LEVERAGING HIGH THROUGHPUT TRANSCRIPTOME SEQUENCING TO CHARACTERIZE ALTERNATIVE POLYADENYLATION ACROSS SPECIES

A Dissertation
Presented to the Faculty of the Weill Cornell Graduate School
of Medical Sciences
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Solomon Shenker
August 2016

LEVERAGING HIGH THROUGHPUT TRANSCRIPTOME SEQUENCING TO CHARACTERIZE ALTERNATIVE POLYADENYLATION ACROSS SPECIES

Solomon Shenker, Ph.D.
Cornell University 2016

The recent advance in high-throughput sequencing (HTS) technologies creates enormous opportunity for discovery, concomitant with new challenges for data analysis and interpretion. A major application of HTS includes studies of how the transcriptome is modulated at the levels of gene expression and RNA processing, and how these events are related to cellular identity, environment, and/or disease status. To understand the impact of alternative polyadenylation (APA) events on post-transcriptional gene regulation, I have analyzed deep mammalian RNA-seq data using conservative criteria, and identified thousands of genes that utilize substantially extended novel 3'UTRs in mouse and human. Global tissue comparisons revealed that APA events generating these extensions were most prevalent in the brain. Collectively, these extensions contain thousands of conserved miRNA binding sites, and are strongly enriched for many well-studied neural miRNAs. Altogether, these revised 3'UTR annotations greatly expand the scope of post-transcriptional regulatory networks in mammals. This work further highlights opportunities to improve methods to leverage RNA-seq for 3'UTR annotation and identification of differential APA events. Existing assembly strategies often fragment long 3'UTRs, and importantly, none of the algorithms can be used to infer tandem 3'UTR isoforms directly from RNA-seq data. Consequently, it is often not possible to identify patterns of APA using existing assembly and differential expression testing workflows. To remedy these limitations, I developed a new method for transcript assembly, Isoform Structural Change Model

(IsoSCM) that incorporates change-point analysis to improve the 3'UTR annotation process. Through evaluation on simulated and experimental data sets, I demonstrate that IsoSCM annotates 3' termini with higher sensitivity and specificity than can be achieved with existing methods. I highlight the utility of IsoSCM by demonstrating its ability to recover known patterns of tissue-regulated APA. The methodology encapsulated by IsoSCM will facilitate future efforts for 3'UTR annotation and genome-wide studies of the breadth, regulation, and roles of APA leveraging RNA-seq data. Finally, I describe CrossBrowse, a multi-species genome browser, and use several examples to illustrate how the visualizations generated by CrossBrowse inform comparative data analysis.

BIOGRAPHICAL SKETCH

Sol Shenker completed his undergraduate degree in Biochemistry with a minor in Computer Science at McGill University. After graduating he joined the Tri-Institutational Program in Computational Biology and Medicine. There, he joined the laboratory of Dr. Eric Lai at Memorial Sloan-Kettering Cancer Center, where he explored multiple facets of alternative cleavage and polyadenylation across species. After graduating Sol will join the laboratory of Dr. Levi Garraway, as a Computational Biologist at the Broad Institute.

This dissertation is dedicated to my family, my wife Tamara, and children Valerie and Max, who provide a constant source of wisdom, inspiration, and support, my mother, without whom I would not be here today, and in loving memory of my father who passed away from cancer on March 27th, 2016.

ACKNOWLEDGEMENTS

I feel fortunate to have had the opportunity to benefit from Dr. Eric Lai's mentorship throughout my graduate years. His thorough approach, unbridled curiosity, and dedication provide an inspiring model. I would like to thank all the members of the Lai lab, especially Peter Smibert, Pedro Miura, Jakub Orzechowski Westholm, Piero Sanfilippo, Sonali Majumdar, and Celia Andreu-Agullo, with whom I collaborated throughout on various aspects of alternative polyadenylation. I would also like to thank Jaaved Mohammed, Jiayu Wen, Jakub Orzechowski Westholm, Erik Ladewig, Brian Joseph, Nicolas Robine, and Jeffrey Vedanayagaam for feedback, advice, and discussions throughout.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Highthroughput sequencing reveals transcriptome diversity

Ongoing advances in high-throughput sequencing (HTS) technologies are transforming the face of modern biology. Indeed, coupled with creative methods for library generation, HTS enables the interrogation of genome sequencing, epgigenomic modification, genomic structure, transcriptome, epitranscriptome, on a scale that was technically impossible less than a decade ago [116].

First generation catalogs describing transcript structures were generated from Sanger sequencing of EST and cDNA sequences [1]. While the read length available with these methods is considerably longer than high-throughput short read sequencing technologies used frequently today, similar approaches are used to incorporate these data to extend and refine gene annotations. Initial strategies for transcript identification followed an "assemble-first" strategy[113], wherein cloned DNA sequences are assembled in the absence of a reference genome. The availability of high quality genomes enabled the development of "align-then-assemble" strategies, that leverage the availability of a reference sequence to simplify the assembly process [50]. While recent technological advances in HTS technologies have greatly increased the throughput of DNA-sequencing, these same strategies form the basis of transcript assembly methods developed for HTS platforms [140][43]. Among the numerous applications, HTS datasets can be mined to identify new genes and long non-coding

RNAs (lncRNA) [18], novel classes of transcription [146], and identify alternative RNA processing events [19][67][104][133][145]. However, due to the ambiguities arising from multi-mapping, biased sampling, and overlapping transcription units, these data also require careful interpretation [90].

Production of a mature RNA transcript requires proper execution of a sequence of processing reactions. The last step in the transcription of a mRNA is polyadenylation, the process that forms the mature 3'end of Pol-II transcripts. The polyadenylation reaction occurs co-transcriptionally [83], and is carried out by a complex of proteins. The core components of this complex are CPSF, CFI, CFII and CSTF, PAP are the minimal set required to reconstite the cleavage and polyadenylation reaction[135], although investigation of interaction partners suggest that up to ∼80 additional proteins can interact with the core polyadenylation machinery [125]. The core polyadenylation machinery recognizes cis-sequence elements in the nascent transcript that comprise the poly-A signal (PAS). The most prominant PAS features are a core hexamer, usually A(A/U)UAAA, and a downstream G/U rich sequence (See Figure 1). However, the requirement for these exact motifs is not strict, and sequence mining has identified additional associated cis-elements [42][12][138].

While polyadenylation is an obligate step in generating a stable mRNA transcript, an individual gene can utilize multiple polyadenylation sites, diverifying transcript structure through a process known as alternative polyadenylation (APA) [12][154]. According to some estimates, 95% of human multiexon genes express two or more alternative isoform [104]. Similarly, a large proportion of genes show variation in 3'UTR isoform expression, with estimates of genes subject to APA rangeing between

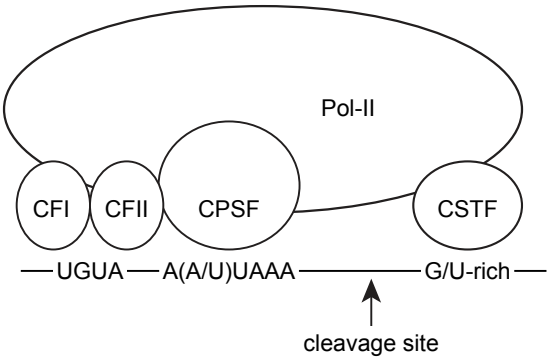


Figure 1: The polyadenyation machinery is composed of four core components, CPSF, CFI, CFII, and CSTF that interact with Pol-II and cis-elements to recognize a poly-A site. The PAS hexamer, A(A/U)UAAA is bound by CPSF, a G/U-rich downstream element is bound by CSTF, and a UGUA motif is bound by CFI.

50-70% [149][138][28]. Recent advances in microfluidics and low-input RNA-seq has revealed that relatively homogenous populations of cells exhibit multimodal expression patterns of alternative splice and polyadenylation isoforms, apparent at the level of single cells [122][143].

Interestingly, 3'UTR expression patterns demonstrate coherent trends when compared between cellular states. For example, differential expression of 3'UTR isoforms is observed to correlate development, tissue, proliferation, neuronal activity, environmental stress, immune cell activation [61][78][152][35][154][121]. While initial observation of global APA phenomena were made using micro-arrays [154], the proliferation of HTS data provides additional oportunities to gain insight into alternative polyadenylation [128, 90]. While conventional RNA-seq is widely used to assay gene expression and alternative splice isoforms, it does not directly capture the location of transcript boundaries. For this reason, specialized protocols to clone the transcript sequences immediately upstream of polyadenylation sites are favorable when precise information about cleavage site position is desired [78]. Notably, even conven-

tial RNA-seq can contain reads spanning the polyadenylation site, allowing a small amount of high resolution information to be extracted from these datasets [107].

1.2 Role of APA in post-transcriptional regulation

The 3'UTR is an important site of post-transcriptional regulation, and is capable of modulating gene activing by influencing transcript stability, translational efficiency, and subcellular localization. These effects are mediated by interactions between primary sequence elements or secondary structures in the 3'UTR and trans-factors such as miRNAs and RNA binding proteins[45]. The importance of post-transcriptional regulation is exemplified by the *C. elegans* germline, where tests of 3'UTR reporters recapitulate cell type specific expression patterns of 24 genes, independent of the endogenous promoter[87].

Although there are examples of factors that are capable of stabilizing bound mRNAs [3][41][66], the regulatory effects of many factors, such as miRNAs and AU-rich elements are thought to have a negative regulatory effect on transcript expression. For this reason, alternative polyadenylation events that result in expression of shortener 3'UTR isoforms are generally thought to increase transcript stability and protein expression[82][121][150], although in specific contexts it has been shown that this dogma does not always hold [47][129]. Moreover, isoforms differing by only a few nucleotides can have different stability [47]. It has been suggested that regulated expression APA isoforms provides a mechanism for ubiquitously transcribed genes to achieve tissue-specific expression via altered post-transcriptional regulation [73].

Of the various roles ascribed to 3'UTRs, their effect on transcript stability has been most thoroughly explored. Using pulse-chase experiments nascent transcripts can be labeled with nucleotide analogs. By measuring transcript abundance as a function of time, the stability of transcript isoforms can be estimated. Expressed transcripts display a wide variation in transcript stability, with measured half-lifes ranging between minutes and hours [114]. By halting transcription, References[150] observed that transcripts with longer 3'UTRs are less stable. More recently, References[129] extended 3'UTRs show a slight but signifant lower stability in 3T3 cells, and References[47] observed no correlation between 3'UTR length and stability in yeast. Beyond 3'UTR length, References[143] examined the importance of additional factors to transcript stability, and showed that the presence of miRNA target sites is most predictive of half-life. In specific cases, the regulatory role of alternative isoforms has been linked with specific cis-elements and trans-factors. For example, deletion of miRNA target sites in the long 3'UTR isoform could explain a portion of the difference in protein expression of constructs using short or long 3'UTR isoforms [121][82]. In a second case, PUF3 RIP is observed to be enriched for isoforms with shorter half-life, and the stability of these isoforms is modified by PUF3 KO [47].

Poor correlation between mRNA and protein levels suggests that additional features beyond transcript abundance influence protein output [65]. An investigation of 200 candidate features of transcripts identified 3'UTR length being among the most strongly correlated with protein abundance in yeast, wherein transcripts with longer 3'UTRs have a lower protein output [144]. Similarly, short 3'UTR isoforms are enriched in the polyribosome fraction in HEK293 cells [132]. However, a clear

relationship between 3'UTR length and protein output does not always exist. In 3T3 cells, longer 3'UTR isoforms display a small but significant association with higher translational efficiency, and comparison between tandem 3'UTR isoforms did not reveal a significant difference in translational efficiency [129]. As the authors acknowledge, it is possible that the regulatory impact of tandem 3'UTR isoforms could be stronger outside of 3T3 cells [129].

Beyond quantitative modulation of transcript stability or translation, alternative processing events can have an effect on expression through quality control pathways [69][40]. Recognition of early polyadenylation signals within the coding sequence can create truncated and non-functional transcripts that are targeted for degradation by non-sense mediated decay (NMD) pathway. Aberrant transcripts are detected by the presence of a premature termination codon (PTC), recognized when the coding sequence is terminated by a stop codon more than ∼50nt upstream of the terminal splice junction [69]. It is estimated that up to 1/3 of alternative splice isoforms generate transcripts that are targeted for NMD [69]. In addition to recognition of PTCs, evidence from cell lines suggests that transcripts with longer 3'UTR are also competent to trigger NMD [126][56].

1.3 Molecular mechanisms regulating APA

APA patterns are identified by comparing isoform abundance between conditions, however there are competing models for how these steady state patterns are established [92]. While differential APA patterns could arise via mechanisms that regulate

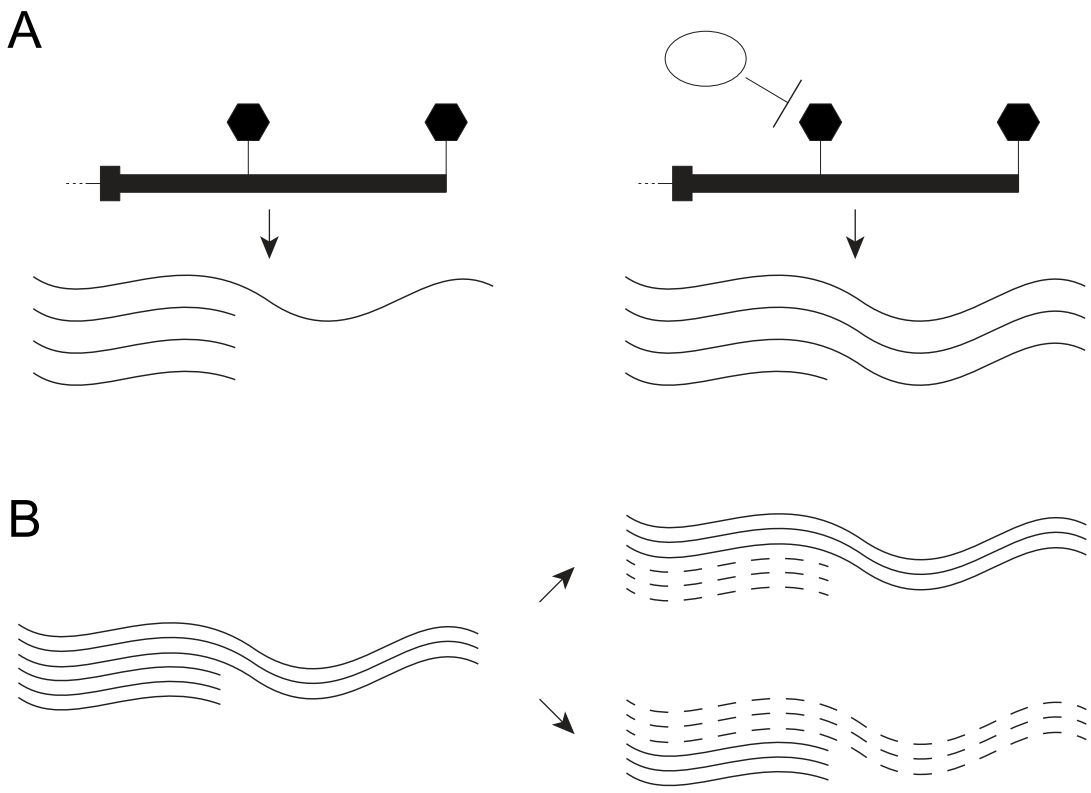


Figure 2: Models for APA regulation (A) Condition specific efficiency of polyadenylation site usage, during transcription (B) Condition specific transcript stability results in differential accumulation of isoforms

polyadenylation site selection, it is also plausible that cell specific regulation of transcript stability plays a role (See Figure 2).

So far, mechanistic studies of the concerted shifts in APA patterns have mostly focused on models for polyadenylation site selection. Some of the first evidence for the role of polyadenylation site selection comes from identification of specific motifs enriched at polyadenylation sites that have biased usage between tissues [154]. Consistent with the polyadenylation machinery being involved, knockdown of CFI_m68 induces shift to proximal isoforms [79]. A concurrent study identified showed that both CFI_m68 and CFI_m25 induce usage of proximal isoforms in HEK293 cells [46].

More recently, profiling of gene expression across cancers has revealed that tumors with abrogated CFI_m25 expression display dramatic 3'UTR shortening[81]. An RNAi screen of 489 RNA binding proteins identified loss of PABPN1 to induce global shift to proximal polyadenylation sites[59]. Conversely, RBBP6, a novel polyadenylation factor identified to interact with known polyadenylation factors, was shown to cause usage of distal polyadenylation sites upon knockdown [29]. In flies, mutants of the RNA binding protein *elav* fail to express extended neural 3'UTR isoforms, while ectopic overexpression of *elav* is sufficient to express extended 3'UTR in an ectopic setting [54]. Tethering experiments suggest *elav* exerts its effect by binding and supressing recognition of proximal polyadenylation signals [54]. In addition to models of RBP competition with the polyadenylation machinery, kinetic mechanisms have also been proposed to affect PAS selection [38].

To evaluate support for the alternative model, where transcript stability determines differential 3'UTR isoform at steady state (See Figure 2), References[129] profiled stability of transcripts using specific 3'ends. By comparing steady state and stability measuresments [129] concluded that differential accumulation of 3'UTR isoforms occurs primarily through differential selection cleavage site in 3T3 cells, and not through transcript stability. However, as the authors note, 3T3 cells are a derived in-vitro system, and they may not be an appropriate model for regulatory impact of APA isoforms observed in-vivo. Along this line, References[10] re-examined this question in an in-vivo context by profiling transcript stability over the course of sperm development. In contrast with the observations from 3T3 cells, References[10] observed UPF to cause selective degradation of extended 3'UTR isoforms in the male

germline.

1.4 APA and disease

The relationship between polyadenylation and disease has long been appreaciated; over 30 years ago, a mutation disrupting a polyadenyation site in beta-globin locus was linked to alpha-thalassaemia [52]. Since then, altered polyadenylation has been associated with a number of diseases, ranging from hematological, immunogical, neurological, endocrine and oncological disorders [26]. Disruption of functional poly-A sites, as in the case typified by alpha-thalasemia, is recurrent across several disease studies [26]. However, other mechanisms such as aberrant poly-A tail length [25], altered isoform expression patterns [82], gain-of-function gene truncation [98], and mis-expression of polyadenylation machinery factors [136] also link polyadenylation with disease etiology.

Over the last few years there has been mounting interest in the role of APA plays in cancer. CCND1 is an example of a oncogene that is dramatically stabilized by 3'UTR truncation[147]. A point mutation in the CCND1 locus generates a novel PAS that removes 1.5kb of 3'UTR sequence, generating a transcript with higher stability and protein output [147][82]. References[105] show that mutation of U2AF1 causes increased recognition of distal PA site of ATG7, resulting in autophagy defect and leading to transformation. These examples suggest that APA could play a more general role in cancer, and promote disease progression through the de-regulation of cancer drivers. Consistent with this idea, GO enrichment of genes with shortened

3'UTRs between cancer versus normal cell lines identified enrichment of mitotic cell-cycle related terms [37].

Beyond a potentially causal role in disease progression, differential APA patterns can be used to predict survival characteristics of histologically indistinguishable lymphoma subtypes with 74% accuracy, underscoring their diagnostic potential [127]. More recently, it has been shown that 3'UTR isoform usage can be used to classify triple negative breast cancer from healthy tissue, and shows prognostic power over 10-year relapse survival study[2].

1.5 Overview of dissertation

1.5.1 Expanding mamallian 3'UTR annotations using RNA-seq

As mentioned above, the proliferation of HTS datasets provides ample opportunity to perform retrospective analysis of published datasets and revisit new questions. Motivated by our prior observations of 3'UTR lengthening in the fly nervous system [128], I investigated tissue APA patterns mouse and human using a compendium of published RNA-seq datasets in Chapter 2. Thousands of gene models in mouse and human could be revised to include 3'UTR extensions supported by RNA-seq data. Examination of sequence conservation patterns and measurements of RNA:protein interactions suggest that these 3'UTR extensions contain elements important for post-transcriptional regulation. Surprisingly, the extended 3'UTR isoforms overlap significantly with recently reported lincRNAs from multiple studies, illustrating the challenges associated with interpreting short-read sequencing data, and highlighting opportunities to ex-

tend existing methods for transcript assembly.

1.5.2 Change-point analysis identifies APA using RNA-seq

In Chapter 3 I formalize these extensions in the method IsoSCM, a transcript assembly tool designed to address the challenges of inferring 3'UTR models from RNA-seq data. In addition to handling the often discontinuous coverage patterns that can cause long 3'UTRs to be fragmented by naive assembly approaches, IsoSCM incorporates statistical change-point detection to enable annotation of tandem 3'UTR isoforms, a functionality that was not provided by existing assembly algorithms. In addition to improving the quality of 3'UTR annotations, change-point analysis also provides a natural framework for detecting differential polyadenylation events between multiple samples. Importantly, these attributes enable IsoSCM to analyze APA in a *de novo* setting, where accurate annotations are not available.

1.5.3 Visualizing genomics data across species

As a next step, I sought to gain insight into the evolutionary dynamics of APA by examining differential polyadenylation patterns between species. While cross-species experiments are becoming more and more common, there is no generic solution for visualizing genomics data across multiple reference sequences. In Chapter 4 I introduce CrossBrowse, a standalone application for visualizing high-throughput datasets at syntenic loci in two or more genomes. I demonstrate the broad utility of CrossBrowse with examples that span transcriptional and post-transcriptional gene regulation. Of note, I illustrate how visual inspection of data in CrossBrowse can reveal artifacts that

are otherwise not apparent. By leveraging standard genomics data formats, Cross-Browse aims to provide a generic solution for visualization of multi-species genomics experiments.

2 Widespread and extensive lengthening of 3' UTRs in the mammalian brain*

2.1 Attributions

Sol Shenker designed and implemented the pipeline for identifying 3'UTR extensions, performed cross-reference with ncRNAs, analyzed PA-site features, differential expression analysis, analyzed miRNA target site enrichment, and AGO-CLIP analysis. Pedro Miura performed PCR, northern blot analysis, and provided input for the design of the 3'UTR analysis pipeline. Jakub Orzechowski Westholm assisted with the miRNA analysis. Celia Andreu-Agullo performed in situ hybridization experiments.

2.2 Abstract

Remarkable advances in techniques for gene expression profiling have radically changed our knowledge of the transcriptome. Recently, the mammalian brain was reported to express many long intergenic noncoding (lincRNAs) from loci downstream from protein-coding genes. Our experimental tests failed to validate specific accumulation of lincRNA transcripts, and instead revealed strongly distal 3' UTRs generated by alternative cleavage and polyadenylation (APA). With this perspective in mind, we analyzed deep mammalian RNA-seq data using conservative criteria, and identified 2035 mouse and 1847 human genes that utilize substantially distal novel 3' UTRs. Each of these extends at least 500 bases past the most distal 3' termini available in

***P. Miura**, **S. Shenker**, C. Andreu-Agullo, J. O. Westholm, and E. C. Lai. Widespread and extensive lengthening of 3' UTRs in the mammalian brain. *Genome Research*, 23(5):812–825, May 2013.

Ensembl v65, and collectively they add 6.6 Mb and 5.1 Mb to the mRNA space of mouse and human, respectively. Extensive Northern analyses validated stable accumulation of distal APA isoforms, including transcripts bearing exceptionally long 3' UTRs (many >10 kb and some >18 kb in length). The Northern data further illustrate that the extensions we annotated were not due to unprocessed transcriptional run-off events. Global tissue comparisons revealed that APA events yielding these extensions were most prevalent in the mouse and human brain. Finally, these extensions collectively contain thousands of conserved miRNA binding sites, and these are strongly enriched for many well-studied neural miRNAs. Altogether, these new 3' UTR annotations greatly expand the scope of post-transcriptional regulatory networks in mammals, and have particular impact on the central nervous system.

2.3 Introduction

The 3' untranslated regions (3' UTRs) of mRNAs contain cis elements that confer post-transcriptional regulation by RNA-binding proteins (RBPs) and microRNAs (miRNAs) [74]. It is now appreciated that alternative cleavage and polyadenylation (APA) generates tremendous transcript diversity, and the majority of genes have multiple functional polyadenylation (polyA) sites. The dominant class of APA events occurs within terminal exons, causing 3' UTR shortening or lengthening. Global 3' UTR shortening is characteristic of proliferating cells and cancer cells [121, 82], whereas 3' UTR lengthening was reported to occur during embryonic development and differentiation [60].

Microarray analysis of assorted tissues indicated that the mammalian brain broadly utilizes distal 3' UTR species [121, 145]. Deep sequencing of 3' ends of polyadenylated transcripts uncovered hundreds of distal APA events in cultured neurons compared with embryonic stem (ES) cells [124], and the picture was broadened by the recognition of more than 1000 3' UTR extensions in mouse cerebellum [103]. Most recently, data from Drosophila tiling microarrays [53] and tissue-specific RNA-seq [128] revealed central nervous system (CNS)-specific 3' UTR extensions across hundreds of transcripts, indicating a conserved phenomenon for 3' UTR lengthening in the nervous system.

Diverse regulatory consequences of APA in the nervous system have been described. Variation of 3' UTR lengths can alter transcript regulation and stability by inclusion or exclusion of miRNA binding sites [21] or other RBP sites. Global analysis of protein-RNA interactions by high-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation (HITS-CLIP) revealed that NOVA1, a neural RBP best-characterized as a splicing regulator, also has extensive influence on APA [75]. As well, some neural 3' UTR extensions direct localization to dendrites and axons, which can provide spatial specificity and/or facilitate their translation [5, 153, 9].

Advances in expression profiling, from tiling microarrays to RNA-seq, promise to reveal a comprehensive view of the transcriptome. This includes a fuller accounting of protein-coding transcript isoforms as well as noncoding transcripts, such as small RNAs and long intergenic noncoding RNAs (lincRNAs). De novo construction of gene models from high-throughput data has identified a plethora of unannotated exons, which have been inferred to include thousands of lincRNAs. However, accurate recon-

struction of parent transcripts, especially when alternative products emanate from a given locus, remains challenging. For example, initial tiling microarray studies of the Drosophila transcriptome [77] showed that a substantial fraction of novel intergenic transcribed regions actually represented alternative 5' noncoding exons of downstream protein-coding gene models, sometimes located tens of kilobases away. On the other end, we recognized that APA frequently generates unanticipated Drosophila 3' UTRs of exceptional length, also ranging up to tens of kilobases [128].

Here, we reassessed previously studied lincRNAs expressed in mammalian brain [109, 22], and found that many represent stable 3' UTR extensions of upstream protein-coding genes. We then used RNA-seq data to uncover thousands of previously unannotated 3' UTR extensions in mouse and human. The predominant tissue-specific trend was for utilization of 3' UTR extensions in brain, and this has substantial impact on post-transcriptional regulatory networks, including by thousands of conserved miRNA binding sites. These findings strongly revise the scope of mammalian transcriptomes, and highlight that a full appreciation of even their protein-coding gene models remains to be realized.

2.4 Results

2.4.1 Reevaluation of proposed neural mRNA/lincRNA pairs instead reveals distal APA isoforms

Previous searches for evolutionarily constrained long intergenic noncoding RNAs (lincRNAs) noted more than 200 brain-expressed lincRNAs originating downstream from

RefSeq protein-coding genes, and some were spatially coexpressed with their upstream neighbors [109]. Experimental and bioinformatic tests argued against connectivity of these coding/noncoding pairs. Nevertheless, we observed many lincRNAs resided in regions of RNA-seq coverage continuous with upstream genes (e.g., Ago3 [also known as eIF2C3]) (Figure 3A). Consequently, we reevaluated mRNA-lincRNA pairs previously reported as experimentally negative for connectivity [109].

Interestingly, we observed reverse transcriptase-dependent PCR products from post-natal brain that join the annotated mRNAs of Mitf, Gabrb1, Ago3/eIF2C3, Ar, and Rbms1 with their reported downstream lincRNAs (Figure 3B; Figure 4). The only pair not validated was Ube2k-AK045737. However, stranded RNA-seq data revealed transcription of AK045737 exclusively on the opposite strand (Figure 4). This places AK045737 downstream from Pds5a, with intervening spliced reads, and rt-PCR products joined these loci (Figure 3B).

More definitive information on transcript connectivity and alternative isoforms is provided by Northern analysis. We designed paired probes targeting the terminal coding region/proximal 3' UTR of the annotated mRNAs and their proposed downstream noncoding RNAs [109]. We did not detect Mitf1 (data not shown), we but observed robust signals for the rest in cortex or cerebellum (Figure 3C). Notably, the dominant bands detected by mRNA probes were always substantially larger than the RefSeq-annotated transcripts, and these always cohybridized with their downstream lincRNA probes (Figure 3C). Moreover, no distal probes identified shorter bands of lengths predicted for the reported lincRNAs. While these data do not rule out the possibility of distinct lowly expressed lincRNAs, they demonstrate the bulk of stable

Figure 3: A) Experimental strategy to test connectivity between a protein-coding gene and a downstream lincRNA. RNA-seq and polyA-seq evidence in the vicinity of eIF2C3 (also known as Ago3) and the proposed lincRNA AK047638. We designed primers to amplify bridge rt-PCR products and Northern probes, as shown. B) Bridge rt-PCR using adult cerebral cortex RNA connects many protein-coding genes with their proposed downstream neighboring lincRNAs (Ponjavic et al. 2009); note Ar was previously referred to as Adr, and Ube2k was termed Hip2. Note that AK045737 was proposed to be a pair with Ube2k (Ponjavic et al. 2009); however, stranded RNA-seq data revealed that AK045737 is continuous with a spliced exon of Pds5a transcribed from the other strand (see Supplemental Figure S1A-C). C) Northern analysis demonstrates that the predominant transcripts detected by probes for the protein-coding loci assayed in B are codetected by probes against their neighboring downstream lincRNAs. Conversely, we did not detect stable transcripts corresponding to the sizes of the annotated lincRNAs. Northern blots are also shown for ncRNAs described by Mattick and colleagues (Clark et al. 2012) and their protein-coding pairs Etv1 and Paqr9. D) RNA-seq and PolyA-Seq tracks for cases of annotated lincRNAs that appear to be contained with exceptionally long, continuous 3' UTRs of stable mRNAs. E) Northern blots for proposed lincRNAs show a band of exceptional length that is of the same molecular weight as the bands identified by probes corresponding to the upstream protein-coding transcripts. Arrowheads identify dominant bands that correspond to size estimates based on RNA-seq data. Note that the sizes of the bands on the Northern blot are consistent with the RNA-seq evidence-based size estimates. Asterisks denote 28S and 18S ribosomal bands corresponding to 4.7 kb and 1.9 kb, respectively. Ladder information can be found in Supplemental Figure S4. For RNA-seq tracks, probe locations, and gene annotations, see Supplemental Figure S1D.

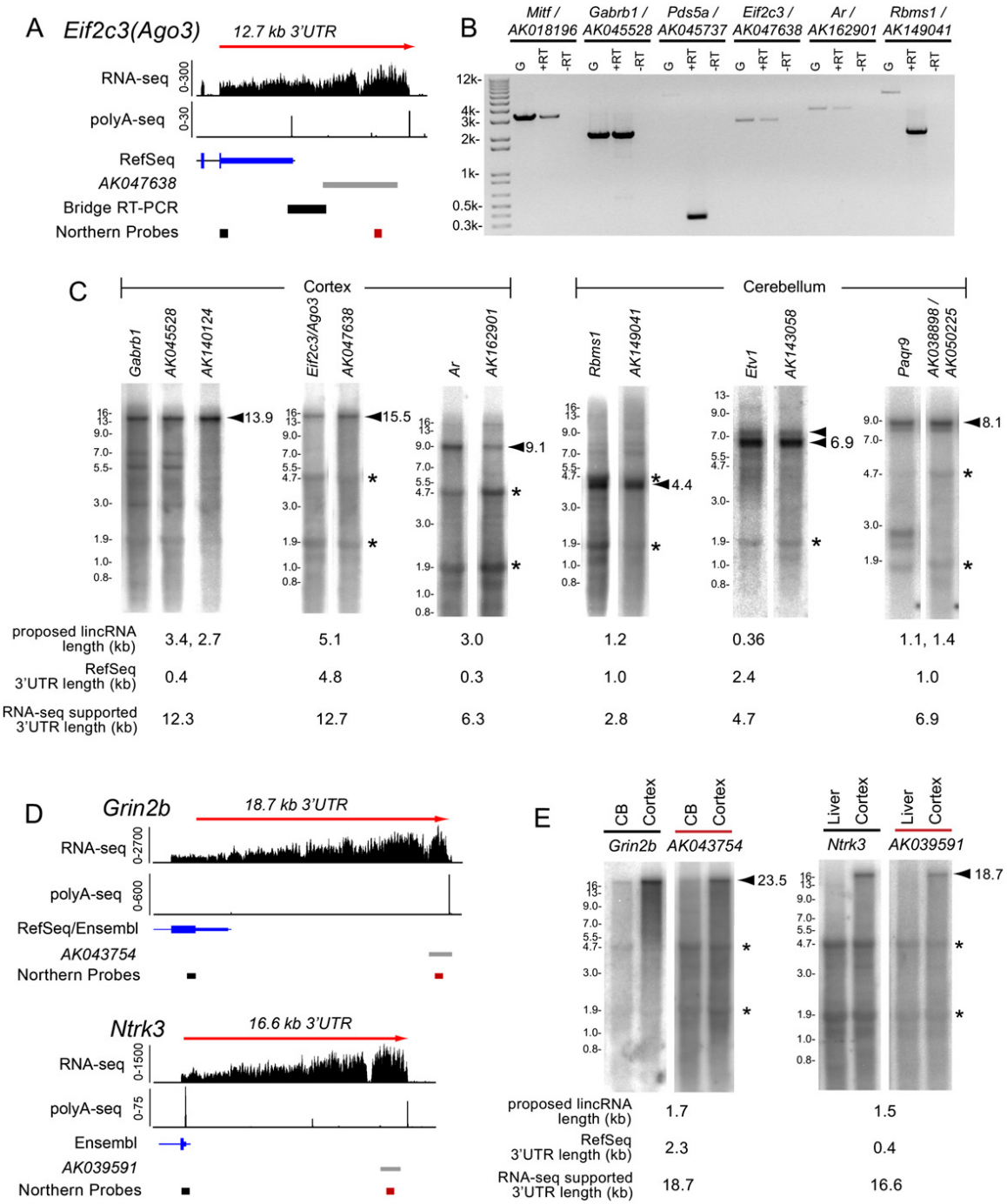

A
Eif2c3(Ago3)
12.7 kb 3'UTR
RNA-seq
polyA-seq
RefSeq
AK047638
Bridge RT-PCR
Northern Probes
B
Mitf / AK018196
Gabrb1 / AK045528
Pds5a / AK045737
Eif2c3 / AK047638
Ar / AK162901
Rbms1 / AK149041
G
+RT
-RT
C
Cortex
Cerebellum
Gabrb1
AK045528
AK140124
Eif2c3/Ago3
AK047638
Ar
AK162901
Rbms1
AK149041
Etv1
AK143058
Paqr9
AK038898 / AK050225
proposed lincRNA length (kb)
3.4, 2.7
5.1
3.0
1.2
0.36
1.1, 1.4
RefSeq 3'UTR length (kb)
0.4
4.8
0.3
1.0
2.4
1.0
RNA-seq supported 3'UTR length (kb)
12.3
12.7
6.3
2.8
4.7
6.9
D
Grin2b
18.7 kb 3'UTR
RefSeq/Ensembl
AK043754
Northern Probes
Ntrk3
16.6 kb 3'UTR
Ensembl
AK039591
E
CB
Cortex
Liver
Grin2b
AK043754
Ntrk3
AK039591
proposed lincRNA length (kb)
1.7
1.5
RefSeq 3'UTR length (kb)
2.3
0.4
RNA-seq supported 3'UTR length (kb)
18.7
16.6

Figure 3

Figure 4: RNA seq and PolyA-seq evidence for 3' UTR extensions of proposed to coding-non-coding pairs. Location of northern probe coordinates and Bridge RT-PCR coordinates are shown. RNAseq evidence is from hippocampus pooled libraries, PolyA-Seq evidence is from pooled tissue libraries [28]. A) AK0457347 was previously designated as a non-coding pair of Hip2/Ube2k. However, stranded RNA-seq data [106] provides evidence that AK0457347 is an unannotated 3' exon spliced from the within the coding region of the Pds5a terminal exon. (Blue: plus strand; Red: minus strand). B) Northern analysis demonstrates that the AK0457347 locus is actually part of a spliced terminal exon of Pds5a. C) The unannotated terminal exon of Pds5a also maps to a genomic loci on Chromosome 17. The northern probe N.397 may cross-hybridize to transcripts expressed at this loci, resulting in a <2kb predicted transcript. D) RNA-seq evidence demonstrating that other annotated lincRNAs downstream of RefSeq protein-coding gene models actually represent 3' UTR extensions. Note that previously unannotated 3' UTR extension regions for Mitf, Eif2c3, Ar and Rbms1 are currently annotated in Ensembl v65.

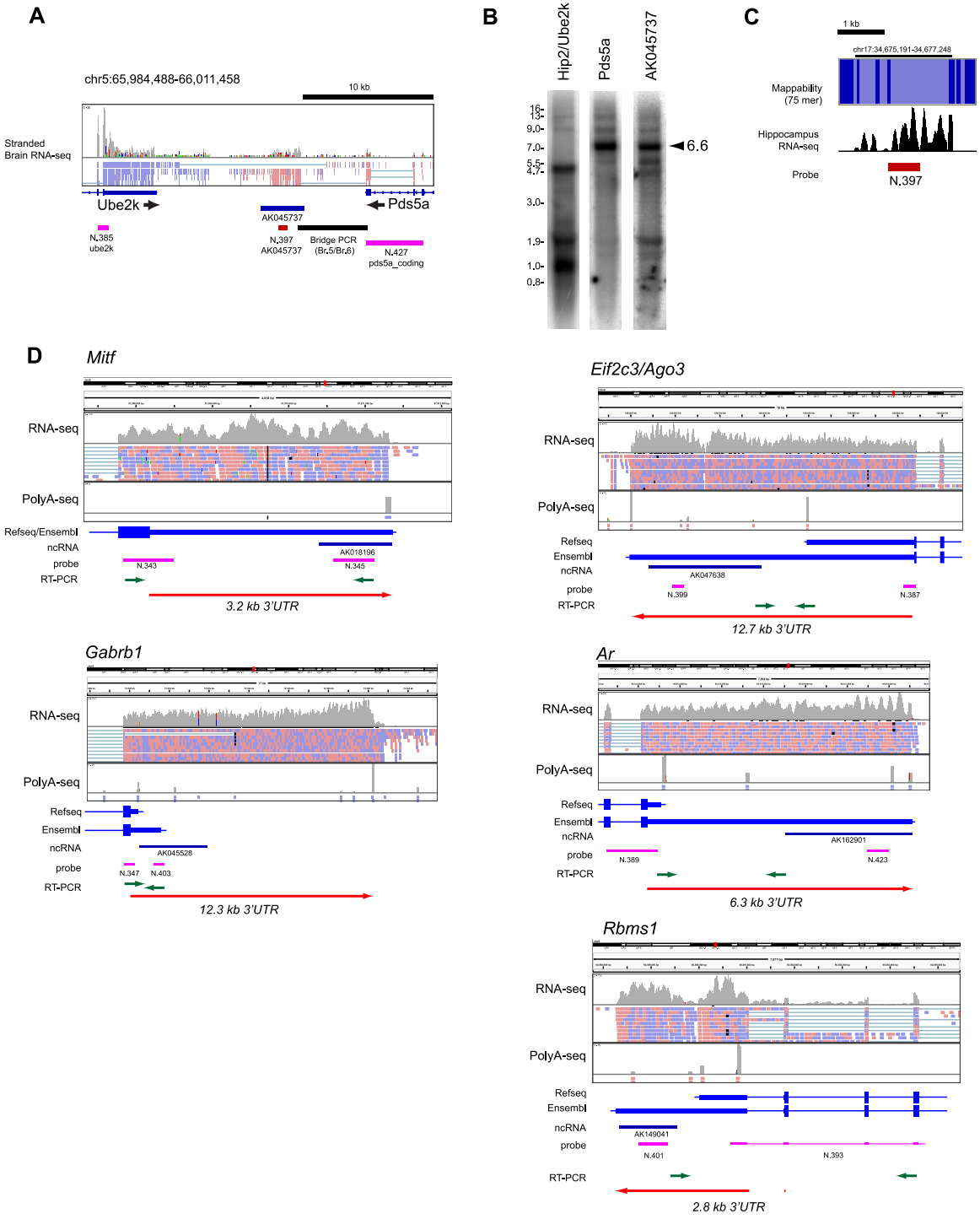

A
chr5:65,984,488-66,011,458
10 kb
Stranded
Brain RNA-seq
Ube2k
Pds5a
AK045737
N.385
ube2k
N.397
AK045737
Bridge PCR
(Br.5/Br.6)
N.427
pds5a_coding
B
Hip2/Ube2k
Pds5a
AK045737
16-
13-
9.0-
7.0-
5.5-
4.7-
3.0-
1.9-
1.0-
0.8-
6.6
C
1 kb
chr17:34,675,191-34,677,248
Mappability
(75 mer)
Hippocampus
RNA-seq
Probe
N.397
D
Mitf
RNA-seq
PolyA-seq
Refseq/Ensembl
ncRNA
probe
RT-PCR
AK018196
N.343
N.345
3.2 kb 3'UTR
Eif2c3/Ago3
RNA-seq
PolyA-seq
Refseq
Ensembl
ncRNA
probe
RT-PCR
AK047638
N.399
N.387
12.7 kb 3'UTR
Gabrb1
RNA-seq
PolyA-seq
Refseq
Ensembl
ncRNA
probe
RT-PCR
AK045528
N.347
N.403
12.3 kb 3'UTR
Ar
RNA-seq
PolyA-seq
Refseq
Ensembl
ncRNA
probe
RT-PCR
AK162901
N.389
N.423
6.3 kb 3'UTR
Rbms1
RNA-seq
PolyA-seq
Refseq
Ensembl
ncRNA
probe
RT-PCR
AK149041
N.401
N.393
2.8 kb 3'UTR

Figure 4

transcripts bearing these noncoding sequences to be 3' UTR extensions of mRNAs.

