BED2RangedDataSeq(testSet)
----------------------------------
```

Next, build a data frame containing the features for the classifier using the function buildFeatureVector. The zebrafish genome from BSgenome is used in this example [206]. For a list of other genomes available through BSgenome, please refer to the BSgenome package documentation [206].

```
----------------------------------
testSet.NaiveBayes = buildFeatureVector(peaks,BSgenomeName =
Drerio, upstream = 40, downstream = 30, wordSize = 6,
alphabet=c("ACGT"), sampleType = "unknown", replaceNAdistance = 30
method = "NaiveBayes", ZeroBasedIndex = 1, fetchSeq = TRUE)
save(testSet.NaiveBayes, file = "test.Rdata")
----------------------------------
```

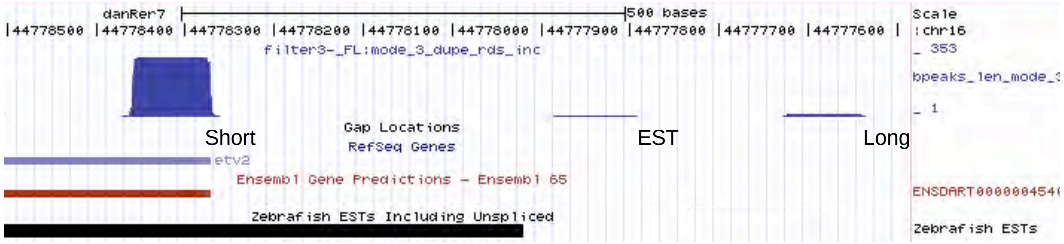
Finally, classify putative polyadenylation sites.

```
----------------------------------
predictTestSet(Ndata.NaiveBayes, Pdata.NaiveBayes, inputFile =
"test.RData",outputFile = "test-predNaiveBayes.tsv", assignmentCutoff =
0.5)
----------------------------------
```

The output file is a tab-delimited file containing the name of the putative polyadenylation sites, the probability that the putative polyadenylation site is false/oligodT internally primed, the probability the putative polyadenylation site if true, the predicted class based on the assignment cutoff and the sequence surrounding the putative polyadenylation site.

APPENDIX V

PAS-SEQ IDENTIFIES NOVEL ETV2 POLYADENYLATION SITES



PAS-Seq identifies novel *etv2* polyadenylation sites. Screen shot from the UCSC genome browser at the *etv2* locus. The peaks at the top represent putative polyadenylation sites. The "short" polyadenylation site corresponds to the annotated *etv2* 3'UTR end in both RefSeq and Ensembl. Two additional putative polyadenylation sites were identified ("EST" and "Long") that do not correspond to any annotations.

References

1. Babich, A., J.R. Nevins, and J.E. Darnell, Jr., *Early capping of transcripts from the adenovirus major late transcription unit.* Nature, 1980. **287**(5779): p. 246-8.
2. Moore, M.J. and N.J. Proudfoot, *Pre-mRNA processing reaches back to transcription and ahead to translation.* Cell, 2009. **136**(4): p. 688-700.
3. Mandel, C.R., et al., *Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease.* Nature, 2006. **444**(7121): p. 953-6.
4. Mandel, C.R., Y. Bai, and L. Tong, *Protein factors in pre-mRNA 3'-end processing.* Cell Mol Life Sci, 2008. **65**(7-8): p. 1099-122.
5. Hirose, Y. and J.L. Manley, *RNA polymerase II is an essential mRNA polyadenylation factor.* Nature, 1998. **395**(6697): p. 93-6.
6. McLauchlan, J., et al., *The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini.* Nucleic Acids Res, 1985. **13**(4): p. 1347-68.
7. Natalizio, B.J., et al., *Upstream elements present in the 3'-untranslated region of collagen genes influence the processing efficiency of overlapping polyadenylation signals.* J Biol Chem, 2002. **277**(45): p. 42733-40.
8. Moreira, A., et al., *Upstream sequence elements enhance poly(A) site efficiency of the C2 complement gene and are phylogenetically conserved.* EMBO J, 1995. **14**(15): p. 3809-19.
9. Weill, L., et al., *Translational control by changes in poly(A) tail length: recycling mRNAs.* Nat Struct Mol Biol, 2012. **19**(6): p. 577-85.
10. Mueller, A.A., T.H. Cheung, and T.A. Rando, *All's well that ends well: alternative polyadenylation and its implications for stem cell biology.* Curr Opin Cell Biol, 2013.
11. Shi, Y., et al., *Molecular architecture of the human pre-mRNA 3' processing complex.* Mol Cell, 2009. **33**(3): p. 365-76.
