Aptardi accurately incorporates expressed polyadenylation sites into sample-specific transcriptomes using a multi-omics deep learning approach

Authors

Ryan Lusk*, Evan Stene, Farnoush Banaei-Kashani, Boris Tabakoff, Katerina Kechris, and Laura M. Saba

Affiliations

Department of Pharmaceutical Sciences, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA

Ryan Lusk, Boris Tabakoff, & Laura M. Saba

Department of Computer Science and Engineering, University of Colorado Denver, Denver, Colorado, USA

Evan Stene & Farnoush Banaei-Kashani

Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA

Katerina Kechris

Abstract

High throughput sequencing technologies – now standard in omics studies – gave rise to rapid advances in bioinformatics to analyze these large datasets. Transcriptome assemblers harness the power of short-read RNA sequencing to assess the expressed transcriptome on a per sample basis. Yet annotation of polyadenylation sites from short-read RNA sequencing alone is a difficult computational task. Other algorithms rooted in DNA sequence predict potential polyadenylation sites; however, *in vivo* expression of a particular site varies based on a myriad of conditions. Here we introduce aptardi (alternative polyadenylation transcriptome analysis from RNA-Seq data and DNA sequence information), which leverages both DNA sequence and RNA sequencing in a machine learning paradigm to predict expressed polyadenylation sites. Specifically, as input aptardi takes DNA nucleotide sequence, genome-aligned RNA-Seq data, and an initial transcriptome. The program evaluates these initial transcripts to identify expressed polyadenylation sites in the biological sample and refines transcript 3' ends accordingly. The average precision of the aptardi model is twice that of a standard transcriptome assembler. In particular, the recall of the aptardi model (the proportion of true polyadenylation sites detected by the algorithm) is improved by over three-fold. Also, the model – trained using the Human Brain Reference RNA commercial standard – performs well when applied to RNA sequencing samples from different tissues and different mammalian species. Finally, aptardi's input is simple to compile and its output is easily amenable to downstream analyses such as quantitation and differential expression.

Main

Alternative polyadenylation (APA) is gene regulation mechanism by which a single gene encodes multiple RNA isoforms with different polyadenylation (polyA) sites[1] (i.e. different transcription stop sites/3' termini). Most APA sites lead to identical protein products but variable 3' untranslated region lengths[2]. APA has been associated with disease through many transcripts displaying APA (e.g. cardiac hypertrophy[3], oculopharyngeal muscular dystrophy[4,5], breast cancer, and lung cancer[6]) and APA in an individual transcript (e.g. Fabry disease[7], amyotrophic lateral sclerosis[8], metachromatic leukodystrophy[9], and facioscapulohumeral muscular dystrophy[10]). Furthermore, differences in expression of APA transcripts have been implicated in diseases[11] and are recognized as risk factors in complex diseases[12]. Indeed, research suggests individual susceptibility to complex diseases is mainly due to variation in gene regulation processes – such as APA – rather than variation in protein coding sequence[13-16]. APA's impact is expected given that it is pervasive, with more than 70% of human genes subjected to APA[17], and also far-reaching, as it modulates mRNA stability, translation, nuclear export, and cellular localization, as well as the localization of the encoded protein[2,18] – often times through differences in microRNA binding availability.

APA patterns are tissue specific[19,20], and "choice" of polyA sites can be influenced by physiological, environment, and disease states[1,21]. This dynamic may explain – at least in part – why polyA sites are often under annotated[22] and, furthermore, why (the often times sparse) prior annotation is typically not relevant to the given set of experimental conditions[23]. As a result, polyA sites often need to be re-defined for the sample(s) of interest to gain insight into

the role of APA in various processes and diseases (e.g. are certain APA transcripts biomarkers of, or therapeutic targets for, a given disease state?). There are three broad sequencing technologies utilized to identify polyA sites: 1) short-read RNA sequencing (RNA-Seq), 2) direct 3' end RNA sequencing, and 3) DNA sequence, but each possesses inherent limitations for sample-specific identification of polyA sites.