Some lincRNAs are substantially separated from known mRNA termini. The lincRNA AK043754 was recently studied in detail [22], and its locus resides 14.9 kb downstream from Grin2b. Examination of hippocampal RNA-seq data [62] revealed continuous coverage from the annotated Grin2b 3' UTR to AK043754 (Figure 3D). Our Northern analysis for AK043754 revealed a single, strong ∼23.5-kb band in cerebral cortex. A probe against Grin2b coding sequence identified the same band; thus, Grin2b expresses an ∼19-kb 3' UTR in brain (Figure 3E). We obtained similar results by Northern analysis of the proposed lincRNA AK039591 [109], which is contained within the 16.6-kb 3' UTR of Ntrk3 (Figure 3D,E). That the same large bands were specifically detected by probes against very proximal and very distal portions of these long 3' UTRs rules out potential alternative events 5' of the stop codon (i.e., retained internal introns, alternative splicing, or alternative promoters). Grin2b and Ntrk3 contain some of the longest stable 3' UTRs ever demonstrated in mammals, highlighting that some "very downstream" lincRNAs can be reinterpreted as 3' UTR extensions.

2.4.2 A bioinformatic pipeline to call 3' UTR extensions using RNA-seq data

None of the 3' UTR extensions analyzed in Figure 3 are present in RefSeq, although many are well-studied genes. We therefore sought to annotate 3' UTR extensions more comprehensively. Searching other databases, we observed that Ensembl currently annotates Ago3/eIF2C3, Ar, and Rbms1 3' UTR extensions; however, neither

database includes the distal 3' UTRs for Gabrb1, Pds5a, Ntrk3, or Grin2b. We consequently used the latest Ensembl version 65 (v65) [36] as a conservative reference for annotating novel 3' UTRs.

We took advantage of deep RNA-seq data from six mouse tissues [62] and reprocessed the raw data to map ∼1.7 billion reads. We initially generated transcript models using Cufflinks [140], but we noticed from browsing its outputs that 3' UTR extents were frequently truncated due to variable read depth, discontinuous coverage, and/or multimapper reads. We therefore developed an alternate approach to identify 3' UTR extensions.

The key features included a sliding window that identified continuously transcribed genomic segments ≥1.0 fragments per kilobase of transcript per million mapped fragments (FPKM) (empirically determined from recall of known, internal, nonalternative exons), followed by judicious merging of adjacent contigs split by lower coverage regions or repeats. To ensure that merging was conservative, we demanded that gaps were bridged by paired-end reads, limited nonrepetitive gaps to <150 nucleotides (nt), and restricted novel extensions from containing >20% of repetitive sequence. We then made extensive efforts to cull potentially ambiguous 3' UTR extensions using many additional filtering steps. We grouped together novel 3' ends from different tissues that were within 30 nt of each other, and extensively confirmed the final calls by visual inspection. Our stringent filtering steps removed some genuine 3' UTR extensions, but we preferred this conservative approach to focus on 3' UTRs of high confidence. The pipeline is described in detail in the Methods.

2.4.3 Analysis of mouse and human RNA-seq data using stringent criteria reveals more than 3850 novel 3' UTR extensions

We compared our 3' UTR calls from six mouse tissues to Ensembl v65 annotations to identify 3' UTR extensions. To focus on substantially novel isoforms, we required Ensembl gene models be extended >500 nt. Although some stable 3' UTR extensions that we validated by Northern failed our conservative annotation pipeline, this analysis strikingly identified 2035 confident 3' UTR extensions over Ensembl v65 mouse gene models. These 3' UTRs comprise ∼6.6 Mb of unannotated sequence and average 4347 nt in length, far above the Ensembl v65 average of 989 nt (Figure 5A). A recent analysis of mouse cerebellum identified many 3' UTR extensions [103], of which 600 remain exclusive of Ensembl v65. Still, our annotations comprised 6.2 Mb sequence beyond these models.

We performed similar analysis of the human transcriptome using the Illumina Body Map 2.0 of 16 tissues, including a pooled stranded data set that confirmed expression directionality. We reprocessed these from the raw data and mapped more than 4 billion reads. We used the same pipeline to annotate 3' UTR extensions, except that we increased expression cutoffs to 1.5 FPKM (based on recall of known nonalternative exons). This analysis identified 1847 confident 3' UTR extensions over Ensembl v65 (Figure 5B), comprising 5.1 Mb of unannotated sequence (Figure 5B).

Altogether, these thousands of confident 3' UTR extensions add ∼11.1 Mb to mouse and human protein-coding gene models. Moreover, we further designated thousands of candidate loci in mouse and human with compelling evidence for exten-

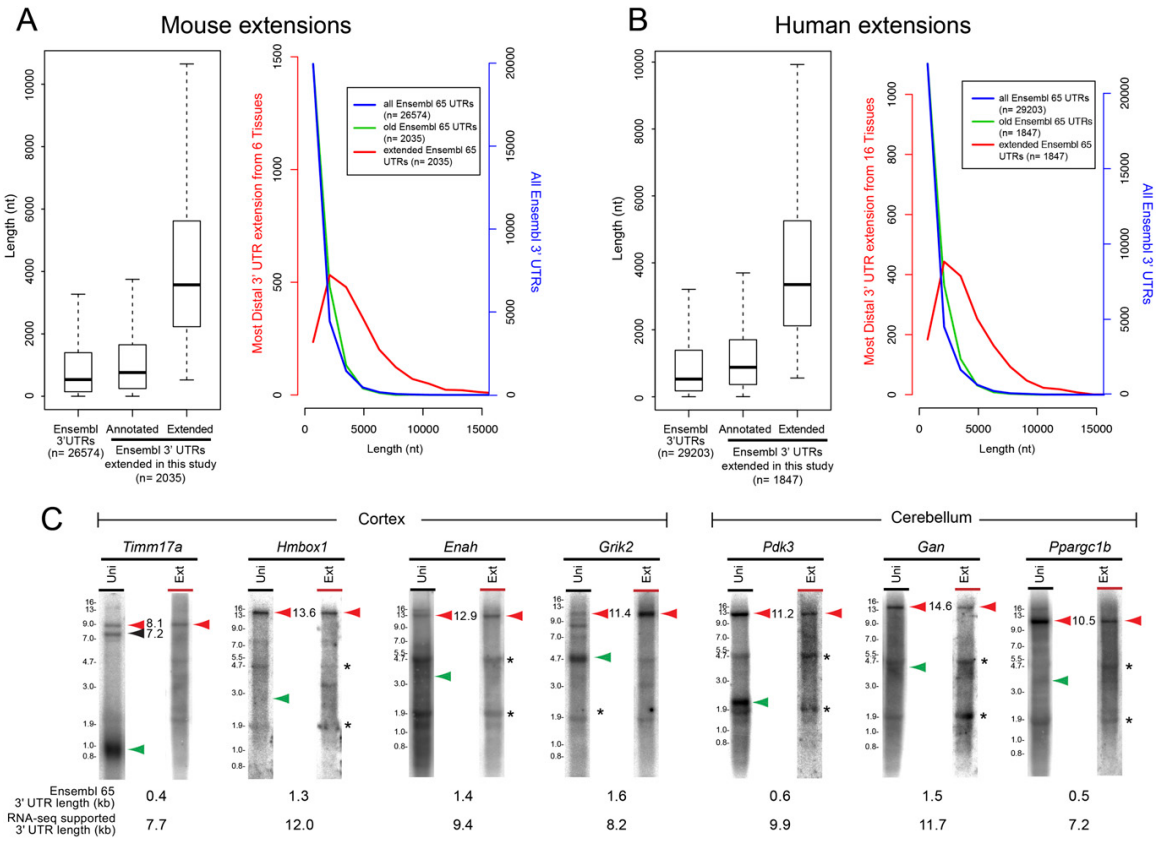


Figure 5: A) Box plot comparing aggregate Ensembl v65 3' UTR lengths (longest annotation per terminal exon) and those Ensembl v65 3' UTRs that were specifically extended in this study. (Right) Histogram of the same data to highlight the abundance of newly annotated long 3' UTRs. B) Analyses similar to A, except plotting known and novel Ensembl v65 human 3' UTRs. C) Northern analysis validates the stable accumulation of many transcripts utilizing very distal polyadenylation signals in cerebellum or cortex, in several cases yielding 3' UTRs >10 kb in length. Green arrowheads indicate predicted mRNA length of Ensembl v65 gene model. Red arrowheads indicate inferred mRNA lengths of novel 3' UTR extension isoforms. Asterisks denote background hybridization to ribosomal RNAs.

sion but failed to meet our full criteria. These included loci bearing 300- to 499-nt extensions relative to Ensembl v65, that were expressed at 0.5-0.99 FPKM, or that had small gaps not bridged by paired-end reads. Thus, the scope of mammalian 3' UTR extensions is likely even larger than we currently annotate.

2.4.4 Experimental support for extended 3' UTR isoforms

We vetted the veracity of 3' UTR extensions by systematic visual inspection, coupled with extensive "gold-standard" Northern evidence. Because commercial RNA ladders only size to 9 kb, we generated a "virtual ladder" composed of endogenous transcripts from 0.8-16 kb (Figure 6). We implemented this by hybridizing mixed probe to stripped blots, a strategy that furthermore reported the integrity of long transcripts.

We focused on genes expressed in the cortex and/or cerebellum. We did not detect specific Kcna4 or Nxph1 transcripts (data not shown), reflecting technical failures or low abundance. However, all proximal 3' UTR probes that detected specific transcripts (e.g., Timm17a, Grik2, Enah, Pdk3, Gan, and Ppargc1b) revealed long species corresponding in length to distal APA isoforms inferred from RNA-seq (Figure 5C). We re-probed for their distal extensions using amplicons separated from the proximal probes by the length of the 3' UTR extension. We consistently detected the same large transcripts with paired proximal and distal probes (Figure 3C,E, Figure 5C), constituting unambiguous evidence for stable mRNAs bearing long 3' UTRs. Moreover, these data comprise strong evidence against possibilities that the underlying RNA-seq data reflect heterogeneous runaway transcription products, pre-mRNA intermediates, or unstable transcripts in the process of being degraded.

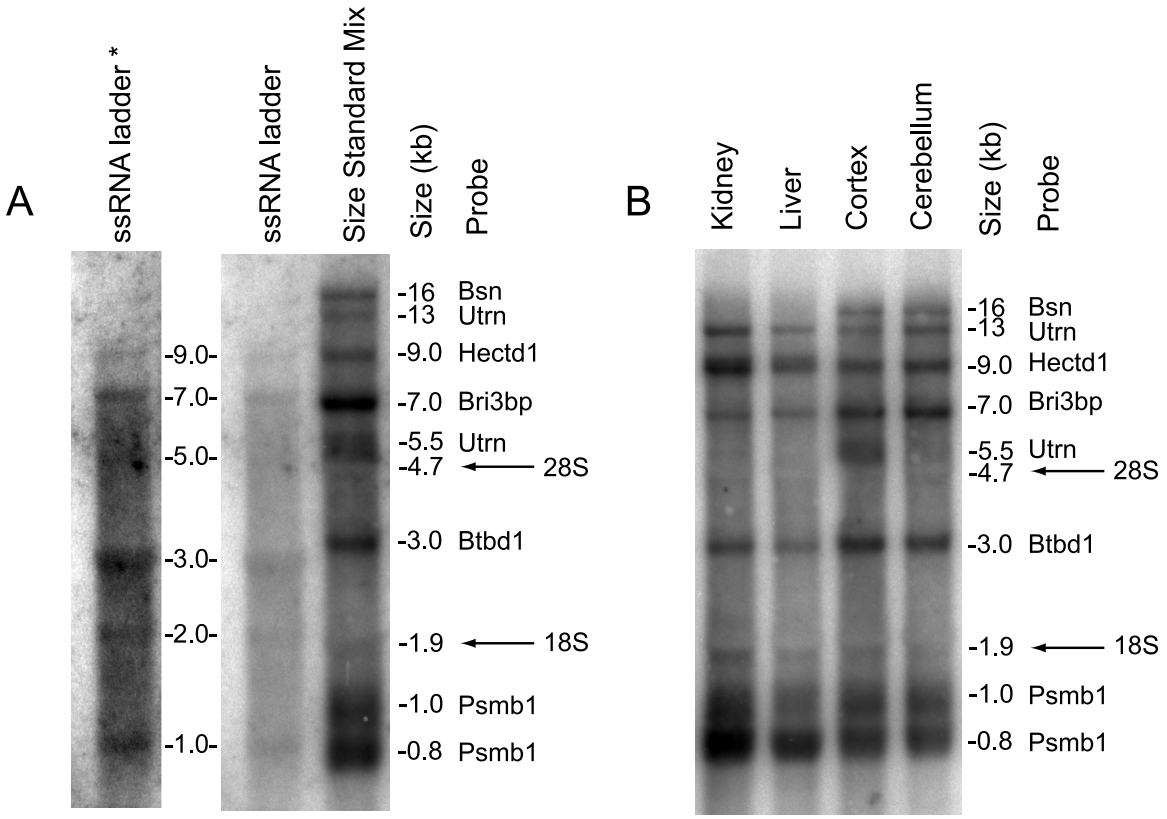


Figure 6: Internal size standards used for northern blots compared to single-stranded RNA ladder. A) Cortex RNA northern blot. B) Multiple tissues northern blot. Note the predicted size for the size standard appears 200 nts longer than the ssRNA ladder due to the presence of endogenous polyA tails.

Several of these genes exhibit >10-kb 3' UTRs, adding to other long extensions validated earlier (Figure 3D). Overall, our high validation rate, using the modestly sensitive Northern technique, reflects the stringency of the annotation pipeline. We also note validation of some loci that did not meet our full bioinformatic criteria. For example, we clearly observed the strongly extended 12-kb 3' UTR of Hmbox1 by Northern, even though it overlaps an annotated alternative last exon and was therefore culled from our pipeline (Figure 5C). This highlights the conservative nature of our annotations, and that our candidate lists undoubtedly contain additional genuine 3' UTR extensions.

2.4.5 Analysis of polyA signals and conservation among novel distal 3' termini

We investigated the characteristics of novel distal 3' termini. Although our pipeline assesses confident 3' UTR extensions, it does not necessarily pinpoint precise 3' ends, especially as RNA-seq protocols undersample near transcript termini. We further noted many instances of likely extensions whose most distal regions did not satisfy our expression cutoff and are thus truncated by our pipeline. We therefore implemented a "dropoff" filter to identify extension calls that coincide with a sharp drop in RNA-seq coverage. We required that two consecutive 100-nt windows downstream from the 3' end call exhibit greater than eightfold reduction in reads, relative to the final 100-nt window of the 3' extension. Since some extensions terminated in repetitive regions, and thus lacked precise 3' end calls, we also culled these from motif analysis. This yielded 691 extended mouse loci comprising 741 distinct 3' ends and 697 extended

human loci totaling 816 distinct 3' ends.

We compared the properties of our novel 3' ends with their annotated Ensembl v65 counterparts. Known 3' termini usually bear canonical polyadenylation signals (PASs) AAUAAA or AUUAAA ∼35 nt upstream of transcript ends, with lower and less specific enrichment for various noncanonical PAS [138]. In addition, U/GU motifs are enriched downstream from known PAS (for motif definitions, see Methods). We observed all of these features in the vicinity of annotated mouse (Figure 7A) and human (Figure 7B) 3' termini, in proportions consistent with previous global analyses of mammalian 3' ends. Curiously, we also observed ∼15% genomically encoded AAAAAA among both mouse and human Ensembl v65 termini, suggesting that some of these annotations may potentially derive from internally primed cDNAs.

Our novel distal 3' UTR extensions exhibited strong positional enrichments of upstream PAS and downstream U/GU motifs, in both mouse (Figure 7C) and human (Figure 7D). In fact, the enrichment of canonical AAUAAA PAS was greater among our novel 3' termini, compared with their Ensembl termini. Moreover, we did not observe enrichment of AAAAAA polymers at our newly defined termini. These observations provided strong support for the quality of our extension annotations. Therefore, while we do not presume that every novel terminus was defined precisely, this set of more than 1500 novel distal termini exhibits motif properties of genuine transcript ends that meet or exceed those of well-annotated Ensembl transcript ends.

We next analyzed the conservation of proximal and distal termini using phastCons values. For both mouse (Figure 7E) and human (Figure 7F), we observed a local spike in conservation 5' to the proximal polyadenylation sites, followed by a drop in conser-

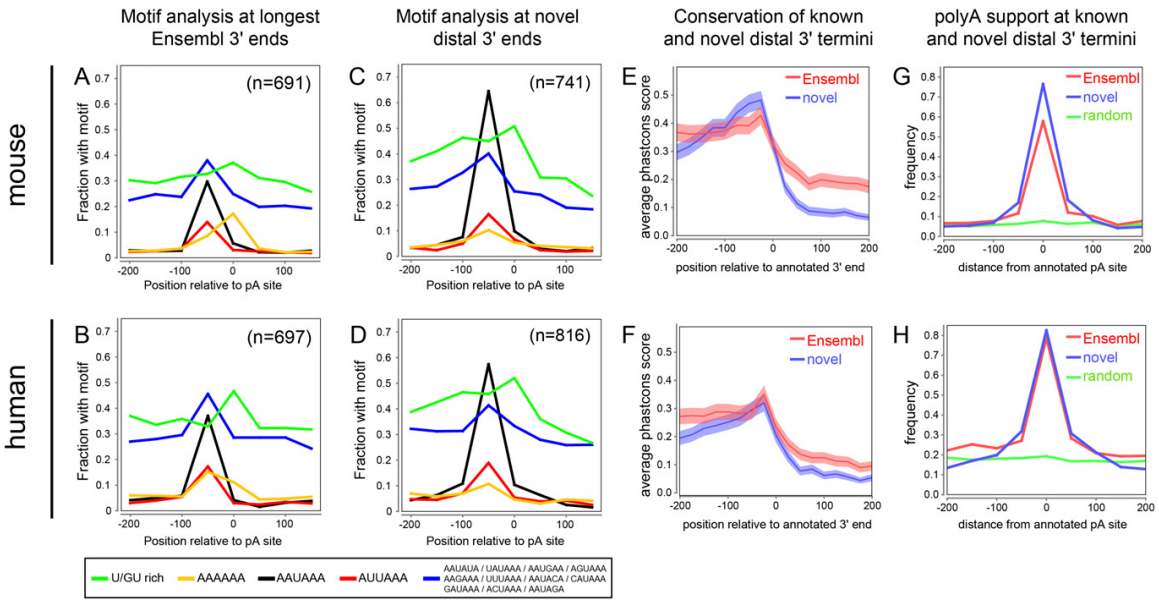


Figure 7: All analyses in this figure concern those mouse (top graphs) and human (bottom graphs) genes whose 3' UTRs were confidently extended in this study. These comprise 691 Ensembl65 mouse gene models for which we precisely annotate 741 novel 3' termini in one or more tissues, and 697 Ensembl65 human genes for which we precisely annotate 816 novel 3' termini in one or more tissues. (A,B) Motif frequency in 50-nt bins in the vicinity of annotated 3' termini. Motifs are listed at bottom, and include the downstream U/GU-rich region that promotes 3' cleavage, the canonical PAS AAUAAA and its most common variant AUUAAA, a panel of low-frequency PAS variants, and genomically encoded hexa-A tracts. As expected for annotated mouse (A) and human (B) 3' termini, there is strong positional enrichment of functional PAS upstream of the polyadenylation site and U/GU downstream. The collection of low-frequency PAS variants exhibits a broad background frequency, with mild enrichment at the normal location of canonical PAS. Unexpectedly, we observed enrichment of A6 at annotated 3' termini, potentially reflecting internal priming events in this collection of curated 3' termini. (C,D) The frequency and positional specificity of PAS and U/GU motifs in our novel mouse (C) and human (D) 3' termini are relatively similar to known termini but lack substantial A6 enrichment at transcript ends. (E,F) Analysis of average phastCons scores in the vicinity of known and newly annotated 3' termini in mouse (E) and human (F) shows that both populations of termini exhibit selective constraint that rises to a peak in the local sequence upstream of 3' termini, and drops sharply in the downstream sequence. Note also that the aggregate conservation of the last ∼500 nt of proximal 3' UTR sequences is higher than that of the distal novel 3' UTR sequences, but the overall level of conservation 3' of our mouse and human extensions drops to background. (G,H) Analysis of location of polyA-seq tags relative to known and newly annotated 3' termini shows a similar positional enrichment at transcript 3' termini. Comparison with a randomly selected set of 3' ends from these transcripts shows no positional enrichment of polyA-seq tags, indicating that our novel annotations include genuine 3' ends.

vation. The local conservation surrounding aggregate distal PAS was comparable to corresponding proximal PAS, indicating their similarly strong evolutionary selection. We also note that the overall conservation 100-500 nt downstream from proximal PAS was higher than downstream from our novel distal PAS. These trends applied to both mouse (Figure 7E) and human (Figure 7F) and are compatible with the scenario that our thousands of novel 3' UTR extensions contain functional cis-regulatory information that distinguishes them from background intergenic sequence.

Finally, we assessed our novel 3' termini for overlap with recent 3'-sequencing of mouse and human transcripts [28]. These data are a valuable resource for the discovery of novel mRNA ends, although the initial study did not systematically annotate these. We reprocessed the polyA-seq data to precisely annotate 3' ends that correspond with the end of RNA-seq coverage. For comparison, we analyzed the extent of polyA-seq support for Ensembl v65 ends of all loci whose models we extended in this study. Nearly 80% of mouse (Figure 7G) and 85% of human (Figure 7H) annotated termini were supported by polyA-seq tags. We did not necessarily expect such a high validation rate a priori, since RNA-seq data do not demarcate transcript ends precisely, and distal 3' UTR extension transcripts often accumulate to lower levels and thus contribute fewer polyA-seq tags than shorter isoforms.

Altogether, these bioinformatic analyses demonstrate that we annotated a large population of functional mammalian transcript termini, adding large expanses of currently unannotated mouse and human genomic sequence to expressed 3' UTR space. Strikingly, these thousands of novel 3' termini exhibit motif and conservation properties that are comparable to known mRNA termini in these well-annotated

genomes.

2.4.6 Tissue-specific 3' UTR lengthening is strongly biased toward neural tissue

We reannotated 3' UTRs from RNA-seq data across a variety of tissues, but until this point, our experimental validation focused on brain. We utilized DEXSeq [7], a statistical approach to detect differential exon usage, to identify tissue-biased APA events. Analysis of 15 pairwise combinations of mouse tissues yielded at least some genes with significant differential expression of 3' UTR extensions between each tissue pair. In addition, many of the novel 3' UTR extensions we annotated were not differentially expressed across tissues. However, we consistently observed that the hippocampus exhibited the highest number and expression of 3' UTR extensions, relative to all other tissues (Figure 9A). We repeated this analysis for 16 human tissues in the Illumina BodyMap 2.0. We show representative comparisons in Figure 9C and provide all 120 pairwise comparisons in Figure 8. These tests recapitulate the pattern observed in mouse, in that the absolute number and relative abundance of the novel 3' UTR extensions are highest in the human brain.

To strengthen the conjecture that preferential usage of unannotated distal polyadenylation sites is a property of neuronal cells, we examined RNA-seq data from mouse ES cells and differentiated neurons derived from these cells [76]. By using the coordinates from DEXSeq analysis of mouse tissues (Figure 9A, top row), we performed DEXSeq expression analysis for ES cells and derived neurons, as well as for mouse embryonic fibroblasts (MEFs). These tests robustly reproduced the pattern of preferential

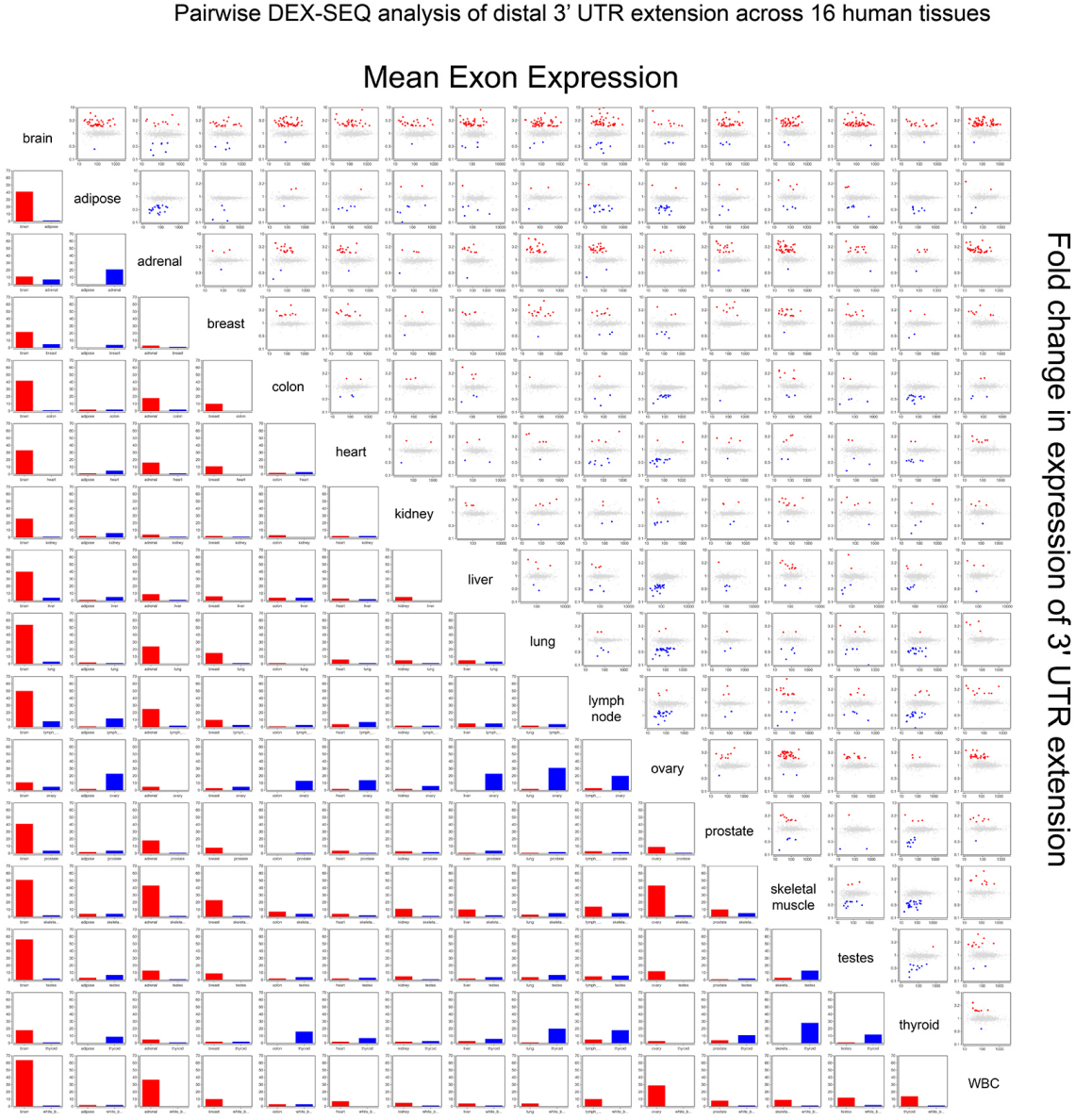


Figure 8: Pairwise DEXSeq analysis of novel 3' UTR extensions annotated in human. In these scatterplots, each point represents the relative expression of a 3' UTR extensions between two tissues. For genes exhibiting a significant (>2 fold, FDR <0.01) difference between the two tissues, the point was colored red if the relative usage is higher in the tissue indicated on the tissue named on the Y-axis, and blue if it was higher in the tissue on the X-axis.

Figure 9: A) Pairwise analysis of tissue-specific preferences of novel mouse 3' UTR extensions using DEXSeq. Each gene is represented as a single point, such that the relative expression of the 3' UTR extension between the pair of tissues (indicated at the left of each row and the bottom of each column) is plotted as the Y-coordinate, and the average expression of the 3' UTR in that pair of tissues is plotted as the X-coordinate. For genes exhibiting a significant (greater than twofold, FDR <0.01) difference between the two tissues the point is colored red if the relative usage is higher in the tissue indicated at the left of the row and blue if it was higher in the tissue indicated at the bottom of the column; all other 3' UTRs are shown in gray. We observed a broad tissue-wide trend toward increased expression of lengthened 3' UTRs in hippocampus, seen as a substantial excess of red points across the top row of tissue comparisons against hippocampus. No particular trend is observed among the other pairwise tissue comparisons. B) Summary of the pairwise analysis of novel 3' UTR extensions annotated in mouse. For each tissue, the set of genes that are detected by DEXSeq to have a higher fold expression of an extended 3' UTR extension compared to at least one other tissue were counted. C) Summary of DEXSeq tissue comparisons of novel 3' UTR extensions in human (for all pairwise scatterplots, see also Supplemental Fig. 8). D) DEXSeq analysis of our novel mouse 3' UTR extensions, assessed in RNA-seq data from mES/neuron/MEF cells. In the scatterplot, mES data are in blue and differentiated neuron data are in red.

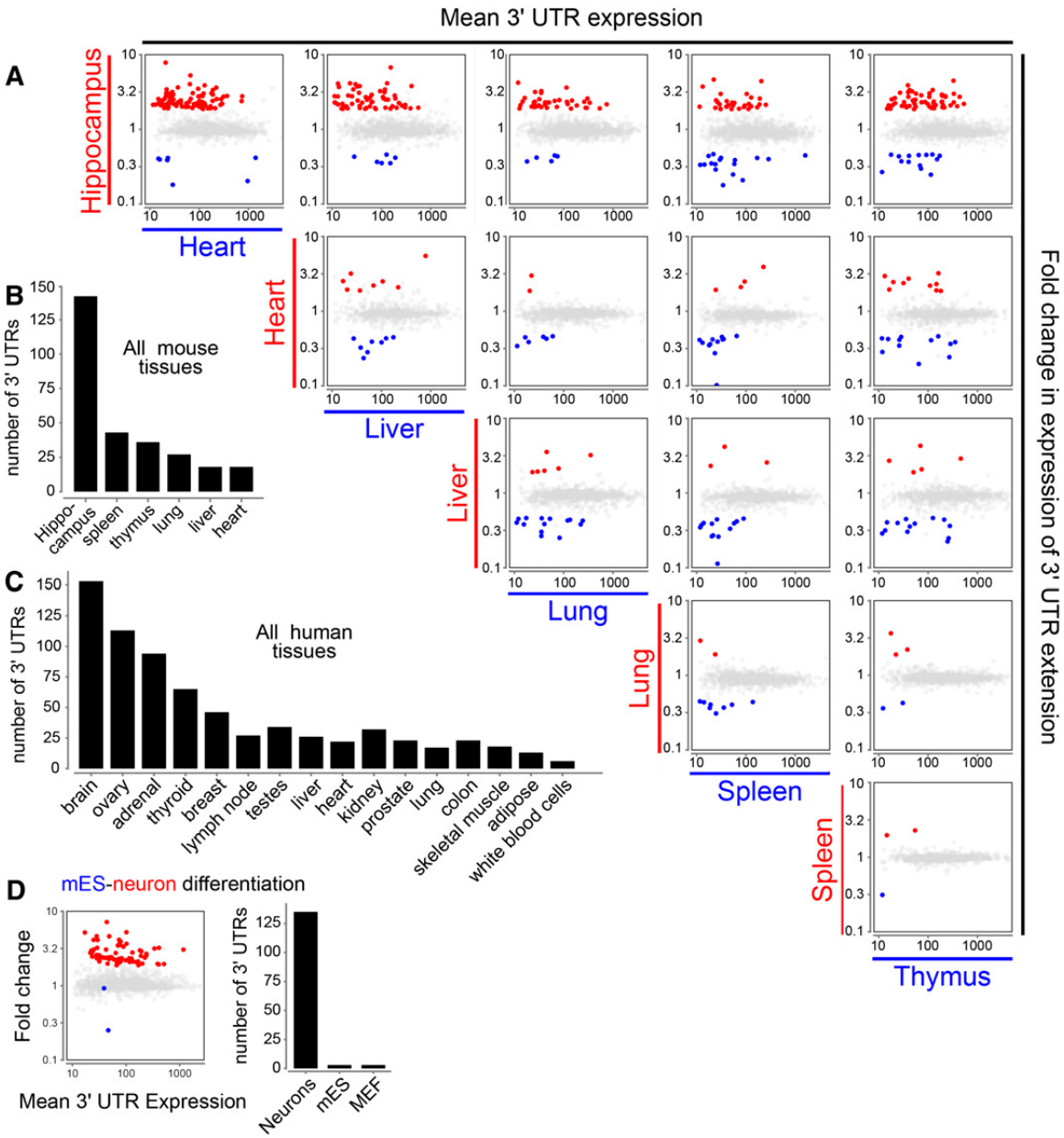

Mean 3' UTR expression
Fold change in expression of 3' UTR extension
A
Hippocampus
Heart
Heart
Liver
Liver
Lung
Lung
Spleen
Spleen
Thymus
B
number of 3' UTRs
All mouse tissues
Hippo campus
spleen
thymus
lung
liver
heart
C
number of 3' UTRs
All human tissues
brain
ovary
adrenal
thyroid
breast
lymph node
testes
liver
heart
kidney
prostate
lung
colon
skeletal muscle
adipose
white blood cells
D
mES-neuron differentiation
Fold change
Mean 3' UTR Expression
number of 3' UTRs
Neurons
mES
MEF

Figure 9

neural expression of novel extensions (Figure 9D).

We confirmed these bioinformatic trends using Northern analysis of mouse kidney, liver, cerebellum, and cortex. As many transcripts whose 3' UTR extensions we validated earlier (Figure 3 and 5) were not necessarily subject to APA, we focused on genes with tissue-specific 3' UTR extensions. Figure 10A illustrates stringent APA analysis using proximal probes and two different extension probes for Ppp1r7, Sod2, and Dnajc15. Their universal probes detected shorter transcripts across the tissue panel and longer transcripts that were brain-specific. In all cases, both sets of extension probes detected exclusively the longer isoforms and only in brain. This was particularly notable for Sod2 and Dnajc15, whose intermediate extension probes detected APA isoforms of intermediate length, which were not detected by the most distal extension probes.

We extended this analysis using paired universal and extension probes for nine other genes exhibiting brain-specific 3' UTR lengthening (Figure 10B; Figure 11). These data broadly support the bioinformatic inference of brain-specific distal APA usage (Figure 9) and further validate the existence of exceptionally long 3' UTRs on stable neural transcript isoforms (e.g., 10.2-kb Sod2 3' UTR and 14-kb Dcun1d5 3' UTR). Altogether, a process of tissue-biased transcript lengthening that generates hundreds of novel, distal 3' UTRs in the nervous system occurs on a global scale in Drosophila, mouse, and human.

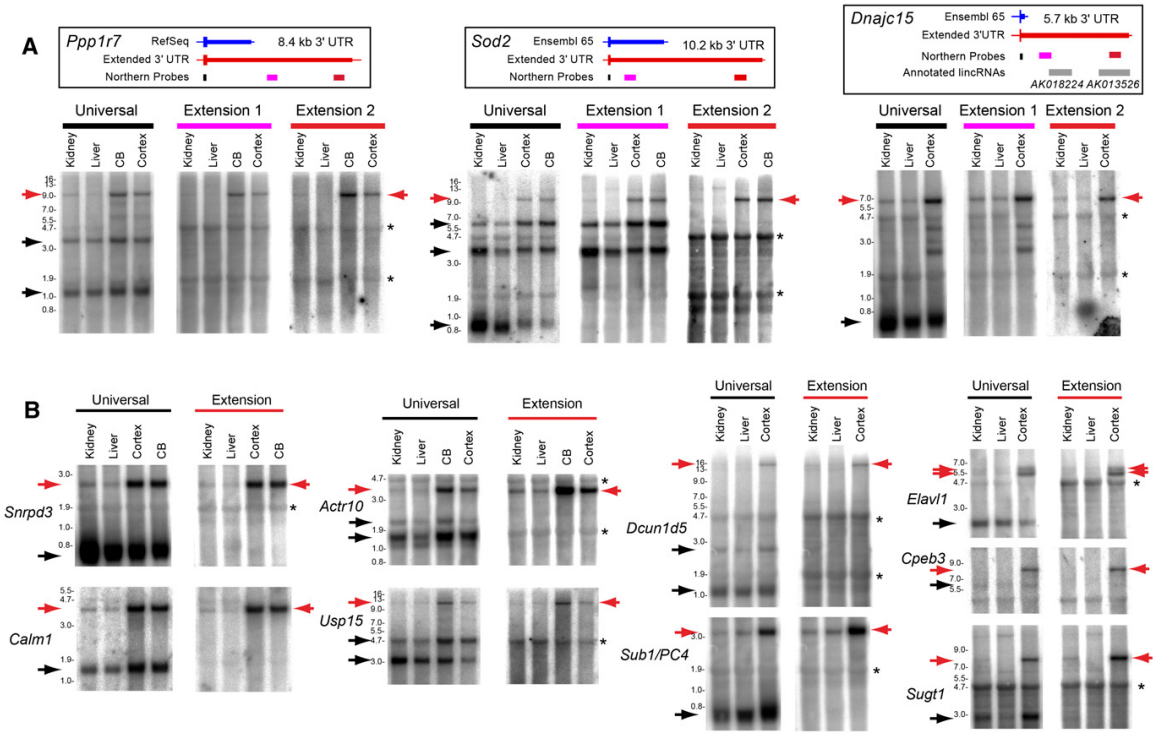


Figure 10: A) Northern analyses that compare universal (proximal) probes with two probes directed against an intermediate and a very distal portion of a 3' UTR extension. The gene models above show the known and newly recognized 3' UTR extensions and locations of Northern probes. In all cases, the universal probes detect broadly expressed transcripts bearing short 3' UTRs as well as longer 3' UTR isoforms that are specific to cerebellum (CB) and/or cortex, while the extension probes detect exclusively the longer 3' UTR isoforms in brain. Note that the intermediate probes (extension 1) for Sod2 and Dnajc15 detect intermediate 3' UTR isoforms that are codetected by their respective universal probes but not by their most distal 3' UTR probes. Asterisks denote cross-hybridization to abundant rRNA bands. B) Additional examples of brain-specific distal APA events validated by Northern blots. Northern analysis using universal Northern probes (black bars) designed to detect all 3' UTR isoforms reveal dominant isoforms used by all tissues examined along with brain-specific long 3' UTR isoforms. Extension probes (red bars) designed to detect the 3' UTR extensions reveal expression only in the brain and not in other tissues. Asterisks denote background hybridization to ribosomal RNAs; (CB) cerebellum.