12. Mangone, M., et al., *The landscape of C. elegans 3'UTRs.* Science, 2010. **329**(5990): p. 432-5.
13. Jan, C.H., et al., *Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs.* Nature, 2010. **469**(7328): p. 97-101.
14. Li, Y., et al., *Dynamic landscape of tandem 3' UTRs during zebrafish development.* Genome Res, 2012. **22**(10): p. 1899-906.
15. Smibert, P., et al., *Global Patterns of Tissue-Specific Alternative Polyadenylation in Drosophila.* Cell Rep, 2012. **1**(3): p. 277-289.
16. Shepard, P.J., et al., *Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq.* RNA, 2011. **17**(4): p. 761-72.

17. Beaudoing, E., et al., *Patterns of variant polyadenylation signal usage in human genes.* Genome Res, 2000. **10**(7): p. 1001-10.
18. Tian, B., et al., *A large-scale analysis of mRNA polyadenylation of human and mouse genes.* Nucleic Acids Res, 2005. **33**(1): p. 201-12.
19. Lin, Y., et al., *An in-depth map of polyadenylation sites in cancer.* Nucleic Acids Res, 2012. **40**(17): p. 8460-71.
20. Martin, G., et al., *Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length.* Cell Rep, 2012. **1**(6): p. 753-63.
21. Mayr, C. and D.P. Bartel, *Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells.* Cell, 2009. **138**(4): p. 673-84.
22. Thomas, C.P., J.I. Andrews, and K.Z. Liu, *Intronic polyadenylation signal sequences and alternate splicing generate human soluble Flt1 variants and regulate the abundance of soluble Flt1 in the placenta.* FASEB J, 2007. **21**(14): p. 3885-95.
23. Yao, C., et al., *Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation.* Proc Natl Acad Sci U S A, 2012. **109**(46): p. 18773-8.
24. Kamasawa, M. and J. Horiuchi, *Identification and characterization of polyadenylation signal (PAS) variants in human genomic sequences based on modified EST clustering.* In Silico Biol, 2008. **8**(3-4): p. 347-61.
25. Chen, F. and J. Wilusz, *Auxiliary downstream elements are required for efficient polyadenylation of mammalian pre-mRNAs.* Nucleic Acids Res, 1998. **26**(12): p. 2891-8.
26. Elkon, R., et al., *E2F mediates enhanced alternative polyadenylation in proliferation.* Genome Biol, 2012. **13**(7): p. R59.
27. Flavell, S.W., et al., *Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection.* Neuron, 2008. **60**(6): p. 1022-38.
28. Takagaki, Y., et al., *The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation.* Cell, 1996. **87**(5): p. 941-52.
29. Danckwardt, S., et al., *Splicing factors stimulate polyadenylation via USEs at non-canonical 3' end formation signals.* EMBO J, 2007. **26**(11): p. 2658-69.
30. Perez Canadillas, J.M. and G. Varani, *Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein.* EMBO J, 2003. **22**(11): p. 2821-30.

31. Venkataraman, K., K.M. Brown, and G.M. Gilmartin, *Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition.* Genes Dev, 2005. **19**(11): p. 1315-27.
32. Chao, L.C., et al., *Assembly of the cleavage and polyadenylation apparatus requires about 10 seconds in vivo and is faster for strong than for weak poly(A) sites.* Mol Cell Biol, 1999. **19**(8): p. 5588-600.
33. Tian, B., Z. Pan, and J.Y. Lee, *Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing.* Genome Res, 2007. **17**(2): p. 156-65.
34. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.
35. Jenal, M., et al., *The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites.* Cell, 2012. **149**(3): p. 538-53.
36. Bava, F.A., et al., *CPEB1 coordinates alternative 3'-UTR formation with translational regulation.* Nature, 2013. **495**(7439): p. 121-5.
37. Licatalosi, D.D., et al., *HITS-CLIP yields genome-wide insights into brain alternative RNA processing.* Nature, 2008. **456**(7221): p. 464-9.
38. Hilgers, V., S.B. Lemke, and M. Levine, *ELAV mediates 3' UTR extension in the Drosophila nervous system.* Genes Dev, 2012. **26**(20): p. 2259-64.
39. Di Giammartino, D.C., K. Nishida, and J.L. Manley, *Mechanisms and consequences of alternative polyadenylation.* Mol Cell, 2011. **43**(6): p. 853-66.
40. Sherstnev, A., et al., *Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation.* Nat Struct Mol Biol, 2012. **19**(8): p. 845-52.
41. Hoque, M., et al., *Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing.* Nat Methods, 2012. **10**(2): p. 133-9.
42. Proudfoot, N.J. and G.G. Brownlee, *3' non-coding region sequences in eukaryotic messenger RNA.* Nature, 1976. **263**(5574): p. 211-4.