Next generation RNA-Seq has become the standard technology to profile the expressed transcriptome. The resulting short reads are used by transcriptome assemblers to produce a genome-scale, sample-specific transcriptome map. Transcriptome assembly has proven a powerful approach to assess the transcriptome, but accurate determination of polyA sites from short read RNA-Seq alone is a known shortcoming[1,24-27]. Unlike splice junctions which can be precisely located via reads that span the junctions, polyA sites are characterized by a gradual drop off in coverage[28]. For assemblers that harness prior annotation to guide the reconstruction, often the annotated polyA site assumed by the assembler is not correct[22]. While many transcriptome assemblers have been developed – each with its own design – to our knowledge none have demonstrated competence at annotating 3' ends. Some assemblers, e.g. Cufflinks[29]/StringTie[22], construct a minimum path RNA-Seq cover to the position where there is zero read coverage to annotate the 3' end of a transcript[28,30]; but, since reads can be derived from precursor mRNA[31], this often results in an overestimation of polyA sites. Others, e.g. Scripture[32], calculate scan statistics above genomic background to define transcript structures, but this approach tends to produce biased estimates of polyA sites and in general is not well-suited for defining 3' ends[33]. Importantly, these strategies are only capable of producing a

single transcript stop site per intron chain structure, which tends to be the distal polyA site, thereby missing imbedded proximal polyA sites, i.e. APA isoforms[28]. The challenge of accurately identifying polyA sites is apparent to both the developers and those evaluating assemblers by way of allowing for error at 3' end predictions when assessing accuracy[22,34].

Acknowledging the challenges of annotating polyA sites and design shortcomings of transcriptome assemblers to do so, researchers have developed supplemental tools to characterize APA dynamics from RNA-Seq. Chen et al.[35] provided a comprehensive critical review of these methods which we will briefly highlight here. There are three main methods for characterizing polyA sites and/or quantifying APA dynamics. Those that require *a priori* annotated polyA sites, e.g. MISO[36], QAPA[37], and PARQ[38], cannot identify *de novo* polyA sites. Others such as Kleat[39] and ContextMap 2[40] utilize reads with strings of adenosines not derived from a DNA template, i.e. polyA tails. However, studies have demonstrated that polyA reads are extremely scarce in RNA-Seq data[41,42], resulting in low sensitivity and missing more weakly expressed polyA sites. Finally, those that consider fluctuations in read coverage near the 3' ends of transcripts, e.g. DaPars[43], APAtrap[44], and TAPAS[45], are largely interested in single gene APA switching and/or quantifying differential APA usage between two groups of samples rather than producing a complete transcriptome. Also – as noted by Chen et al.[35] – these tools are not user-friendly; specific input formats are required and outputs are not readily integrable into downstream studies.

An alternative approach is to directly capture 3' ends of mRNA with sequencing technology e.g. PolyA-Seq[46], 3' READS[47], PAS-Seq[48], etc. (see Shi[17], Elkon et al.[49], and Ji et al.[50] for a complete review). These methods are extremely accurate at characterizing the genomic locations of polyA sites; however, whereas RNA-Seq data are widely available, 3' sequencing data represents only a small fraction of available sequencing data and is costly and labor intensive to produce[28,35].

A final category of algorithms have sought to capitalize on the wealth of research connecting specific strings of DNA nucleotides, or DNA sequence elements, to polyadenylation (see Tian and Gaber[51] for a detailed review). Most of these methods, e.g. DeepPASTA[52], Omni-PolyA[53], and Conv-Net[54], deploy machine learning but conspicuously do not consider *in vivo* expression.

To overcome current limitations, we introduce aptardi (alternative polyadenylation transcriptome analysis from RNA-Seq data and DNA sequence information). Aptardi leverages the information afforded by DNA nucleotide sequence information (from the appropriate reference genome) and RNA-Seq, as well as the predilection of transcriptome assemblers to accurately characterize splice junctions, in a use-all-data, multi-omics approach to create a modified, sample-specific transcriptome that includes information on expressed polyA sites (Fig. 1). Specifically, harnessing the power of (supervised) machine learning, we trained aptardi to detect polyA sites from DNA nucleotide sequence and RNA-Seq read coverage by training on polyA sites identified by 3' sequencing. Using what it learned, aptardi makes predictions from DNA sequence and RNA-Seq alone, alleviating the burden of generating 3' sequencing data. The

program evaluates initial transcripts in the input original transcriptome to identify expressed polyA sites in the biological sample and refines transcript 3' ends accordingly and outputs its results to a modified transcriptome (as a General Feature Format [GTF] file). Additionally, aptardi's input is simple to compile and its output is easily amenable to downstream analyses such as quantitation and differential expression.

Results.

Construction of multi-omics model for identification of polyadenylation sites.