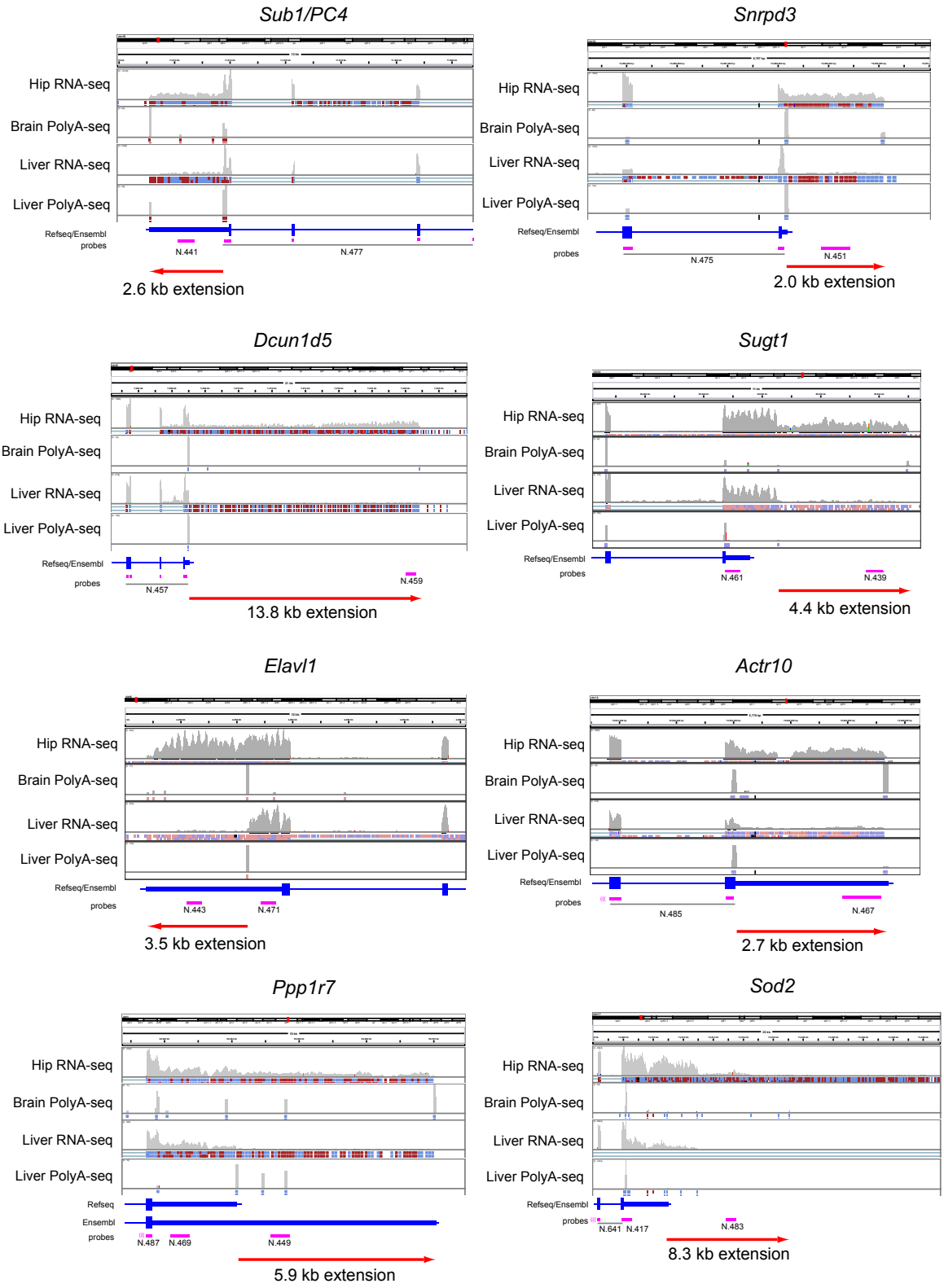


Figure 11: Gene models, RNA-seq and PolyA-seq tracks for several neural APA isoforms that have Northern support shown in Figure 10.

2.4.7 In situ hybridization validates expression of neural-specific 3' UTR extensions

We used in situ hybridization of E13.5 mice to assess the spatial expression patterns of several genes undergoing neural 3' UTR extension. We were particularly interested if paired proximal and distal probes ever detected differential patterns, as observed in Drosophila [53, 128].

Nedd4l encodes an ubiquitin ligase whose substrates include receptors and kinases in several signaling pathways. By using paired universal and extension probes, we observed expression of both short and long 3' UTR isoforms in brain Northerns (Figure 12A,B). In situ analysis using a Nedd4l universal probe detected expression in brain and dorsal root ganglion (DRG) (Figure 12C). The extended 3' UTR isoform probe similarly detected expression in brain; however, no signals were obtained in PNS (Figure 12D).

Figure 12: A) RNA-seq data for Nedd4l indicate an alternative 4.9-kb-long 3' UTR isoform that includes a proposed lincRNA AK038898. B) Northern blotting demonstrates that an AK038898 probe detects the long 3' UTR isoform of Nedd4l. C) A Nedd4l universal probe detects expression in both brain and dorsal root ganglia (DRG). D) A probe directed against the very distal portion of the Nedd4l 3' UTR extension detects only brain expression. E) RNA-seq data for Tcf4 indicate the existence of a 3' UTR extension of the annotated gene model, with preferential expression in hippocampal data. F) Tissue Northern blot using a distal probe confirms the existence of a discrete band expressed in brain that corresponds to a 3' UTR extension isoform. G) In situ hybridization to a probe in the common 3' coding exon detects Tcf4 predominantly in the CNS and the intervertebral discs; (LV) lateral ventricle. A whole-embryo cross-section is shown at left, and the regions boxed are enlarged at right. H) The Tcf4 3' UTR extension probe only detects expression in the brain. I) RNA-seq data for Rspo3 indicate a candidate 3' UTR extension, although this level of expression (0.22 FPKM) was below our cutoff for genome-wide calls of 3' UTR extensions. J) A universal Rspo3 probe predominantly detects CNS expression in the cortex and hem, as well as PNS expression in the spinal cord, mainly in dorsal root ganglia (DRG). K) The intermediate Rspo3 extension probe hybridizes specifically to the cortical hem. L) A probe directed against the very low abundance Rspo3 extension region similarly detects expression in cortical hem.

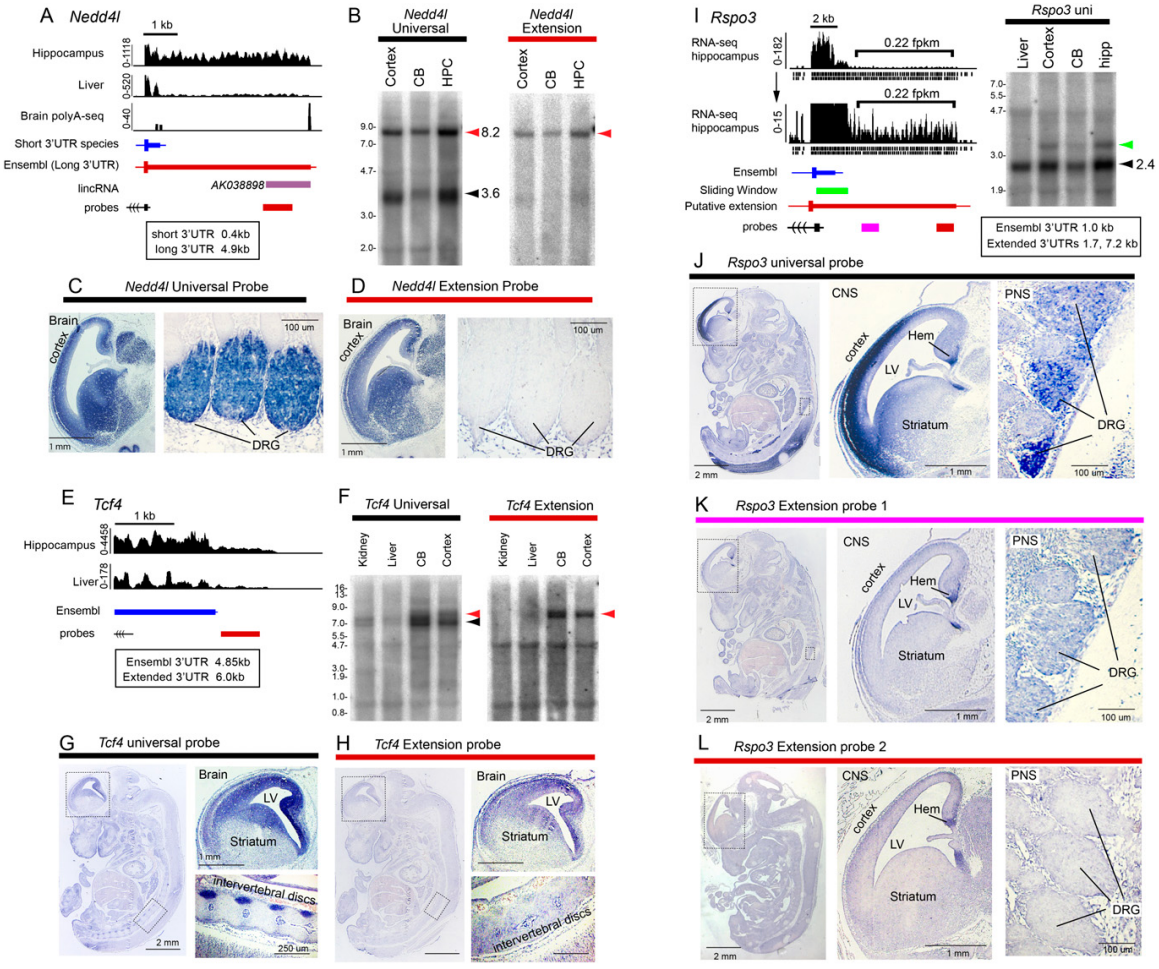

A Nedd4l
1 kb
Hippocampus
Liver
Brain polyA-seq
Short 3'UTR species
Ensembl (Long 3'UTR)
lincRNA
AK038898
probes
short 3'UTR 0.4kb
long 3'UTR 4.9kb
B Nedd4l Universal
Nedd4l Extension
Cortex
CB
HPC
8.2
3.6
C Nedd4l Universal Probe
D Nedd4l Extension Probe
Brain
cortex
DRG
100 um
1 mm
E Tcf4
Ensembl
Ensembl 3'UTR 4.85kb
Extended 3'UTR 6.0kb
F Tcf4 Universal
Tcf4 Extension
Kidney
Liver
G Tcf4 universal probe
H Tcf4 Extension probe
LV
Striatum
Intervertebral discs
I Rspo3
2 kb
RNA-seq hippocampus
0.22 fpkm
Sliding Window
Putative extension
Rspo3 uni
Ensembl 3'UTR 1.0 kb
Extended 3'UTRs 1.7, 7.2 kb
J Rspo3 universal probe
CNS
PNS
Hem
K Rspo3 Extension probe 1
L Rspo3 Extension probe 2

Figure 12

We next analyzed Tcf4, a transcription factor in the Wnt pathway. Recent studies proposed differential stability of TCF4 and its downstream lincRNAs [24]. However, RNA-seq data indicate continuous transcription downstream from Tcf4 (Figure 12E). Northern analysis using a probe to the downstream unannotated region supported the presence of a 3' UTR extension of Tcf4 and did not detect shorter ncRNAs (Figure 12F). The extended APA isoform was spatially restricted, since a Tcf4 universal probe detected expression in forebrain and intervertebral discs (Figure 12G), whereas the extended isoform was only found in forebrain (Figure 12H).

Finally, we analyzed Rspo3, which encodes a thrombospondin type 1 repeat family protein, members of which are also involved in Wnt signaling. Our pipeline predicted a ∼700-nt extension of its annotated 3' UTR. However, visual inspection of RNA-seq data revealed a potential 3' UTR extension of ∼7.2 kb that was below our expression cutoff (0.22 FPKM) (Figure 12I) and was therefore excluded from our confident bioinformatic list. Northern analysis validated the short 3' UTR extension, but not its substantially longer counterpart. Nevertheless, in situ hybridization revealed discrete and distinct expression of the extended isoform. The universal probe detected expression broadly in the pallium, cortical hem, spinal cord, and DRG (Figure 12J). In contrast, the Rspo3 3' UTR extension probe hybridized exclusively to the cortical hem and did not reveal PNS expression (Figure 12K). A probe against the distal unique sequence of the 3' UTR extension revealed the same patterns (Figure 12L), suggesting the existence of a stable isoform not detected by Northern. The restricted spatial pattern of Rspo3 extensions explains its seemingly low expression in total hippocampal RNA and emphasizes that our computational identification of thousands

of 3' UTR extensions still underestimate the magnitude of this phenomenon.

2.4.8 Novel 3' UTR extensions harbor thousands of conserved miRNA target sites

We sought regulatory implications of this large network of 3' UTR extensions. We assessed all 7-mers for evidence of conservation above background among our 2035 mouse 3' UTR extensions. Notably, the motifs with highest signals and highest numbers of conserved instances corresponded to seeds for miRNAs with well-described neural functions [134], including let-7, miR-124, miR-9, miR-96, miR-125, and miR-137 (Figure 14A). We asked whether the enrichment for neural miRNA seeds was a property of coherent targeting or mutual exclusion. We segregated the 2035 extensions for those detected in hippocampus, and compared their site properties to the rest. This analysis showed that extensions expressed in hippocampus bore the majority of neural miRNA target sites, indicating that neural 3' UTR extensions are utilized to confer regulation by neural-specific miRNAs.

We subsequently performed a directed search of the unannotated mouse extensions for seed matches conserved between rodents and primates, restricting this to mammalian-conserved miRNAs. We identified nearly 4000 conserved miRNA target sites, which substantially extend the scope of post-transcriptional regulatory networks in mammals (Figure 14A, inset). Even though the list of human genes with novel extensions overlapped only partially with mouse, de novo analysis revealed mostly the same neural miRNA seeds among their best-conserved 7-mers (Figure 13).

The presence of conserved miRNA binding sites on sense strands downstream

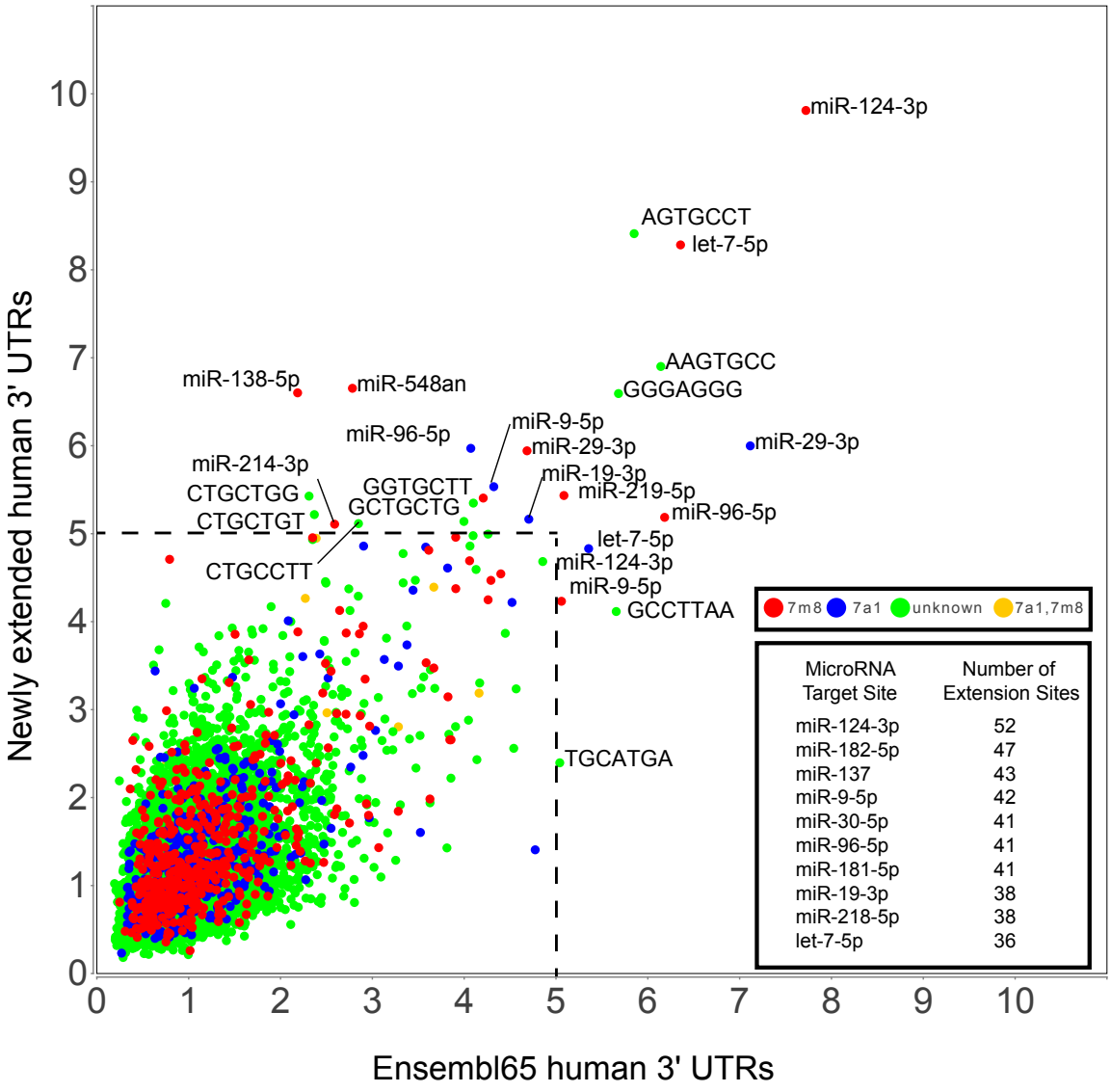


Figure 13: Signal to background ratio (S:B) of 7-mers found in the proximal 3' UTR annotations compared to the novel extended 3' UTR region annotated in human.

Figure 14: A) Signal-to-background ratio (S:B) of 7-mers found in the proximal 3' UTR annotations compared with the novel extended 3' UTR region annotated in mouse from all tissues analyzed. Note that target sites for several well-characterized neural miRNAs are found among the most well-conserved 7mers in both proximal and novel extended 3' UTR regions, including miR-124, miR-137, miR-9, let-7, miR-96, and miR-125. Supplemental Figure S10 demonstrates that the signal for neural miRNA seed matches is driven by genes with neural-expressed 3' UTR extensions. B) Analysis of seed matches to mammalian-conserved miRNAs, that are present among mouse 3' UTR extensions that lack companion expression evidence for an orthologous 3' UTR extension in human (top graph) or that do have such experimental evidence for a human extension (bottom graph). The proportion of conserved miRNA binding sites is much higher among genes with evidence for a conserved 3' UTR extension. C) Regions surrounding miRNA target sites located in proximal (in blue) and novel distal 3' UTR mouse extensions (in red) show enrichment of Ago HITS-CLIP tags over background. The signal:background (S:B) of clip tags at let-7 and miR-124 seed matches is actually higher in the novel 3' UTR extension regions.

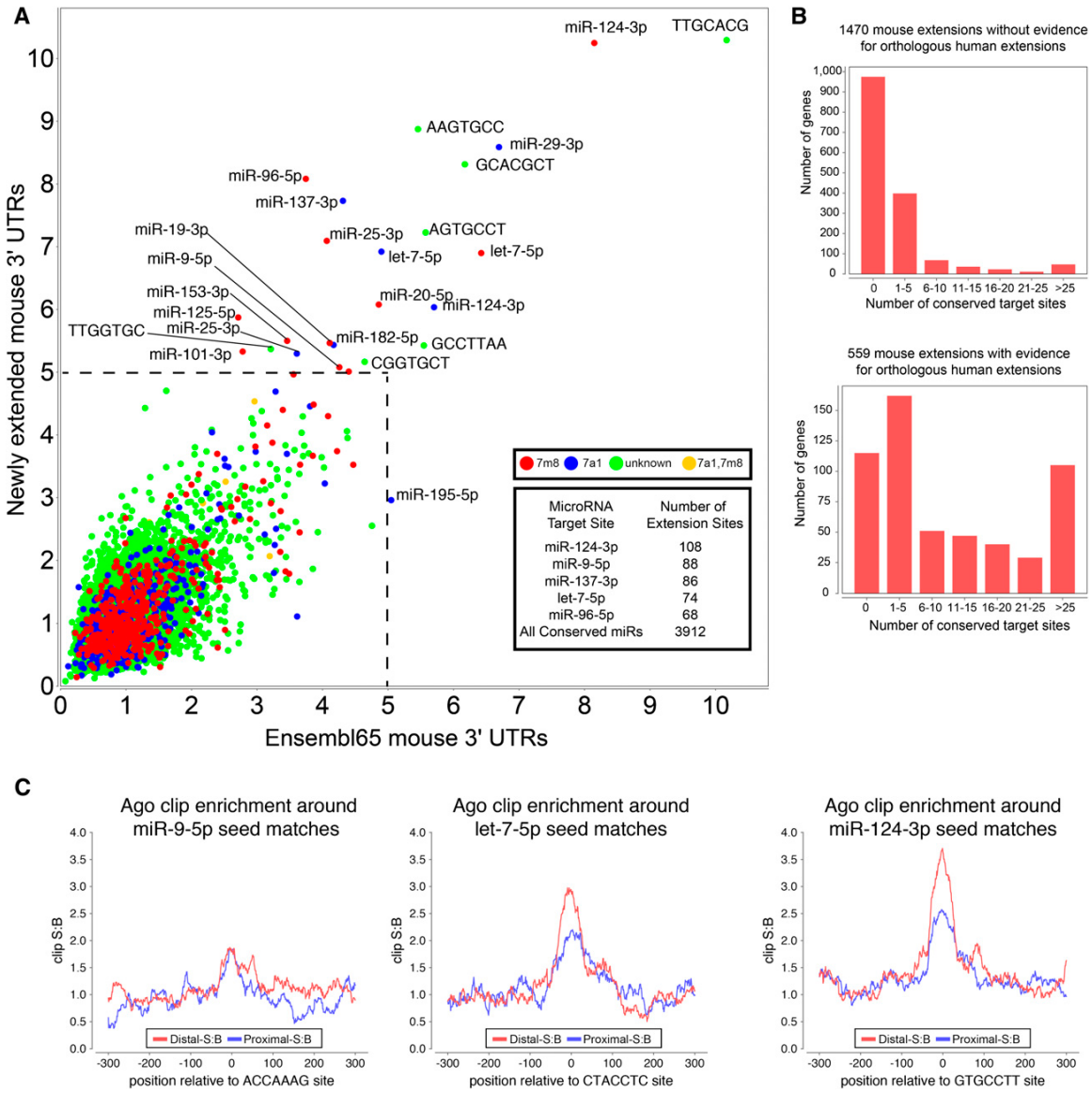

A
B
C
Newly extended mouse 3' UTRs
Ensembl65 mouse 3' UTRs
miR-124-3p
TTGCACG
AAGTGCC
miR-29-3p
GCACGCT
miR-96-5p
miR-137-3p
miR-19-3p
miR-25-3p
AGTGCCT
miR-9-5p
let-7-5p
let-7-5p
miR-153-3p
miR-20-5p
miR-125-5p
miR-124-3p
TTGGTGC
miR-25-3p
miR-182-5p
GCCTTAA
miR-101-3p
CGGTGCT
miR-195-5p
7m8
7a1
unknown
7a1,7m8
MicroRNA Target Site
Number of Extension Sites
miR-124-3p 108
miR-9-5p 88
miR-137-3p 86
let-7-5p 74
miR-96-5p 68
All Conserved miRs 3912
1470 mouse extensions without evidence for orthologous human extensions
Number of genes
Number of conserved target sites
559 mouse extensions with evidence for orthologous human extensions
Ago clip enrichment around miR-9-5p seed matches
Ago clip enrichment around let-7-5p seed matches
Ago clip enrichment around miR-124-3p seed matches
clip S:B
Distal-S:B
Proximal-S:B
position relative to ACCAAAG site
position relative to CTACCTC site
position relative to GTGCCTT site

Figure 14

from annotated gene models serves as corroborating evidence for 3' UTR extensions. For the 2035 novel mouse extensions, 559 have an orthologous extension in human (469 of which are annotated in Ensembl v65, and 96 of which are newly identified here). The remaining 1470 loci are tentatively supported only in mouse, although at least some are associated with "downstream" human RNA-seq evidence that did not meet the 1.5 FPKM threshold. We asked whether the conserved miRNA sites preferentially partitioned into mouse 3' UTR extensions that did, or did not, have an orthologous human extension. We observed highly significant ($P < 1$ E-15, binomial test) enrichment of conserved miRNA binding sites among genes that currently share experimental evidence for neural 3' UTR extensions in both mouse and human (Figure 14B). Nevertheless, a population of "mouse-only" extensions harbor miRNA binding sites conserved in human (in some cases more than 25 sites), indicating that the catalog of neural 3' UTR APA events is still not complete.

To further address the functionality of miRNA binding sites in neural 3' UTR extensions, we queried Ago binding sites in mouse brain using published HITS-CLIP data [21]. In this study, some Ago-bound tags were noted downstream from certain gene models, and inferred to represent potentially unannotated 3' UTR sequences. We surveyed our 3' UTR extensions and observed robust signals for Ago binding at miRNA seeds in both proximal and extended 3' UTRs. Exemplar seeds enriched in Ago-CLIP tags in extended 3' UTRs included miR-124-3p, let-7-5p, and miR-9-5p (Figure 14C). While a lower fraction of 7-mers overlapped Ago-CLIP tags in the extended portion of UTR, the ratio of Ago-IP frequency at these sites to background UTR Ago-IP was actually higher in extended regions. The lower CLIP frequency may

be due to differential isoform abundance, since CLIP tags are sampled more frequently from abundant mRNAs, and shorter isoforms tend to accumulate to higher levels than the distal extension isoforms. Nevertheless, the Ago HITS-CLIP data strongly support that these novel neural 3' UTR extensions confer substantial regulation by many mammalian neural miRNAs.

2.5 Discussion

2.5.1 Extensive usage of highly distal APA is a well-conserved feature of the nervous system

Genes expressed in the nervous system contain longer 3' UTRs, on average, compared with other tissues [130, 145, 115]. Moreover, a number of transcripts have been noted to undergo alternative polyadenylation (APA) in the nervous system, yielding longer 3' UTRs in their neural isoforms. Very recently, the Drosophila central nervous system was recognized to utilize novel distal APA sites across hundreds of transcripts, often generating 3' UTRs of exceptional length [53, 128]. Analysis of mammalian brain and cultured neurons similarly provides evidence of distal APA events [103, 124].

In this study, we substantially increase the number and magnitude of neural distal APA events in the mammalian brain, including the accumulation of a multitude of stable mRNAs bearing exceptionally long 3' UTRs (nearly 20 kb in length). In a day and age where the amount of "extragenic" transcription that contributes to discrete, stable transcripts remains hotly contested [110, 142, 23], it is striking that we can add 6.6 Mb and 5.1 Mb of currently unannotated sequence to the confident mRNA space

of the mouse and human transcriptomes, respectively. These transcript extensions have substantial impact on post-transcriptional networks, especially those mediated by the particularly extensive collection of neural 3' UTR extensions.

We find that short and long tissue-specific 3' UTRs exhibit marked usage of canonical PAS hexamers (AAUAAA or AUUAAA) and downstream U/GU-rich elements, although it is striking that our plethora of novel, distal 3' ends exhibit motif properties that are slightly stronger than their Ensembl-annotated counterparts (Figure 7A-D). Still, Northern analysis of mouse brain tissue revealed that although a longer 3' UTR is used, the shorter 3' UTR continues to be expressed, often at relatively high levels. This contrasts with the "switch-like" behavior of many Drosophila genes to dominantly express the distal 3' UTR in heads and dissected CNS [128]. At present, it is unclear if this reflects that the mammalian brain does not bypass proximal PAS as efficiently, or whether this reflects differences in cellular composition in samples analyzed. The mammalian brain may contain a higher proportion of glial cells and lower proportion of neurons, compared with Drosophila heads, which may potentially comprise a higher proportion of neurons. Our analysis of RNA-seq data from ES cells differentiated into neurons showed more robust switching to distal 3' UTR isoforms (Figure 9D) than observed in the tissue comparisons (Figure 9B,C).

2.5.2 Challenges for accurate transcript assembly from RNA-seq data

The exponential rise in the throughput of next-generation sequencing has outpaced many aspects of its analysis and interpretation. With respect to the task of assembling transcripts from shards of RNA sequence, solutions such as ERANGE, Cuf-

flinks, and Trinity have proven effective and are widely used. Nevertheless, even current ENCODE project efforts to annotate the human transcriptome, undertaken by the HAVANA and GENCODE subgroups (http://www.sanger.ac.uk/research/-projects/vertebrategenome/havana/), acknowledge that substantial manual curation is required to provide confident gene models from RNA-seq data.

Our study highlights the substantial extent to which bioinformatics must advance to fully exploit the power of RNA-seq. Although Cufflinks was invaluable for initial processing of RNA-seq data, we found numerous truncations of clear 3' UTR extensions evident from visual inspection. A major issue is that current conceptions of the transcriptome do not generally include the broad possibility for large processed exons, yet we show that hundreds of continuous 3' UTRs ranging from 5-25 kb are encoded by the Drosophila, mouse and human genomes. In general, understanding the connectivity of transcripts from RNA-seq data remains challenging. Major impediments include the highly nonlinear representation of reads across individual transcripts, the difficulty of deconvolving overlapping and/or alternative transcripts of unequal abundance, and the existence of intervening multi-mapping sequences, which can be common in 3' UTRs that include fragments of repetitive elements. It is hoped that some of these issues may be ameliorated as technologies for direct and/or long read sequencing improve.

We supported the veracity of our computational 3' UTR inferences with extensive experimental analysis. While >10-kb 3' UTRs are rare in current annotations and only a handful have been experimentally validated, we provide Northern support for many 3' UTRs in this size range in this study alone. Notably, in all cases

where purported lincRNAs reside in our 3' UTR extensions, our extensive Northern studies failed to reveal evidence for lincRNA transcripts of annotated lengths. We do not formally exclude that some of these 3' UTR extensions might be processed into shorter RNAs [85]; for example, multiple pathways digest 3' UTR segments into endo-siRNAs or even piRNAs [101]. Nevertheless, the parsimonious interpretation of our studies is that alternative 3' UTR extensions of stable mRNAs remain a strongly under-appreciated aspect of the mammalian protein-coding transcriptome. This has consequences for interpreting lincRNAs, which are well-documented to exhibit tissue-specific expression, to be conserved, and to be associated with phenotypes when depleted. In fact, all of these are also characteristics of 3' UTRs, and our studies provide extensive evidence that simply being distant from an annotated gene model is not a reliable predictor of being transcribed independently (Figure 3, 5 and 10). Our studies suggest that additional loci currently annotated as lincRNAs may actually correspond to unannotated 3' UTR extensions. For example, we provide Northern evidence that lincRNAs described by Mattick and colleagues [24] can be detected as stable 3' UTR extensions of Etv1, Paqr9, and Tcf4 mRNAs (Figure 3C, Figure 12F).

We are confident that the myriad and unexpected roles for long noncoding RNAs are just beginning to be unraveled [119, 64, 39, 49]. At the same time, our findings serve as a reminder that establishing transcript connectivity and full-length structures from tiling array and RNA-seq data are not trivial operations, but present and ongoing challenges for transcriptome studies.

2.6 Methods

2.6.1 RNA preparation

Adult male ICR (CD-1) mice (Taconic) were euthanized by CO2 overdose. Total RNA from dissected brain samples was extracted using RNeasy lipid tissue kit (Qiagen). Other tissues were extracted using TRIzol (Invitrogen). Poly(A)+ RNA was prepared from total RNA using Oligotex mRNA kit.

2.6.2 Northern analysis and RT-qPCR

For Northern analysis, 1.5-2 μg of poly(A)+ RNA was denatured using glyoxal, and electrophoresis was performed using 1% agarose BPTE gels and blotted and probed as described. Internal size standards were prepared using a mix of probes against several highly expressed genes with known sizes (Figure 6). Note that the migration of the lower bands at 0.8, 1, and 3 kb differ by $\sim$200 nt from commercially single-stranded RNA ladders (New England Biolabs) due to the presence of endogenous polyA tails.

For cDNA preparation, reverse transcription was performed using superscript III reverse transcriptase (Invitrogen) on DNase I (Ambion) treated total RNA. End-point PCR was performed using Taq DNA polymerase (New England Biolabs) with 55°C-62°C annealing temperatures and 28-35 cycles. To rule out amplification of genomic DNA, first-strand synthesis was performed using control reactions lacking reverse transcriptase (-RT). qPCR was performed using SYBR green PCR mastermix (Qiagen).

2.6.3 Next-generation sequencing data sets analyzed

Mouse RNA-seq data (GSE30617) generated by the Wellcome Trust Sanger Institute included six pooled libraries from the hippocampus, spleen, heart, lung, thymus, and liver [62]. We also analyzed data from mouse ES cells differentiated into neurons (GSE27866) [76]. Wiggle tracks from the mouse ENCODE project were downloaded from the ENCODE DCC (http://hgdownload.cse.ucsc.edu/goldenPath/-mm9/encodeDCC/) and merged into a single track [120]. Human RNA-seq data were from the Illumina Human Body Map 2.0 Project (GSE30611), which includes non-stranded RNA-seq libraries from 16 tissues, as well as stranded RNA-seq data for a mixture of these tissues. We used 3'-seq data (SRP007359/GSE3019) from the brain, kidney, liver, muscle, and testis for both mouse and human [28]. All RNA-Seq data were mapped to the human (hg19) and mouse (mm9) genomes using TopHat with the default parameters [140].

2.6.4 Identification of 3' UTR extensions

As a first step to identifying candidate 3' UTR extensions, we ran a sliding window across the genome to identify all continuously transcribed regions. We defined a 100-nt window to be continuously transcribed if more than 80/100 positions were covered by more than 10 reads. This criterion was applied at single-nucleotide increments across the genome, and the overlapping segments were merged. We provisionally made further refinement by merging neighboring regions separated by <150 nt of nonrepetitive sequence. We also merged expressed windows separated by repetitive elements (as

identified by RepeatMasker). The resulting segmentation was subjected to extensive filtering that aggressively culled potentially ambiguous extension cases, which might not be distinguished from intragenic transcription, overlapping transcripts, retained introns, etc.

In brief, we identified continuously transcribed regions that overlap annotated stop codons, and subjected them to the following criteria: (1) The identified segment overlaps only a single gene. (2) The extended region does not overlap any exons, only either intronic or intergenic space. (3) The called 3' UTR overlaps only a single stop codon. (4) There is at least 500 nt of nontranscribed space before the next gene or exon. (5) The gaps bridged between neighboring segments in an extension accounts for <20% of the extension, and these gaps are spanned by at least one paired-end read. (6) The extension is expressed above 1.0 FPKM in mouse and 1.5 FPKM in human. The requirement of 1 FPKM (1.5 FPKM Illumina bodymap data) was selected by comparing the accuracy of the sliding window for annotating known exon boundaries at different FPKM cutoffs and selecting a cutoff with >90% sensitivity within 100 nt. (7) The extension increases current Ensembl v65 gene models by >500 nt. (8) Less than 20% of reads from available stranded libraries derive from the opposite strand. (9) For any extension containing spliced reads, the percentage of reads supporting splicing across either junction account for <20% of reads mapped across that genomic location.

To identify continuous genomic regions that are transcribed above a minimum level, we used a 100 nt sliding window, advanced across the genome in single nucleotide increments. Using this strategy, we required that at least 80/100 positions

in that window are covered at a depth of at least 10 reads. Depending on the depth to which each tissue was sequenced, a minimum depth of 10 reads at 80/100 positions corresponded to a read density of 0.5-1.0 FPKM. All overlapping windows were merged into larger contiguously transcribed segments. Visual inspection of the resulting segmentation revealed that regions of high coverage were occasionally broken by regions of low-coverage due to non-unique regions of the genome. For this reason, we consolidated neighboring segments separated by less than 150nt of non-repetitive sequence, provided that the gap was bridged by at least one paired-end read.

The resulting transcribed segments were cross-referenced with Ensembl 65 gene models to identify 3'UTRs with extensions that were not annotated. Since intragenic transcription and overlapping transcripts are well-documented, and RNA-seq is unable to resolve overlapping transcripts (distinguishing intronic 3'UTR extension from retained introns), we implemented a series of criteria to cull ambiguous cases. Each identified extension could not overlap with any annotated exons, including exons of adjacent genes, and alternative exons of the gene model being extended. To be conservative, we not only excluded strict overlaps, but also required a minimum of 500 nt of genomic space that is not expressed (defined as the complement of the expressed segments identified by the sliding window), before the next annotated gene or exon. Therefore, our 3'UTR annotations are extensions from exons exclusively into unannotated intronic or intergenic space.

We further ensured that extensions were attributed to transcripts on the correct strand by requiring at least 80% of reads from available stranded RNA-seq data to derive from the corresponding strand. To prevent ambiguous extension calls, we

required that less than 20% of an extension could be comprised of regions that cannot be uniquely mapped to. Finally, to focus on substantially novel APA isoforms, we required Ensembl gene models to be extended by at least 500 nt. We grouped together novel 3' end calls from different tissues that were within 30 nt of each other to report a set of non-redundant 3'UTR extensions.

While the above pipeline enabled the confident 3' UTR extensions, visual inspection revealed that it does not necessarily pinpoint the precise 3' ends. Many instances of likely extensions that did not precisely identify the 3' end appeared truncated because the most distal portion of the 3'UTR did not satisfy our expression cutoff. To select a subset of confident 3'UTR extensions for which the annotation likely represents the precise 3' ends we further filtered the identified extensions. To remove cases in which the RNA-seq coverage gradually falls below the established minimum level level of expression, we required that the level of coverage at the 3' end exhibited a sharp "drop-off" when comparing the regions upstream and downstream of the extension 3' end. To implement this we required that the two consecutive 100 nt windows downstream of the 3' end call exhibit >8-fold reduction in reads, relative to the final 100 nt window of the 3' extension. If a repetitive element overlapped the 3' end annotation, the call was removed. Ends within 30 nt passing "drop-off" criteria in more than 1 tissue were merged and the median coordinate was used for analysis.