43. Sheets, M.D., S.C. Ogg, and M.P. Wickens, *Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro.* Nucleic Acids Res, 1990. **18**(19): p. 5799-805.
44. Keller, W., et al., *Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3' processing signal AAUAAA.* EMBO J, 1991. **10**(13): p. 4241-9.

45. Derti, A., et al., *A quantitative atlas of polyadenylation in five mammals.* Genome Res, 2012. **22**(6): p. 1173-83.
46. Takagaki, Y. and J.L. Manley, *RNA recognition by the human polyadenylation factor CstF.* Mol Cell Biol, 1997. **17**(7): p. 3907-14.
47. Brown, K.M. and G.M. Gilmartin, *A mechanism for the regulation of pre-mRNA 3' processing by human cleavage factor Im.* Mol Cell, 2003. **12**(6): p. 1467-76.
48. Nunes, N.M., et al., *A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence.* EMBO J, 2010. **29**(9): p. 1523-36.
49. Hon, C.C., et al., *Quantification of stochastic noise of splicing and polyadenylation in Entamoeba histolytica.* Nucleic Acids Res, 2013. **41**(3): p. 1936-52.
50. Graber, J.H., et al., *Genomic detection of new yeast pre-mRNA 3'-end-processing signals.* Nucleic Acids Res, 1999. **27**(3): p. 888-94.
51. van Helden, J., M. del Olmo, and J.E. Perez-Ortin, *Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals.* Nucleic Acids Res, 2000. **28**(4): p. 1000-10.
52. Ozsolak, F., et al., *Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation.* Cell, 2010. **143**(6): p. 1018-29.
53. Shen, Y., et al., *Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation.* Nucleic Acids Res, 2008. **36**(9): p. 3150-61.
54. Salisbury, J., K.W. Hutchison, and J.H. Graber, *A multispecies comparison of the metazoan 3'-processing downstream elements and the CstF-64 RNA recognition motif.* BMC Genomics, 2006. **7**: p. 55.
55. Retelska, D., et al., *Similarities and differences of polyadenylation signals in human and fly.* BMC Genomics, 2006. **7**: p. 176.
56. Ulitsky, I., et al., *Extensive alternative polyadenylation during zebrafish development.* Genome Res, 2012. **22**(10): p. 2054-66.
57. Fox-Walsh, K., et al., *A multiplex RNA-seq strategy to profile poly(A+) RNA: application to analysis of transcription response and 3' end formation.* Genomics, 2011. **98**(4): p. 266-71.
58. Graber, J.H., et al., *In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species.* Proc Natl Acad Sci U S A, 1999. **96**(24): p. 14055-60.
59. Murthy, K.G. and J.L. Manley, *The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation.* Genes Dev, 1995. **9**(21): p. 2672-83.

60. Jenny, A., H.P. Hauri, and W. Keller, *Characterization of cleavage and polyadenylation specificity factor and cloning of its 100-kilodalton subunit.* Mol Cell Biol, 1994. **14**(12): p. 8183-90.
61. Legendre, M. and D. Gautheret, *Sequence determinants in human polyadenylation site selection.* BMC Genomics, 2003. **4**(1): p. 7.
62. Hu, J., et al., *Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation.* RNA, 2005. **11**(10): p. 1485-93.
63. MacDonald, C.C., J. Wilusz, and T. Shenk, *The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location.* Mol Cell Biol, 1994. **14**(10): p. 6647-54.
64. Chen, F., C.C. MacDonald, and J. Wilusz, *Cleavage site determinants in the mammalian polyadenylation signal.* Nucleic Acids Res, 1995. **23**(14): p. 2614-20.
65. Shen, Y., et al., *Transcriptome dynamics through alternative polyadenylation in developmental and environmental responses in plants revealed by deep sequencing.* Genome Res, 2011. **21**(9): p. 1478-86.
66. Wu, X., et al., *Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation.* Proc Natl Acad Sci U S A, 2011. **108**(30): p. 12533-8.
67. Fu, Y., et al., *Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing.* Genome Res, 2011. **21**(5): p. 741-7.
68. Sandberg, R., et al., *Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites.* Science, 2008. **320**(5883): p. 1643-7.
69. Ji, Z., et al., *Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development.* Proc Natl Acad Sci U S A, 2009. **106**(17): p. 7028-33.
70. Liu, D., et al., *Systematic variation in mRNA 3'-processing signals during mouse spermatogenesis.* Nucleic Acids Res, 2007. **35**(1): p. 234-46.
71. Hilgers, V., et al., *Neural-specific elongation of 3' UTRs during Drosophila development.* Proc Natl Acad Sci U S A, 2011. **108**(38): p. 15864-9.