The initial dataset used for developing the aptardi model was derived from Human Brain Reference[55] (HBR) RNA using Illumina's TruSeq stranded mRNA sample preparation kit to generate 100 base, paired end reads[56]. The transcriptome reconstruction contained 113,923 transcripts (excluding those from scaffold chromosomes) with 94,369 unique transcript termini, and the corresponding PolyA-Seq data contained 94,322 polyA sites in this sample. Throughout this manuscript we refer to the polyA sites identified from PolyA-Seq[46] as "true" polyA sites to distinguish them from polyA sites predicted by a computational algorithm, but we acknowledge that there are false negatives and false positives among the PolyA-Seq derived polyA sites. After processing (see Methods) and integration with the PolyA-Seq data – generated from the same HBR RNA – 70,748 transcript models with zero to 50 true polyA sites per transcript model were used for learning and evaluating the aptardi prediction model. The modified 3' terminal exons of these transcript models were binned into 100 base increments for machine learning, and 14 RNA-Seq features, 12 DNA sequence-related features, and one feature derived from the original transcriptome were calculated for each bin.

Here we present results for the canonical polyA signal (5'-AATAAA-3') feature as an example of a feature derived from DNA sequence (Fig. 2a). This feature was independently associated with the presence of a polyA site ($\chi^2 =$ 80,837, p-value < 0.0001). Of the 100 base bins that possessed a true polyA site via PolyA-Seq, over half also possessed this signal. In contrast, only approximately 10% of bins that were not a polyA site had the signal. This enrichment was observed for all binary features, i.e. all the DNA sequence features and the original transcriptome end location feature. Likewise, the distribution of the quantitative RNA-Seq features, e.g. the inter-bin RNA-Seq features (see Supplementary Information for more details; Fig. 2b) differed based on the presence or absence of a true polyA site although no one feature distinguishes the true polyA sites perfectly. Furthermore, features derived from DNA sequence and RNA-Seq were independent of one another across omics type but were often correlated within an omics category (Fig. 2c).

When aptardi was built using only features derived from RNA-Seq or only features derived from DNA sequence, the average precision (AP) in the testing dataset was significantly greater than simply relying on the polyA sites identified in the original transcriptome (Fig. 2d). Furthermore, when the RNA-Seq features and the DNA sequence features were combined, the multi-omics model had higher AP than either single-omics model (multi-omics AP = 0.58, DNA-only AP = 0.41, RNA-only AP = 0.44). Using a specific prediction threshold (probability > 0.5), the precision in the multi-omics model (0.74) increased from the DNA-only model (0.64) but only modestly increased from the RNA-only model (0.72); however, the recall dramatically improved

compared to both single-omics models (multi-omics recall = 0.38, DNA-only recall = 0.18, RNA-only = 0.24; Fig. 2e). In addition, performance results were consistent across five random splits of the data for training/validation/testing.

Evaluation of the generalizability of aptardi.

To evaluate the generalizability of the aptardi prediction model, we asked two questions: 1) does the performance of the aptardi prediction model, built on the HBR dataset, remain consistent across diverse datasets, and 2) are the performances of prediction models built on alternative datasets comparable to the aptardi prediction model (built from the HBR dataset)? To answer these questions, we analyzed four alternative datasets. These datasets were chosen because they had sufficient similarities and differences to assess the applicability of the aptardi prediction model (Supplementary Table 1). Namely, an additional Human Brain Reference RNA dataset was included that was derived from the same Human Brain Reference RNA sample but processed and sequenced in another laboratory (2nd HBR), and this laboratory also produced another dataset we included from Universal Human Reference (UHR) RNA[57]. To include a cross-species comparison and to examine similar tissue across two genetically different individuals, we also used data derived from two inbred rat strains; the congenic Brown Norway strain with polydactyly-luxate syndrome (BN-Lx/Cub; BNLx) and the spontaneously hypertensive rat strain (SHR/OlaIpcv; SHR). All true polyA sites were derived from 3' sequencing PolyA-Seq data; true polyA sites for the HBR and UHR datasets were from the same corresponding RNA, whereas the true polyA sites for the two rat datasets were derived from Sprague Dawley rat brain RNA[46].

We first examined whether users can confidently apply the aptardi prediction model, built from the HBR dataset, on their own datasets (i.e. on a dataset not used to train the model) by comparing its performance on the four alternative datasets not used to train the model. The AP of the HBR-based aptardi prediction model across the four other datasets ranged from 0.55 to 0.63, whereas the AP of this HBR aptardi prediction model on its own HBR dataset was 0.65 (Fig. 3; orange bars). Specifically, its performance on the other human RNA samples (2nd HBR and UHR) only differed in AP by two percentage points (AP = 0.63 for each), whereas on the BNLx and SHR rat brain datasets the HBR-based aptardi prediction model performed more modestly (AP = 0.55 for each). The major differences between the two rat datasets and the HBR dataset include species, strandedness of the library preparation (rat samples were unstranded), and the inexact matching between RNA-Seq and PolyA-Seq RNA sources.