2.6.5 Comparison with recent annotations of 3' UTR extensions

Recently, 1460 3' UTR extensions were annotated from mouse cerebellum [103]. The majority of these extensions (860 genes) are now contained within 3' UTR annota-

tions in Ensembl v65, which was the starting point for our analysis. However, as Pal et al annotations are not systematically included in Ensembl gene models, we analyzed the degree to which our annotations overlap. Of our set of 2035 non-overlapping extensions from mouse, 487 extensions that overlap those from Pal and colleagues. Of this 487, 277 terminate within 500 nt of their annotation (which we broadly considered to be the same extension call), 55 are fully contained within a Pal annotation (meaning that the expression of their distal, downstream region did not meet our annotation criteria), and 154 were longer than a Pal annotation. Of the 154 overlapping loci for which our annotation was longer, we annotate 851 kb as confident extensions to mRNA. The 1548 extensions that do not overlap with their annotated extensions comprise 5.35 Mb over Ensembl v65. Therefore, our study extends the mouse transcriptome far beyond this recent study [103].

2.6.6 Analysis of polyadenylation site features

Although gene models might bear confidently extended 3' UTRs, the genomic regions that satisfied expression FPKM cutoffs did not always correspond to a clean 3' terminus. In cases where lower-level transcription continued for some distance past the called end, our sliding window truncated the 3' terminus. Alternatively, a called end might terminate in a repetitive element, and thus no specific 3' end was identified. Such loci, although confidently extended, are not germane for the analysis of 3' terminal motifs.

Visual inspection of RNA-seq data on the IGV Browser indicated that bioinformatically called 3' ends were likely to be precise when there was a substantial dropoff

in expression emanating from the downstream genomic sequence. We selected those 3' termini that did not overlap an annotated repetitive element, and exhibited greater than eightfold coverage dropoff between the 100 nt upstream of the annotated 3' end and both of two subsequent 100-nt windows downstream from the 3' end.

We centered on each of these novel 3' termini and analyzed their genomic vicinity for canonical polyadenylation signals (PAS) AAUAAA or its closest variant AUUAAA, as well as noncanonical polyadenylation motifs defined previously [138]. The U/GU-rich motif was defined as six consecutive U and/or G, with at least three Us (corresponding to 22 distinct 6-mers).

We assessed the positional enrichment of published polyA-seq tags [28] around our novel extended 3' UTRs. We counted the fraction of ends that have at least one 3' sequencing read in 50-nt bins 200 nt upstream of and downstream from the annotated polyadenylation site. We estimated the background frequency of 3' sequencing tag enrichment by sampling random points distributed uniformly between the longest Ensembl v65 annotation and the location of our confident 3' end annotation.

2.6.7 Post-transcriptional regulatory motif analysis

MULTIZ multiple species alignments projected onto the human (hg19) and mouse (mm9) genomes were downloaded from the UCSC Genome Browser. Signal-to-noise ratios of all 7-mers were calculated according to the method previously described [145]. We considered a 7-mer to be conserved if it was aligned without mismatches or gaps between human, mouse, rat, dog. The conservation frequency was calculated by dividing the number of conserved instances of a 7-mer by the total number of

occurrences of that sequence. The signal to background ratio was calculated for each 7-mer by dividing the conservation frequency by the average conservation frequency of a set of at least 10 control sequences. Control sequences exhibited matched GC content and occurred within twofold frequency of the query 7-mer in the selected regions. The 7-mers were cross-referenced with mature miRNA seed sequences from miRBase.

We also downloaded FASTQ files of 130-kD AGO-CLIP data [21] from http://ago.-rockefeller.edu/rawdata.php, and mapped them to mm9 using Bowtie with default parameters. For each 7-mer matching a miRNA 7m8 target site, the 300 nt upstream of and downstream from that sequence were queried for overlapping Ago-CLIP reads. The fraction of sequences that overlapped a CLIP tag at each nucleotide position relative to the seed match were counted and compared with the background rate of clip frequency around 7-mers with similar GC content. The signal:background ratio was calculated at each nucleotide relative to the target site by dividing the fraction of sequences overlapped by an Ago-CLIP read at each position by the average fraction of Ago-CLIP reads overlapping GC-matched control 7-mers in the same genomic regions.

2.6.8 Third party software used

We used picard liftOver (http://picard.sourceforge.net/) to identify orthologous mouse and human positions, SAM-JDK to query BAM files, apache-commons and R to perform statistical calculations, BigWig (http://code.google.com/p/bigwig/) to parse phastCons conservation scores, and JFreeChart and R for plotting.

2.6.9 Mouse in situ hybridization

E13.5 embryos were fixed in 4% paraformaldehyde overnight. The following day, embryos were washed with PBS, dehydrated, and paraffin embedded. Sagittal 7-μm embryo sections were processed using a Leica microtome. We prepared DIG-labeled antisense RNA probes according to the method previously described [128] and performed in situ hybridization according to the method previously described [15].

2.7 Acknowledgments

We thank the many researchers who made their deep sequencing data available for this study. P.M. was supported by a fellowship from the Canadian Institutes of Health Research. S.S. was supported by the Tri-Institutional Training Program in Computational Biology and Medicine. C.A-A. was supported by a NYSTEM post-doctoral fellowship. Work in E.C.L.'s group was supported by the Burroughs Wellcome Fund, the Starr Cancer Consortium (I3-A139), and the NIH (R01-GM083300 and RC2-HG005639).

3 Change-point analysis improves 3'UTR annotation using RNA-seq and provides an effective framework for detecting differential APA*

3.1 Attributions

Sol Shenker designed and implemented the IsoSCM assembly method, evaluated performance of methods on simulated and experimental datasets, and analysis of tissue expression patterns. Pedro Miura performed northern blot analysis. Sol Shenker, Pedro Miura, and Piero Sanfilippo visually inspected IsoSCM output.

3.2 Abstract

Major applications of RNA-seq data include studies of how the transcriptome is modulated at the levels of gene expression and RNA processing, and how these events are related to cellular identity, environmental condition, and/or disease status. While many excellent tools have been developed to analyze RNA-seq data, these generally have limited efficacy for annotating 3' UTRs. Existing assembly strategies often fragment long 3' UTRs, and importantly, none of the algorithms in popular use can apportion data into tandem 3' UTR isoforms, which are frequently generated by alternative cleavage and polyadenylation (APA). Consequently, it is often not possible to identify patterns of differential APA using existing assembly tools. To address these limitations, we present a new method for transcript assembly, Isoform Structural Change Model (IsoSCM) that incorporates change-point analysis to improve

***S. Shenker**, P. Miura, P. Sanfilippo, and E. C. Lai. IsoSCM: improved and alternative 3' UTR annotation using multiple change-point inference. *RNA*, 21(1):14–27, Jan 2015.

the 3' UTR annotation process. Through evaluation on simulated and genuine data sets, we demonstrate that IsoSCM annotates 3' termini with higher sensitivity and specificity than can be achieved with existing methods. We highlight the utility of IsoSCM by demonstrating its ability to recover known patterns of tissue-regulated APA. IsoSCM will facilitate future efforts for 3' UTR annotation and genome-wide studies of the breadth, regulation, and roles of APA leveraging RNA-seq data. The IsoSCM software and source code are available from our website https://github.com/-shenkers/isoscm.

3.3 Introduction

The recent astonishing advances in RNA-sequencing (RNA-seq) technologies have spurred the development of numerous methods to exploit these data for diverse applications, such as inferring transcript structures, differential gene expression, and alternative RNA processing. These analyses require models of transcript structure that serve as a foundation for transcriptome analysis, and to quantify differences in read mapping between conditions [80]. While transcript structures were historically inferred from full-length cDNA sequences [1], the wealth of data from RNA-seq experiments has greatly expanded transcriptome annotation pipelines more recently [30, 18].

Strategies for transcript assembly can be categorized based on their dependence on a reference genome sequence. Genome-independent (de novo) approaches typically construct a De Brujin graph representing reads sharing compatible subsequences,

then use heuristics to decompose this graph to recover full-length transcript sequences [80]. Genome-dependent approaches decompose the transcript assembly problem into smaller subproblems by first mapping the reads to the reference sequence, and then constructing gene models that are consistent with the aligned reads ([27]; [94]; [151]; [48]; [139]). Since these strategies operate in the absence of any reference gene models, they are termed ab initio transcript assembly methods.

In addition to transcript assembly, a separate class of methods attempts to quantify the relative abundance of transcripts and isoforms using RNA-seq data ([34]; [71],[72]; [13]; [55]). Since alternatively processed gene products are frequently encoded in overlapping genomic locales, and RNA-seq reads typically cannot be unambiguously assigned to overlapping transcripts, these methods attempt to construct parsimonious models for transcript abundance that are consistent with observed read counts. Importantly, these approaches cannot be applied if a set of reference models is not available, and may provide inaccurate quantification if these models are not complete.

While the problem of full-length transcript assembly from short reads is in general difficult and underdetermined [13], current methods have particular difficulties in annotating 3' terminal exons correctly. In contrast to the boundaries of internal exons, which can be identified precisely by virtue of reads that span splice junctions, terminal exon boundaries only evince themselves in RNA-seq data as the position at which read coverage decreases. To identify terminal boundaries the popular transcriptome assembly tool Cufflinks constructs a minimum path cover from compatible RNA-seq reads, resulting in a single terminal exon annotation that extends from the

last splice acceptor site to the position where there is zero read coverage [139]. Since low-frequency polyadenylation site read-through events are captured in RNA-seq experiments, this strategy will often result in the terminal 3' exon boundary being extended beyond the dominantly used polyadenylation site. Cufflinks implements a post hoc trimming process to mitigate this problem. Another well-utilized transcriptome assembly tool, Scripture, identifies transcribed segments of the genome using scan statistics. Here, the terminal exon boundary is determined as the location where the statistic calculated within a window overlapping the terminal exon drops below the genome-wide significance threshold [48].

Neither of these strategies is specifically designed to identify the set of positions at which the level of RNA-seq coverage transitions from high-to-low coverage, and neither is capable of generating more than one terminal exon annotation. Consequently, the terminal exon annotations built from RNA-seq data can be inaccurate and incomplete. This point was emphasized during a recent comparative assessment of 14 different algorithms for transcript assembly and exon identification, conducted as part of the RNA-seq Genome Annotation Assessment Project (RGASP) [131]. In particular, the outputs for transcript termini from all of the algorithms tested were sufficiently inaccurate that a relaxed criteria exon correctness was used that evaluated only on the 5' boundary of the 3' terminal exon [131]. This highlighted the need for improved methods to identify transcript termini from RNA-seq data.

The 3' untranslated region (3' UTR) is an important location of post-transcriptional regulation, and deregulation of 3' end formation has medical relevance [31]. In recent years, there has been increasing appreciation that most genes are subject to

alternative cleavage and polyadenylation (APA) to yield multiple 3' UTR isoforms, and that APA is frequently modulated in tissue-specific, state-specific, or environmentally responsive manner [92]. Although various specialized techniques have been developed to sequence the 3' ends of transcripts [31], conventional RNA-seq methods remain the methodology of choice for most laboratories, and the amount of existing RNA-seq data is far greater than for 3'-seq data. Thus, it would be highly desirable to improve the accuracy with which we can infer 3' UTR boundaries and alternative isoform expression from RNA-seq data.

Recently, we used tissue-specific RNA-seq data to refine terminal exon models in the human and mouse genomes [91]. We encountered unique challenges when refining terminal exons, which motivated the implementation of a specialized annotation process for 3' terminal exons. For example, while transposable element insertions are strongly selected against in the coding exons of a gene, UTRs are less constrained and harbor thousands of repetitive elements genome wide [20]. Nonuniform read coverage arising from sequence specific [51, 68] as well as positional biases [16], and uncertain allocation of multimapping reads [94] can cause artificial local gaps in RNA-seq coverage, making it difficult to annotate full-length 3' UTR models using next-generation sequencing data alone (Figure 15A). We previously developed an ad hoc procedure to bridge short gaps in RNA-seq coverage when known sources of sequencing bias could be identified. This enabled us to extend thousands of 3' UTR models in the extensively annotated mouse and human genomes. A substantial proportion of these extensions show tissue-specific expression patterns, and identify a previously unappreciated potential for post-transcriptional regulation in the extended genes [91].

Figure 15: A) Genome browser view of Kif3a illustrates how regions of low-coverage RNA-seq can result in fragmented 3' UTR assemblies reported by Cufflinks and Scripture. There is an annotated repetitive element coinciding with the position of the gap, which could explain the observed decrease in coverage at this location, due to ambiguous read mappability. B) Hdlbp illustrates a gene with tandem polyadenylation sites that is not completely annotated using either Cufflinks or Scripture. The Ensembl 73 annotation contains transcript models with alternative short/long 3' UTR isoforms, and the relative abundance of these isoforms is reflected by the pattern of RNA-seq depth. However, neither Cufflinks nor Scripture is able to assemble the short isoform; both exclusively report the long isoform. C) The maximum marginal likelihood segmentation is computed using a dynamic programming algorithm that recursively computes Q(t), the likelihood of the optimal segmentation of the subsequence of observations yt:n, given that there is a change point at position t - 1. This graphic illustrates how the value of Q(t) at each position in the sequence is decomposed into a component representing the likelihood of sequences of the data starting at position t (red) and the maximum likelihood segmentation of the remainder of the data (blue), and a component representing the likelihood that there are no change points in the sequence yt:n (green). RNA-seq data are shown for the Ict1 gene, which utilizes tandem alternative 3' ends, both of which are supported by 3'-seq data. D) Toy data are used to illustrate the effect of constraints on the segmentation process. Above, the scatter plot displays simulated coverage data, with a local region of low coverage indicated by the bracket at top. Below, vertical lines indicate location of change points identified using the standard and constrained formulations of the change-point inference algorithm, where red lines indicate locations of decreased coverage and green lines indicate the locations of increased coverage. The unconstrained solution identifies both the local dip and the most distal change point, while the constrained solution reports the most likely configuration of change points that conforms to the requirement for monotonically decreasing coverage in sequential segments.

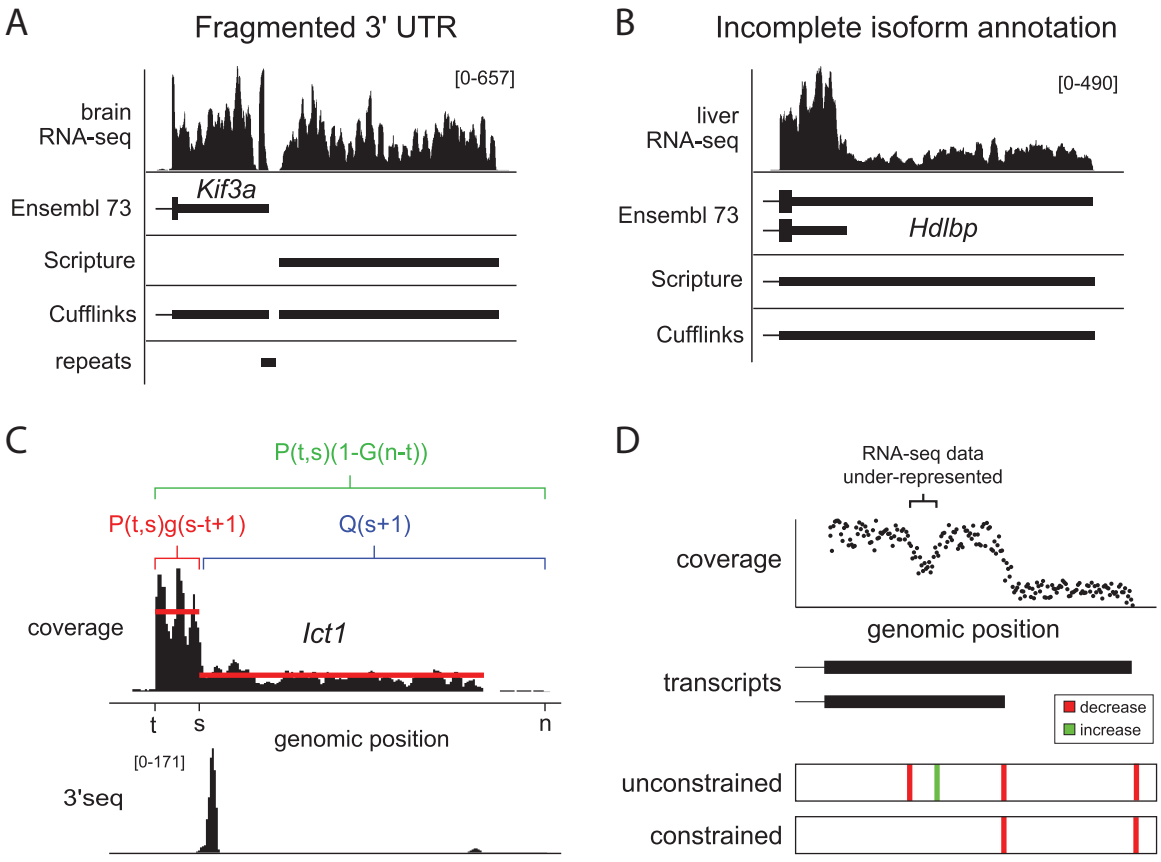

A
Fragmented 3' UTR
[0-657]
brain
RNA-seq
Kif3a
Ensembl 73
Scripture
Cufflinks
repeats
B
Incomplete isoform annotation
[0-490]
liver
RNA-seq
Ensembl 73
Hdlbp
Scripture
Cufflinks
C
P(t,s)(1-G(n-t))
P(t,s)g(s-t+1)
Q(s+1)
coverage
Ict1
t
s
n
genomic position
[0-171]
3'seq
D
RNA-seq data
under-represented
coverage
genomic position
transcripts
decrease
increase
unconstrained
constrained

Figure 15

We formalize this procedure here, extending our previous work for terminal exon annotation, using a segmentation approach that integrates long-range patterns of RNA-seq coverage to identify polyadenylation sites with greater sensitivity and specificity than existing methods. More importantly, we demonstrate its utility for identifying complex patterns of tandem polyadenylation site usage that are inaccessible with conventional annotation strategies. We implement our approach as the standalone program Isoform Structural Change Model (IsoSCM), which is available from our website (https://github.com/shenkers/isoscm).

3.4 Results

3.4.1 Transitions in coverage depth identify 3' UTR boundaries

RNA-seq protocols sample reads from across transcript bodies, approximately uniformly, although with certain biases [94]. Existing approaches use minimum path coverage [139], or a scan statistic [48], to identify transcribed segments, and annotate at most one 3' boundary for each terminal exon, typically the longest isoform compatible with the reads. Since the longest isoform will not in general reflect the dominant 3' UTR isoform used by a gene, Cufflinks uses a heuristic post-assembly processing step to trim terminal exon annotations to a prespecified fraction of the average level of coverage. While such a strategy will identify high abundance short isoforms at a subset of loci, a single trimming parameter will not result in optimal annotations genome wide. Moreover, these strategies tend to generate incomplete 3' UTR assemblies because they cannot capture tandem terminal exon isoforms that are

coexpressed in a given sample, as illustrated in Figure 15B.

Given the unique challenges associated with transcript assembly within 3' UTRs, and to address the limitations of existing tools, we developed a more expressive framework for transcript assembly that incorporates information from the patterns of read coverage into the process of UTR boundary definition. If we assume that sequenced reads are distributed approximately uniformly across the transcript, the boundaries of transcription will be marked by a change in the level of coverage. In instances where a shorter exon is nested within a longer exon, there can still be a significant number of reads aligning downstream from the shorter isoform, creating a "step-like" pattern of coverage at the boundary of the nested exon model. For example, RNA-seq data for the Hdlbp and Ict1 genes show such drop-offs within their 3' UTRs, indicative of tandem APA events (Figure 15B,C).

To identify terminal exon boundaries, we thus seek critical points ("change points") that mark transitions in RNA-seq coverage. Previously, segmentation approaches were used to identify transcript boundaries from tiling microarray probe intensities [57], and while these change points have been described in RNA-seq data [95], no existing RNA-seq ab initio transcript assembly tool fully leverages this information to annotate 3' UTR boundaries. To fill this gap, we adapt multiple change-point inference to the problem of 3' UTR isoform identification.

3.4.2 Inference for multiple change-point problems

To implement change-point inference, we made use of a Bayesian framework for change-point inference established previously [33]. For a sequence of n observations

$y_{1:n}$=$y_1, \ldots, y_n$ representing the level of coverage at sequential genomic positions, we consider all possible combinations of m change points $\tau_1, \ldots, \tau_m$ where $0 < \tau_i < n$, $\tau_i < \tau_{i+1}$, and $0 \leq m < n$, such that the jth segment pertains to the observed level of coverage between two successive change points. We assume that the level of coverage observed at each position within a segment are independent samples from a common probability distribution $f(x|\theta)$, parameterized by θ, with prior distribution $\pi(\theta)$. To model the expected length of a segment, we define a probability mass function $g(t)$ for the length t of the genomic segment between two successive change points, with a cumulative mass function $G(t) = \sum_{s=1}^{t} g(s)$. Given this probability model, the goal of change-point inference is to identify the set of change points that maximizes the marginal likelihood of the data, i.e., the set of change points that "best explain" the observed pattern of read coverage.

To achieve this, we implemented a dynamic programming algorithm that recursively calculates the maximum marginal likelihood solution for nested subsequences of the observations. To begin, for all indices in a subsequences $y_{t:s}$, such that $0 \leq t \leq s \leq n$ we calculate a marginal likelihood $P(t, s)$ that the observations $y_{t:s}$ were sampled from a common distribution as the integral of the joint data likelihood over the possible parameter values within that segment:

$$P(t, s) = \int \prod_{i=t}^{s} f(y_i|\theta)\pi(\theta)d\theta \tag{1}$$

The most likely segmentation using this model is defined recursively in terms of the likelihood of the current segment (Figure 15C, red bracket), and the likelihood of

the remainder of the data (Figure 15C, blue bracket), and alternately, the likelihood of the remainder of the data coming from a single segment (Figure 15C, green bracket). This requires the construction of two tables $Q(t)$ and $R(t)$ of length n, indexed by t, the location of the start of the current segment. $Q(t)$ stores the maximum marginal likelihood of a segmentation of the subsequence of observations from $y_{t:n}$, given a change point at $t-1$, while $R(t)$ stores the index of the next change point in the maximum marginal likelihood segmentation, given a change point at $t-1$. These tables can be recursively computed in $O(n^2)$ operations using the formulas:

$$Q(t) = \max_{s \in (t,n)} \begin{cases} P(t,s)Q(s+1)g(s-t+1) & \text{if } s < n \\ P(t,s)(1-G(n-t)) & \text{if } s = n \end{cases} \tag{2}$$

$$R(t) = \text{argmax}_{s \in (t,n)} \begin{cases} P(t,s)Q(s+1)g(s-t+1) & \text{if } s < n \\ P(t,s)(1-G(n-t)) & \text{if } s = n \end{cases} \tag{3}$$

The sequence of change points can be recovered by performing a trace-back through table $R(t)$; the location of the first change point τ_1 is given by $R(1)$, and subsequent change points are given by $\tau_{j+1} = R(\tau_j)$, while $\tau_j < n$. A more detailed derivation of these equations along with proofs is given previously [33].

3.4.3 Constraining the location of change points

Although the change point model is able to tolerate a degree of variation within each segment, real RNA-seq data contain sequencing biases that can cause the sampling of reads across the transcript body to deviate from a uniform distribution. These

biases typically cause short segments of the transcript to be sequenced at a lower frequency, resulting in a local drop in the level of coverage. The conventional (unconstrained) implementation of the change-point detection procedure identifies these local in-homogeneities in coverage as a segment with a distinct coverage distribution. However, in the context of annotating terminal exon boundaries we wish to disregard these local aberrations as they do not correspond to terminal exon boundaries. In order to minimize the effect of local biases on transcript model inference we introduce additional constraints to the change-point identification procedure.

To distinguish local changes from change points that mark a sustained decrease in the level of coverage, we restrict the set of identified change points to conform to a pattern of monotonically decreasing coverage over sequential segments. Solutions satisfying this requirement are achieved by discarding any configurations of change points in which the fold change in the level of coverage between two neighboring segments is less than a specified fold-ϕ.

This constraint is implemented by modifying the recursions in Equations 2 and 3. To do so, we introduce the functions $M(t, s)$ to represent a point estimator of the level of coverage for observations $y_{t:s}$, and $C(s)$ to represent the maximum estimated coverage of the most likely segmentation of observations $y_{s:n}$, respectively. We enforce this constraint by assigning zero probability to all configurations in which the estimated coverage of a downstream segment exceeds the estimated coverage of an upstream segment. Formally, this requires replacing the unconstrained recursions (Equations 2 and 3), with constrained versions (Equations 4 and 5).

$$Q_c(t) = \begin{cases} 0 & \text{if } M(t,s) < C(s+1) \\ Q(t) & \text{if } M(t,s) \geq C(s+1) \end{cases} \tag{4}$$

$$R_c(t) = \begin{cases} 0 & \text{if } M(t,s) < C(s+1) \\ R(t) & \text{if } M(t,s) \geq C(s+1) \end{cases} \tag{5}$$

Here, a parameterization $\phi = 1$ corresponds to a requirement that the level of coverage of sequential segments are strictly decreasing, while $\phi = 2$ the level of coverage of sequential segments drops at least twofold at each change point. The consequences of including these constraints are illustrated for a toy example in Figure 15D. In this scenario, a hypothetical 3' UTR has relatively uniform coverage, except for a local region (Figure 15D, bracket), where read coverage is underrepresented. When the unconstrained implementation of change-point detection procedure is applied to identify polyadenylation sites within the last exon, it detects two locations where the coverage drops, one upstream of the underrepresented segment, and the other at the polyadenylation site of the hypothetical exon. While this accurately reflects the observed pattern of coverage, for the purpose of polyadenylation site annotation, we are only interested in the location of the second change point.

Application of the constrained segmentation procedure has the desired effect; a change point is reported at the position of the polyadenylation site, while the change point caused by the local drop in coverage is omitted. While this limits the ability to identify arbitrary combinations of change points, reads sampled from tandem termi-

nal exon isoforms are expected to produce a "step-like" coverage pattern. Thus, this constraint is intended to promote sensible inference of change-point location in situations where read distributions deviate from a theoretically uniform sampling across the transcript body.

3.4.4 Implementing change-point detection for 3' UTR annotation

In order to implement the described change-point inference algorithm, one needs to specify the distributions $f(y_i|\theta)$, $\pi(\theta)$, and $g(t)$. While the depth of sequencing coverage is a discrete counting process and would typically be modeled using a Poisson distribution, it has been suggested that read counts from sequencing data are overdispersed, and are more appropriately modeled by a Negative Binomial distribution [6]. Therefore, we model $f(y_i|\theta) \sim \mathrm{NB}(p, r)$, and use an uninformative conjugate prior $\pi(p) \sim \beta(1, 1)$, which allows for analytical integration of Equation 1.

While the procedure described above efficiently identifies optimal sets of change points, these do not by themselves provide meaningful transcript models. For this reason the change points need to be interpreted in the context of the exon that they occur. For example, a transition from high-to-low coverage (over increasing genomic coordinates) would suggest the presence of a polyadenylation site on the plus strand, and vice versa for a gene on the minus strand. To enable application of change-point detection in an ab initio setting, prior to segmentation, covered genomic segments and spliced reads are used to identify regions that will be searched for change points. The output, including inferred overlapping terminal exons models, is reported in a compact splice graph. This graph identifies exon boundaries and connections between

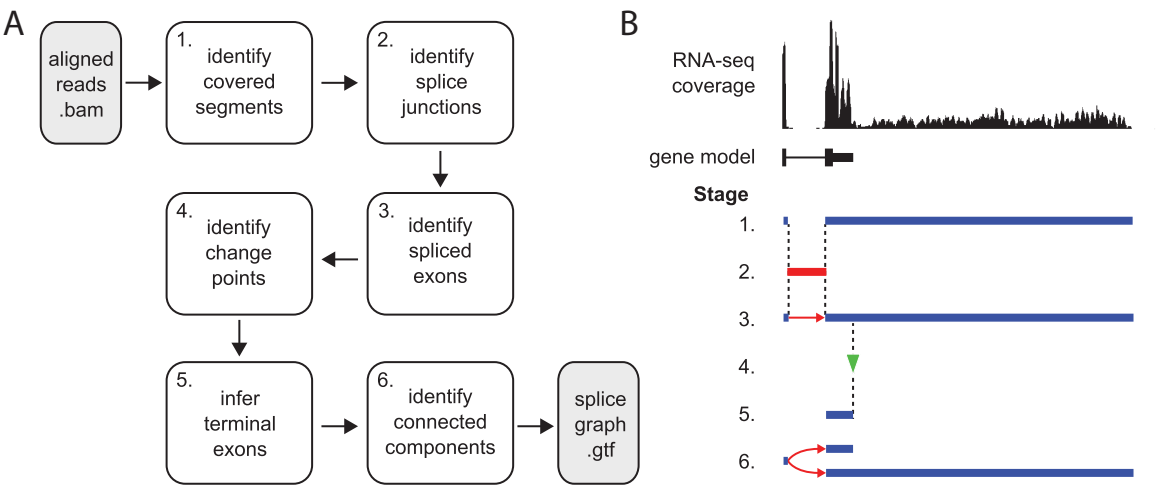


Figure 16: A) The sequence of steps IsoSCM uses to build a splice graph are illustrated as a flow chart, starting from a set of reads that have been mapped to a genome using splice-aware alignment software. B) The assembly operations from the flow chart at left are illustrated for a prototypical gene with tandem 3' ends. These steps are (1) identification of segments of the genome where continuous read coverage is observed; if paired end reads information is available mate-pairs will be used to scaffold segments together. Locations with expected low coverage (i.e., repetitive elements or low-complexity sequence) can also be used to scaffold segments separated by gaps. (2) The location of splice junctions is recovered from the mapped reads; (3) the boundaries of spliced exons are inferred by intersecting continuously transcribed regions with the positions of splice junctions; (4) change points in the level of coverage are identified, using the constrained segmentation procedure; (5) terminal exon structure is inferred from change-point location; (6) the assembled splice-graph is reported, where exons are labeled as elements of a common transcription unit if they are connected either by spliced reads, or occupy overlapping genomic segments.

exons supported by spliced reads. The stages of the IsoSCM annotation algorithm are illustrated in Figure 16.

3.4.5 Method evaluation: simulated data

To assess the benefit of incorporating coverage-based segmentation for terminal exon annotation, we compared the performance of IsoSCM with two widely used *ab initio* reference based annotation tools: Cufflinks [139] and Scripture [48]. We prepared a test set using simulated data, where the transcript structures underlying the sequenc-

ing data are known a priori (illustrated in Supplemental Figure 17A). This test set was comprised of 14,263 nonoverlapping genes, based on transcript models obtained from the mouse Ensembl 73 release. For each gene, we selected a single-transcript isoform, and generated a new isoform by randomly truncating the terminal exon at a random position that was at least 150 nt from both the 5' and 3' boundaries of the original exon.

As we expect the accuracy of assembled models to relate to sequencing depth, we measured the performance of each method over a range of coverage levels. For each pair of isoforms we simulated reads sampled uniformly across the body of the transcript, such that the aggregate density of reads over exonic segments was 2000 reads per kilo base (kb). This library was recursively subsampled such that evaluations spanned a range of read densities from 5 to 2000 reads/kb. For each simulated data set, we used IsoSCM, Cufflinks, and Scripture to assemble a set of predicted transcript models. For each assembly we assessed the correctness of 3' UTR predictions by classifying assembled terminal exons as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) with respect to the reference ground truth transcript set.

Since it is difficult to predict nested terminal exon boundaries with nucleotide-level accuracy from RNA-seq data, and there is value in an annotation that is "close" to the true transcript model, we use a classification scheme that allows a degree of flexibility. In contrast to the relatively generous assessment scheme used by the RGASP project [131], which required only the spliced 5' boundary of the terminal exon to match the reference and disregards the predicted 3' end, we require both 5'

Figure 17: A) A graphic overview illustrates the framework for evaluating assembly methods using simulated data. B) The level of coverage for a simulated gene is plotted for a range of read densities. At low read density the location of the tandem polyadenylation sites is not apparent. The position of the nested boundary becomes evident as sequencing depth is increased. C) Classification labels are derived from reference transcript models (black). The region encompassing the terminal exon is broken into target segments with a radius w base pairs from a reference 3'boundary. Segments that overlap a 3'boundary are considered positives (green), while segments that do not overlap a 3'boundary are considered negatives (red). Assembled models (gray) are evaluated by determining consistency with these segments. A positive segment that overlaps a predicted 3'boundary is considered a TP, and otherwise is considered a FN. Similarly, a negative segment that does not overlap a predicted 3'boundary is considered a TN and otherwise a FP.

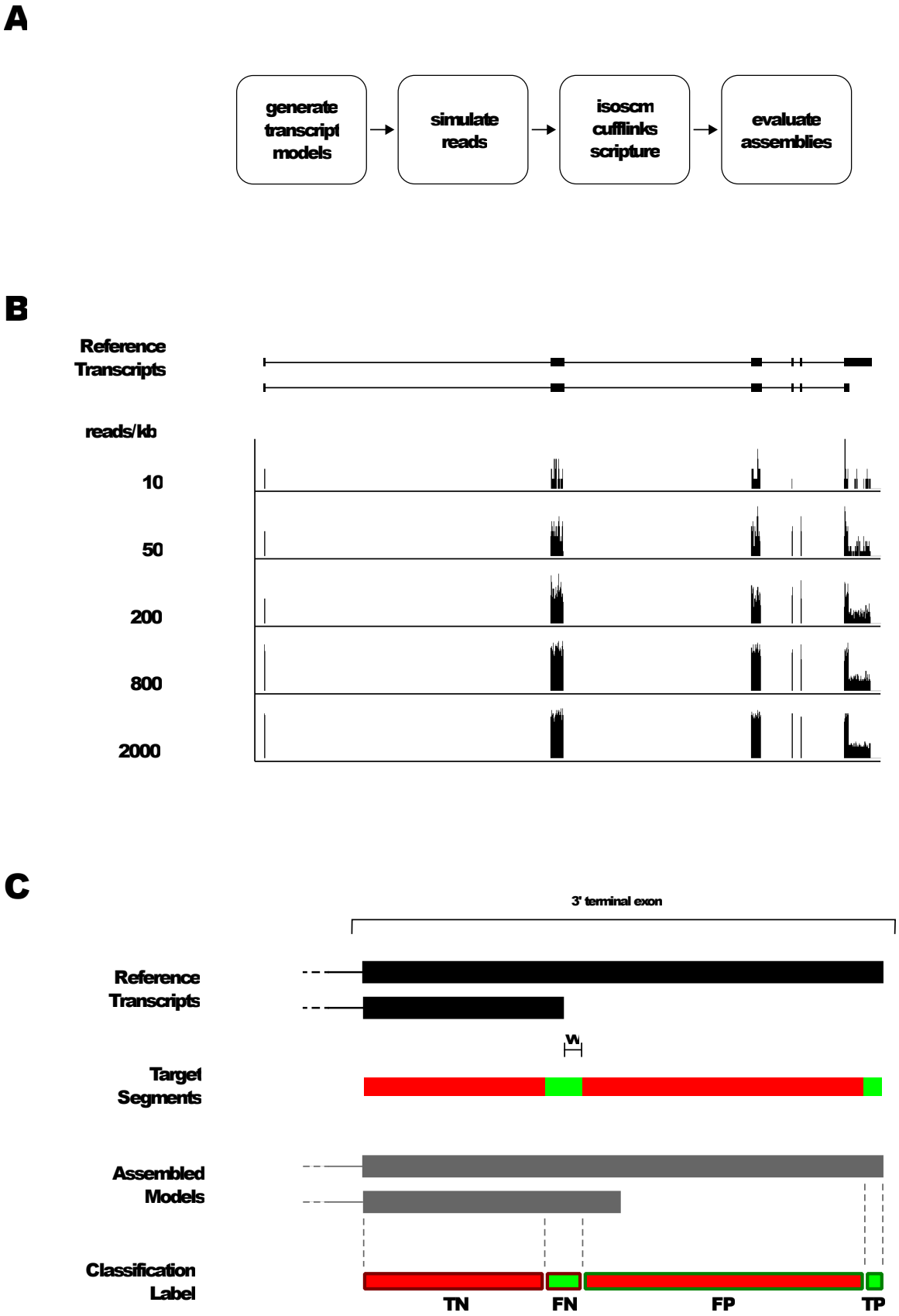

A
generate transcript models
simulate reads
isoscm cufflinks scripture
evaluate assemblies
B
Reference Transcripts
reads/kb
10
50
200
800
2000
C
3' terminal exon
Reference Transcripts
w
Target Segments
Assembled Models
Classification Label
TN
FN
FP
TP

Figure 17

and 3' boundaries to be similar to reference models. We classified predictions that are within a small (100 nt) distance of a reference model to be correct by dividing the interval spanning the terminal exon into segments that are positive or negative for polyadenylation, such that assemblies can be judged for their consistency with these segments (Figure 17C). The genomic segments within 100 nt of a reference 3' ends were counted as TPs if an assembled 3' end fell within that segment, and were otherwise counted as FN. Likewise, segments $>$ 100 nt from a reference end were counted as FP if an assembled 3' end fell within that segment, and TN otherwise. Using these counts we calculated the PPV = TP/TP + FP, TPR = TP/TP + FN, FPR = FP/FP + TN, NPV = TN/TN+ FN, for each assembly.