72. Miura, P., et al., *Widespread and extensive lengthening of 3' UTRs in the mammalian brain.* Genome Res, 2013.

73. Rhinn, H., et al., *Alternative alpha-synuclein transcript usage as a convergent mechanism in Parkinson's disease pathology.* Nat Commun, 2012. **3**: p. 1084.
74. Kumar, V., et al., *Robbins and Cotran pathologic basis of disease*. 7th ed. 2005, Philadelphia: Elsevier Saunders. xv, 1525 p.
75. Cohen, O.S., et al., *Transcriptomic analysis of postmortem brain identifies dysregulated splicing events in novel candidate genes for schizophrenia.* Schizophr Res, 2012. **142**(1-3): p. 188-99.
76. Bennett, C.L., et al., *A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA-->AAUGAA) leads to the IPEX syndrome.* Immunogenetics, 2001. **53**(6): p. 435-9.
77. Orkin, S.H., et al., *Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene.* EMBO J, 1985. **4**(2): p. 453-6.
78. Losekoot, M., et al., *Homozygous beta+ thalassaemia owing to a mutation in the cleavage-polyadenylation sequence of the human beta globin gene.* J Med Genet, 1991. **28**(4): p. 252-5.
79. Whitelaw, E. and N. Proudfoot, *Alpha-thalassaemia caused by a poly(A) site mutation reveals that transcriptional termination is linked to 3' end processing in the human alpha 2 globin gene.* EMBO J, 1986. **5**(11): p. 2915-22.
80. Harteveld, C.L., et al., *A novel polyadenylation signal mutation in the alpha 2-globin gene causing alpha thalassaemia.* Br J Haematol, 1994. **87**(1): p. 139-43.
81. Stacey, S.N., et al., *A germline variant in the TP53 polyadenylation signal confers cancer susceptibility.* Nat Genet, 2011. **43**(11): p. 1098-103.
82. Wiestner, A., et al., *Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival.* Blood, 2007. **109**(11): p. 4599-606.
83. Gautheret, D., et al., *Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering.* Genome Res, 1998. **8**(5): p. 524-30.
84. Zhang, H., J.Y. Lee, and B. Tian, *Biased alternative polyadenylation in human tissues.* Genome Biol, 2005. **6**(12): p. R100.
85. Zhang, H., et al., *PolyA_DB: a database for mammalian mRNA polyadenylation.* Nucleic Acids Res, 2005. **33**(Database issue): p. D116-20.

86. Lee, J.Y., et al., *PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes.* Nucleic Acids Res, 2007. **35**(Database issue): p. D165-8.
87. Beaudoing, E. and D. Gautheret, *Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data.* Genome Res, 2001. **11**(9): p. 1520-6.
88. Ara, T., et al., *Conservation of alternative polyadenylation patterns in mammalian genes.* BMC Genomics, 2006. **7**: p. 189.
89. Graber, J.H., G.D. McAllister, and T.F. Smith, *Probabilistic prediction of Saccharomyces cerevisiae mRNA 3'-processing sites.* Nucleic Acids Res, 2002. **30**(8): p. 1851-8.
90. Brockman, J.M., et al., *PACdb: PolyA Cleavage Site and 3'-UTR Database.* Bioinformatics, 2005. **21**(18): p. 3691-3.
91. Salamov, A.A. and V.V. Solovyev, *Recognition of 3'-processing sites of human mRNA precursors.* Comput Appl Biosci, 1997. **13**(1): p. 23-8.
92. Tabaska, J.E. and M.Q. Zhang, *Detection of polyadenylation signals in human DNA sequences.* Gene, 1999. **231**(1-2): p. 77-86.
93. Liu, H., et al., *An in-silico method for prediction of polyadenylation signals in human sequences.* Genome Inform, 2003. **14**: p. 84-93.
94. Cheng, Y., R.M. Miura, and B. Tian, *Prediction of mRNA polyadenylation sites by support vector machine.* Bioinformatics, 2006. **22**(19): p. 2320-5.
95. Akhtar, M.N., et al., *POLYAR, a new computer program for prediction of poly(A) sites in human sequences.* BMC Genomics, 2010. **11**: p. 646.
96. Hajarnavis, A., I. Korf, and R. Durbin, *A probabilistic model of 3' end formation in Caenorhabditis elegans.* Nucleic Acids Res, 2004. **32**(11): p. 3392-9.
97. Ji, G., et al., *Predictive modeling of plant messenger RNA polyadenylation sites.* BMC Bioinformatics, 2007. **8**: p. 43.
98. Ji, G., et al., *A classification-based prediction model of messenger RNA polyadenylation sites.* J Theor Biol, 2010. **265**(3): p. 287-96.