Also for the four alternative datasets, we built dataset-specific prediction models and compared their performance on their own dataset to the performance of the HBR-based aptardi prediction model on the given dataset to demonstrate the robustness of the machine learning pipeline used to build the aptardi prediction model. For all four datasets, the increase in AP when the same dataset for training the prediction model is used for evaluating the prediction model (as opposed to the performance of the HBR aptardi prediction model on the same given dataset) was minimal, i.e. less than or equal to 2 percentage points (Fig. 3). Furthermore, the similarity (within two percentage points) of the AP between the training, testing, and analysis (i.e. not merging transcripts; see Methods) sets demonstrate the aptardi prediction model is not prone to overfitting and, when these intra- and inter- model/dataset comparisons with the

HBR dataset were extended to each of the four datasets, similar results were achieved (Supplementary Fig. 1).

Improvement of 3' end annotation in the transcriptome map by aptardi.

The overarching goal of aptardi is to yield an updated, sample/experiment-specific transcriptome map from the original transcriptome with more accurately annotated 3' ends of expressed polyadenylated transcripts. As such, aptardi was primarily benchmarked by comparing how it improved upon the reconstruction generated by the popular assembler StringTie[22]. The StringTie assembly also incorporated Ensembl[58] (v.99) annotation, which helps guide its reconstruction – especially at 3' ends. Note that aptardi outputs all original transcript structures from the input transcriptome (i.e. original transcriptome) in addition to those novel annotations identified by the program.

Of the 113,923 transcripts in the original transcriptome from the HBR sample (i.e. StringTie used the HBR sample RNA-Seq data and existing Ensembl annotation to generate the original transcriptome), only 39,842 (35%) had a 3' terminus that corresponded to a true polyA site (+/-100 bases). When the aptardi prediction model was incorporated, 27,853 transcript annotations were added to the original transcriptome where the polyA site/3' terminus differed from its original transcript structure. Of these additional 27,853 transcripts, 22,846 (82%) matched the location of a true polyA site, meaning the majority of aptardi transcript structures incorporated into the original transcriptome had accurate polyA site annotation (Fig. 4). Furthermore, the confusion matrix of predictions made by the aptardi prediction model on

each 100 base increment (i.e. bin) improved the true positive to false positive ratio compared to the original transcriptome (produced by StringTie) while simultaneously decreasing the number of false negatives in favor of true negatives (Supplementary Fig. 2)

We also compared aptardi to TAPAS[45] – identified by Chen et al.[35] as the top performer for characterizing APA from RNA-Seq. The aptardi pipeline identified the 3' termini correctly for 62,688 transcripts compared to 22,804 transcripts using TAPAS. Although the number of transcripts with a false positive 3' terminus was higher in the aptardi modified transcriptome compared to TAPAS[45] due to annotations from the original transcriptome, the positive predictive value was higher for the aptardi modified transcriptome because it added many more true positive than false positive 3' termini to the original transcriptome (aptardi modified transcriptome = 0.44, TAPAS = 0.31). Finally, the aptardi pipeline captured more unique true polyA sites (as identified by PolyA-Seq) compared to both TAPAS and the original transcriptome (aptardi modified transcriptome = 29,327, TAPAS = 25,180, original transcriptome = 23,685).

Aptardi identifies novel sample-specific transcripts missed by current transcriptome reconstruction methods.

We next sought to ascertain if aptardi could identify APA transcripts observed in a previous study where differential APA expression was induced by knocking down the cleavage and polyadenylation machinery CFIm25[59]. In this study, the authors experimentally confirmed expression of short APA transcript isoforms after CFIm25 knockdown for three genes capable of

undergoing APA[60,61] – *CCND1*, *DICER1*, and *TIMP2* – and used DaPars[43] to computationally estimate the locations of polyA sites.