These metrics are plotted as a function of sequencing depth in Figure 18A,B and Figure 19. In these simulations, IsoSCM identifies the terminal exon boundaries with a PPV comparable to Cufflinks and Scripture at all sequencing depths, indicating that the change-point inference procedure does not inappropriately increase the number of false positives compared with methods that attempt to assemble only a single 3' UTR isoform. In contrast, comparisons of the TPR achieved by each method highlight the fact that the models built by Cufflinks and Scripture were unable to annotate tandem terminal exons. Even as the number of available reads is increased toward infinity, the maximum TPR achieved by these methods was 0.5. A priori, we expect that Cufflinks and Scripture cannot exceed this level of sensitivity since the gene models they build are not expressive enough to represent tandem overlapping terminal exons, and one out of every two 3' terminal exons in the test set is nested within a longer exon. In contrast, given sufficient data to statistically identify changes in the level of

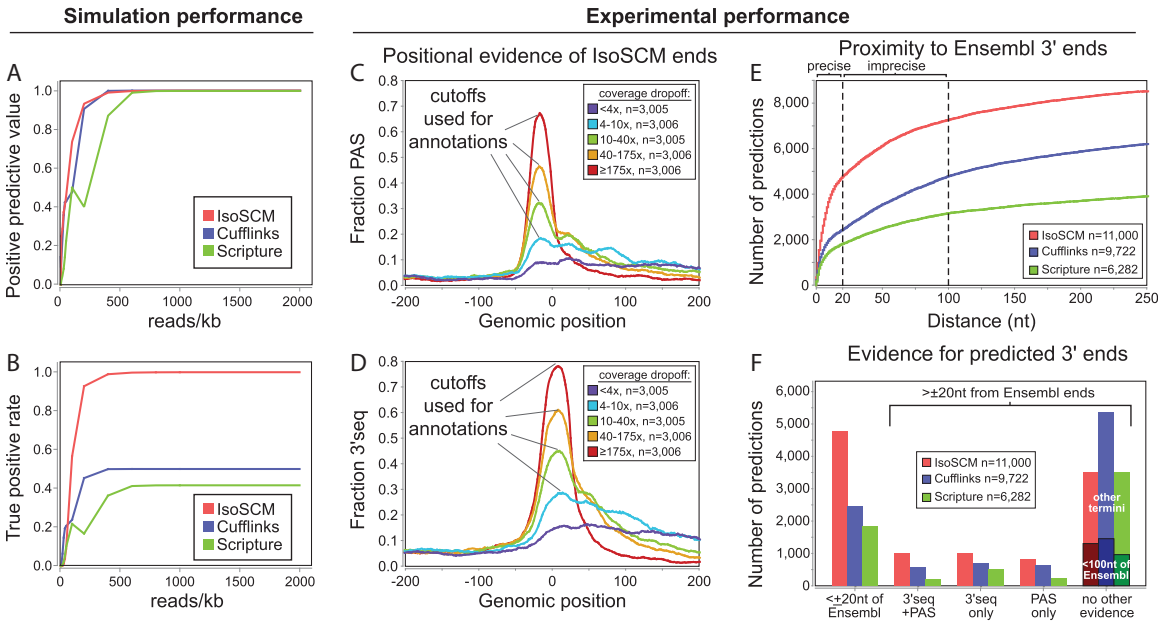


Figure 18: (A-B) Simulated RNA-seq data were used to assess predictive positive value (A) and true positive rate (B) of IsoSCM, Cufflinks, and Scripture outputs for 3' end prediction. We generated a set of 14,263 nonoverlapping gene models that contain transcripts with nested 3' terminal exons, and used these as a reference set of "true" transcripts. These metrics were calculated for simulated sequencing depths ranging from 5 to 2000 reads/kb. (C-D) We assessed the positional accuracy of IsoSCM outputs across a range of change-point magnitudes. We partitioned these events into quintiles (with n = 3005 termini in each group), corresponding to bins of <4x, 4-10x, 10-40x, 40-175x, and ≥175x drop-offs in read coverage. For each group the fraction of predicted termini with either canonical polyadenylation signals (PAS, AATAAA, or ATTAAA) or 3'-seq tags within 20 nt are shown at each position relative to the predicted boundary. Based on signals for appropriate positional enrichment, we utilized the top four cutoffs for running IsoSCM. E) Proximity of IsoSCM, Cufflinks, and Scripture terminal outputs relative to Ensembl 3' end annotations. The cumulative number of annotations at each distance to the closest Ensembl 3' end is plotted. Based on apparent inflection points (dashed lines), we categorize annotations within 20 nt of Ensembl as precise annotations, and ones between 20 and 100 nt as imprecise matches to reference models. F) Validation of IsoSCM, Cufflinks, and Scripture terminal outputs within ±20-nt windows of various types of supporting evidence. Ends were initially assigned, if possible, to Ensembl models, and then checked for proximity to PAS and/or 3'-seq tags. Of the remaining termini with "no evidence," many would be validated using a relaxed 100-nt window. As the largest numbers of these correspond to "imprecise" calls of Ensembl ends (see E), we marked their numbers as sub-bars in the "no evidence" category. Regardless of the type or types of evidence considered, IsoSCM yields the largest numbers of validated termini without inflating numbers of unvalidated predictions.

coverage, IsoSCM is able to achieve perfect TPR for the test set (Figure 18A).

We explored TPR, PPV, NPV, FPR measurements further, by subdividing the aggregate performance on the entire test set into component corresponding to proximal and distal isoforms exclusively (Figure 19). The methods performed comparably for the task of distal isoform identification, and the performance gain achieved by IsoSCM corresponds to an increased number of correct annotations of nested isoforms. To explore the relationship between sequencing depth and annotation accuracy we repeated the evaluation, requiring that predictions be within 10 nt of the reference transcript to be classified as TP. Under these conditions, we observed qualitatively the same pattern; only IsoSCM is able to reconstruct the nested isoforms correctly (Figure 20). However, while a sequencing depth of 200 reads/kb was sufficient to achieve a TPR = 0.9 and FPR = 0.1 with 100-nt resolution, this level of performance required over 1000 reads/kb at the 10-nt resolution. We note that all methods fail to reliably assemble low expressed transcripts, and that the advantages of IsoSCM are more fully realized when there are sufficient data to distinguish overlapping isoforms. Nevertheless, IsoSCM performs equally or better to Cufflinks and Scripture over a range of sequencing depths.

3.4.6 Method evaluation: real RNA-seq data

Of course, we do not expect experimentally generated RNA-seq data to be as well-behaved as simulated data. To verify that the benefits observed in our simulation experiment are achieved when IsoSCM is applied to real RNA-seq data, we utilized published RNA-seq data comprised of nine mouse tissues analyzed in triplicates [86].

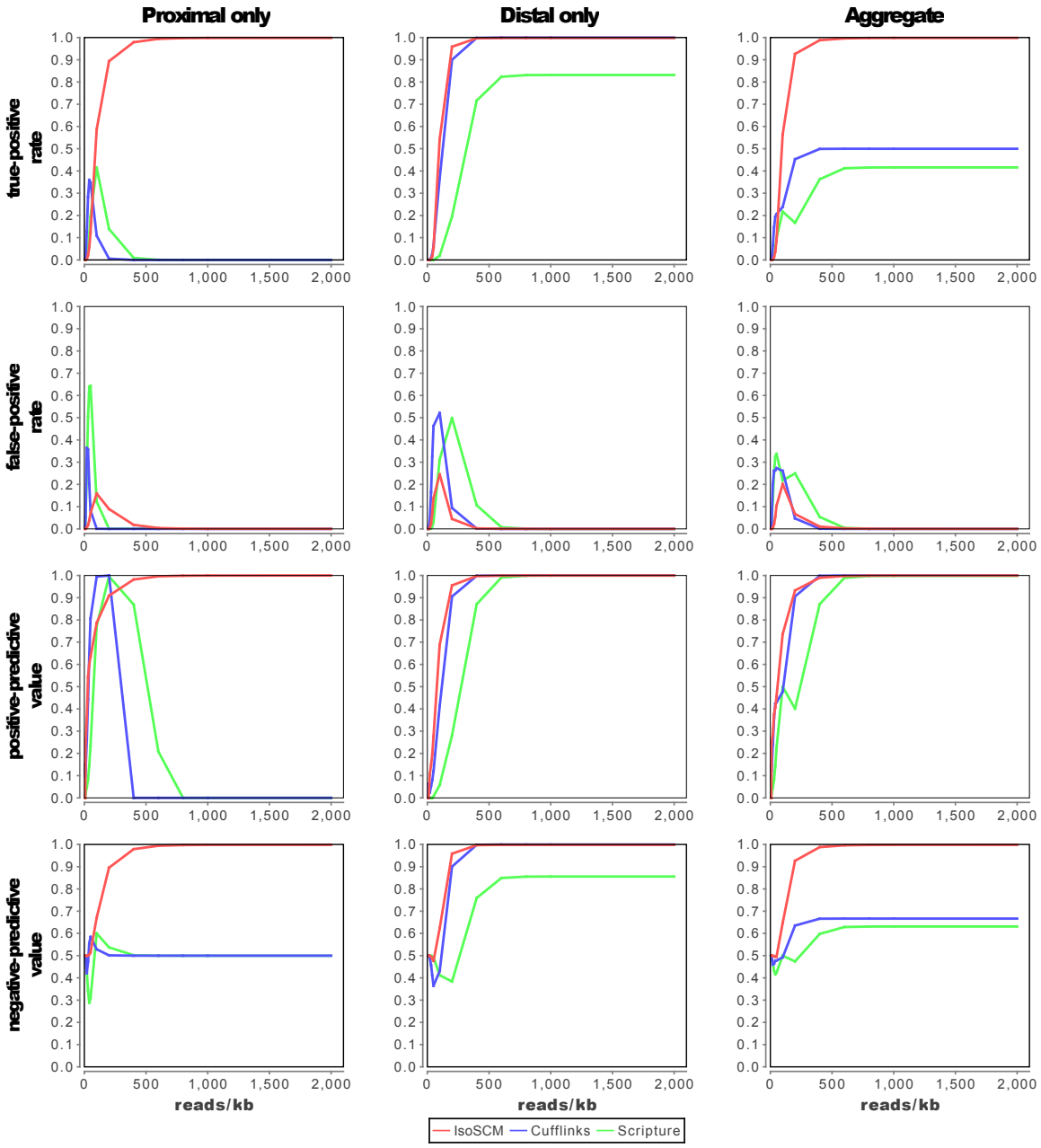


Figure 19: TPR, FPR, PPV, and NPV are plotted for simulated transcripts over a range of 5-2000 reads/kb, evaluated using a 100nt resolution. In addition to considering aggregate performance for all simulated 3'ends, we separately consider the separate contribution for predictions of proximal and distal 3' UTR isoforms exclusively. While performance for prediction of distal ends improves for all methods with increasing sequencing depth, the ability of Cufflinks and Scripture to predict proximal 3' UTR isoforms deteriorates with increasing sequencing depth.

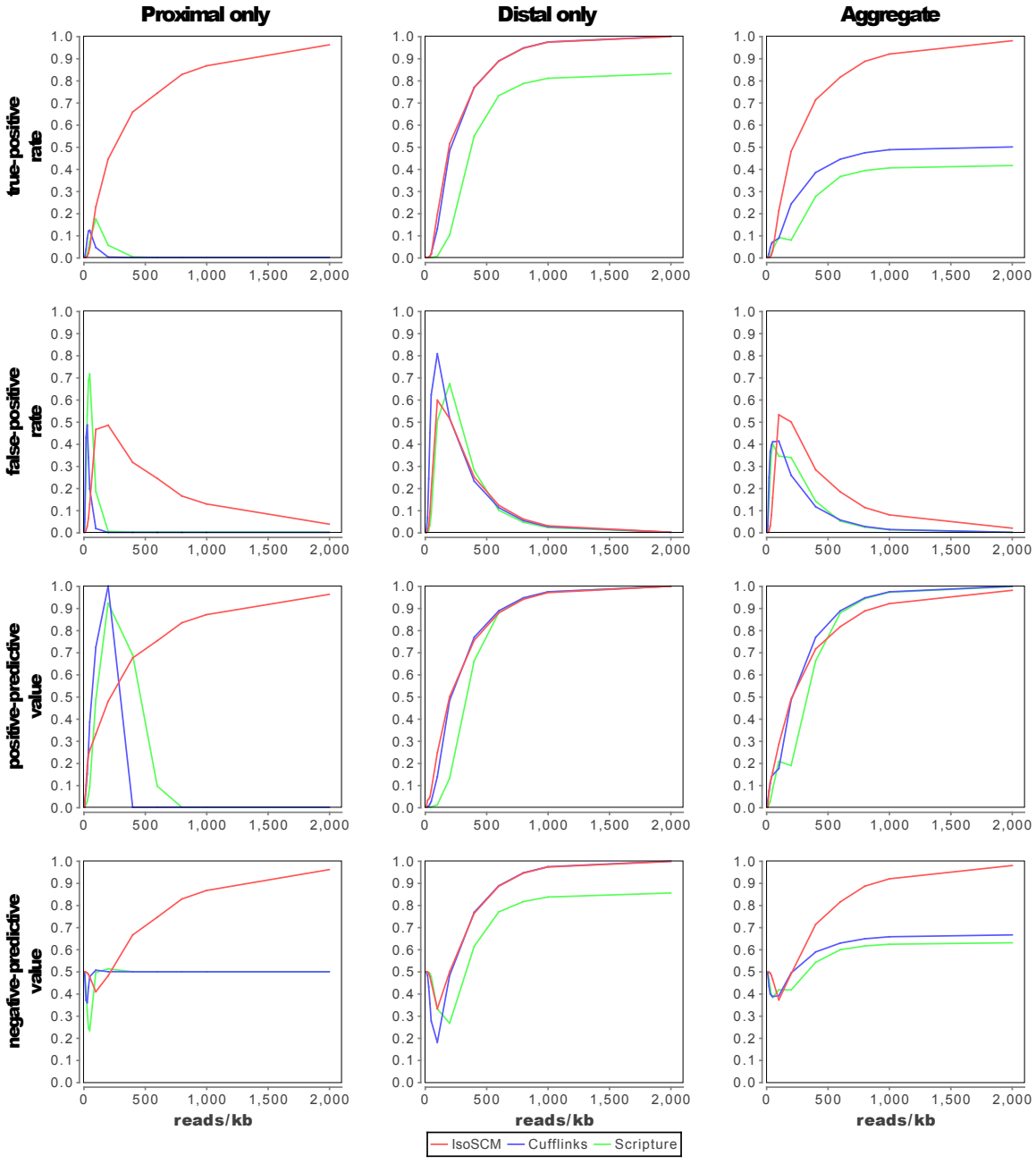


Figure 20: As in Figure 19, TPR, FPR, PPV, and NPV are plotted for simulated transcripts over a range of 5-2000 reads/kb, except using a 10nt window resolution.

In contrast to the simulation experiment, the ground truth set of transcripts expressed in each sample is not known a priori. As a first measure of model correctness, we compared the terminal boundaries of constructed transcript models with those in the Ensembl 73 reference annotation. However, as existing 3' UTR reference annotations are incomplete [91], we expect that sole comparisons to the Ensembl reference to inflate FPR estimates. Therefore, we sought additional orthogonal sources of evidence to support predicted polyadenylation sites. Directed 3'-sequencing (3'-seq) protocols permit experimental mapping of polyadenylation sites, and we used an atlas of 3'-seq data from multiple mouse tissues [28]. As well, cleavage and polyadenylation sites have long been known to be associated with a polyadenylation signal (PAS) hexamer [112], most frequently either AATAAA or ATTAAA, located $\sim$21 upstream of the cleavage site. While such motifs are not sufficient to specify 3' end formation, we could assay PAS enrichment in the vicinity of novel 3' end annotations as another measure of their quality.

We first sought to evaluate change-point cutoffs that delivered appropriate accuracy for genuine 3' termini. We tested this using mouse brain (SRR594393), and partitioned predicted ends into quintiles by the magnitude of the change point. We then evaluated the positional enrichment of PAS and 3'-seq data supporting these groups, to examine the genomic precision with which IsoSCM called 3' termini. We observed robust positional enrichments of PAS and 3'-seq tags in the appropriate locations relative to IsoSCM annotations (Figure 18C,D), demonstrating that these sources of evidence could be used to evaluate de novo predictions. As expected, annotated 3' termini associated with the sharpest drops in coverage identified genuine

polyadenylation sites with greatest accuracy, while positional support for PAS and 3'-seq tags was distributed more broadly for weaker change points. We focused on models for which the coverage drops at least fourfold in order to balance sensitivity and specificity of the predictions.

We then compared the performance of the different transcript assembly methods on the mouse brain RNA-seq reads, from which 11000 3' terminal exon models were constructed by IsoSCM, 9722 by Cufflinks, and 6282 by Scripture. Upon examining the distribution of the distance to the closest Ensembl 3' end for each method (Figure 18E) we observed that all three methods generate two populations of predictions distinguished by the precision with which they identify Ensembl 3' ends. The inflection point at $\sim$20 nt identifies a population of predictions that capture annotated termini quite precisely, while a second inflection point at $\sim$100 nt identifies a set of predictions that are localized in a wider window around annotated 3' ends (Figure 18E, "imprecise"). For many types of genome-wide analysis, i.e., when assessing general trends of transcript isoform expression, nucleotide-precise annotations of 3' termini are not necessary. Nevertheless, as the narrow 20-nt window enables the merits of predicted models to be evaluated stringently, we assessed how frequently each method produced 3' termini that were validated by polyadenylation site features at a distance of $\pm$20 nt (Figure 18F).

Of the total set of predictions, IsoSCM made 4785 annotations that were $\pm$20 nt of Ensembl 3' ends, compared with 2450 by Cufflinks and 1827 by Scripture. In addition, each method annotated thousands of termini not supported by reference models. Among these, IsoSCM identified 2718 novel termini supported by 3'-seq

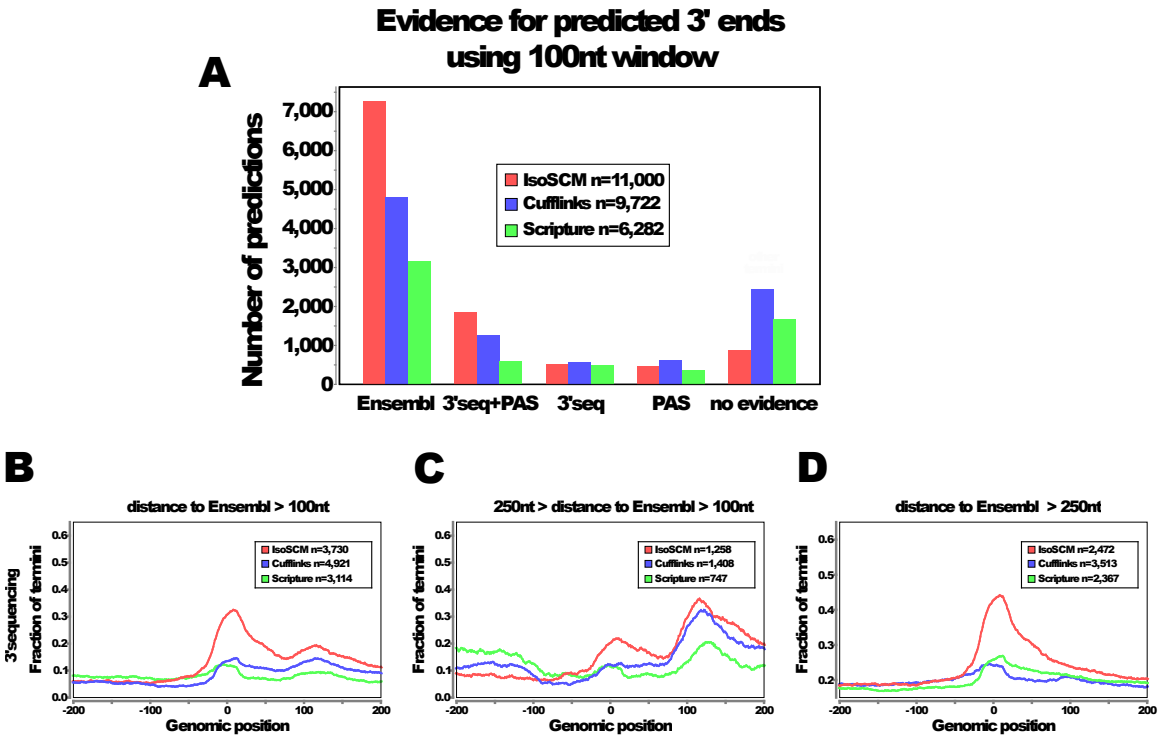


Figure 21: A) Validation of IsoSCM, Cufflinks, and Scripture terminal outputs within +100 nt windows of various types of supporting evidence. Ends were initially assigned, if possible, to Ensembl models, and then checked for proximity to PAS and/or 3'-seq tags. At this resolution IsoSCM also has the largest aggregate number of validated 3' ends, and the lowest proportion of unvalidated 3' ends. B) Positional enrichment of features examined for all non-Ensembl ends predicted by each method, defined as being located >100 nt from an Ensembl terminus. These ends appear compromised by inclusion of these "Ensembl-imprecise" termini, which create an offset peak of evidence at 100nt downstream, the boundary of the window utilized to define novel ends. To clarify this pattern, we partitioned these 3'ends into those that are "Ensembl-imprecise" (250 >distance to Ensembl $\geq$ 100) in (C), or "substantially novel" (distance to Ensembl >250) in (D). While 20% of the 1,258 "Ensembl-imprecise" ends reported by IsoSCM are supported by 3'-seq evidence at position-0, over 30% have 3'-seq evidence 100nt downstream. Similar patterns of out-of-position enrichment are seen for "Ensembl-imprecise" predictions made by Cufflinks and Scripture. On the other hand, predictions that are "substantially novel" are predominantly enriched for these features at the expected position.

tags, PAS, or both features, within 20 nt, compared with 1914 identified by Cufflinks and 941 identified by Scripture (Figure 18F). These evaluations demonstrate the strong performance of IsoSCM on real RNA-seq data using a stringent cutoff. Finally, a substantial portion of the remaining termini annotated by each method ("no evidence") could be considered validated if one utilized a broader window, e.g., 100 nt. Since the majority of these events can be defined as being "imprecise" captures of Ensembl ends, we have noted their numbers within the "no evidence" category of Figure 18F, and using the relaxed 100-nt criteria in Figure 21A. A full accounting of the termini annotated by the three algorithms using both 20- and 100-nt windows is given in Table 1. Overall, we find that IsoSCM correctly annotates more 3' UTRs while simultaneously reporting the fewest unvalidated annotations, regardless of the window size used for model evaluation.

Table 1: Aggregate and method specific counts of the number of 3'end predictions made by each method that are supported by Ensembl, PAS, or 3'seq in 20 and 100nt windows are reported.

		Validation of all predicted ends								
		20nt < distance to feature				20nt < distance to feature < 100				> 100nt any feature
		Ensembl	3'Seq&PAS	3'Seq	PAS	Ensembl	3'Seq&PAS	3'Seq	PAS	unvalidated
method	IsoSCM	4785	990	1009	719	1304	623	320	369	881
	Cufflink	2450	580	710	624	1454	573	397	487	2447
	Scripture	1827	190	522	229	962	265	323	296	1668
		Validation of method-specific ends								
	IsoSCM	1284	395	453	265	506	325	196	185	506
	Cufflink	171	96	179	183	229	215	253	329	2062
	Scripture	165	50	208	98	299	153	258	240	1491

3.4.7 Enrichment of PAS features at predicted 3' ends

To gain further insight into the resolution of the predictions made by each method, we investigated positional enrichments of PAS signals, 3'-seq reads, and conservation profiles around predicted 3' termini. Considering all predictions of each method in aggregate, we see PAS signals enriched ∼21-nt upstream, 3'-seq tags at position 0, and a conservation profile characteristic of polyadenylation sites for all three methods (Figure 22A-C). However, when we compare the relative frequency of evidence at these positions, we see that ends predicted by IsoSCM are supported 22.6%–23.1% more frequently by 3'-seq (Figure 22A) and 19.9%–20.4% more frequently by a PAS (Figure 22B) than either Cufflinks or Scripture. The 3' ends annotated by IsoSCM also exhibited greater overall conservation, as assessed by PhastCons (Figure 22C).

Close inspection of these analyses reveal differential biases of the methods. In particular, consideration of the tail of "imprecise" PAS and 3'-seq predictions showed that they are biased to be upstream of annotations provided by Scripture, while they are more likely to be downstream from Cufflinks and IsoSCM predictions (Figure 22A,B). This suggests that Scripture has some tendency to overextend transcript models, whereas IsoSCM and Cufflinks are slightly more likely to truncate 3' UTRs. The overextension bias observed for Scripture is a known consequence of the scan statistic used by this method to define the boundaries of transcribed regions [57]. We might expect a bias toward slight truncation using IsoSCM and Cufflinks if the regions directly upstream of the polyadenylation site are underrepresented in input reads, perhaps due to a cloning, sequencing, or mapping bias. Nevertheless, the robust

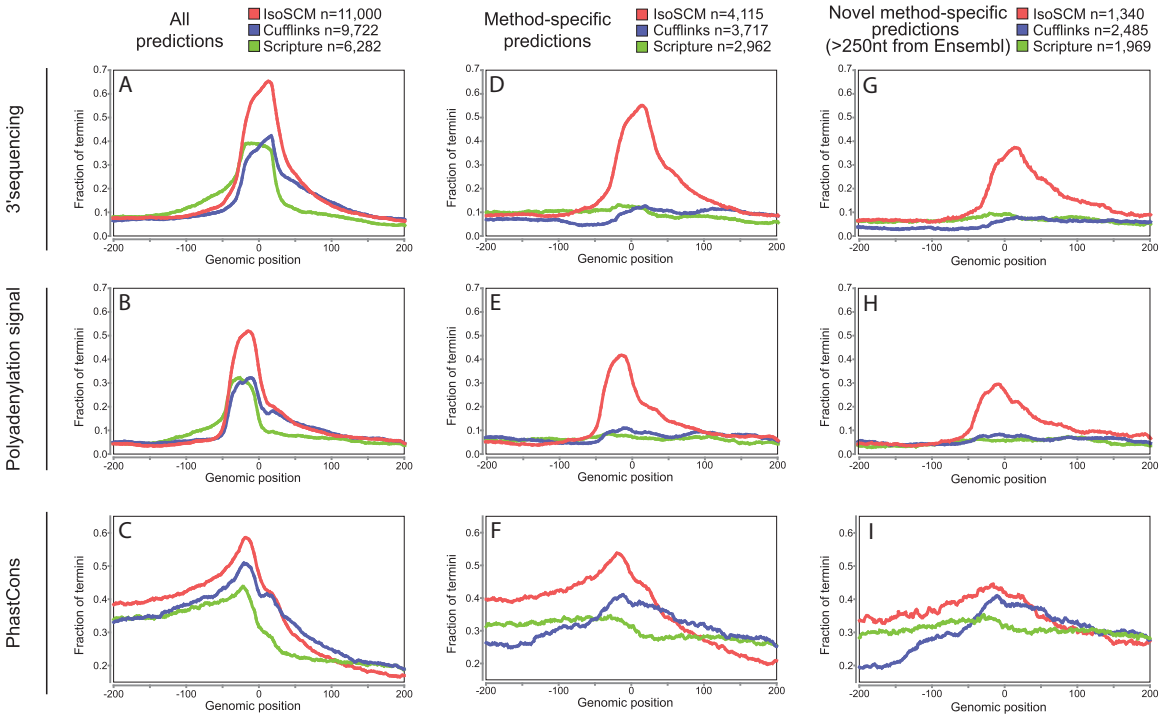


Figure 22: (A-C) Analysis of all 3' ends reported by each method shows the frequency of 3'-seq evidence (A), canonical PAS (B), and genomic conservation (C) relative to predicted termini. The fraction of predicted 3' end sites with the indicated type of support within 20 nt is shown at each position relative to the predicted boundary. The peak position of the PAS and 3'-seq tags are at their characteristic locations, about -21 and 0, respectively, while average PhastCons show characteristic peaked conservation at -21. All of these features are most robust for IsoSCM termini. (D-F) Positional enrichment of features examined for all method-specific ends, defined as termini that are >100 nt away from those generated by the other methods. Here, IsoSCM is the only method with substantial enrichment of 3'-seq evidence (D) and PAS (E) in the appropriate locations. The IsoSCM-specific annotations also exhibit the most robust conservation signature at these termini (F). (G-I) Positional enrichment of features examined for all method-specific ends that are "substantially novel", defined as being located >250 nt from Ensembl termini and >100 nt away from other method predictions. Again, IsoSCM is the only method that displays the expected enrichment of polyadenylation site features for this subset of predicted 3' ends.

peaks at positions -21 and 0, for PAS and 3'-seq evidence, respectively, indicate that IsoSCM accurately localizes polyadenylation sites in the majority of its predictions, and that its performance is substantially higher than Cufflinks or Scripture.

We next judged the qualities of the substantial populations of 3' end annotations that were specific to each method. Of the 11000 annotations of 3' ends, IsoSCM identified 4115 termini that were ¿100 nt away from annotations made by either Cufflinks or Scripture, and these annotations exhibit enrichment for polyadenylation site features that is comparable to the aggregate set of predictions for the three methods (Figure 22D-F). In contrast, analysis of 3' end annotations specific to Cufflinks (n = 3717) or Scripture (n = 2962) outputs show little or no positional enrichment of 3'-seq tags (Figure 22D) or PAS (Figure 22E) at the appropriate locations. Therefore, IsoSCM identifies thousands of sites that in aggregate exhibit the expected features of 3' ends, and are not reported by either Cufflinks or Scripture. Reciprocally, the ends that are uniquely provided by the other methods are not supported by similar evidence.

Finally, we were interested to assess the quality of substantially novel, method-specific termini. As mentioned, all methods report a population of ends that localize near Ensembl termini, but do not identify cleavage sites precisely (Figure 18C). We found that positional assessments of bulk novel ends was compromised by inclusion of these "Ensembl-imprecise" termini, which create offset peaks of evidence at the boundary of the window utilized (Figure 21B-D). To provide clarity to these comparisons, we differentiated the population of "substantially novel" ends from imprecise annotations, by investigating predictions located ¿250 nt from Ensembl annotations.

At this distance, IsoSCM identified 2472 termini, compared with 3513 by Cufflinks and 2367 by Scripture (Figure 21D). To highlight the differences between methods, we focused on "substantially novel" annotations that were specific to each method. This yielded sets of IsoSCM-specific (n = 1340), Cufflinks-specific (n = 2485), and Scripture-specific (n = 1969) ends that were ¿100 nt from each other. Analysis of these novel 3' termini showed that only the IsoSCM-specific ends were enriched for 3'-seq tags (Figure 22G) and PAS (Figure 22H), whereas the other sets of method-specific ends were not distinguishable from background. As well, the IsoSCM-specific novel ends exhibited the highest level of genomic conservation (Figure 22I).

3.4.8 Robust performance of IsoSCM across data sets

To assess if the performance differences we observed are representative, we used the framework for evaluating transcript assemblies from the simulation experiment to compare the performance of each method for 26 other mouse tissue RNA-seq data sets [86]. To provide a compact visualization of the evaluation, we used evidence of Ensembl terminal exons, 3'-seq, and PAS within ±20 nt to define sets of true and false predicted 3' ends and plotted the differences in TPR and PPV obtained using IsoSCM and Cufflinks or Scripture in Figure 23A.

While there is sample-to-sample variability in these estimated performance metrics, IsoSCM consistently achieved a TPR that is at least 12.2% higher than the other methods, without compromising the PPV. Additional metrics of predictive performance for each method, including evaluations at both the 20-nt and 100-nt resolution, are provided in Tables 2 and 3. Concrete examples illustrating how IsoSCM

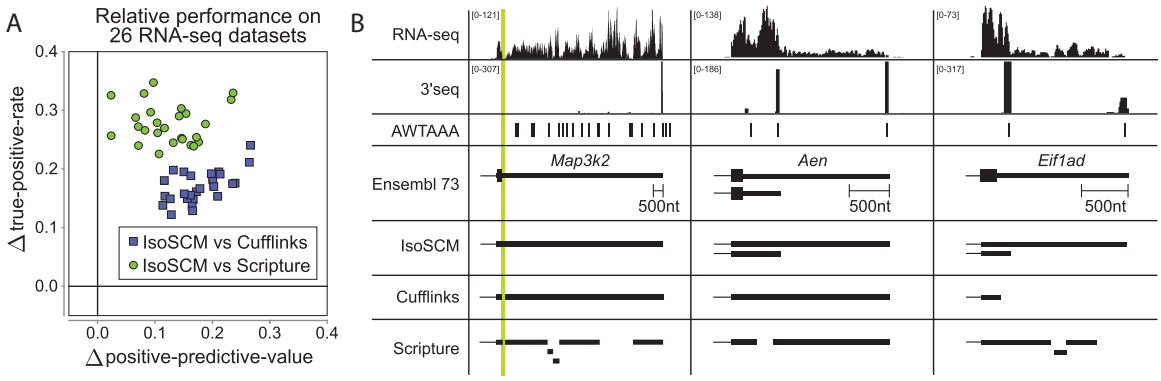


Figure 23: A) We used 26 RNA-seq data sets from nine mouse tissues to assess the ability of each method to correctly annotate 3' terminal exons. For this analysis, we estimated the true positive rate and positive predictive value of each method by its ability to recapitulate Ensembl reference termini, or otherwise be supported by 3'-seq and/or PAS (as described in Fig. 3). The sensitivity and specificity of Cufflinks and Scripture was compared with IsoSCM by subtracting the value of each metric from the value obtained using IsoSCM, such that each point represents the relative performance of a pair of methods on a single sample. Each comparison involved evaluation of 9237-22,955 3' exon annotations (median 18,737), depending on the number expressed in each sample. For every data set comparison, the TPR and PPV of IsoSCM annotations exceed those of Cufflinks or Scripture. B) The genes Map3k2, Aen, and Eif1ad illustrate scenarios where Cufflinks and Scripture assemble gene models that are less optimal than those reported by IsoSCM. Map3k2 is a gene where a gap in RNA-seq coverage fragments the 3' UTR of models reported by Cufflinks and Scripture, due to a repetitive element (green bar). IsoSCM is able to scaffold the complete 3' UTR together. Aen and Eif1ad are examples of genes whose tandem polyadenylation sites are missed by Cufflinks and Scripture, but are captured by IsoSCM. IsoSCM correctly annotates the short and long 3' UTR isoforms of both genes by identifying a change point in the level of RNA-seq coverage. In these examples, Cufflinks and Scripture annotate at most one isoform correctly, and sometimes neither. AWTAAA track indicates genomic matches to the two most common PAS, AATAAA and ATTAAA; note that many such instances are not actually functional PAS.

reduces the number of incorrectly truncated 3' UTRs and identifies additional 3' UTR isoforms that are missed by Cufflinks and Scripture are given in Figure 23B.

Table 2: Estimated TPR, FPR, PPV and NPV are reported for each method for each of the 26 RNA-seq samples by assessing assembled models with Ensembl 73 reference transcript models, PAS, and 3'seq evidence using 20nt criteria.

sample_id	TPR			FPR			PPV			NPV		
	Iso.	Cuff.	Scrip.	Iso.	Cuff.	Scrip.	Iso.	Cuff.	Scrip.	Iso.	Cuff.	Scrip.
SRR594393	0.3896	0.2284	0.1437	0.1402	0.1751	0.1118	0.7365	0.5675	0.5637	0.5835	0.5153	0.5077
SRR594394	0.4396	0.2445	0.1629	0.1204	0.1796	0.1087	0.7859	0.5779	0.6012	0.6095	0.5192	0.5143
SRR594395	0.4265	0.2314	0.1389	0.1388	0.1498	0.062	0.7557	0.6085	0.6925	0.5988	0.5237	0.5198
SRR594396	0.4041	0.2228	0.1593	0.1492	0.1957	0.1057	0.7315	0.534	0.6028	0.5866	0.507	0.5139
SRR594397	0.429	0.2624	0.175	0.1251	0.1759	0.1144	0.7753	0.6002	0.6062	0.6036	0.5262	0.5162
SRR594398	0.406	0.208	0.0773	0.1508	0.1386	0.0414	0.7305	0.6017	0.6527	0.5869	0.5194	0.5079
SRR594399	0.4115	0.2636	0.1709	0.1338	0.1826	0.1164	0.7558	0.5923	0.5964	0.5939	0.5245	0.5143
SRR594400	0.4626	0.2223	0.1444	0.0945	0.1701	0.0962	0.8312	0.5679	0.6016	0.6263	0.5149	0.5124
SRR594401	0.3864	0.1944	0.1608	0.1831	0.2078	0.1191	0.6796	0.4846	0.5757	0.5698	0.4945	0.5108
SRR594402	0.3773	0.2014	0.0985	0.1426	0.2105	0.0591	0.7267	0.4903	0.6262	0.578	0.4958	0.5094
SRR594403	0.4584	0.3089	0.164	0.1064	0.1605	0.0844	0.8125	0.6593	0.6615	0.6214	0.5471	0.5214
SRR594404	0.4182	0.2482	0.115	0.1404	0.2038	0.0749	0.7498	0.5505	0.6071	0.595	0.5129	0.5097
SRR594405	0.3976	0.2443	0.1279	0.1394	0.2132	0.0762	0.7415	0.5354	0.6281	0.5868	0.5086	0.5129
SRR594406	0.4196	0.283	0.1295	0.1298	0.1871	0.0779	0.7647	0.6033	0.6257	0.5986	0.53	0.5131
SRR594407	0.3909	0.2622	0.1293	0.134	0.1882	0.0717	0.7459	0.5835	0.6445	0.5857	0.5224	0.5145
SRR594408	0.376	0.2011	0.1102	0.161	0.2289	0.0674	0.7013	0.4691	0.6219	0.572	0.4897	0.5103
SRR594409	0.374	0.2187	0.1341	0.1999	0.2248	0.0957	0.6528	0.4943	0.5847	0.5598	0.4968	0.5096
SRR594410	0.3623	0.2244	0.1056	0.1537	0.1551	0.0493	0.7035	0.593	0.683	0.5686	0.5197	0.5136
SRR594411	0.4824	0.294	0.1351	0.1008	0.1465	0.0493	0.828	0.6687	0.7338	0.6333	0.5458	0.5221
SRR594412	0.4487	0.3266	0.152	0.1024	0.1477	0.0577	0.815	0.6897	0.7258	0.6182	0.5573	0.525
SRR594413	0.4089	0.2553	0.1701	0.1435	0.1527	0.1255	0.7415	0.6273	0.5771	0.59	0.5305	0.5114
SRR594414	0.398	0.2489	0.0724	0.151	0.1652	0.0304	0.7262	0.6027	0.7059	0.5836	0.5248	0.5095
SRR594415	0.4032	0.212	0.1507	0.1592	0.2068	0.1121	0.718	0.5077	0.575	0.5834	0.5002	0.5097
SRR594416	0.4354	0.2551	0.1057	0.1214	0.1265	0.087	0.7828	0.6697	0.5498	0.6075	0.5384	0.5038
SRR594417	0.4023	0.1909	0.1515	0.1457	0.2131	0.1057	0.7352	0.474	0.5905	0.587	0.4916	0.5117
SRR594418	0.4057	0.2481	0.1335	0.1543	0.1823	0.0699	0.7255	0.5777	0.6577	0.586	0.5196	0.5163

Table 3: Estimated TPR, FPR, PPV and NPV are reported for each method for each of the 26 RNA-seq samples by assessing assembled models with Ensembl 73 reference transcript models, PAS, and 3'seq evidence using 100nt criteria.

sample_id	TPR			FPR			PPV			NPV		
	Iso.	Cuff.	Scrip.	Iso.	Cuff.	Scrip.	Iso.	Cuff.	Scrip.	Iso.	Cuff.	Scrip.
SRR594393	0.5276	0.388	0.2438	0.0318	0.0686	0.0631	0.946	0.8568	0.8033	0.6597	0.5901	0.5396
SRR594394	0.5666	0.404	0.2623	0.0268	0.0748	0.0593	0.9572	0.8512	0.824	0.6795	0.5945	0.5463
SRR594395	0.5854	0.3972	0.2101	0.0295	0.0517	0.0244	0.9547	0.8909	0.9014	0.6878	0.5969	0.5376
SRR594396	0.556	0.399	0.2601	0.0337	0.0732	0.0555	0.9462	0.8529	0.833	0.6717	0.5917	0.5454
SRR594397	0.5665	0.4308	0.2869	0.0273	0.0633	0.0565	0.9567	0.8785	0.8436	0.6786	0.6075	0.5545
SRR594398	0.564	0.3521	0.12	0.033	0.0543	0.0205	0.9477	0.8729	0.861	0.677	0.5797	0.5126
SRR594399	0.5575	0.4394	0.2813	0.0264	0.0632	0.0585	0.9573	0.8809	0.8365	0.6741	0.611	0.5519
SRR594400	0.5724	0.3686	0.2407	0.0181	0.0773	0.0522	0.971	0.8349	0.8303	0.6839	0.5793	0.5405
SRR594401	0.5443	0.3517	0.2678	0.0598	0.0928	0.0627	0.9065	0.8016	0.8198	0.6594	0.5677	0.5457
SRR594402	0.5216	0.3746	0.1588	0.0355	0.0865	0.0347	0.9396	0.8209	0.8289	0.6558	0.5798	0.5202
SRR594403	0.581	0.4658	0.2445	0.0254	0.0581	0.0438	0.9604	0.8945	0.8552	0.6873	0.6249	0.5446
SRR594404	0.5692	0.4294	0.1905	0.0316	0.0738	0.0457	0.9502	0.8606	0.8156	0.6795	0.6049	0.5265
SRR594405	0.5414	0.426	0.1975	0.038	0.0781	0.0399	0.9379	0.8527	0.84	0.664	0.6021	0.5299
SRR594406	0.5695	0.4599	0.1993	0.0273	0.066	0.0433	0.9567	0.8807	0.8297	0.681	0.6204	0.5303
SRR594407	0.5353	0.4409	0.2049	0.0268	0.0619	0.0353	0.955	0.8831	0.8604	0.6636	0.6124	0.5333
SRR594408	0.5413	0.374	0.1728	0.0396	0.1013	0.0382	0.9355	0.7968	0.8279	0.6635	0.5748	0.5227
SRR594409	0.5522	0.3892	0.2093	0.072	0.0923	0.0522	0.8908	0.8176	0.81	0.6609	0.5829	0.5299
SRR594410	0.52	0.3809	0.156	0.033	0.0619	0.0269	0.9435	0.867	0.8601	0.6555	0.5886	0.5212
SRR594411	0.6152	0.4386	0.1913	0.0205	0.0646	0.024	0.9695	0.8782	0.8944	0.7058	0.611	0.5322
SRR594412	0.5852	0.4852	0.2198	0.0171	0.0534	0.023	0.9733	0.9065	0.9106	0.6897	0.633	0.5401
SRR594413	0.5326	0.3965	0.2701	0.0583	0.0643	0.0621	0.9064	0.8673	0.8219	0.6551	0.5938	0.5478
SRR594414	0.5529	0.4169	0.1103	0.0354	0.0579	0.0171	0.9433	0.8848	0.873	0.6692	0.6023	0.5088
SRR594415	0.5544	0.3813	0.2557	0.0384	0.0894	0.0584	0.9387	0.819	0.8227	0.6706	0.5813	0.544
SRR594416	0.5632	0.3802	0.1755	0.0399	0.0517	0.0442	0.9375	0.8865	0.8084	0.6742	0.5902	0.5218
SRR594417	0.5435	0.3535	0.249	0.0341	0.1013	0.0568	0.9442	0.7874	0.8231	0.6659	0.567	0.5419
SRR594418	0.5554	0.4149	0.2063	0.0459	0.0681	0.0332	0.9279	0.8664	0.8685	0.6685	0.5996	0.5338

3.4.9 Application of IsoSCM to identify tissue-differential tandem APA events

To illustrate a practical application, we next applied change-point inference to examine differential usage of 3' UTR isoforms between different tissues. In the previous section, we demonstrate that the IsoSCM segmentation strategy is effective for annotating terminal transcript boundaries within an individual sample; however, the quantification of differential isoform usage requires an accounting of all transcripts expressed among the samples being compared. When generating a single annotation that is representative of two samples, one could attempt to merge annotations assembled independently from each sample. However, it is not obvious what the optimal strategy is for combining annotations from separate samples, while preserving alternative events. For example, the cuffmerge program, available as part of the Cufflinks suite, merges transcript models that share overlapping and compatible chains of introns, and during this process discards shorter terminal exon annotations, reporting only the longest 3' exon. While such an approach could be used to identify the longest 3' UTR isoform, it discards potential alternative events, and is thus inappropriate for assessing tissue-differential tandem APA events.