99. Wilkening, S., et al., *An efficient method for genome-wide polyadenylation site mapping and RNA quantification.* Nucleic Acids Res, 2013.
100. Beck, A.H., et al., *3'-end sequencing for expression quantification (3SEQ) from archival tumor samples.* PLoS One, 2010. **5**(1): p. e8768.
101. Ozsolak, F., et al., *Direct RNA sequencing.* Nature, 2009. **461**(7265): p. 814-8.
102. Jan, C.H., et al., *Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs.* Nature, 2011. **469**(7328): p. 97-101.

103. Nam, D.K., et al., *Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription.* Proc Natl Acad Sci U S A, 2002. **99**(9): p. 6152-6.
104. Aaronson, J.S., et al., *Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data.* Genome Res, 1996. **6**(9): p. 829-45.
105. Kawamoto, S., et al., *BodyMap: a collection of 3' ESTs for analysis of human gene expression information.* Genome Res, 2000. **10**(11): p. 1817-27.
106. Sprague, J., et al., *The Zebrafish Information Network: the zebrafish model organism database.* Nucleic Acids Res, 2006. **34**(Database issue): p. D581-5.
107. Wang, L., R.D. Dowell, and R. Yi, *Genome-wide maps of polyadenylation reveal dynamic mRNA 3'-end formation in mammalian cell lineages.* RNA, 2013. **19**(3): p. 413-25.
108. Kotsiantis, S.B., I.D. Zaharakis, and P.E. Pintelas, *Machine learning: a review of classification and combining techniques.* Artificial Intelligence Review, 2006. **26**(3): p. 159-190.
109. Mitchell, T.M., *Chapter 1: Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression*. 2010. p. 1-17.
110. Domingos, P. and M. Pazzani, *On the optimality of the simple Bayesian classifier under zero-one loss.* Machine Learning, 1997. **29**(2-3): p. 103-130.
111. Wang, Q., et al., *Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.* Appl Environ Microbiol, 2007. **73**(16): p. 5261-7.
112. Wei, W., S. Visweswaran, and G.F. Cooper, *The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data.* J Am Med Inform Assoc, 2011. **18**(4): p. 370-5.
113. Yang, M.C., et al., *Whole breast lesion detection using naive bayes classifier for portable ultrasound.* Ultrasound Med Biol, 2012. **38**(11): p. 1870-80.
114. Risau, W. and I. Flamme, *Vasculogenesis.* Annu Rev Cell Dev Biol, 1995. **11**: p. 73-91.
115. De Val, S. and B.L. Black, *Transcriptional control of endothelial cell development.* Dev Cell, 2009. **16**(2): p. 180-95.
116. Sharrocks, A.D., *The ETS-domain transcription factor family.* Nat Rev Mol Cell Biol, 2001. **2**(11): p. 827-37.

117. Lelievre, E., et al., *The Ets family contains transcriptional activators and repressors involved in angiogenesis.* Int J Biochem Cell Biol, 2001. **33**(4): p. 391-407.

118. Dejana, E., A. Taddei, and A.M. Randi, *Foxs and Ets in the transcriptional regulation of endothelial cell differentiation and angiogenesis.* Biochim Biophys Acta, 2007. **1775**(2): p. 298-312.

119. Hollenhorst, P.C., D.A. Jones, and B.J. Graves, *Expression profiles frame the promoter specificity dilemma of the ETS family of transcription factors.* Nucleic Acids Res, 2004. **32**(18): p. 5693-702.

120. Liu, F. and R. Patient, *Genome-wide analysis of the zebrafish ETS family identifies three genes required for hemangioblast differentiation or angiogenesis.* Circ Res, 2008. **103**(10): p. 1147-54.

121. Maroulakou, I.G., T.S. Papas, and J.E. Green, *Differential expression of ets-1 and ets-2 proto-oncogenes during murine embryogenesis.* Oncogene, 1994. **9**(6): p. 1551-65.

122. Pham, V.N., et al., *Combinatorial function of ETS transcription factors in the developing vasculature.* Dev Biol, 2007. **303**(2): p. 772-83.

123. Stiegler, P., et al., *The c-ets-1 proto-oncogenes in Xenopus laevis: expression during oogenesis and embryogenesis.* Mech Dev, 1993. **41**(2-3): p. 163-74.

124. Kappel, A., et al., *Role of SCL/Tal-1, GATA, and ets transcription factor binding sites for the regulation of flk-1 expression during murine vascular development.* Blood, 2000. **96**(9): p. 3078-85.