For each the control and knockdown RNA-Seq dataset, the aptardi modified transcriptome was compared to the original transcriptome, which contained both Ensembl annotations and sample-specific expressed transcripts identified through StringTie reconstruction. In the control RNA-Seq dataset, neither aptardi nor the original transcriptome identified a shorter APA transcript for *CCND1* in agreement with the original study design; in the knockdown treatment RNA-Seq dataset, only aptardi recapitulated the short APA isoform (Fig. 5a), demonstrating its sensitivity to sample-specific data and its ability to improve upon current annotation methods. Likewise, only aptardi identified the proximal APA transcript for *DICER1* (Fig. 5b). For *TIMP2*, multiple transcript isoforms are annotated in Ensembl[58], and StringTie[22] retained all these transcripts in its reconstruction. In contrast, aptardi annotated a novel short APA transcript only in the treatment consistent with Masamha et al.[59], again demonstrating its sample-specific sensitivity (Fig. 5c). Finally, the locations of the proximal transcripts for these genes identified by aptardi were similar to the original study (*CCND1*: aptardi = chr11: 69,651,917, original study = chr11: 69,651,578; *DICER1*: aptardi = chr11: 95,090,264, original study = chr11: 95,090,400; *TIMP2*: aptardi = chr17: 78,855,465, original study = chr17: 78,855,601).

Evaluation of the use of aptardi in a differential expression pipeline.

The influence of aptardi on differential expression analysis was evaluated using the BNLx and SHR rat brain datasets by evaluating transcripts identified as differentially expressed between

strains with the aptardi modified transcriptome (p-value <= 0.001) but not the original transcriptome derived from the StringTie/Ensembl pipeline (p-value >= 0.001). Note that since aptardi incorporates new transcripts into annotation, expression levels of existing transcripts can also change, i.e. transcripts present in both the aptardi modified transcriptome and the original transcriptome may be identified as differentially expressed in one and not the other. A total of 1,166 out of 32,348 transcripts and 918 out of 28,329 transcripts expressed above background were differentially expressed (p-value <= 0.001) using the aptardi modified transcriptome and original transcriptome, respectively. A total of 40 transcripts that could be associated with an Ensembl gene symbol were differentially expressed in the aptardi modified transcriptome but not in the original transcriptome although they had identical structures, including 3' ends, in both (i.e. original transcriptome transcripts). Furthermore, 54 novel aptardi transcripts that could be associated with a gene symbol were differentially expressed and NOT measured/identified in the original transcriptome (Supplementary Table 2). The RNA-Seq read coverage for six of these novel aptardi transcripts are depicted in Fig. 6. For *Unc79* (Fig. 6a), *Sf3b1* (Fig. 6b), *Ptn* (Fig. 6c) and *Ap3b1* (Fig. 6d) the original transcript was differentially expressed in the aptardi modified transcriptome but not the original transcriptome, and for *Zdhhc22* (Fig. 6e) and *RGD1559441* (Fig. 6f) the novel aptardi transcript was differentially expressed. The RNA-seq read coverage across these genes support the presence of the novel aptardi transcripts and differential expression of the various isoforms between strains. Moreover, these results demonstrate that aptardi is capable of identifying both shortening and lengthening events – e.g. four of the six genes were annotated with a shorter transcript by aptardi and two of the six a longer one – as well as identifying novel isoforms across a broad

range of RNA-Seq coverage depths; the peak coverage value for each gene ranged from approximately 200 to 8,000.

Discussion

Aptardi leverages the information afforded by both DNA sequence and short-read RNA-Seq to accurately annotate the polyA sites of expressed transcripts in a biological sample. We first established the applicability of aptardi by showing that 1) a prediction model derived from a single dataset performed well on datasets that differ on technical issues and even species, 2) the process of training the prediction model is generalizable across different types of RNA-Seq data/DNA sequence, and 3) the algorithm is not prone to overfitting. Namely, we showed that the aptardi prediction model provided for users (built from the HBR dataset) performs equally well on RNA-Seq datasets derived from different library preparations, organisms, and with different RNA sequencing depths. We note that aptardi performed modestly worse on the BNLx and SHR datasets and hypothesize this is because the true polyA sites were derived from the Sprague Dawley rat instead of the specific rat strain. This is supported by the fact that prediction models built from the BNLx and SHR datasets and tested on these same datasets performed similarly to the aptardi prediction model built on the HBR data (Supplementary Fig. 1). However, we cannot rule out the possibility that this is due to the unstranded RNA-Seq for these datasets and/or different species. Moreover, the comparable results when models were built and evaluated on a single dataset versus models applied to different datasets from those used to train the model suggest the data processing pipeline/prediction models are generalizable. Finally, the similarity of the average precision estimates between training and

testing sets demonstrate that aptardi is not prone to overfitting. Overall, these results indicate aptardi can be broadly used.