Conveniently, the change-point detection framework described above can be extended naturally to the assembly of tandem 3' UTR boundaries from multiple samples simultaneously. To do this, we define Pjoint, the joint likelihood of k independent samples,

$$P_{joint} = \prod_{i=1}^{k} P_i(t, s) \tag{6}$$

where P_i is the marginal likelihood of the ith sample. Thus, by replacing P in Equations 2 and 3 with P_{joint}, the segmentation procedure will identify the configuration of change points that maximize the joint marginal likelihood of all the samples simultaneously. Importantly, this modification enables identification of polyadenylation sites that are specific to one condition, in addition to sites that are common among the samples being compared.

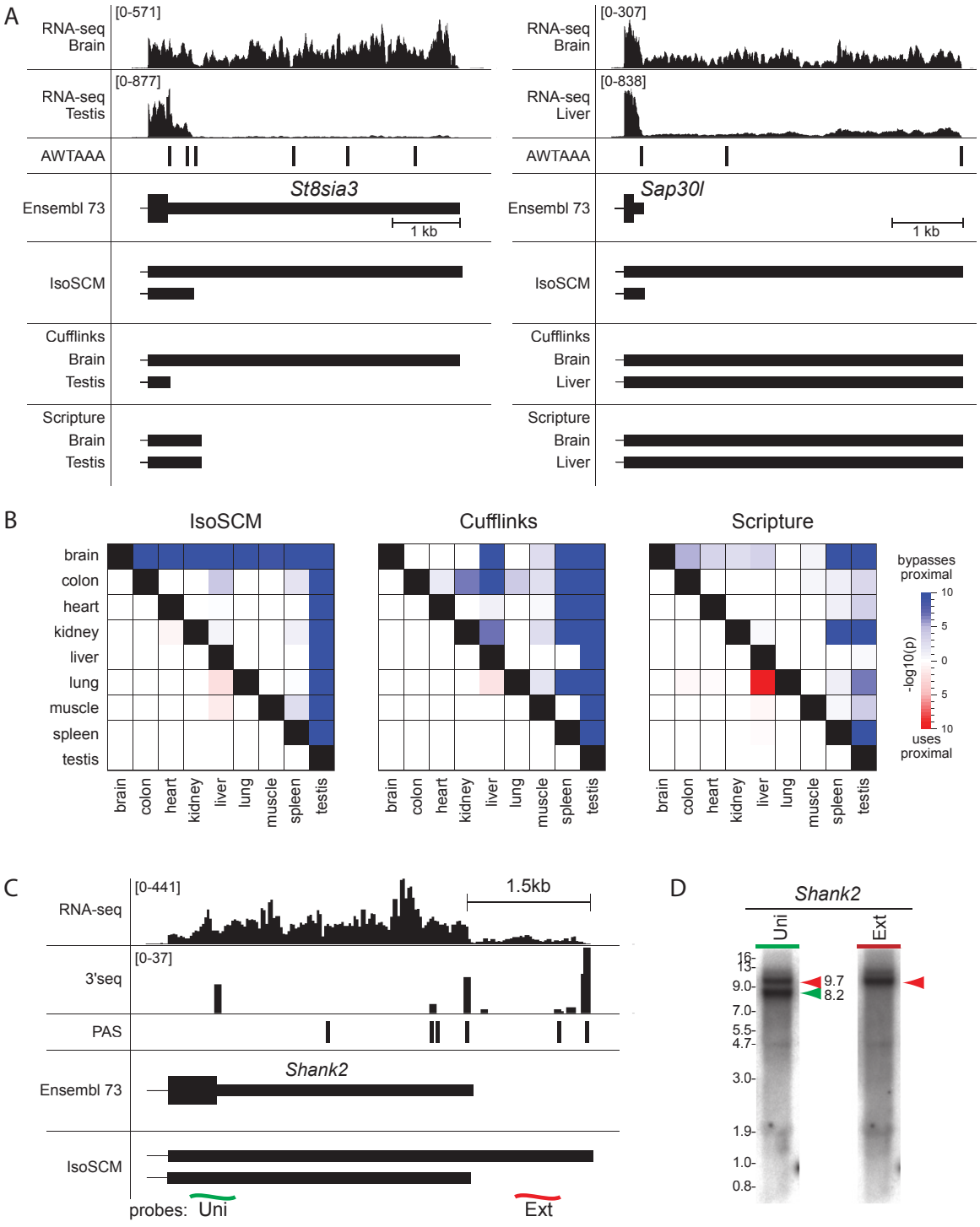
Previously, we and others observed that the nervous system accumulates many transcripts with longer 3' UTR isoforms than in other tissues, whereas the testis expresses many transcripts with relatively shorter 3' UTR isoforms ([154]; [58]; [28]; [128]; [141]; [91]). Using the IsoSCM framework, we reexamined these observations by analyzing differential polyadenylation site usage in nine mouse tissues. By performing the joint segmentation for each pair of tissues, we identified a set of change points representing tandem polyadenylation events. For each of these events, we estimated a polyadenylation-site-usage index in each condition, representing the relative frequency with which a particular polyadenylation site is used. By calculating the difference in this index between a pair of conditions, we identified polyadenylation events with differential usage patterns among tissues. St8sia3 and Sap30l typify transcripts IsoSCM identifies as being differentially polyadenylated between a pair of tissues, and are illustrated in Figure 24A. By identifying global differences in the distribution of the polyadenylation-site-usage statistic at proximal sites between tissue pairs, we

assessed whether there were systematic patterns of altered 3' UTR length. Indeed, we observed clear signatures in which the termini annotated by IsoSCM were broadly extended in brain relative to all other tissues, whereas they were generally shorter in testis relative to all other tissues (Figure 24B).

To emphasize the importance of using accurate terminal exon annotations when quantifying tandem terminal exon expression patterns, we repeated the analysis using the transcript assemblies of Cufflinks and Scripture in place of the models generated by IsoSCM. As these methods do not provide a means to assemble transcript models considering two samples simultaneously, we used the union of terminal exon models from tissues being compared to define the set of tandem polyadenylation events identified by these methods. As shown in Figure 24B, there is only modest agreement between patterns identified using models built by Cufflinks and Scripture and IsoSCM. Enhanced usage of proximal polyadenylation sites in the testis was reliably recovered using Cufflinks. This is likely because the abrupt truncation events observed in testis can be captured by the Cufflinks assembly procedure, as is the case for St8sia3 (Figure 24A). However, neither Cufflinks nor Scripture were able to assemble models that consistently captured the pattern of increased abundance of extended 3' UTRs when comparing brain with other tissues (Figure 24B).

Cufflinks and Scripture seek only to annotate the longest terminal isoform that is consistent with the reads, and in brain we observe many cases where both isoforms are expressed and only their relative usage changes between tissues. In such cases, it is expected that the short 3' UTR isoform will not be captured by these methods. This is illustrated by the gene Sap30l in Figure 24A, for which both a short and

Figure 24: A) Instances of tissue-differential APA identified by IsoSCM are illustrated St8sia3 and Sap30l. IsoSCM models generated by the joint segmentation procedure are compared with models assembled by Cufflinks and Scripture analyzing each tissue independently. In these examples, Cufflinks is able to recover switches in tandem polyadenylation site usage for St8sia3 when comparing independent assemblies of the data sets, but does not detect the more nuanced difference with Sap30l as the short was not assembled. Scripture does not generate alternative models in either case. Since IsoSCM uses change-point detection for 3' UTR boundary annotation, IsoSCM assembles both isoforms correctly, and is able to consistently capture these patterns. The AWTAAA track indicates genomic matches to the two most common PAS, AATAAA and ATTAAA; note that many such instances are not actually functional PAS. B) Exon models built using IsoSCM were used to identify global patterns in differential 3' UTR usage between nine mouse tissues. Events arising from overlapping tandem 3' exons were used to estimate the relative usage of 3' ends in each tissue. For all genes that are expressed in a pair of tissues, the difference in relative usage of a polyadenylation site is computed. To assess whether there are global trends toward 3' UTR lengthening (or shortening) between tissues, the observed distribution of differences in 3' end usage is compared with a null model that lengthening (or shortening) events are equally likely to occur in either sample. The 36 pairwise comparisons are represented in a matrix, with the tissues being compared labeled at the left of each row, and the bottom of each column. A cell shaded blue indicates that the tissue labeled at the left of the row tends to have lower usage of proximal 3' ends than the tissue labeled at the bottom of the column, while red shading indicates the opposite pattern. Here, we see that the brain tends to use polyadenylation sites that are more distal, while the testis tends to use more proximal polyadenylation sites. C) Neural APA event detected for Shank2; the alternative 3' UTRs annotated by IsoSCM are supported by 3'-seq data. Probes were designed against a universal region present in both isoforms (green), and a region exclusive to the extended isoform (red). D) Sequential Northern blotting shows that the universal probe hybridizes to two bands with estimated lengths of 8.2 and 9.7 kb, while the extension probe hybridizes only with the 9.7-kb band. These bands are consistent with the 3' exon annotations assembled by IsoSCM, and likely correspond to the transcript ENSMUST00000105900 (8169-nt long) as well as an isoform bearing a 1.5-kb 3' UTR extension.


A
RNA-seq Brain
[0-571]
RNA-seq Testis
[0-877]
AWTAAA
St8sia3
Ensembl 73
1 kb
IsoSCM
Cufflinks
Brain
Testis
Scripture
Brain
Testis
RNA-seq Brain
[0-307]
RNA-seq Liver
[0-838]
AWTAAA
Sap30l
Ensembl 73
1 kb
IsoSCM
Cufflinks
Brain
Liver
Scripture
Brain
Liver
B
IsoSCM
Cufflinks
Scripture
brain
colon
heart
kidney
liver
lung
muscle
spleen
testis
bypasses proximal
-log10(p)
uses proximal
C
RNA-seq
[0-441]
1.5kb
3'seq
[0-37]
PAS
Shank2
Ensembl 73
IsoSCM
probes: Uni
Ext
D
Shank2
Uni
Ext
16
13
9.0
7.0
5.5
4.7
3.0
1.9
1.0
0.8
9.7
8.2

long isoform are expressed in brain and liver, but there is a higher proportion of isoforms bearing the extended 3' UTR in the brain. In this case, using Cufflinks or Scripture this alternative event is not identified, since neither method annotates the short isoform correctly.

We previously provided extensive molecular validation of neural 3' UTR lengthening events of a similar quality to those identified by the IsoSCM pipeline, although those APA events were previously cataloged by a combination of visual annotation and partially automated procedures [128, 91]. Along these lines, we tested a 3' UTR extension for Shank2 identified by IsoSCM (Figure 24C) by Northern blot. We designed double stranded DNA probes that either hybridize to the region common to both transcripts ("uni"), or that exclusively recognizes the extended 3' UTR isoform ("ext"). Based on the predicted size of the full-length transcript differing only in the 3' boundary of the terminal exon, we expect full-length transcripts of length 8.2 and 9.7 kb. Northern blotting showed that the common probe hybridize to two dominant bands that are consistent with these lengths, while extension-specific probe hybridizes exclusively to the longer RNA species (Figure 24D), confirming our inference of tandem terminal exon boundaries from stepped patterns of RNA-seq coverage. Overall, these analyses are consistent with previously described patterns of tissue-specific APA, and suggest that the methodology encapsulated by IsoSCM can be readily applied in other settings to gain new insights into alternative 3' UTR isoform regulation.

3.5 Conclusions

The inference of transcript structures from short RNA-seq reads is a complex problem, particularly for the identification of 3' UTR boundaries. As we have illustrated, confounding factors such as the presence of repetitive sequences, overlapping coexpressed isoforms, and nonuniform coverage patterns make the identification of 3' UTR boundaries from RNA-seq data a nontrivial problem. These issues are exacerbated by the absence of a framework to incorporate observed patterns of read depth into the transcript assembly process. To address these challenges and the limitations of existing tools, we developed IsoSCM. By benchmarking IsoSCM against state of the art methods for transcript assembly on simulated and experimentally generated RNA-seq data sets, we have demonstrated by several measures that the performance of IsoSCM is superior to existing tools for 3' UTR reconstruction. Although evaluations indicate substantial benefits of change-point inference, we reiterate that the advantages of IsoSCM are manifest where sufficient levels of isoform expression exist, and that it does not correctly identify 3' UTR isoforms if the level of coverage over exonic segments is not proportional to isoform abundance. These are indeed general challenges for ab initio transcript assemblers. Nevertheless, we show that IsoSCM consistently identifies thousands of functionally supported 3' termini in excess of Cufflinks and Scripture, and reduces the number of false predictions reported.

The improvement in sensitivity and accuracy of reconstructed transcript models gained by using IsoSCM has implications beyond transcriptome annotation. A major application of RNA-seq technologies is for the inference of alternative isoform reg-

ulation, and these inferences require accurate gene models. By analyzing a panel of RNA-seq samples from different tissues, we demonstrate how previously observed patterns of APA are faithfully recapitulated by IsoSCM. Moreover, by extending the constrained change-point detection procedure we developed for single-sample annotation to the joint analysis of two samples, we obtain an effective method to detect and annotate differential APA in an ab initio setting. Notably, these patterns are not recovered using existing tools for transcript reconstruction, because these methods are by design not expressive enough to annotate tandem 3' UTR isoforms.

While this work was in revision, Wagner and colleagues reported their algorithm DaPars and used it to identify genes with altered polyadenylation patterns following CFIm25 knockdown and in glioblastoma [81]. While their work also exploits change points in RNA-seq data, it provides only a subset of the functionalities of IsoSCM, making direct comparisons of the methods difficult. A significant advantage of IsoSCM is that it operates in an ab initio setting, and its accuracy does not depend on the quality and completeness of a reference annotation. Even for the well-annotated mouse transcriptome, dominant isoforms are missing (i.e., the extended isoform of Sap30l) (Figure 24A), and could only be analyzed using ab initio methods such as IsoSCM. For many species, especially nonmodel organisms, 3' UTR models are incomplete or absent, precluding the application of DaPars to these transcriptomes. The generality of IsoSCM will enable change-point analysis to be applied to APA in a wider scope of problems.

While IsoSCM utilizes only RNA-seq to identify 3' ends, there also exist specialized protocols for high-throughput mapping of 3' UTR boundaries by direct cloning

and sequencing of 3' ends. Clearly such methods provide greater and more direct information on 3' ends. However, as these strategies are technically more complicated than conventional RNA-seq protocols, the available 3'-seq data are only a small fraction of the aggregate RNA-seq data that are available for diverse organisms. We present IsoSCM as a methodological advance for ab initio 3' end identification that can take advantage of existing RNA-seq data and extend its analysis beyond current algorithms. The refinement of 3' UTR boundaries provided by IsoSCM provides greater accuracy and more systematic accounting of 3' end processing, and enables diverse studies of differential APA using the wealth of RNA-seq data now available.

3.6 Materials and Methods

3.6.1 RNA-seq simulations

A graphic overview of the simulation framework is given in Figure 17A. To generate a set of reference transcript models with tandem polyadenylation events we selected 14,263 nonoverlapping transcript models that contain a terminal exon at least 500-nt long from the Ensembl 73 annotation for RNA-seq simulations. Within the last exon of each transcript model, a truncated isoform was generated by uniformly sampling a position at least 150 nt from both 5' and 3' exon boundaries, and generating a new model sharing the complete upstream exon structure, while bearing a terminal exon truncated to this sampled position. Unstranded 50-nt single end reads were simulated by sampling alignment start positions uniformly across the body of the transcript until a target depth of 2000 reads/kb of exonic sequence (94,962,730 reads

total) was reached. This library was recursively subsampled to generate a test set spanning the range of depths (5, 7.5, 10, 20, 30, 40, 50, 100, 200, 400, 600, 800, 1000, 2000) reads/kb (Figure 17B).

3.6.2 Evaluation of transcript models

For each method, we assessed the quality of 3' terminal exon assemblies of annotated protein coding genes. After identifying assembled 3' terminal exon models for each method that shared a 5' splice boundary with the terminal coding exon of an expressed gene from the Ensembl 73 annotation, we assessed the extent to which that annotation is supported by various types of evidence. For each predicted 3' end we assessed whether an Ensembl 73 annotation, 3'-seq tag, or polyadenylation signal were detected within a short distance of the predicted boundary of the 3' UTR. We assessed evidence supporting assembled exon model using both a stringent (20 nt) and relaxed (100 nt) windows. Annotations from one method that were within 100 nt of an annotation made by another method were defined to be common between those methods, while annotations that were ¿100 nt from an annotation made by any other method were considered to be method-specific ends. The number of distinct 3' ends was counted, such that a 3' end that was common between two terminal exon models was only counted once.

To calculate the positional enrichment of polyadenylation signals and 3'-seq data, all annotations of terminal exons of known protein coding genes were aligned with their 3' boundary at position 0. At each position in a 200-nt window upstream of and downstream from each annotation, we calculated the fraction of sequences

that have either a PAS or 3'-seq evidence within 20 nt of that position. For the PAS we considered the two most frequent motifs, AATAAA and ATTAAA, and at the terminally aligned position of a 3'-seq read was used as the position for 3'-seq evidence.

3.6.3 Transcript assembly

Cufflinks 2.2.0 was downloaded from http://cufflinks.cbcb.umd.edu/downloads/cufflinks-2.2.0.Linux_x86_64.tar.gz, and was run with the default parameters, except that –library-type fr-firststrand was provided to indicate the strandedness of the sequencing data, and the –overlap-radius was set to 100 bp.

The scripture-beta2.jar was downloaded from ftp://ftp.broadinstitute.org/pub/-papers/lincRNA/scripture-beta2.jar. Scripture was run on each chromosome independently, with the default parameters except that 48 GB of memory was allocated, as smaller allocations resulted in "out of memory" errors. The resulting bed files from assemblies for each chromosome were concatenated to form the final assembly for each sample.

3.6.4 Data sets analyzed

Sequencing fastq files with RNA-seq (GSE41637) [86] and 3'-seq GSE30198 [28] were downloaded from GEO. Reference genome sequences were downloaded from Ensembl, ftp://ftp.ensembl.org/pub/release-73/fasta/. RNA-seq was mapped to the genome using Tophat2 with default parameters, except that an Ensembl 73 reference annotation was provided with the -G option, and the –segment-length was set to 20. Bowtie

was used to map the 3'-seq data, with the default parameters.

3.6.5 Identification of tissue differential APA events

For each pair of tissues, a joint segmentation of the two samples was computed. For each change point identified between a pair of conditions, we calculated the magnitude of change in the level of coverage in each condition. From these values a usage score (U) was calculated as $U = 1 - (cov_{dn}/cov_{up})$, where cov_{up} and cov_{dn} are point estimates of the level of read density in the segments upstream of and downstream from the change point, respectively. The differential usage between conditions A and B was calculated as the difference in the estimated usage at that change point, $U_{\Delta AB} = U_A - U_B$.

3.6.6 Northern analysis

We used PCR to amplify universal and proximal regions of the predicted isoforms of Shank2 using these primer sets.

shank2_uni; GGACCTCTTTGGCTTGAACC

shank2_uni; CTATGGCAGCCTCTGAGACC

chr7:151606740-151607287

shank2_ext2; GGGAGCAGAAGACTGAGTGG

shank2_ext2; CAGCATCATCAGGACAGTGG

chr7:151610801-151611426

Random primed radiolabeled probes were generated using these templates. RNA was isolated from mouse brains, and sequentially hybridized with extension and uni-

versal probes as described previously [91].

3.6.7 Software availability

IsoSCM transcript assembly is implemented as a standalone Java program, available at https://github.com/shenkers/isoscm. Using the "assemble" keyword IsoSCM will assemble the mapped reads in a BAM file into a splice graph, identify nested terminal exons boundaries using the constrained segmentation procedure, and report the resulting models in GTF format. Functionality to enumerate splice isoforms from the assembled splice graph can be accessed by the keyword "enumerate". Pairwise comparison of tandem isoform usage can be performed using the "compare" keyword, which reports the relative usage of change points in each sample in a tabular format. Complete documentation is available from the IsoSCM website (https://github.com/-shenkers/isoscm).

3.6.8 Acknowledgments

S.S. was supported by the Tri-Institutional Training Program in Computational Biology and Medicine. P.M. was supported by a fellowship from the Canadian Institutes of Health Research. Work in E.C.L.'s group was supported by the Burroughs Wellcome Fund and the National Institute of General Medical Sciences of the National Institutes of Health (National Institute of Neurological Disorders and Stroke: R01-NS074037 and R01-NS083833).

4 CrossBrowse: A versatile genome browser for visualizing comparative experimental data*

4.1 Attributions

Sol Shenker designed and implemented CrossBrowse, and performed analysis of CTCF syntenic binding. Jaaved Mohammed suggested interesting miRNA based on his previous analysis and provided tracks of small-RNA sequencing across species. The small RNA sequencing libraries were prepared by Alex Flynt. Piero Sanfilippo generated 3'seq and RNA-seq samples used for the example of poly-A site dynamics.

4.2 Abstract

The recent beyond-exponential growth in diverse collections of deep sequencing datasets creates enormous opportunities for discovery, concomitant with new challenges for displaying and interpreting these data. Notably, the availability of scores of whole genome sequences in multiple species clades enables comparative studies of functional elements. However, current genome browsers do not permit effective visualization of multigenome experimental data. Here, we present CrossBrowse, a standalone desktop application for displaying and browsing cross-species genomic datasets. We utilize data standards and graphic representation of popular browsers, and incorporate an intuitive graphical visualization of genome synteny that facilitates and drives human interrogation of comparative data. Our platform permits users with minimal informatics capacity to select arbitrary sets of genomes for display, upload and configure

*__S. Shenker__, and E. C. Lai. CrossBrowse: A versatile genome browser for visualizing comparative experimental data. *Submitted*.

multiple datasets, and interact with vertebrate-sized genomic datasets in real-time. We illustrate the utility of CrossBrowse with interrogation of comparative invertebrate and mammalian datasets that provide insights into diverse aspects of transcriptional and post-transcriptional regulation. Of note, we show examplars of both preservation and divergence of functional elements that cannot be inferred from sequence alignments alone. Moreover, we demonstrate how inspection of primary data using CrossBrowse exposes an artifact in a typical strategy for assigning species-specific functional elements, and drives the implementation of an improved computational strategy. We anticipate that CrossBrowse will greatly foster user-based discovery within multispecies genomic datasets, and inform their bioinformatic interpretation.

4.3 Introduction

Ongoing advances in next-generation sequencing (NGS) technologies, coupled with creative assays that interrogate diverse aspects of genomes and transcriptomes, are transforming the face of modern biology[116, 108]. Indeed, many laboratories are now heavily reliant on techniques such as whole exome and whole genome sequencing, GWAS, ChIP-seq and genomic structure analyses, long and short RNA-seq, and so forth, all of which barely existed or were non-existent a decade ago. Many expedient software packages exist to process genomic and transcriptome data, and generate summary statistics of their properties. Nevertheless, human curation (i.e., "sequence gazing") is still integral to effective and rigorous analysis. Not only is it prudent to provide reality checks regarding overall conclusions, the human eye is often instru-

mental in first recognition of novel genomic phenomena that are not appropriately handled by existing software packages. In turn, such insights can fuel subsequent computational work.

As an example from our experience, while transcripts were annotated through a standard pipeline in the modENCODE project[18], our visual recognition of long, neural 3' UTR extensions that were not assembled as continuous exons led us to perform the initial annotation of the transcriptome via end-to-end manual browsing of the Drosophila genome[128]. The broad foundation we gained by human curation was critical to develop our de novo transcriptome assembly and analysis method IsoSCM[123], which enables effective analysis of alternative polyadenylation trends from RNA-seq data. Thus, effective data visualization can facilitate human-driven identification of novel genomic patterns that would be otherwise overlooked by existing analysis tools, and guide the development of new analytical strategies.

The burgeoning amount of comparative genomic datasets creates new challenges in genomics visualization. One of the most-widely used public genomic resources is the UCSC Genome Browser, a portal for real-time user interactions with multiple genome alignments. In addition, a multitude of genomewide datasets are publicly hosted by the UCSC Genome Browser, and one can also link to private datasets. However, a clear limitation is that the UCSC platform does not permit comparison of experimental data from multiple species, nor visualization of genome rearrangements beyond linear relationships. While the UCSC comparative assembly hub introduce snake tracks to represent structural genome changes[97], substantial issues remain to navigate and interpret multi-genome datasets.

The democratization of next-generation sequencing means that it is now straightforward to generate genomewide datasets in historically non-model organisms, and public deposits of such cross-species data now present largely untapped opportunities for discovery. Moreover, the advent of CRISPR/Cas9 genome engineering now means that a much broader range of species are experimentally and genetically tractable. Thus, genomics data can not only help interpret the genomes of "classic" model organisms, but serve as the basis for experimental designs in new species. Nevertheless, while cross-species datasets provide rich resources for bioinformatics investigations, they are currently largely inaccessible to biologists and laboratories lacking significant computational expertise.

While some approaches have begun to address this problem, none of the available cross-species browsers satisfactorily deal with the challenges of current genomics data. Combo was one of the first comparative browsers[32], but it does not handle NGS data or permit more than two genomes to be displayed. Sybil[118] handles only bacterial-sized genomes, displays only annotated genes, and does not support NGS formats. mGSV[117] is a web-based browser through which one must upload data prior to visualization, and is consequently slow and not secure. In addition, it limits the user to loading only one data track, and uses a custom upload format that can render data configuration tedious. Finally, GBrowse_syn[84] requires a dedicated Linux server, and significant command line usage to modify session configuration, limiting breadth of its utilization. In addition, all of these packages share the significant limitation of showing only a fixed coarse-grain representation of synteny relationships, and thus lack dynamic visualization as the user zooms between multi-gene and the nucleotide

level resolutions. This is a signature feature of the UCSC Genome Browser and the Integrated Genome Viewer that today's biologists have come to expect. Reciprocally, the UCSC Genome Browser focuses on local, linear alignment relationships, at the expense of visualizing interrupted segments (indels) and larger genomic rearrangements. Finally, while web-based genome browsers have advantages with respect to aggregating datasets and a means to share a session among collaborators, this design can place inordinate amount of strain on server, degrading responsiveness and user experience.

We sought to address all of these limitations in CrossBrowse, our standalone desktop application for cross-species data visualization for vertebrate-sized genomes. To create a tool familiar to users of existing genome browsers, we took advantage of data standards and graphical representation established by first generation genome browsers, and combined it with an intuitive graphical representation of genome synteny relationships and multispecies experimental data. We demonstrate the utilities of CrossBrowse for analyzing transcriptional and post-transcriptional regulation, and highlight its attributes and capacities that permit bench scientists to interact with and make discoveries from cross-species genomics data.

4.4 Results and Discussion

A genome browser with adaptive granularity visualization across syntenic loci We designed a browser to meet the aforementioned challenges for multi-species visualization of aligned genomes and experimental datasets. Here, we describe key features of

CrossBrowse, especially with regard to issues not satisfactorily handled by existing browsers.

CrossBrowse generates the synteny visualization using whole genome sequence alignments. These allow CrossBrowse to translate coordinate positions between reference genomes, the same functionality provided by the liftOver tool from the UCSC genome browser. CrossBrowse reads whole genome alignments from CHAIN-format files that can be downloaded from the UCSC genome browser. The window is composed of multiple stacked "traditional" genome browsers, which are connected by graphics that illustrate syntenic relationships. These are represented as colored "blocks" that indicate boundaries of homologous regions. Unaligned sequences are left white, but the coloring scheme easily allows one to identify when such regions reside within a larger syntenic region. When sequences reside on opposite strands with respect to the reference genome, the block will reflect this inversion.

Since there can be extensive homology between related genomes, CrossBrowse divides each longer block of homology into a series of smaller blocks, so that the synteny visualization adapts to the viewed coordinates. The synteny visualization reacts to user interaction as the mouse moves over the view area, allowing for the rapid assessment of information at orthologous positions, even in the context of structural genome changes. Syntenic regions are shaded to facilitate the location of orthologous positions across genomes. The coloring scheme is stable with respect to the displayed coordinates, thereby maximizing visual continuity as the view is adjusted. We note that tools such as Sybil and mGSV and Gbrowse_syn generate conceptually similar graphical motifs to visualize synteny. However, only CrossBrowse can represent syn-

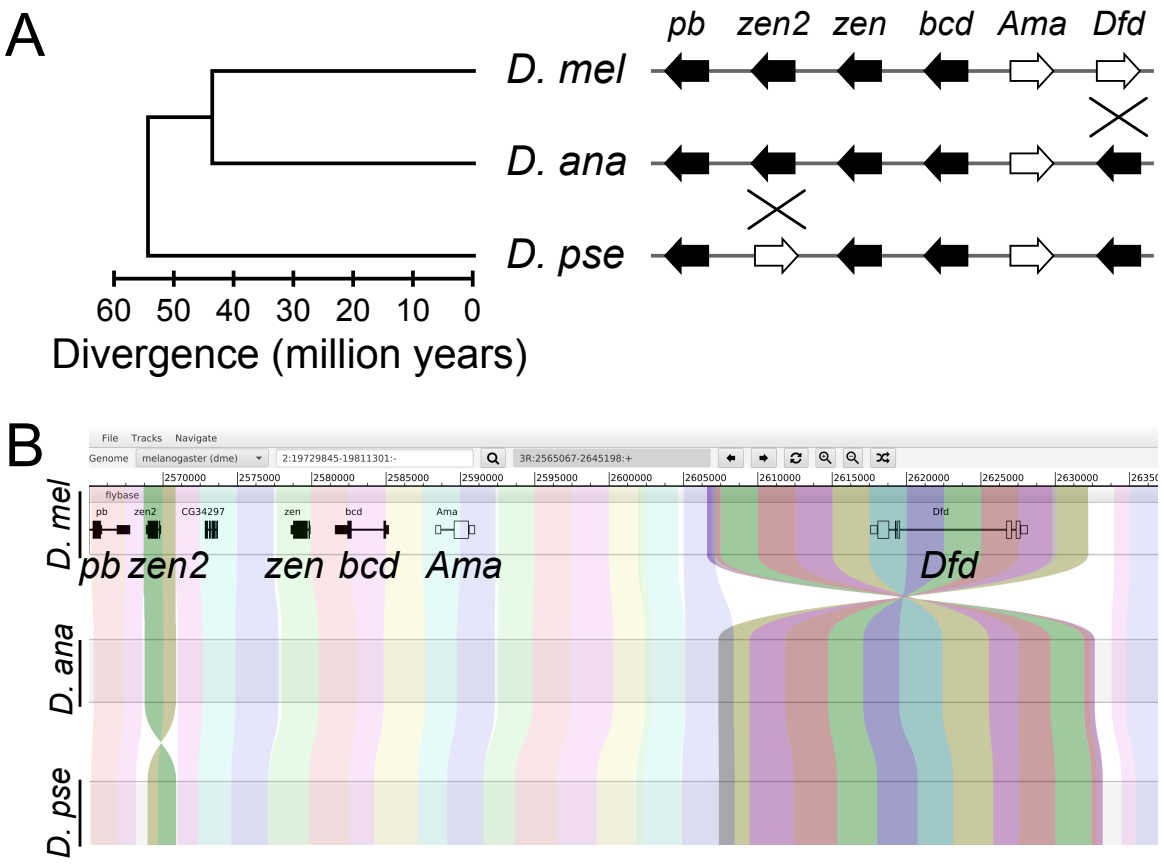


Figure 25: (A) Graphical representation of structural changes within the Antennapedia-Complex locus across the radiation of D. melanogaster, D. ananassae, and D. pseudoobscura. The separation of D. pseudoobscura from D. melanogaster and D. ananassae is accompanied by inversion of zen2. Likewise, the speciation of D. melanogaster and D. pseudoobscura is accompanied by an additional inversion of the Dfd locus. (B) A CrossBrowse screenshot illustrates how these structural rearrangements occurring at the Antennapedia-Complex are readily visualized using CrossBrowse. A track displaying gene structure is displayed in the melanogaster genome. The location of syntenic genome sequences are indicated by ribbons that span the individual genome views. Structural inversions relative to neighboring genomes are represented as twists in the ribbon. Darker shading is used to highlight the location of the structural changes.

teny with a granularity that adapts to the user's view. Thus, CrossBrowse can be readily be applied to investigate features ranging from megabase windows all the way down to single nucleotide resolution.

We illustrate the utility of synteny visualization using homeobox gene re-arrangements observed in the Drosophila Antennapedia-Complex. Hox genes are essential to establish segmental identities during development and exhibit striking correspondence between their relative expression domains along the anterior-posterior axis and their linear order along the genome. The linear order of Hox complexes is generally conserved, but at least seven structural re-arrangements are observed across the Drosophilid phylogeny[99]. Such re-arrangements are not evident when using popular interfaces such as the UCSC genome browser or IGV, and are often rendered manually (Figure 25A). However, using CrossBrowse, one can easily zoom in and out of this complex genomic region and observe the phylogenetic relationships of their changing gene orientations in a graphically intuitive manner (Figure 25B).

4.4.1 An intuitive and facile interface for concurrent exploration of multiple genomes

CrossBrowse implements navigation features and graphical standards established by existing browsers, thereby providing a familiar user environment. As seen in the example CrossBrowse session (Figure 26), the top of the interface displays Session, Track, and Navigation menus. The Session menu provides an interface for the user to configure which genomes are displayed in the browser, and pairwise alignments between them. The Track menu is used to load tracks in the browser, and adjust how

they are displayed. The Track View is partitioned into separated stacks of Tracks when the CrossBrowse session is configured to display multiple genomes. The Navigation menu provides access to functionality for selecting views of syntenic regions between reference sequences, access navigation history, refresh tracks, and adjust the zoom level.

We incorporated features that facilitate genome browsing in the context of multiple genomes. For example, coordinate conversion is tightly integrated with the browser, and is accessed through the toolbar by selecting Navigate?liftOver. The source genome for the liftOver can be selected from the drop down menu at the top left, and the list of target regions identified are displayed in the list below. Selecting one of the target regions and pressing "liftOver" will set the coordinates in the target genome based on the region of synteny with the window displayed in the source genome (Figure 26).

One issue with CHAIN files is that the existing tools for coordinate conversion (which are necessary to build the overlay visualization) are designed for batch queries. Given a set of intervals in one genome they read the entire CHAIN file, from start to finish, and report the translated coordinates for all query segments that can be mapped to the target genomes. This process is too slow to support interactive cross-species data visualization, since the delays in refreshing each view of adjusted genome coordinates are unacceptable for browsing data. To enable low latency coordinate conversion, CrossBrowse uses TABIX indexing[70] to accelerate access to CHAIN file genome alignments. The UCSC genome browser hosts CHAIN files generated with respect to a popular reference organisms (e.g., relative to human, mouse, fly).

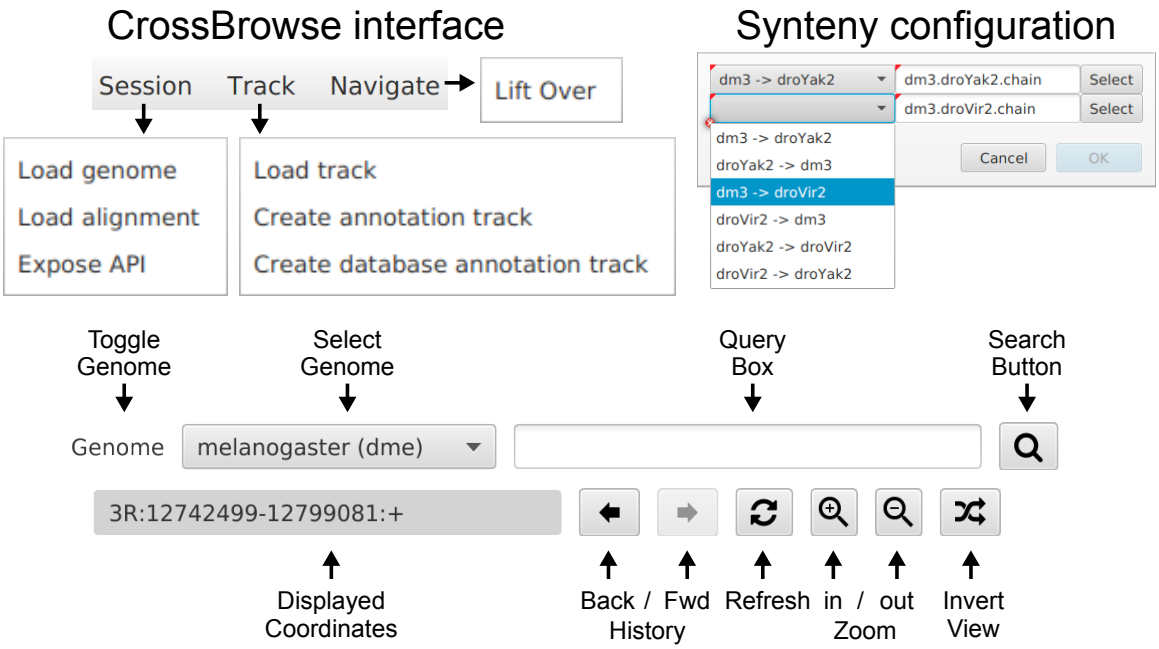


Figure 26: Configuration of the browser is easily achieved using a graphical interface. Below, the navigation bar is annotated with the function of each button. The popup at right demonstrates the procedure for loading whole genome alignment CHAIN files.