125. Wakiya, K., et al., *A cAMP response element and an Ets motif are involved in the transcriptional regulation of flt-1 tyrosine kinase (vascular endothelial growth factor receptor 1) gene.* J Biol Chem, 1996. **271**(48): p. 30823-8.

126. Iljin, K., et al., *Role of ets factors in the activity and endothelial cell specificity of the mouse Tie gene promoter.* FASEB J, 1999. **13**(2): p. 377-86.

127. Teruyama, K., et al., *Neurophilin-1 is a downstream target of transcription factor Ets-1 in human umbilical vein endothelial cells.* FEBS Lett, 2001. **504**(1-2): p. 1-4.

128. Spyropoulos, D.D., et al., *Hemorrhage, impaired hematopoiesis, and lethality in mouse embryos carrying a targeted disruption of the Fli1 transcription factor.* Mol Cell Biol, 2000. **20**(15): p. 5643-52.

129. Barton, K., et al., *The Ets-1 transcription factor is required for the development of natural killer cells in mice.* Immunity, 1998. **9**(4): p. 555-63.

130. Wei, G.H., et al., *Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo.* EMBO J, 2010. **29**(13): p. 2147-60.
131. Hollenhorst, P.C., L.P. McIntosh, and B.J. Graves, *Genomic and biochemical insights into the specificity of ETS transcription factors.* Annu Rev Biochem, 2011. **80**: p. 437-71.
132. Wei, G., et al., *Ets1 and Ets2 are required for endothelial cell survival during embryonic angiogenesis.* Blood, 2009. **114**(5): p. 1123-30.
133. Ferdous, A., et al., *Nkx2-5 transactivates the Ets-related protein 71 gene and specifies an endothelial/endocardial fate in the developing embryo.* Proc Natl Acad Sci U S A, 2009. **106**(3): p. 814-9.
134. Lee, D., et al., *ER71 acts downstream of BMP, Notch, and Wnt signaling in blood and vessel progenitor specification.* Cell Stem Cell, 2008. **2**(5): p. 497-507.
135. Sumanas, S. and S. Lin, *Ets1-related protein is a key regulator of vasculogenesis in zebrafish.* PLoS Biol, 2006. **4**(1): p. e10.
136. Gomez, G., et al., *Identification of vascular and hematopoietic genes downstream of etsrp by deep sequencing in zebrafish.* PLoS One, 2012. **7**(3): p. e31658.
137. Koyano-Nakagawa, N., et al., *Etv2 is expressed in the yolk sac hematopoietic and endothelial progenitors and regulates Lmo2 gene expression.* Stem Cells, 2012. **30**(8): p. 1611-23.
138. Wong, K.S., et al., *Identification of vasculature-specific genes by microarray analysis of Etsrp/Etv2 overexpressing zebrafish embryos.* Dev Dyn, 2009. **238**(7): p. 1836-50.
139. Ginsberg, M., et al., *Efficient direct reprogramming of mature amniotic cells into endothelial cells by ETS factors and TGFbeta suppression.* Cell, 2012. **151**(3): p. 559-75.
140. Wareing, S., et al., *The Flk1-Cre-Mediated Deletion of ETV2 Defines Its Narrow Temporal Requirement During Embryonic Hematopoietic Development.* Stem Cells, 2012. **30**(7): p. 1521-31.
141. Bertrand, J.Y., et al., *Definitive hematopoiesis initiates through a committed erythromyeloid progenitor in the zebrafish embryo.* Development, 2007. **134**(23): p. 4147-56.
142. Shestopalov, I.A. and J.K. Chen, *Spatiotemporal control of embryonic gene expression using caged morpholinos.* Methods Cell Biol, 2011. **104**: p. 151-72.
143. Ouyang, X., et al., *Versatile synthesis and rational design of caged morpholinos.* J Am Chem Soc, 2009. **131**(37): p. 13255-69.

144. Hayashi, M., et al., *Endothelialization and altered hematopoiesis by persistent Etv2 expression in mice.* Exp Hematol, 2012. **40**(9): p. 738-750 e11.

145. Nicoli, S., et al., *MicroRNA-mediated integration of haemodynamics and Vegf signalling during angiogenesis.* Nature, 2010. **in press**.

146. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2004. **116**(2): p. 281-97.

147. Nicoli, S., et al., *miR-221 is required for endothelial tip cell behaviors during vascular development.* Dev Cell, 2012. **22**(2): p. 418-29.

148. Gomez, G.A., et al., *Discovery and characterization of novel vascular and hematopoietic genes downstream of etsrp in zebrafish.* PLoS One, 2009. **4**(3): p. e4994.

149. Heo, I., et al., *Lin28 mediates the terminal uridylation of let-7 precursor MicroRNA.* Mol Cell, 2008. **32**(2): p. 276-84.