However, bidirectional coordinate conversion requires these alignments to be available in both directions, which would require two CHAIN files. To simplify configuration and allow users to easily use publicly available CHAIN files, CrossBrowse includes the functionality for deriving inverted CHAINs, obviating the need for the user to perform unnecessary pre-processing.

A further limitation of existing tools for multi-species visualization is the difficulty for configuration. While the session for many popular single genome browsers can be conveniently configured via an intuitive graphical user interface, existing multi-species browsers require the user to install the browser on a web server. These browsers require substantial command line usage to configure and install the browser (GBrowse_syn[84]) and/or convert data into browser-specific formats (mGSV[117]). In contrast, CrossBrowse provides a graphical interface, making the browser more

readily accessible to users lacking the time or technical expertise to install and configure a session. The only requirement for running CrossBrowse is a Java virtual machine (JVM) which is readily available on all operating systems.

Upon launch, the user is presented with an empty browser session, within which a user can load genomes, CHAIN files, and data tracks from the menu bar. Once configured, the synteny display is synchronized as the user navigates directly to a specific genomic coordinate or using text search against loaded annotations. A detailed text description that illustrates the process of configuring and navigating a multi-species session is provided at the CrossBrowse homepage.

4.4.2 Practical applications of CrossBrowse

Coding sequences are often broadly conserved, but strong constraints on other genomic elements can be used to infer their underlying functionality and utility (e.g., TF binding sites, miRNA and RBP binding sites, etc.). While existing browsers (e.g. UCSC genome browser) permit facile visualization of conserved sequences alongside experimental data from one organism, they do not permit simultaneous visualization of multispecies data. This is needed to distinguish whether utilization of a conserved functional element is similar, or perhaps divergent between species. A greater challenge is that many genomic elements are under modest positional constraint and/or difficult to infer directly from primary genome sequence alone. Thus, it can be difficult to distinguish functional elements that are functionally conserved but are defined by different primary sequences, from those that are truly species-specific and evolutionarily divergent. We illustrate the general utility of CrossBrowse with analyses of

invertebrate and mammalian genomic data, which demonstrate insights into diverse aspects of transcriptional and post-transcriptional regulation.

4.4.3 Evolutionary dynamics of enhancers and insulators

Facile browsing provides any bench scientist the ability to evaluate general trends in genomewide comparative data, as well as to navigate specifically to genes of particular interest for functional study. To illustrate the capability of CrossBrowse to handle mammalian-scale datasets, we downloaded ChIP-seq data for several chromatin features (H3K27ac, P300, TFAP2A) used to identify divergent enhancers between human and chimp[111]. By navigating to the COL13A1 locus, the resulting visualization readily illustrates the human-biased enhancer noted in the prior work (Figure 27, red box). Moreover, one can identify a weakly human-biased enhancer (yellow box) as well as a strongly chimp-biased enhancer (blue box). This emphasizes the ease with which CrossBrowse can be configured for retrospective analysis of published datasets.

In comparative genomic studies, it is typical to designate one genome as a reference, and translate the coordinates of all measurements from all experiments made in non-reference genomes into a common coordinate system of the reference genome. This "coordinate translation" between regions with sequence similarity is performed using tools such as liftOver and CrossMap[63, 156], using pre-calculated whole genome alignments. While this approach works well for comparing features that occur in well-aligned and clearly orthologous regions, care must be taken to design the analysis to properly account for events that occur in regions without an easily identified homologous sequence. Although such events are typically considered in bulk to be

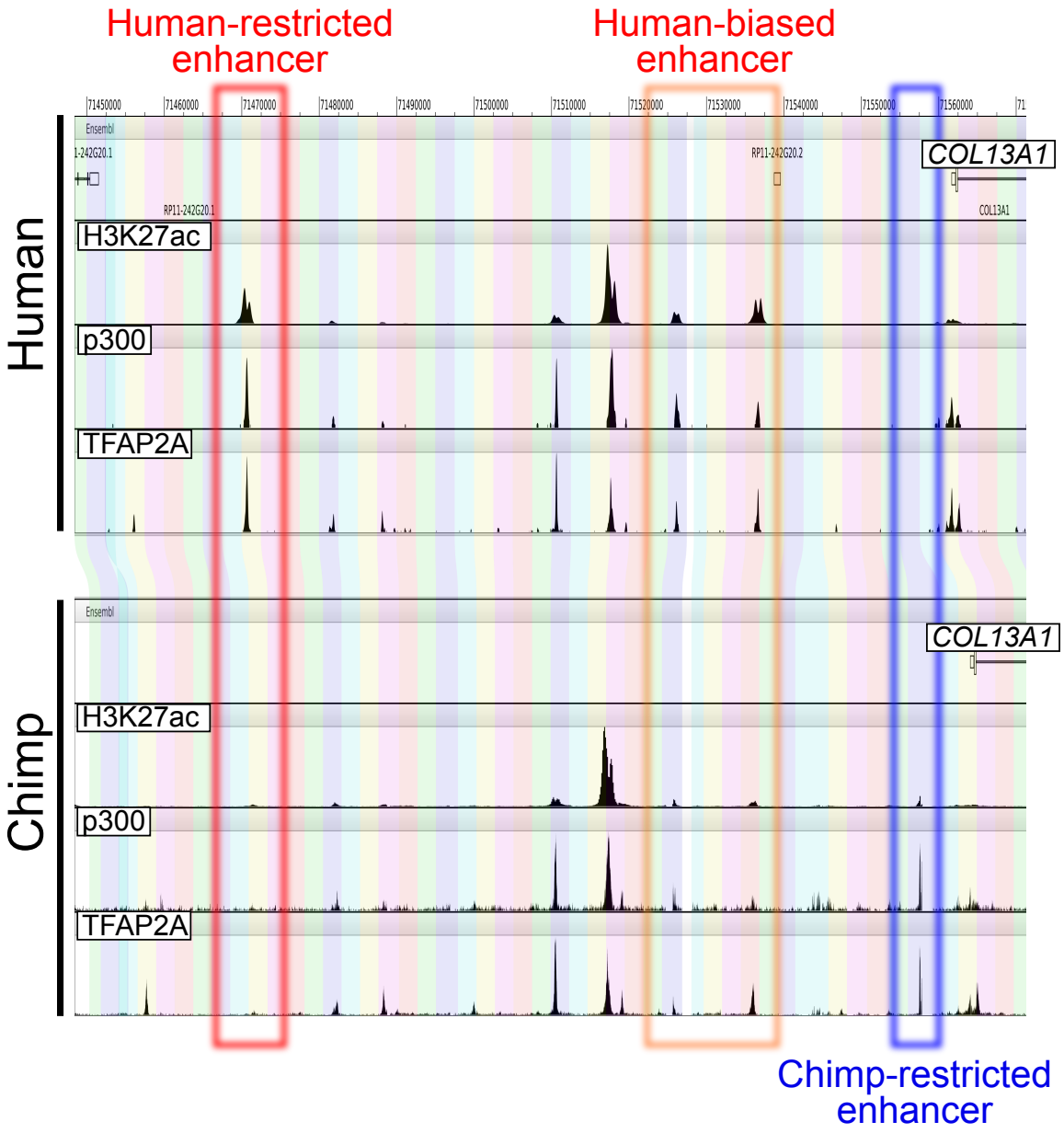


Figure 27: Screenshot of a CrossBrowse session that allows interrogation of species-specific enhancer evolution at mammalian COL13A1 reported by Wysocka and colleagues[111]. Here, we plot WIG tracks of their H3K27ac, P300, and TFAP2A ChIP-seq data from human and chimp. The red box highlights the human specific enhancer earlier reported[111], and the blue arrow indicates the presence of a chimp biased enhancer.

species-specific, the direct analysis of coordinate translated measurements can create artifacts of interpretation.

We illustrate this case using comparative data for the insulator DNA-binding protein CTCF, analyzed across 4 Drosophila species[99]. This work identified conserved- and species-specific CTCF binding events, and concluded that while CTCF DNA binding motif is conserved, the location of binding events between species evolves dynamically. In particular, liftOver coordinate-translation of CTCF ChIP-seq data at Abd-B was used to identify cases of CTCF binding in D. melanogaster that were lost in D. pseudoobscura[99] (Figure 28, yellow box). However, comparative visualization of the CTCF ChIP-seq data aligned to each cognate genome in CrossBrowse reveals that CTCF binds at a similar genomic position in both species (Figure 28, red box). This region is anchored by flanking syntenic sequences, even though the CTCF binding itself occurs within a region that was not well-aligned. This situation can be classified as a "false-positive" from the liftOver pipeline, since CTCF binds in an analogous genomic position in these species relative to the genes its insulator activity should affect.

Given that coordinate translation of sequencing data can introduce such false-positive calls, we were interested to assess its impact genomewide, and to develop a refined analysis strategy. In particular, we sought to distinguish true species-specific events from ones that occur in regions of locally low sequence identity, whose overall genomic correspondence was anchored by nearby flanking orthologous regions (as observed in the Abd-B locus). To do so, we expanded the binding peak called from the ChIP-seq data by 500bp on each side before performing coordinate translation

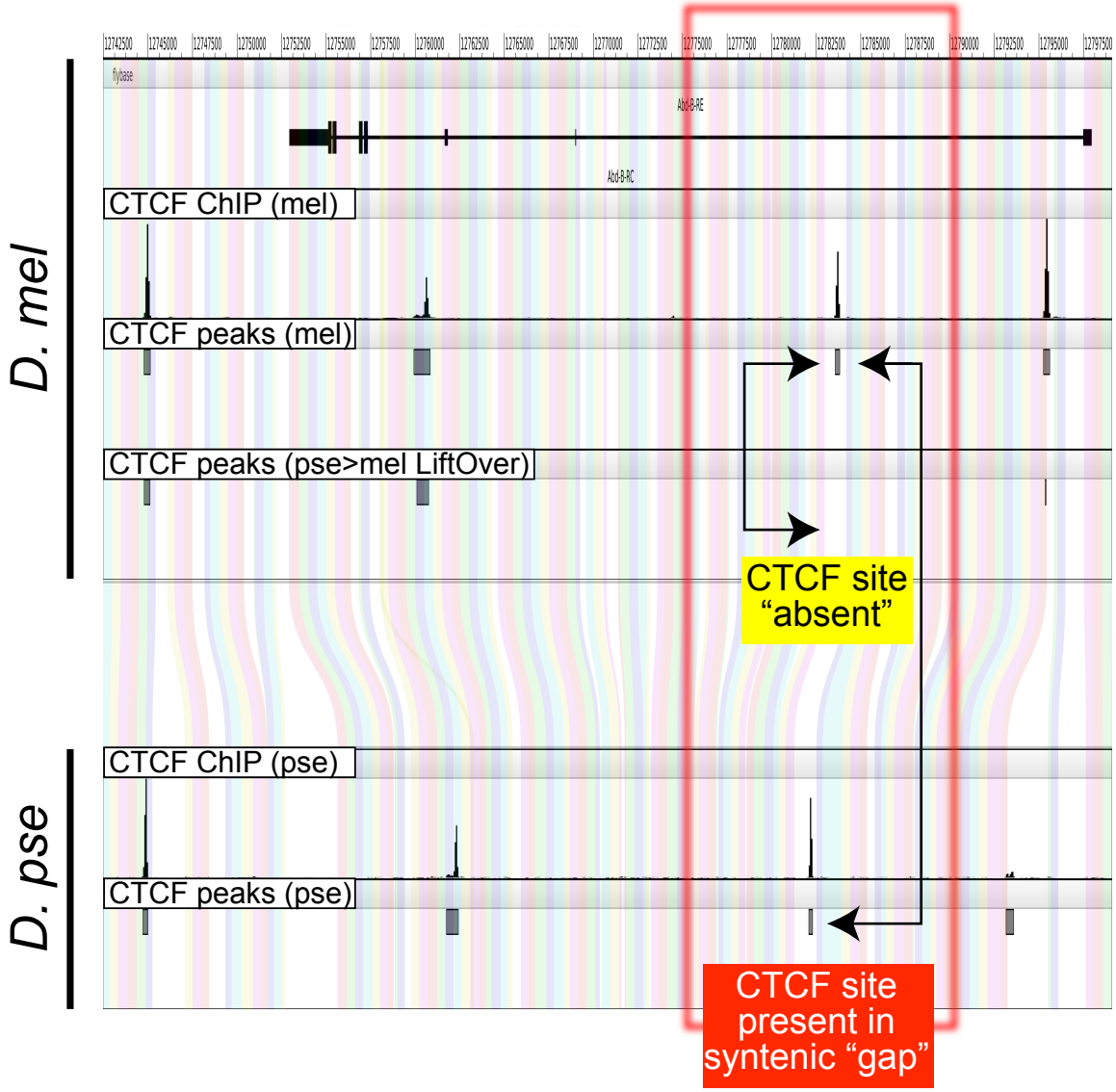


Figure 28: Screenshot of a CrossBrowse session comparing CTCF sites assayed across multiple Drosophilid species. The Abd-B locus was originally reported to harbor evolutionary divergent CTCF binding by White and colleagues[99]. The red box highlights a CTCF binding in D. melanogaster that is not directly aligned in D. pseudoobscura, and is thus lost in the standard liftOver translation. However, visual affirmation of their synteny is easily established via proximal flanking regions.

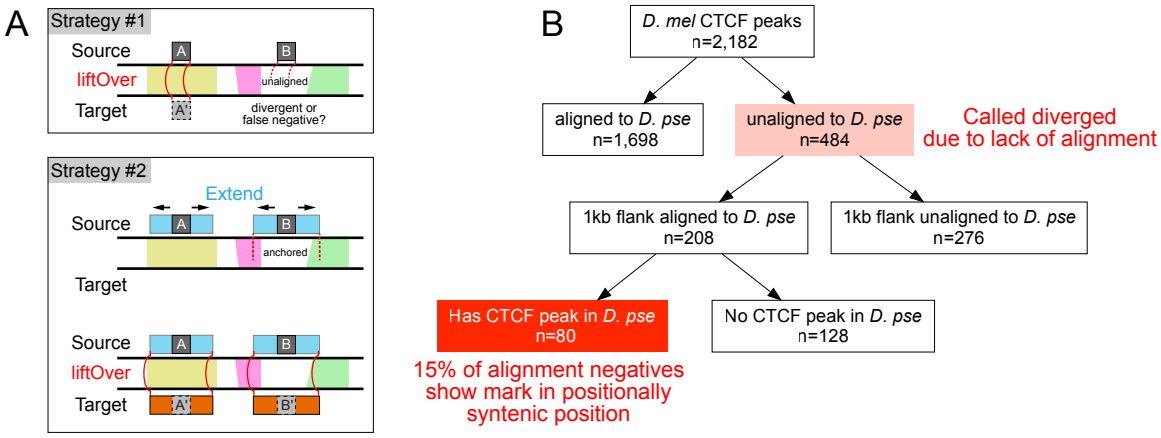


Figure 29: (A) Comparison of strategies to ascertain conservation and divergence of functional elements. Strategy number 1 was used in the original CTCF study[99], and is typically employed in cross-species studies. ChIP-seq peak coordinates were directly translated using liftOver when a sequence alignment is available (element "A", left), while peaks residing within segments that are locally unaligned are discarded (element "B", right). As informed by CrossBrowse visualization, we implemented the modifications shown in Strategy number 2. First, CTCF peaks were extended by a fixed distance (500 bp) on either side (blue rectangles). If extended peaks could be anchored in flanking aligned regions, coordinate translation was performed to identify homologous regions (red connectors). This strategy allows one to query both elements A and B for functional conservation in the experimental data. (B) Flowchart for reanalysis of comparative CTCF ChIP-seq data[99]. Of 484 events in D. melanogaster originally called as absent in D. pseudoobscura, due to lack of liftOver coordinates, we were able to rescue nearly half of them as residing in clearly orthologous regions based on their genomic flanks. About 40% of these regions (comprising 16.5% of total alignment negatives in the D. melanogaster $\mapsto$ D. pseudoobscura liftOver), indeed harbored conserved CTCF binding in the latter species. We vetted all 80 locations by manual inspection in CrossBrowse.

(Figure 29). Using this strategy we were able to identify >200 CTCF binding events in D. melanogaster that failed standard liftOver, but where the flanking genomic DNA can be used to identify the syntenic position in the D. pseudoobscura genome (Figure 29). Of these, browsing confirmed CTCF binding within the syntenic segment in ∼40% of the "rescued" genomic regions (80 loci, comprising 15% of the total CTCF peaks).

We emphasize that manual validation is extremely cumbersome using existing

browser solutions; i.e., necessitating the identification of orthologous regions and navigating to the respective data in separate browsers. These maneuvers are sufficiently onerous that they usually preclude systematic manual assessment of species-specific features. Our detection of a substantial number of falsely positive functional divergences highlights how visual browsing of multispecies genomic data in CrossBrowse can inform computational approaches for improved comparative analyses.

4.4.4 Evolutionary dynamics of alternative polyadenylation

In addition to visualizing the evolution of DNA-based functional elements, CrossBrowse has diverse utilities for analysis of transcriptome elements. RNA-seq has been used to study the evolutionary conservation of alternative splice isoforms[11, 86, 17] and alternative polyadenylation (APA) events[123, 28]. We previously described tissue-specific utilization of 3' UTR isoforms[90, 128], including broad trends for expression of shorter 3' UTR isoforms in testis and longer 3' UTR isoforms in neural tissues.

Despite the qualitative similarity of these tissue-specific patterns, proximal polyadenylation sites often lack clear signatures for primary sequence conservation, as we observed in Drosophila[128]. To better understand how the observed primary sequence contributes to alternative isoform expression between species, we analyzed RNA-seq and 3'-seq data from head and testis of D. yakuba, which is only 6 million years diverged from D. melanogaster (Sanfilippo and Shenker et al, in preparation). Examination of differential isoform expression patterns using CrossBrowse reveals different modes of 3' UTR expression pattern evolution between species.

Figure 30: Shown are examples of how diverse types of transcriptome data can be interpreted using CrossBrowse to identify evolutionary divergence in functional elements. (A, B) Screenshots of a CrossBrowse session loaded with total RNA-seq and 3'-seq data from heads and testes of D. melanogaster and D. yakuba; note that 3' UTRs are not annotated in existing D. yakuba gene models. These panels illustrate loci utilizing tissue-specific alternative polyadenylation that are evolving in distinct manners between these species. (A) Top, in the case of CG2201, the proximal testis PAS is lost in D. yakuba. Below, zooming to the nucleotide level shows that the canonical testis PAS used in D. melanogaster is diverged in D. yakuba (red box, asterisk). (B) The higher level view at Pde1c shows that both species incur testis-specific 3' UTR shortening, but the color-coded alignments indicate that non-orthologous termini are utilized. Below, zooming to the nucleotide level shows that D. yakuba two mutations and a single nucleotide deletion relative to D. melanogaster have disrupted the canonical PAS (red box, asterisk), but that a less frequently utilized D. melanogaster PAS located downstream is converted into a dominantly used testis PAS in D. yakuba (blue box). This conserves the overall pattern of testis 3' UTR shortening via different cis-regulatory sequences. (C, D) Screenshots of a CrossBrowse session loaded with small RNA-seq data from male bodies. (C) Examining orthologous miRNAs dme-mir-960, dsi-mir-960, and dse_335 reveals a difference in the relative accumulation of the 5p vs 3p arms between D. melanogaster, D. simulans, and D. sechellia, respectively. (D) An indel occurring between homologous sequences in D. pseudoobscura and D. persimilis expresses miRNA dps_3417.

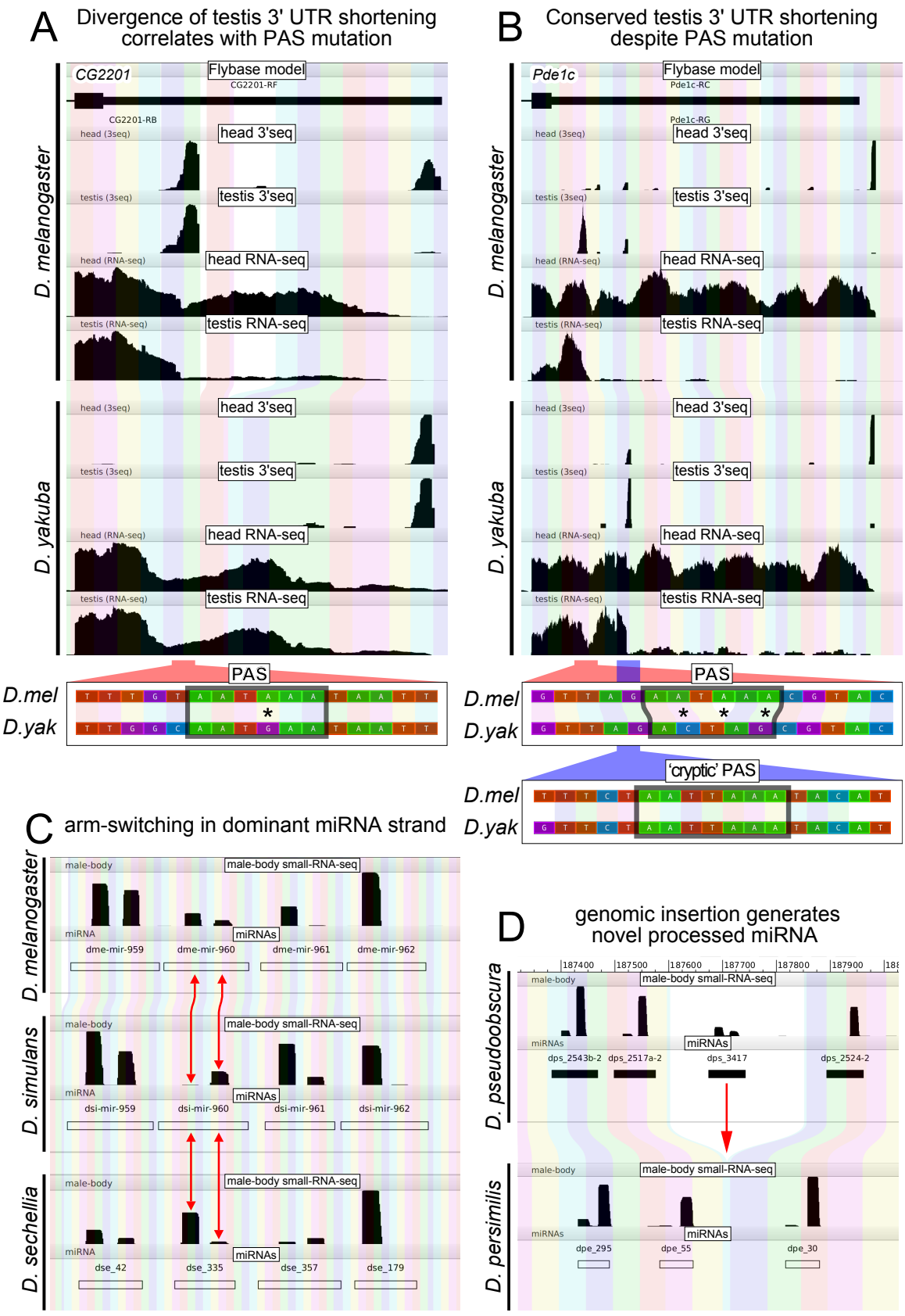

A
Divergence of testis 3' UTR shortening correlates with PAS mutation
B
Conserved testis 3' UTR shortening despite PAS mutation
D. melanogaster
D. yakuba
CG2201
Flybase model
CG2201-RF
CG2201-RB
Pde1c
Pde1c-RC
Pde1c-RG
head (3seq)
head 3'seq
testis (3seq)
testis 3'seq
head (RNA-seq)
head RNA-seq
testis (RNA-seq)
testis RNA-seq
PAS
D.mel
D.yak
'cryptic' PAS
C
arm-switching in dominant miRNA strand
D. melanogaster
D. simulans
D. sechellia
male-body
male-body small-RNA-seq
miRNA
miRNAs
dme-mir-959
dme-mir-960
dme-mir-961
dme-mir-962
dsi-mir-959
dsi-mir-960
dsi-mir-961
dsi-mir-962
dse_42
dse_335
dse_357
dse_179
D
genomic insertion generates novel processed miRNA
D. pseudoobscura
D. persimilis
187400
187500
187600
187700
187800
187900
dps_2543b-2
dps_2517a-2
dps_3417
dps_2524-2
dpe_295
dpe_55
dpe_30

Figure 30

In Figure 30A, the CG2201 locus provides an example where loss of the testis-specific polyadenylation in D. yakuba results in loss of highly differential 3' UTR expression between tissues observed in D. melanogaster. We find it remarkable to find such divergent mRNA processing patterns over recent evolutionary history. By contrast, Pde1c illustrates a case where despite the fact that both fly species exhibit 3' UTR shortening in testis relative to head, the actual sites of proximal polyadenylation are not at orthologous positions (Figure 30B). This type of "divergent, yet functionally similar" gene regulation (akin to the CTCF example, Figure 28) would have been difficult to appreciate without simultaneous browsing of data mapped to each cognate genome.

By zooming to the primary sequence, we observe that alteration of polyadenylation events at both Pde1c and CG2201 are accompanied by mutations in a proximally located PAS. Notably, there exists an alignable proximal PAS in D. yakuba CG2201 bearing only a single nucleotide change relative to its ortholog (Figure 30A). While one might have inferred this to represent a weak variant site, inspection in CrossBrowse makes it plainly evident that this altered PAS is essentially non-functional. On the other hand, the multiple changes to the D. yakuba Pde1c proximal PAS can more easily be imagined to render it inactive (Figure 30B). Interestingly, inspection of the primary sequence of the testis-specific polyadenylation event in D. yakuba reveals usage of a putatively cryptic PAS that is unmasked by loss of the upstream polyadenylation signal (Figure 30B), thereby preserving the pattern of testis 3' UTR shortening. Notably, this signal is embedded within an identical sequence block in D. melanogaster, but does not yield any functional 3' termini in this species.

Overall, these examples highlight different modes of functional divergence between otherwise very related species. Importantly, these evolutionary changes in functional elements would have been difficult to infer from primary sequence alone, and might otherwise require extensive bioinformatics to appreciate the diversity of evolutionarily changing patterns. Instead, they can be effectively uncovered and classified by human interaction with multiple types of cross-species deep sequencing data via CrossBrowse.

4.4.5 Direct visualization of alterations in miRNA locus structure and processing

As another application in the post-transcriptional realm, we illustrate the utility of CrossBrowse for interpreting miRNA evolution. While miRNA loci were originally studied with respect to conserved arrangement and processing, presumably reflecting their orchestration of conserved target networks, the advent of deep sequencing of small RNAs across multiple species clades has permitted evolutionarily divergent aspects of miRNA pathway to be studied[93, 88, 14].

We loaded CrossBrowse sessions with small RNA datasets from closely related groups of Drosophilids, namely D. melanogaster/simulans/sechellia, and D. pseudoobscura/persimilis (Mohammed and Flynt et al, in preparation), and browsed them for evidence of evolutionary lability in miRNA content and/or processing. Based on our recent appreciation of accelerated divergence in testis-restricted miRNA clusters[93], we were interested to browse their behavior using our platform.

Candidate events of miRNA emergence can be categorized bioinformatically, but these require conscientious manual vetting as many fortuitous degradation products

might masquerade as miRNAs[14]. However, it is extremely tedious using existing genome browsers to gain confidence that these represent genuine birth events in a specific species, and if so, what genomic alterations lead to their emergence. By simply inspecting the raw data in CrossBrowse, we could identify cases of "arm-switching" in the dominant accumulation of 5p vs 3p hairpin species within a a testis cluster in species of the melanogaster subclade (Figure 30C). This alteration is expected to alter the functional output of a miRNA locus[102]. More dramatically, we could also intuitively visualize a genomic alteration in a testis cluster that results in the recent birth of a miRNA locus within an individual species in the obscura subclade (Figure 30D).

Overall, such observations can be the basis of directed computational analyses, whose outputs can then be manually validated using CrossBrowse. Such cycles of computational and human assessment comprise an efficient workflow to interpret cross-species deep sequencing data.

4.5 Conclusion: a generic tool for integrating cross-species datasets

The proliferation of public genomic experiments across many species enables opportunities for comparisons between species. Unfortunately, most existing browsers are difficult to configure, do not adequately visualize synteny and rearrangements between genomes, are not capable of displaying experimental data between species, and are not scalable to vertebrate-sized genomes. These limitations inhibit users, especially

data producers who may not have substantial computational expertise, to quickly ask new questions using genomic datasets. We address these issues with CrossBrowse, an easily configurable comparative browser that supports standard genomics data formats and permits intuitive visualization of synteny and structural changes across the full gamut of genomic windows of interest. Data access routines are performed in background threads to maximize responsiveness of the user interface. This browser is implemented in Java, can be easily be installed and run on any platform with a JVM, and does not require internet access or data upload, allowing maximum flexibility for data with privacy requirements. We anticipate that CrossBrowse will find broad utilization, and help foster direct interrogation of cross-species data by bench scientists with limited computational expertise. Indeed, recent years have shown that several basic features of regulatory networks, RNA processing, and gene expression in the well-studied organisms are actively evolving, or have proven to be atypical or derived in some fashion. This highlights the importance of a broader cross-species foundation for experimental biology in the future.

4.6 Methods

4.6.1 Algorithm for constructing synteny representation

Whole genome alignments provide a means of mapping coordinates between two genome assemblies, from large regions down to nucleotide resolution. To generate an adaptive representation of these mapping functions, we down-sample the mapping function to render a visual representation of primary sequence alignment that is ap-

propriate to the scale of the view. The synteny visualization is generated from a graph of the genomes represented in the session, where the genomes represent the nodes of the graph, and the whole genome alignment represent edges between two genomes. Since CrossBrowse represents synteny between neighboring pairs of genomes, the order in which they are stacked in the session induces an ordering between genomes. The top genome in the browser is labeled as the "source" genome, and the bottom genome is labeled as the "sink". The source genome determines the parameters of the synteny visualization. The base-pair width of the segment displayed in this view determines the resolution of the syntenic blocks. The resolution is determined to be the maximum size such that at least the minimum width of a syntenic block in the source genome is 20 pixels. For example, if the browser window is 1000 pixels across, and displaying a region that is 100 bp long, 2 pixels will be the width allocated to each nucleotide, and one syntenic block will represent 10 nucleotides in the source genome.

The granularity of the source genome establishes interval partitions, whose register is aligned such that syntenic blocks line up with "round" coordinates in the source genome (1, 5, 10, 25, 50, . . .). This partitioning establishes the "base" of each syntenic block. Syntenic blocks are constructed using a recursive algorithm that is applied to each interval in the source genome. For each interval, whole genome alignments are queried to project the coordinates of source sequence into the coordinate system of the next genome in the stack. The source coordinates are added to a stack, and the coordinates in the target genome become the new source interval. This process is repeated until coordinate translation either reaches the "sink" genome, or there are

no remaining segments for which a valid coordinate translation exists.

There are a number of edge cases that must be considered when performing these projections. For a given query interval, the whole genome alignment may not allow coordinate translation for a subsequence (or entirety) of the queried segment. If these gaps occur on the boundary of a syntenic block, this will have the effect of "narrowing" the syntenic block, and must be propagated along the sequence of preceding syntenic segments to faithfully represent the boundaries segments that are syntenic across all displayed genomes. Additionally, a single segment can map to multiple locations in a target genome. In this manner, the nascent syntenic blocks form a tree, where the source interval forms the parent node, which will projected onto one or more "child" segments in the target genome.

Once the syntenic intervals composing each syntenic block are assembled, graphical parameters are from the GUI, including the relative vertical space allocated to each genome, are integrated with parameters of the view indicating whether each genome should be drawn in a left-to-right or inverted coordinate space, to establish the coordinates used to render each synteny block. Additional genomes can be stacked if desired, and their relationship will be visualized by the above logic.

4.6.2 Deriving CHAIN files to support rapid coordinate translation

CrossBrowse represents whole genome alignments using uses CHAIN files, which users can easily access from the UCSC genome browser. CHAIN files have a polarity, allowing for coordinate queries from a source genome to be translated into the coordinate space of a target genome. To enable facile exploration of multiple genomes simul-

taneously, CrossBrowse derives CHAIN files with the necessary polarity to support rapid queries in either direction between any pair of genomes with no additional user interaction. Additionally, whole genome alignments are often generated with a "star" topology (i.e. all organisms versus human), such that it is necessary to make multiple queries each time coordinates are between two species.

Represented as a graph where the genomes represent the nodes and whole genome alignments the edges between nodes, CrossBrowse includes the functionality to derive missing edges. To do this requires two fundamental operations, CHAIN inversion and CHAIN composition. Inversion is trivial to implement, and is essentially achieved by swapping the 'source' and 'target' fields of the chain, and ordering the chains with respect to the target genome. Given genomes A, B, and C, the composition $f \circ g$, takes two input alignments $f : A \mapsto B$ and $g : B \mapsto C$ and outputs the mapping from $A \mapsto C$. Addition is implemented by taking the boundaries of each aligned block in the source genome, and recursively translating its coordinates into the space of the 'sink' genome.

With these two fundamental operations in hand, we use a greedy algorithm to derive missing edges from a graph of genomes and their alignments. To begin, all available chains are inverted. Second, all pairs of nodes (genomes) are enumerated, and pairs for which no alignment exists are identified. For each pair without an alignment we identify the shortest path between those two nodes using existing edges in the graph, and tabulate the frequency of all 3-node subsequences (triplets) in the set of shortest paths. Taking the triplet that is common to the most missing paths, we apply composition on the two CHAIN files that form the edges between the three

nodes, and the composition is inverted to generate the reverse edge. The length of all paths that contain this triplet reduced by one by the composition operation. This procedure is recursively repeated applied there are no nodes that have a path length of greater than two between them.

4.6.3 Indexing CHAIN alignments

The UCSC CHAIN format is used to represent whole genome alignments. However, in its original form, the CHAIN format is not amenable to random access queries. TABIX indexing can be used to index many commonly used data formats. To enable TABIX indexing CrossBrowse converts conventional CHAIN format files to a row-oriented GFF format file. The resulting GFF file is compressed, indexed, and subsequently queried using the TABIX API provided by HTSJDK.

4.6.4 Analysis of chromatin features

We downloaded human/chimp ChIP-seq data for chromatin features from Gene Expression Omnibus (GEO) accession GSE70751[111], and CTCF ChIP-seq data from Drosophila species from GSE24449[99]. We also utilized CTCF peak calls annotated in D. melanogaster and D. pseuodoobscura, respectively (in files GSM602326_dmel_peak_regions.bed and GSM602335_dpse_peak_regions.bed).

We compared two strategies for determining the presence of orthologous CTCF binding sites. The first approach replicated the previously reported scheme[99], in which we used liftOver to translate the coordinates of the 2,182 D. melanogaster peak calls to the D. pseuodoobscura genome. Sites that could not be translated

were marked as melanogaster-specific. We intersected the remaining sites with the 2,332 CTCF binding intervals called in D. pseudoobscura, to annotated the overlap as conserved CTCF sites.

In the second strategy, we created two new intervals for each D. melanogaster CTCF peak call by alternately extending the original interval by 500 nt on either side. The coordinates of each extended interval were translated between assemblies. If a pair of intervals was (1) translated to the same reference sequence, (2) spanned a genomic interval no more than two-fold longer than the original melanogaster interval, and (3) in a relative orientation between the two was consistent with that observed for melanogaster, then they were considered to represent orthologous regions. As noted, this procedure rescues >200 D. melanogaster CTCF sites from being unaligned in D. pseuodoobscura, to having orthologous genomic assignments. We assessed these for experimental evidence of CTCF binding in both species, paying especial attention to manually assess all "rescued" conserved sites.

4.6.5 Software Availability

CrossBrowse was implemented in Java. Executables and source code are available from https://github.com/shenkers/ComparativeBrowser/releases. An online document with detailed instructions for configuration and use are available at https://-github.com/shenkers/ComparativeBrowser/wiki.

4.7 Acknowledgements

We thank Piero Sanfilippo for access to 3'-end sequencing data, Alex Flynt and Jaaved Mohammed for observations on miRNA evolution, and Brian Joseph and Jeffrey Vedanayagam for manuscript comments. S.S. was supported by the Tri-Institutional Training Program in Computational Biology and Medicine. Work in E.C.L.'s group was supported by the National Institutes of Health (R01-NS074037, R01-NS083833 and R01-GM083300) and MSK Core Grant P30-CA008748.

5 Summary

5.1 Regulatory implications of differential 3'UTR isoforms

Through our analysis of deep RNA-seq datasets we have revealed 3'UTR extensions in fly, mouse, and human. Our observations based on sequence conservation and CLIP experiments suggest that one functions these 3'UTR extensions might play is to provide a context specific mechanism to modulate a gene's exposure to regulation by miRNAs. However there are many open questions regarding the importance of tissue specific alternative 3'UTR isoforms, and in particular, their role in the nervous system. While we focused on miRNA targeting in our studies, there may be additional elements in the 3'UTR that are modulated by alternative polyadenylation. Supporting the existence of this regulatory "dark-matter", recent studies have shown that only a fraction of the elements with regulatory activity in a 3'UTR represent miRNA target sites [100, 148]. This suggests that alternative 3'UTR isoforms we observe between tissues may affect gene regulation through currently unappreciated mechanisms.

In addition to affecting gene activity by modulating transcript stability and protein expression, the 3'UTR can also affect subcellular transcript localization [8]. Given the unique highly polarized morphology of cells in the nervous system, it is tempting to speculate that alternative 3'UTR isoforms provide a mechanism to localize these transcripts. Lending credence to this hypothesis, tandem 3'UTR isoforms of BDNF are known to be differentially localized in neurons [4], and a recent study has shown that alternative last exon isoforms are frequently differentially localized to neural pro-

jections [137]. It is plausible that the 3'UTR extensions we identified could regulate the transcripts they affect by one or more of these mechanisms, and our catalogue of tissue specific 3'UTR isoforms provides a setting to explore the regulatory significance of these events in more detail.

5.2 Utility of changepoint analysis

The recent proliferation of RNA-seq experiments provides abundant opportunities to make discoveries through retrospective analysis of published datasets. While detection of alternative polyadenylation events in RNA-seq data was traditionally difficult without high qualtity reference models, the IsoSCM algorithm facilitates the extraction of alternative polyadenylation patterns, allowing investigators to leverage existing RNA-seq datasets to gain insight into the regulation of alternative polyadenylation. For example, our preliminary analysis of RNA from cell bodies and peripheral axons [89] display distinct enrichment of neural 3'UTR isoforms, and among FACS purified neural cell types [155], neurons express the highest levels of 3'UTR extensions.

Additionally, while the change-point model used by IsoSCM was developed for the identification of differential tandem 3'UTR isoforms, it has equal power to detect tandem 5'UTR isoforms. To this end, we have identified hundreds of tandem 5'UTR isoforms with enriched accumulation in the fly testis using IsoSCM. Interesetingly, the testis is observed to be a tissue with frequent gene turnover [157], and existence of additional elements capable of driving gene expression could explain part of this phenomenon. While we have not developed these observations further at this point,

these examples illustrate that IsoSCM has applications outside the scope of detecting tissue differential alternative polyadenylation events.

5.3 Insights from comparative analysis

Through our comparative analysis of polyadenylation patterns across fly species using CrossBrowse we have identified different modes of polyadenylation site evolution. Interestingly, polyadenylation sites are in general under purifying selection, many instances of species specific polyadenylation site usage can also be observed, suggesting that the transcriptional process is somewhat labile. We have used CrossBrowse to illustrate observations relating changes in genomic sequence to differences in polyadenylation site activity. Moreover, this strategy could be extended to relate genomic changes to tissue differential polyadenylation site usage. While patterns identified by this strategy are purely correlative, the development of methods for efficient editing of the fly genome using CRISPR [44], makes testing their causality a practical possibility.

As we illustrate with visualizations several diverse data-types, the utility of CrossBrowse extends beyond its application to polyadenylation site usage. Indeed, the technical difficulty of comparing experimental data across genomes has likely limited the broad usage of multi-species datasets, and CrossBrowse provides a platform to begin using these datasets to their full potential. So far, we have only used CrossBrowse to visualize data with respect to a species' reference genome. However, a hallmark of cancers is genomic instability [96], and thus the genome of a cancer cell

is in general different from the germline it is derived from. For this reason, it is worth exploring whether there is any utility in using CrossBrowse to visualize re-arranged cancer genomes.

5.4 Conclusion

Through analysis of RNA-seq data we have shown that coherent patterns of biased polyadenylation site usage between emerge tissues, in a pattern that is consistent between fly, mouse, and human. In the process we have revised the 3'UTR models of thousands of genes, revealing an expanded repertoire of apparently functional miRNA target sites in these genes, and identifying challenges to interpreting short-read sequencing data. To address the limitations of existing tools we developed IsoSCM, and demonstrated its advantage for 3'UTR annotation and detection of differential 3'UTR expression. Motivated by our desire to understand evolutionary dynamics of alternative polyadenylation, we developed CrossBrowse to facilitate visualization of multi-species genomics datasets. Beyond the utility of IsoSCM and CrossBrowse to our own investigation of tissue alternative polyadenylation, these tools provide a framework for future studies to readily apply our methodology to explore new questions.

REFERENCES

[1] M. D. Adams, A. R. Kerlavage, R. D. Fleischmann, R. A. Fuldner, C. J. Bult, N. H. Lee, E. F. Kirkness, K. G. Weinstock, J. D. Gocayne, and O. White. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*, 377(6547 Suppl):3–174, Sep 1995.

[2] H. B. Akman, M. Oyken, T. Tuncer, T. Can, and A. E. Erson-Bensan. 3'UTR shortening and EGF signaling: implications for breast cancer. *Hum. Mol. Genet.*, 24(24):6910–6920, Dec 2015.

[3] M. Allen, C. Bird, W. Feng, G. Liu, W. Li, N. I. Perrone-Bizzozero, and Y. Feng. HuD promotes BDNF expression in brain neurons via selective stabilization of the BDNF long 3'UTR mRNA. *PLoS ONE*, 8(1):e55718, 2013.

[4] J. J. An, K. Gharami, G. Y. Liao, N. H. Woo, A. G. Lau, F. Vanevski, E. R. Torre, K. R. Jones, Y. Feng, B. Lu, and B. Xu. Distinct role of long 3' UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons. *Cell*, 134(1):175–187, Jul 2008.

[5] J. J. An, K. Gharami, G. Y. Liao, N. H. Woo, A. G. Lau, F. Vanevski, E. R. Torre, K. R. Jones, Y. Feng, B. Lu, and B. Xu. Distinct role of long 3' UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons. *Cell*, 134(1):175–187, Jul 2008.

[6] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol.*, 11(10):R106, 2010.

[7] S. Anders, A. Reyes, and W. Huber. Detecting differential usage of exons from RNA-seq data. *Genome Res.*, 22(10):2008–2017, Oct 2012.

[8] C. Andreassi and A. Riccio. To localize or not to localize: mRNA fate is in 3'UTR ends. *Trends Cell Biol.*, 19(9):465–474, Sep 2009.

[9] C. Andreassi, C. Zimmermann, R. Mitter, S. Fusco, S. De Vita, S. Devita, A. Saiardi, and A. Riccio. An NGF-responsive element targets myo-inositol monophosphatase-1 mRNA to sympathetic neuron axons. *Nat. Neurosci.*, 13 (3):291–301, Mar 2010.

[10] J. Bao, K. Vitting-Seerup, J. Waage, C. Tang, Y. Ge, B. T. Porse, and W. Yan. UPF2-Dependent Nonsense-Mediated mRNA Decay Pathway Is Essential for Spermatogenesis by Selectively Eliminating Longer 3'UTR Transcripts. *PLoS Genet.*, 12(5):e1005863, May 2016.

[11] N. L. Barbosa-Morais, M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C. M. Misquitta-Ali, M. D. Wilson, P. M. Kim, D. T. Odom, B. J. Frey, and B. J. Blencowe. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–1593, Dec 2012.

[12] E. Beaudoing, S. Freier, J. R. Wyatt, J. M. Claverie, and D. Gautheret. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.*, 10(7): 1001–1010, Jul 2000.

[13] J. Behr, A. Kahles, Y. Zhong, V. T. Sreedharan, P. Drewe, and G. Ratsch. MITIE: Simultaneous RNA-Seq-based transcript identification and quantification in multiple samples. *Bioinformatics*, 29(20):2529–2538, Oct 2013.

[14] E. Berezikov, N. Liu, A. S. Flynt, E. Hodges, M. Rooks, G. J. Hannon, and E. C. Lai. Evolutionary flux of canonical microRNAs and mirtrons in Drosophila. *Nat. Genet.*, 42(1):6–9, Jan 2010.

[15] S. Blaess, G. O. Bodea, A. Kabanova, S. Chanet, E. Mugniery, A. Derouiche, D. Stephen, and A. L. Joyner. Temporal-spatial changes in Sonic Hedgehog expression and signaling reveal different potentials of ventral mesencephalic progenitors to populate distinct ventral midbrain nuclei. *Neural Dev*, 6:29, 2011.

[16] R. Bohnert and G. Ratsch. rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res.*, 38(Web Server issue):W348–351, Jul 2010.

[17] D. Brawand, M. Soumillon, A. Necsulea, P. Julien, G. Csardi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, F. W. Albert, U. Zeller, P. Khaitovich, F. Grutzner, S. Bergmann, R. Nielsen, S. Paabo, and H. Kaessmann. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348, Oct 2011.

[18] J. B. Brown, N. Boley, R. Eisman, G. E. May, M. H. Stoiber, M. O. Duff, B. W. Booth, J. Wen, S. Park, A. M. Suzuki, K. H. Wan, C. Yu, D. Zhang, J. W. Carlson, L. Cherbas, B. D. Eads, D. Miller, K. Mockaitis, J. Roberts, C. A. Davis, E. Frise, A. S. Hammonds, S. Olson, S. Shenker, D. Sturgill, A. A. Samsonova, R. Weiszmann, G. Robinson, J. Hernandez, J. Andrews, P. J. Bickel, P. Carninci, P. Cherbas, T. R. Gingeras, R. A. Hoskins, T. C. Kaufman, E. C. Lai, B. Oliver, N. Perrimon, B. R. Graveley, and S. E. Celniker. Diversity and dynamics of the Drosophila transcriptome. *Nature*, 512(7515):393–399, Aug 2014.

[19] J. C. Castle, C. Zhang, J. K. Shah, A. V. Kulkarni, A. Kalsotra, T. A. Cooper, and J. M. Johnson. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nat. Genet.*, 40(12): 1416–1425, Dec 2008.

[20] C. Chen, T. Ara, and D. Gautheret. Using Alu elements as polyadenylation sites: A case of retroposon exaptation. *Mol. Biol. Evol.*, 26(2):327–334, Feb 2009.

[21] S. W. Chi, J. B. Zang, A. Mele, and R. B. Darnell. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–486, Jul 2009.

[22] R. A. Chodroff, L. Goodstadt, T. M. Sirey, P. L. Oliver, K. E. Davies, E. D. Green, Z. Molnar, and C. P. Ponting. Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biol.*, 11 (7):R72, 2010.

[23] M. B. Clark, P. P. Amaral, F. J. Schlesinger, M. E. Dinger, R. J. Taft, J. L. Rinn, C. P. Ponting, P. F. Stadler, K. V. Morris, A. Morillon, J. S. Rozowsky, M. B. Gerstein, C. Wahlestedt, Y. Hayashizaki, P. Carninci, T. R. Gingeras, and J. S. Mattick. The reality of pervasive transcription. *PLoS Biol.*, 9(7): e1000625; discussion e1001102, Jul 2011.

[24] M. B. Clark, R. L. Johnston, M. Inostroza-Ponta, A. H. Fox, E. Fortini, P. Moscato, M. E. Dinger, and J. S. Mattick. Genome-wide analysis of long noncoding RNA stability. *Genome Res.*, 22(5):885–898, May 2012.

[25] E. K. Crawford, J. E. Ensor, I. Kalvakolanu, and J. D. Hasday. The role of 3' poly(A) tail metabolism in tumor necrosis factor-alpha regulation. *J. Biol. Chem.*, 272(34):21120–21127, Aug 1997.

[26] A. Curinha, S. Oliveira Braz, I. Pereira-Castro, A. Cruz, and A. Moreira. Implications of polyadenylation in health and disease. *Nucleus*, 5(6):508–519, 2014.

[27] F. Denoeud, J. M. Aury, C. Da Silva, B. Noel, O. Rogier, M. Delledonne, M. Morgante, G. Valle, P. Wincker, C. Scarpelli, O. Jaillon, and F. Artiguenave. Annotating genomes with massive-scale RNA sequencing. *Genome Biol.*, 9(12): R175, 2008.

[28] A. Derti, P. Garrett-Engele, K. D. Macisaac, R. C. Stevens, S. Sriram, R. Chen, C. A. Rohl, J. M. Johnson, and T. Babak. A quantitative atlas of polyadenylation in five mammals. *Genome Res.*, 22(6):1173–1183, Jun 2012.

[29] D. C. Di Giammartino, W. Li, K. Ogami, J. J. Yashinskie, M. Hoque, B. Tian, and J. L. Manley. RBBP6 isoforms regulate the human polyadenylation machinery and modulate expression of mRNAs with AU-rich 3' UTRs. *Genes Dev.*, 28(20):2248–2260, Oct 2014.

[30] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar,

P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, and T. R. Gingeras. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, Sep 2012.

[31] R. Elkon, A. P. Ugalde, and R. Agami. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.*, 14(7):496–506, Jul 2013.

[32] R. Engels, T. Yu, C. Burge, J. P. Mesirov, D. DeCaprio, and J. E. Galagan. Combo: a whole genome comparative browser. *Bioinformatics*, 22(14):1782–1783, Jul 2006.

[33] Fearnhead and Paul. Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213, June 2006. ISSN 0960-3174. doi: 10.1007/s11222-006-8450-8. URL `http://dx.doi.org/10.1007/s11222-006-8450-8`.

[34] J. Feng, W. Li, and T. Jiang. Inference of isoforms from short sequence reads. *J. Comput. Biol.*, 18(3):305–321, Mar 2011.

[35] S. W. Flavell, T. K. Kim, J. M. Gray, D. A. Harmin, M. Hemberg, E. J. Hong, E. Markenscoff-Papadimitriou, D. M. Bear, and M. E. Greenberg. Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection. *Neuron*, 60(6):1022–1038, Dec 2008.

[36] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. K. Kahari, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H. S. Riat, G. R. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y. A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernandez-Suarez, J. Harrow, J. Herrero, T. J. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, and S. M. Searle. Ensembl 2012. *Nucleic Acids Res.*, 40(Database issue):84–90, Jan 2012.

[37] Y. Fu, Y. Sun, Y. Li, J. Li, X. Rao, C. Chen, and A. Xu. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res.*, 21(5):741–747, May 2011.

[38] B. Fusby, S. Kim, B. Erickson, H. Kim, M. L. Peterson, and D. L. Bentley. Coordination of RNA Polymerase II Pausing and 3' End Processing Factor Recruitment with Alternative Polyadenylation. *Mol. Cell. Biol.*, 36(2):295–303, Jan 2015.

[39] A. V. Gendrel and E. Heard. Fifty years of X-inactivation research. *Development*, 138(23):5049–5055, Dec 2011.

[40] R. Gilat and D. Shweiki. A novel function for alternative polyadenylation as a rescue pathway from NMD surveillance. *Biochem. Biophys. Res. Commun.*, 353(2):487–492, Feb 2007.

[41] H. Goodarzi, H. S. Najafabadi, P. Oikonomou, T. M. Greco, L. Fish, R. Salavati, I. M. Cristea, and S. Tavazoie. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature*, 485(7397):264–268, May 2012.

[42] J. H. Graber, C. R. Cantor, S. C. Mohr, and T. F. Smith. In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species. *Proc. Natl. Acad. Sci. U.S.A.*, 96(24):14055–14060, Nov 1999.

[43] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29(7):644–652, Jul 2011.

[44] S. J. Gratz, A. M. Cummings, J. N. Nguyen, D. C. Hamm, L. K. Donohue, M. M. Harrison, J. Wildonger, and K. M. O'Connor-Giles. Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease. *Genetics*, 194(4): 1029–1035, Aug 2013.

[45] A. Grimson, K. K. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, 27(1):91–105, Jul 2007.

[46] A. R. Gruber, G. Martin, W. Keller, and M. Zavolan. Cleavage factor Im is a key regulator of 3' UTR length. *RNA Biol*, 9(12):1405–1412, Dec 2012.

[47] I. Gupta, S. Clauder-Munster, B. Klaus, A. I. Jarvelin, R. S. Aiyar, V. Benes, S. Wilkening, W. Huber, V. Pelechano, and L. M. Steinmetz. Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA-protein interactions. *Mol. Syst. Biol.*, 10:719, 2014.

[48] M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander, and A. Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, 28 (5):503–510, May 2010.

[49] M. Guttman, J. Donaghey, B. W. Carey, M. Garber, J. K. Grenier, G. Munson, G. Young, A. B. Lucas, R. Ach, L. Bruhn, X. Yang, I. Amit, A. Meissner, A. Regev, J. L. Rinn, D. E. Root, and E. S. Lander. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, 477(7364):295–300, Sep 2011.

[50] B. J. Haas, A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Smith, L. I. Hannick, R. Maiti, C. M. Ronning, D. B. Rusch, C. D. Town, S. L. Salzberg, and O. White. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.*, 31(19):5654–5666, Oct 2003.

[51] K. D. Hansen, S. E. Brenner, and S. Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.*, 38(12): e131, Jul 2010.

[52] D. R. Higgs, S. E. Goodbourn, J. Lamb, J. B. Clegg, D. J. Weatherall, and N. J. Proudfoot. Alpha-thalassaemia caused by a polyadenylation signal mutation. *Nature*, 306(5941):398–400, 1983.

[53] V. Hilgers, M. W. Perry, D. Hendrix, A. Stark, M. Levine, and B. Haley. Neural-specific elongation of 3' UTRs during Drosophila development. *Proc. Natl. Acad. Sci. U.S.A.*, 108(38):15864–15869, Sep 2011.

[54] V. Hilgers, S. B. Lemke, and M. Levine. ELAV mediates 3' UTR extension in the Drosophila nervous system. *Genes Dev.*, 26(20):2259–2264, Oct 2012.

[55] D. Hiller and W. H. Wong. Simultaneous isoform discovery and quantification from RNA-seq. *Stat Biosci*, 5(1):100–118, May 2013.

[56] J. R. Hogg and S. P. Goff. Upf1 senses 3'UTR length to potentiate mRNA decay. *Cell*, 143(3):379–389, Oct 2010.

[57] W. Huber, J. Toedling, and L. M. Steinmetz. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, 22(16):1963–1970, Aug 2006.

[58] C. H. Jan, R. C. Friedman, J. G. Ruby, and D. P. Bartel. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature*, 469(7328):97–101, Jan 2011.

[59] M. Jenal, R. Elkon, F. Loayza-Puch, G. van Haaften, U. Kuhn, F. M. Menzies, J. A. Oude Vrielink, A. J. Bos, J. Drost, K. Rooijers, D. C. Rubinsztein, and R. Agami. The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. *Cell*, 149(3):538–553, Apr 2012.

[60] Z. Ji and B. Tian. Reprogramming of 3' untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PLoS ONE*, 4(12):e8419, 2009.

[61] Z. Ji, W. Luo, W. Li, M. Hoque, Z. Pan, Y. Zhao, and B. Tian. Transcriptional activity regulates alternative cleavage and polyadenylation. *Mol. Syst. Biol.*, 7: 534, 2011.

[62] T. M. Keane, L. Goodstadt, P. Danecek, M. A. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, M. Goodson, N. A. Furlotte, E. Eskin, C. Nellaker, H. Whitley, J. Cleak, D. Janowitz, P. Hernandez-Pliego, A. Edwards, T. G. Belgard, P. L. Oliver, R. E. McIntyre, A. Bhomra, J. Nicod, X. Gan, W. Yuan, L. van der Weyden, C. A. Steward, S. Bala, J. Stalker, R. Mott, R. Durbin, I. J. Jackson, A. Czechanski, J. A. Guerra-Assuncao, L. R. Donahue, L. G. Reinholdt, B. A. Payseur, C. P. Ponting, E. Birney, J. Flint, and D. J. Adams. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–294, Sep 2011.

[63] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Res.*, 12(6):996–1006, Jun 2002.

[64] A. M. Khalil, M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B. E. Bernstein, A. van Oudenaarden, A. Regev, E. S. Lander, and J. L. Rinn. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, 106(28):11667–11672, Jul 2009.

[65] T. Kislinger, B. Cox, A. Kannan, C. Chung, P. Hu, A. Ignatchenko, M. S. Scott, A. O. Gramolini, Q. Morris, M. T. Hallett, J. Rossant, T. R. Hughes, B. Frey, and A. Emili. Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell*, 125(1):173–186, Apr 2006.

[66] K. Kristjansdottir, E. A. Fogarty, and A. Grimson. Systematic analysis of the Hmga2 3' UTR identifies many independent regulatory sequences and a novel interaction between distal sites. *RNA*, 21(7):1346–1360, Jul 2015.

[67] T. Kwan, E. Grundberg, V. Koka, B. Ge, K. C. Lam, C. Dias, A. Kindmark, H. Mallmin, O. Ljunggren, F. Rivadeneira, K. Estrada, J. B. van Meurs, A. Uitterlinden, M. Karlsson, C. Ohlsson, D. Mellstrom, O. Nilsson, T. Pastinen, and J. Majewski. Tissue effect on genetic control of transcript isoform variation. *PLoS Genet.*, 5(8):e1000608, Aug 2009.

[68] N. F. Lahens, I. H. Kavakli, R. Zhang, K. Hayer, M. B. Black, H. Dueck, A. Pizarro, J. Kim, R. Irizarry, R. S. Thomas, G. R. Grant, and J. B. Hogenesch. IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol.*, 15(6):R86, 2014.

[69] B. P. Lewis, R. E. Green, and S. E. Brenner. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. U.S.A.*, 100(1):189–192, Jan 2003.

[70] H. Li. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, 27(5):718–719, Mar 2011.

[71] J. J. Li, C. R. Jiang, J. B. Brown, H. Huang, and P. J. Bickel. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proc. Natl. Acad. Sci. U.S.A.*, 108(50): 19867–19872, Dec 2011.

[72] W. Li, J. Feng, and T. Jiang. IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. *J. Comput. Biol.*, 18(11):1693–1707, Nov 2011.

[73] S. Lianoglou, V. Garg, J. L. Yang, C. S. Leslie, and C. Mayr. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.*, 27(21):2380–2396, Nov 2013.

[74] D. D. Licatalosi and R. B. Darnell. RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.*, 11(1):75–87, Jan 2010.

[75] D. D. Licatalosi, A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. W. Chi, T. A. Clark, A. C. Schweitzer, J. E. Blume, X. Wang, J. C. Darnell, and R. B. Darnell. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469, Nov 2008.

[76] F. Lienert, F. Mohn, V. K. Tiwari, T. Baubec, T. C. Roloff, D. Gaidatzis, M. B. Stadler, and D. Schubeler. Genomic prevalence of heterochromatic H3K9me2 and transcription do not discriminate pluripotent from terminally differentiated cells. *PLoS Genet.*, 7(6):e1002090, Jun 2011.

[77] J. R. Manak, S. Dike, V. Sementchenko, P. Kapranov, F. Biemar, J. Long, J. Cheng, I. Bell, S. Ghosh, A. Piccolboni, and T. R. Gingeras. Biological function of unannotated transcription during the early development of Drosophila melanogaster. *Nat. Genet.*, 38(10):1151–1158, Oct 2006.

[78] M. Mangone, A. P. Manoharan, D. Thierry-Mieg, J. Thierry-Mieg, T. Han, S. D. Mackowiak, E. Mis, C. Zegar, M. R. Gutwein, V. Khivansara, O. Attie, K. Chen, K. Salehi-Ashtiani, M. Vidal, T. T. Harkins, P. Bouffard, Y. Suzuki, S. Sugano, Y. Kohara, N. Rajewsky, F. Piano, K. C. Gunsalus, and J. K. Kim. The landscape of C. elegans 3'UTRs. *Science*, 329(5990):432–435, Jul 2010.

[79] G. Martin, A. R. Gruber, W. Keller, and M. Zavolan. Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Rep*, 1(6):753–763, Jun 2012.

[80] J. A. Martin and Z. Wang. Next-generation transcriptome assembly. *Nat. Rev. Genet.*, 12(10):671–682, Oct 2011.

[81] C. P. Masamha, Z. Xia, J. Yang, T. R. Albrecht, M. Li, A. B. Shyu, W. Li, and E. J. Wagner. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature*, 510(7505):412–416, Jun 2014.

[82] C. Mayr and D. P. Bartel. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138(4): 673–684, Aug 2009.

[83] S. McCracken, N. Fong, K. Yankulov, S. Ballantyne, G. Pan, J. Greenblatt, S. D. Patterson, M. Wickens, and D. L. Bentley. The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature*, 385 (6614):357–361, Jan 1997.

[84] S. J. McKay, I. A. Vergara, and J. E. Stajich. Using the Generic Synteny Browser (GBrowse_syn). *Curr Protoc Bioinformatics*, Chapter 9:Unit 9.12, Sep 2010.

[85] T. R. Mercer, M. E. Dinger, C. P. Bracken, G. Kolle, J. M. Szubert, D. J. Korbie, M. E. Askarian-Amiri, B. B. Gardiner, G. J. Goodall, S. M. Grimmond, and J. S. Mattick. Regulated post-transcriptional RNA cleavage diversifies the eukaryotic transcriptome. *Genome Res.*, 20(12):1639–1650, Dec 2010.

[86] J. Merkin, C. Russell, P. Chen, and C. B. Burge. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, 338(6114):1593–1599, Dec 2012.

[87] C. Merritt, D. Rasoloson, D. Ko, and G. Seydoux. 3' UTRs are the primary regulators of gene expression in the C. elegans germline. *Curr. Biol.*, 18(19): 1476–1482, Oct 2008.

[88] J. Meunier, F. Lemoine, M. Soumillon, A. Liechti, M. Weier, K. Guschanski, H. Hu, P. Khaitovich, and H. Kaessmann. Birth and expression evolution of mammalian microRNA genes. *Genome Res.*, 23(1):34–45, Jan 2013.

[89] A. Minis, D. Dahary, O. Manor, D. Leshkowitz, Y. Pilpel, and A. Yaron. Subcellular transcriptomics-dissection of the mRNA composition in the axonal compartment of sensory neurons. *Dev Neurobiol*, 74(3):365–381, Mar 2014.

[90] P. Miura, S. Shenker, J. O. Westholm, C. Andreu-Agullo, and E. C. Lai. Widespread Lengthengin of 3'UTRs in the Mammalian Brain. *Submitted*, Jul 2012.

[91] P. Miura, S. Shenker, C. Andreu-Agullo, J. O. Westholm, and E. C. Lai. Widespread and extensive lengthening of 3' UTRs in the mammalian brain. *Genome Res.*, 23(5):812–825, May 2013.

[92] P. Miura, P. Sanfilippo, S. Shenker, and E. C. Lai. Alternative polyadenylation in the nervous system: to what lengths will 3' UTR extensions take us? *Bioessays*, 36(8):766–777, Aug 2014.

[93] J. Mohammed, D. Bortolamiol-Becet, A. S. Flynt, I. Gronau, A. Siepel, and E. C. Lai. Adaptive evolution of testis-specific, recently evolved, clustered miRNAs in Drosophila. *RNA*, 20(8):1195–1209, Aug 2014.

[94] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5(7): 621–628, Jul 2008.

[95] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, Jun 2008.

[96] S. Negrini, V. G. Gorgoulis, and T. D. Halazonetis. Genomic instability–an evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.*, 11(3):220–228, Mar 2010.

[97] N. Nguyen, G. Hickey, B. J. Raney, J. Armstrong, H. Clawson, A. Zweig, D. Karolchik, W. J. Kent, D. Haussler, and B. Paten. Comparative assembly hubs: web-accessible browsers for comparative genomics. *Bioinformatics*, 30(23):3293–3301, Dec 2014.

[98] T. K. Ni and C. Kuperwasser. Premature polyadenylation of MAGI3 produces a dominantly-acting oncogene in human breast cancer. *Elife*, 5, 2016.

[99] X. Ni, Y. E. Zhang, N. Negre, S. Chen, M. Long, and K. P. White. Adaptive evolution and the birth of CTCF binding sites in the Drosophila genome. *PLoS Biol.*, 10(11):e1001420, 2012.

[100] P. Oikonomou, H. Goodarzi, and S. Tavazoie. Systematic identification of regulatory elements in conserved 3' UTRs of human transcripts. *Cell Rep*, 7(1): 281–292, Apr 2014.

[101] K. Okamura. Diversity of animal small RNA pathways and their biological utility. *Wiley Interdiscip Rev RNA*, 3(3):351–368, 2012.

[102] K. Okamura, M. D. Phillips, D. M. Tyler, H. Duan, Y. T. Chou, and E. C. Lai. The regulatory activity of microRNA* species has substantial influence on microRNA and 3' UTR evolution. *Nat. Struct. Mol. Biol.*, 15(4):354–363, Apr 2008.

[103] S. Pal, R. Gupta, H. Kim, P. Wickramasinghe, V. Baubet, L. C. Showe, N. Dahmane, and R. V. Davuluri. Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res.*, 21(8):1260–1272, Aug 2011.

[104] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, 40(12):1413–1415, Dec 2008.

[105] S. M. Park, J. Ou, L. Chamberlain, T. M. Simone, H. Yang, C. M. Virbasius, A. M. Ali, L. J. Zhu, S. Mukherjee, A. Raza, and M. R. Green. U2AF35(S34F) Promotes Transformation by Directing Aberrant ATG7 Pre-mRNA 3' End Formation. *Mol. Cell*, 62(4):479–490, May 2016.

[106] D. Parkhomchuk, T. Borodina, V. Amstislavskiy, M. Banaru, L. Hallen, S. Krobitsch, H. Lehrach, and A. Soldatov. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.*, 37(18):e123, Oct 2009.

[107] J. K. Pickrell, A. A. Pai, Y. Gilad, and J. K. Pritchard. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.*, 6(12):e1001236, 2010.

[108] A. M. Plocik and B. R. Graveley. New insights from existing sequence data: generating breakthroughs without a pipette. *Mol. Cell*, 49(4):605–617, Feb 2013.

[109] J. Ponjavic, P. L. Oliver, G. Lunter, and C. P. Ponting. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet.*, 5(8):e1000617, Aug 2009.

[110] C. P. Ponting and T. G. Belgard. Transcribed dark matter: meaning or myth? *Hum. Mol. Genet.*, 19(R2):R162–168, Oct 2010.

[111] S. L. Prescott, R. Srinivasan, M. C. Marchetto, I. Grishina, I. Narvaiza, L. Selleri, F. H. Gage, T. Swigut, and J. Wysocka. Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell*, 163(1):68–83, Sep 2015.

[112] N. J. Proudfoot and G. G. Brownlee. 3' non-coding region sequences in eukaryotic messenger RNA. *Nature*, 263(5574):211–214, Sep 1976.

[113] J. Quackenbush, J. Cho, D. Lee, F. Liang, I. Holt, S. Karamycheva, B. Parvizi, G. Pertea, R. Sultana, and J. White. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.*, 29(1):159–164, Jan 2001.

[114] M. Rabani, R. Raychowdhury, M. Jovanovic, M. Rooney, D. J. Stumpo, A. Pauli, N. Hacohen, A. F. Schier, P. J. Blackshear, N. Friedman, I. Amit, and A. Regev. High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell*, 159(7):1698–1710, Dec 2014.

[115] D. Ramskold, E. T. Wang, C. B. Burge, and R. Sandberg. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.*, 5(12):e1000598, Dec 2009.

[116] J. A. Reuter, D. V. Spacek, and M. P. Snyder. High-throughput sequencing technologies. *Mol. Cell*, 58(4):586–597, May 2015.

[117] K. V. Revanna, D. Munro, A. Gao, C. C. Chiu, A. Pathak, and Q. Dong. A web-based multi-genome synteny viewer for customized data. *BMC Bioinformatics*, 13:190, 2012.

[118] D. R. Riley, S. V. Angiuoli, J. Crabtree, J. C. Dunning Hotopp, and H. Tettelin. Using Sybil for interactive comparative genomics of microbes on the web. *Bioinformatics*, 28(2):160–166, Jan 2012.

[119] J. L. Rinn, M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, E. Segal, and H. Y. Chang. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129(7):1311–1323, Jun 2007.

[120] K. R. Rosenbloom, C. A. Sloan, V. S. Malladi, T. R. Dreszer, K. Learned, V. M. Kirkup, M. C. Wong, M. Maddren, R. Fang, S. G. Heitner, B. T. Lee, G. P. Barber, R. A. Harte, M. Diekhans, J. C. Long, S. P. Wilder, A. S. Zweig, D. Karolchik, R. M. Kuhn, D. Haussler, and W. J. Kent. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.*, 41(Database issue):56–63, Jan 2013.

[121] R. Sandberg, J. R. Neilson, A. Sarma, P. A. Sharp, and C. B. Burge. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science*, 320(5883):1643–1647, Jun 2008.

[122] A. K. Shalek, R. Satija, X. Adiconis, R. S. Gertner, J. T. Gaublomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, J. J. Trombetta, D. Gennert, A. Gnirke, A. Goren, N. Hacohen, J. Z. Levin, H. Park, and A. Regev. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–240, Jun 2013.

[123] S. Shenker, P. Miura, P. Sanfilippo, and E. C. Lai. IsoSCM: improved and alternative 3' UTR annotation using multiple change-point inference. *RNA*, 21 (1):14–27, Jan 2015.

[124] P. J. Shepard, E. A. Choi, J. Lu, L. A. Flanagan, K. J. Hertel, and Y. Shi. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*, 17(4):761–772, Apr 2011.

[125] Y. Shi, D. C. Di Giammartino, D. Taylor, A. Sarkeshik, W. J. Rice, J. R. Yates, J. Frank, and J. L. Manley. Molecular architecture of the human pre-mRNA 3' processing complex. *Mol. Cell*, 33(3):365–376, Feb 2009.

[126] G. Singh, I. Rebbapragada, and J. Lykke-Andersen. A competition between stimulators and antagonists of Upf complex recruitment governs human nonsense-mediated mRNA decay. *PLoS Biol.*, 6(4):e111, Apr 2008.

[127] P. Singh, T. L. Alley, S. M. Wright, S. Kamdar, W. Schott, R. Y. Wilpan, K. D. Mills, and J. H. Graber. Global changes in processing of mRNA 3' untranslated regions characterize clinically distinct cancer subtypes. *Cancer Res.*, 69(24): 9422–9430, Dec 2009.

[128] P. Smibert, P. Miura, J. O. Westholm, S. Shenker, G. May, M. O. Duff, D. Zhang, B. D. Eads, J. Carlson, J. B. Brown, R. C. Eisman, J. Andrews, T. Kaufman, P. Cherbas, S. E. Celniker, B. R. Graveley, and E. C. Lai. Global Patterns of Tissue-Specific Alternative Polyadenylation in Drosophila. *Cell Rep*, 1(3):277–289, Feb 2012.

[129] N. Spies, C. B. Burge, and D. P. Bartel. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.*, 23(12):2078–2090, Dec 2013.

[130] A. Stark, J. Brennecke, N. Bushati, R. B. Russell, and S. M. Cohen. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, 123(6):1133–1146, Dec 2005.

[131] T. Steijger, J. F. Abril, P. G. Engstrom, F. Kokocinski, T. J. Hubbard, R. Guigo, J. Harrow, P. Bertone, J. F. Abril, M. Akerman, T. Alioto, G. Ambrosini, S. E. Antonarakis, J. Behr, P. Bertone, R. Bohnert, P. Bucher, N. Cloonan, T. Derrien, S. Djebali, J. Du, S. Dudoit, P. Engstrom, M. Gerstein, T. R. Gingeras, D. Gonzalez, S. M. Grimmond, R. Guigo, L. Habegger, J. Harrow, T. J. Hubbard, C. Iseli, G. Jean, A. Kahles, F. Kokocinski, J. Lagarde, J. Leng, G. Lefebvre, S. Lewis, A. Mortazavi, P. Niermann, G. Ratsch, A. Reymond, P. Ribeca, H. Richard, J. Rougemont, J. Rozowsky, M. Sammeth, A. Sboner, M. H. Schulz, S. M. Searle, N. D. Solorzano, V. Solovyev, M. Stanke, T. Steijger, B. J. Stevenson, H. Stockinger, A. Valsesia, D. Weese, S. White, B. J. Wold, J. Wu, T. D. Wu, G. Zeller, D. Zerbino, and M. Q. Zhang. Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods*, 10(12):1177–1184, Dec 2013.

[132] T. Sterne-Weiler, R. T. Martinez-Nunez, J. M. Howard, I. Cvitovik, S. Katzman, M. A. Tariq, N. Pourmand, and J. R. Sanford. Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res.*, 23(10):1615–1623, Oct 2013.

[133] M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M. L. Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–960, Aug 2008.

[134] A. X. Sun, G. R. Crabtree, and A. S. Yoo. MicroRNAs: regulators of neuronal fate. *Curr. Opin. Cell Biol.*, 25(2):215–221, Apr 2013.

[135] Y. Takagaki, L. C. Ryner, and J. L. Manley. Four factors are required for 3'-end cleavage of pre-mRNAs. *Genes Dev.*, 3(11):1711–1724, Nov 1989.

[136] Y. Takagaki, R. L. Seipelt, M. L. Peterson, and J. L. Manley. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. *Cell*, 87(5):941–952, Nov 1996.

[137] J. M. Taliaferro, M. Vidaki, R. Oliveira, S. Olson, L. Zhan, T. Saxena, E. T. Wang, B. R. Graveley, F. B. Gertler, M. S. Swanson, and C. B. Burge. Distal Alternative Last Exons Localize mRNAs to Neural Projections. *Mol. Cell*, 61 (6):821–833, Mar 2016.

[138] B. Tian, J. Hu, H. Zhang, and C. S. Lutz. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.*, 33(1):201–212, 2005.

[139] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28(5):511–515, May 2010.

[140] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, 7(3):562–578, Mar 2012.

[141] I. Ulitsky, A. Shkumatava, C. H. Jan, A. O. Subtelny, D. Koppstein, G. W. Bell, H. Sive, and D. P. Bartel. Extensive alternative polyadenylation during zebrafish development. *Genome Res.*, 22(10):2054–2066, Oct 2012.

[142] H. van Bakel, C. Nislow, B. J. Blencowe, and T. R. Hughes. Most "dark matter" transcripts are associated with known genes. *PLoS Biol.*, 8(5):e1000371, May 2010.

[143] L. Velten, S. Anders, A. Pekowska, A. I. Jarvelin, W. Huber, V. Pelechano, and L. M. Steinmetz. Single-cell polyadenylation site mapping reveals 3' isoform choice variability. *Mol. Syst. Biol.*, 11(6):812, Jun 2015.

[144] C. Vogel, R. d. e. S. Abreu, D. Ko, S. Y. Le, B. A. Shapiro, S. C. Burns, D. Sandhu, D. R. Boutz, E. M. Marcotte, and L. O. Penalva. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.*, 6:400, Aug 2010.

[145] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, Nov 2008.

[146] J. O. Westholm, P. Miura, S. Olson, S. Shenker, B. Joseph, P. Sanfilippo, S. E. Celniker, B. R. Graveley, and E. C. Lai. Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Rep*, 9(5):1966–1980, Dec 2014.

[147] A. Wiestner, M. Tehrani, M. Chiorazzi, G. Wright, F. Gibellini, K. Nakayama, H. Liu, A. Rosenwald, H. K. Muller-Hermelink, G. Ott, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. Vose, J. O. Armitage, R. D. Gascoyne, J. M. Connors, E. Campo, E. Montserrat, F. Bosch, E. B. Smeland, S. Kvaloy, H. Holte, J. Delabie, R. I. Fisher, T. M. Grogan, T. P. Miller, W. H. Wilson, E. S. Jaffe, and L. M. Staudt. Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival. *Blood*, 109(11):4599–4606, Jun 2007.

[148] E. M. Wissink, E. A. Fogarty, and A. Grimson. High-throughput discovery of post-transcriptional cis-regulatory elements. *BMC Genomics*, 17:177, 2016.

[149] J. Yan and T. G. Marr. Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Res.*, 15(3):369–375, Mar 2005.

[150] E. Yang, E. van Nimwegen, M. Zavolan, N. Rajewsky, M. Schroeder, M. Magnasco, and J. E. Darnell. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.*, 13(8):1863–1872, Aug 2003.

[151] M. Yassour, T. Kaplan, H. B. Fraser, J. Z. Levin, J. Pfiffner, X. Adiconis, G. Schroth, S. Luo, I. Khrebtukova, A. Gnirke, C. Nusbaum, D. A. Thompson, N. Friedman, and A. Regev. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, 106(9): 3264–3269, Mar 2009.

[152] O. K. Yoon, T. Y. Hsu, J. H. Im, and R. B. Brem. Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells. *PLoS Genet.*, 8(8):e1002882, 2012.

[153] D. Yudin, S. Hanz, S. Yoo, E. Iavnilovitch, D. Willis, T. Gradus, D. Vuppalanchi, Y. Segal-Ruder, K. Ben-Yaakov, M. Hieda, Y. Yoneda, J. L. Twiss, and M. Fainzilber. Localized regulation of axonal RanGTPase controls retrograde injury signaling in peripheral nerve. *Neuron*, 59(2):241–252, Jul 2008.

[154] H. Zhang, J. Y. Lee, and B. Tian. Biased alternative polyadenylation in human tissues. *Genome Biol.*, 6(12):R100, 2005.

[155] Y. Zhang, K. Chen, S. A. Sloan, M. L. Bennett, A. R. Scholze, S. O'Keeffe, H. P. Phatnani, P. Guarnieri, C. Caneda, N. Ruderisch, S. Deng, S. A. Liddelow,

C. Zhang, R. Daneman, T. Maniatis, B. A. Barres, and J. Q. Wu. An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J. Neurosci.*, 34(36):11929–11947, Sep 2014.

[156] H. Zhao, Z. Sun, J. Wang, H. Huang, J. P. Kocher, and L. Wang. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, 30(7):1006–1007, Apr 2014.

[157] L. Zhao, P. Saelao, C. D. Jones, and D. J. Begun. Origin and spread of de novo genes in Drosophila melanogaster populations. *Science*, 343(6172):769–772, Feb 2014.