150. Nam, Y., et al., *Molecular basis for interaction of let-7 microRNAs with Lin28.* Cell, 2011. **147**(5): p. 1080-91.

151. Viswanathan, S.R., G.Q. Daley, and R.I. Gregory, *Selective blockade of microRNA processing by Lin28.* Science, 2008. **320**(5872): p. 97-100.

152. Giraldez, A.J., et al., *MicroRNAs regulate brain morphogenesis in zebrafish.* Science, 2005. **308**(5723): p. 833-8.

153. Scott, E.W., et al., *Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages.* Science, 1994. **265**(5178): p. 1573-7.

154. McKercher, S.R., et al., *Targeted disruption of the PU.1 gene results in multiple hematopoietic abnormalities.* EMBO J, 1996. **15**(20): p. 5647-58.

155. Anderson, K.L., et al., *Myeloid development is selectively disrupted in PU.1 null mice.* Blood, 1998. **91**(10): p. 3702-10.

156. Anderson, K.L., et al., *Neutrophils deficient in PU.1 do not terminally differentiate or become functionally competent.* Blood, 1998. **92**(5): p. 1576-85.

157. Chang, T.C., et al., *Lin-28B transactivation is necessary for Myc-mediated let-7 repression and proliferation.* Proc Natl Acad Sci U S A, 2009. **106**(9): p. 3384-9.

158. Johnson, S.M., et al., *RAS is regulated by the let-7 microRNA family.* Cell, 2005. **120**(5): p. 635-47.

159. Thornton, J.E. and R.I. Gregory, *How does Lin28 let-7 control development and disease?* Trends Cell Biol, 2012. **22**(9): p. 474-82.

160. Yu, F., et al., *let-7 regulates self renewal and tumorigenicity of breast cancer cells.* Cell, 2007. **131**(6): p. 1109-23.

161. Veldman, M.B. and S. Lin, *Etsrp/Etv2 is directly regulated by Foxc1a/b in the zebrafish angioblast.* Circ Res, 2012. **110**(2): p. 220-9.
162. Kuehbacher, A., et al., *Role of Dicer and Drosha for endothelial microRNA expression and angiogenesis.* Circ Res, 2007. **101**(1): p. 59-68.
163. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.
164. Westerfield, M., *The zebrafish book : a guide for the laboratory use of zebrafish (Brachydanio rerio)*. 1993, [Eugene, OR]: M. Westerfield. 1 v. (unpaged).
165. Roman, B.L., et al., *Disruption of acvrl1 increases endothelial cell number in zebrafish cranial vessels.* Development, 2002. **129**(12): p. 3009-19.
166. Lawson, N.D. and B.M. Weinstein, *In vivo imaging of embryonic vascular development using transgenic zebrafish.* Dev Biol, 2002. **248**(2): p. 307-18.
167. Covassin, L.D., et al., *A genetic screen for vascular mutants in zebrafish reveals dynamic roles for Vegf/Plcg1 signaling during artery development.* Dev Biol, 2009.
168. Ciruna, B., et al., *Production of maternal-zygotic mutant zebrafish by germ-line replacement.* Proc Natl Acad Sci U S A, 2002. **99**(23): p. 14919-24.
169. Wienholds, E., et al., *The microRNA-producing enzyme Dicer1 is essential for zebrafish development.* Nat Genet, 2003. **35**(3): p. 217-8.
170. Nicoli, S., et al., *MicroRNA-mediated integration of haemodynamics and Vegf signalling during angiogenesis.* Nature, 2010. **464**(7292): p. 1196-200.
171. Villefranc, J.A., J. Amigo, and N.D. Lawson, *Gateway compatible vectors for analysis of gene function in the zebrafish.* Dev Dyn, 2007. **236**(11): p. 3077-87.
172. Ramachandran, R., B.V. Fausett, and D. Goldman, *Ascl1a regulates Muller glia dedifferentiation and retinal regeneration through a Lin-28-dependent, let-7 microRNA signalling pathway.* Nat Cell Biol, 2010. **12**(11): p. 1101-7.
173. Geiss, G.K., et al., *Direct multiplexed measurement of gene expression with color-coded probe pairs.* Nat Biotechnol, 2008. **26**(3): p. 317-25.
174. Hart, D.O., et al., *Selective interaction between Trf3 and Taf3 required for early development and hematopoiesis.* Dev Dyn, 2009. **238**(10): p. 2540-9.

175. Hauptmann, G., *Two-color detection of mRNA transcript localizations in fish and fly embryos using alkaline phosphatase and beta-galactosidase conjugated antibodies.* Dev Genes Evol, 1999. **209**(5): p. 317-21.
176. Kim, S.W., et al., *A sensitive non-radioactive northern blot method to detect small RNAs.* Nucleic Acids Res, 2010. **38**(7): p. e98.
177. Proudfoot, N.J., *Ending the message: poly(A) signals then and now.* Genes Dev, 2011. **25**(17): p. 1770-82.
178. Pauws, E., et al., *Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis.* Nucleic Acids Res, 2001. **29**(8): p. 1690-4.
179. Baldi, P., et al., *Assessing the accuracy of prediction algorithms for classification: an overview.* Bioinformatics, 2000. **16**(5): p. 412-24.
180. Matthews, B.W., *Comparison of the predicted and observed secondary structure of T4 phage lysozyme.* Biochim Biophys Acta, 1975. **405**(2): p. 442-51.
181. Kan, Z., et al., *UTR reconstruction and analysis using genomically aligned EST sequences.* Proc Int Conf Intell Syst Mol Biol, 2000. **8**: p. 218-27.
182. Ulitsky, I., et al., *Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution.* Cell, 2011. **147**(7): p. 1537-50.
183. Murray, E.L. and D.R. Schoenberg, *Assays for determining poly(A) tail length and the polarity of mRNA decay in mammalian cells.* Methods Enzymol, 2008. **448**: p. 483-504.
184. Kimmel, C.B., et al., *Stages of embryonic development of the zebrafish.* Dev Dyn, 1995. **203**(3): p. 253-310.
185. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
186. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
187. Hollander, M. and D.A. Wolfe, *Nonparametric Statistical Methods*. 2nd ed. 1999, New York: John Wiley & Sons.
188. Meyer, D., et al., *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. 2012.
189. Pietrokovski, S., *Searching databases of conserved sequence regions by aligning protein multiple-alignments.* Nucleic Acids Res, 1996. **24**(19): p. 3836-45.

190. R Core Team, *R: A Language and Environment for Statistical Computing*. 2013, R Foundation for Statistical Computing: Vienna, Austria.

191. Bailey, T.L. and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* Proc Int Conf Intell Syst Mol Biol, 1994. **2**: p. 28-36.

192. Martin, G. and W. Keller, *Tailing and 3'-end labeling of RNA with yeast poly(A) polymerase and various nucleotides.* RNA, 1998. **4**(2): p. 226-30.

193. Loke, J.C., et al., *Compilation of mRNA polyadenylation signals in Arabidopsis revealed a new signal element and potential secondary structures.* Plant Physiol, 2005. **138**(3): p. 1457-68.

194. Graveley, B.R., E.S. Fleming, and G.M. Gilmartin, *RNA structure is a critical determinant of poly(A) site recognition by cleavage and polyadenylation specificity factor.* Mol Cell Biol, 1996. **16**(9): p. 4942-51.

195. Caruana, R. and A. Niculescu-Mizil. *An Empirical Comparison of Supervised Learning Algorithms*. in *Proceedings of the 23rd International Conference on Machine Learning*. 2006.

196. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.

197. Goecks, J., A. Nekrutenko, and J. Taylor, *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.* Genome Biol, 2010. **11**(8): p. R86.

198. Giardine, B., et al., *Galaxy: a platform for interactive large-scale genome analysis.* Genome Res, 2005. **15**(10): p. 1451-5.

199. Blankenberg, D., et al., *Galaxy: a web-based genome analysis tool for experimentalists.* Curr Protoc Mol Biol, 2010. **Chapter 19**: p. Unit 19 10 1-21.

200. Nasevicius, A. and S.C. Ekker, *Effective targeted gene 'knockdown' in zebrafish.* Nat Genet, 2000. **26**(2): p. 216-20.

201. Li, W., et al., *The tauCstF-64 polyadenylation protein controls genome expression in testis.* PLoS One, 2012. **7**(10): p. e48373.

202. Wallace, A.M., et al., *Developmental distribution of the polyadenylation protein CstF-64 and the variant tauCstF-64 in mouse and rat testis.* Biol Reprod, 2004. **70**(4): p. 1080-7.

203. Cade, L., et al., *Highly efficient generation of heritable zebrafish gene mutations using homo- and heterodimeric TALENs.* Nucleic Acids Res, 2012. **40**(16): p. 8001-10.

204. Boutet, S.C., et al., *Alternative polyadenylation mediates microRNA regulation of muscle stem cell function.* Cell Stem Cell, 2012. **10**(3): p. 327-36.

205. Hays, W.L., *Statistics*. 1994, Belmont, CA: Wadsworth, a division of Thomson Learning, Inc.

206. Pages, H., *BSgenome: Infrastructure for Biostrings-based genome data packages*.