We next established that incorporating aptardi into current transcriptome reconstruction methods improves annotation of 3' ends. This was done by comparing the aptardi modified transcriptome to the original transcriptome assembled using the power of both existing annotation via Ensembl and taking into consideration RNA-Seq coverage via StringTie. Adding aptardi transcripts increased the number of unique true polyA sites captured by the transcriptome and furthermore increased the ratio of true positive to false positive termini compared to the original transcriptome. Aptardi also outperformed TAPAS in these respects, and TAPAS was previously identified as the top performer for identifying polyA sites from RNA-Seq[35].

Applying aptardi in control and CFIm25 knockdown RNA-Seq data demonstrated its 1) sensitivity to sample-specific expression, 2) ability to identify both shortening and lengthening APA events, 3) competence across a broad range of RNA-Seq coverage depths, and 4) ability to improve upon current reconstruction methods. Of interest, in the control for *TIMP2*, aptardi identified several novel 3' ends close to the annotated distal transcript that were not noted in the original study[59]. The RNA-Seq from the control sample displays uneven coverage in this region, meaning aptardi may have uncovered additional, previously unknown isoforms (Fig. 5c). Of note, the library preparation for these RNA-Seq data were unstranded (unlike the data used to generate aptardi), further supporting aptardi's broad applicability. These results also

highlight potential weaknesses of aptardi. For instance, aptardi incorporates a single novel 3' end into multiple transcripts for each gene listed because it does not distinguish transcripts overlapping the same genomic region. This may be somewhat mitigated by curating a more selective input transcriptome. Here the entire Ensembl annotation was provided, which includes many "pseudo" transcripts (e.g. retained introns, nonsense mediated decay, and isoforms only identified computationally), and these transcripts often overlap manually identified mRNAs. Secondly, aptardi will add novel polyA sites for a transcript regardless of if another transcript isoform from the same gene already has a transcript stop site at the given location, as is the case for *TIMP2*. As a result, it is possible that aptardi incorporates a transcript stop site belonging to a different transcript.

We further examined how incorporating aptardi into downstream transcriptome analyses such as differential expression may alter interpretation of results. We found that multiple isoforms – some of which were already present in the original transcriptome and some of which were novel transcripts identified by aptardi – were differentially expressed between BNLx and SHR recombinant inbred rats only when using the aptardi modified transcriptome. Some of these transcripts are derived from genes that have also been implicated in phenotypes related to the SHR rat, such as greater sensitivity to addictive drugs[62] and increased voluntary ethanol consumption[63]. For instance, expression of *Ptn* – which has verified APA sites[64] – is modulated by amphetamine in rat nucleus accumbens[65]. Furthermore, *Unc79* knockout mice displayed hypersensitivity to ethanol, e.g. increased preference for and consumption of alcohol[63]. Undoubtedly further investigation is needed to elucidate the role of *Ptn and Unc79* APA on

addiction phenotypes, but these preliminary results demonstrate how aptardi may help unravel the genetic architecture of complex diseases. Of note, while Ensembl annotation provides two transcript isoforms for *Ap3b1*, StringTie assembly resulted in inclusion of the longer isoform only, highlighting the difficulty of current assembly methods for identifying 3' ends of transcripts embedded in a longer version. However, aptardi identified a shorter transcript within 100 bases of the original Ensembl annotation for the shorter transcript that is likewise supported by the RNA-Seq data (Fig. 6d).

Also of note, aptardi is easily integrable into existing analyses pipelines. For instance, unlike current supplemental methods designed for APA detection from RNA-Seq, no additional data manipulation is required prior to running the program. The input files are readily available (e.g. reference genome and reference transcriptome) or already generated during the course of transcriptomic analysis (e.g. RNA-Seq data and a reconstructed transcriptome). The output GTF file can be used in the same manner as other annotation files (e.g. those accessed via Ensembl or generated via a transcriptome assembler such as StringTie). Moreover, the program can be seamlessly integrated into a single operation with upstream transcriptome assembly and downstream analyses (e.g. quantitation) via piping to make for streamlined analysis. Finally, there is also the option of constructing a new prediction model using the aptardi architecture, which increases the breadth of its applicability to diverse data sources.

An additional perceived limitation may be that aptardi makes predictions on 100 base bins (for positive predictions, it annotates the transcript stop site as the 3' most base in the given 100

base bin). However, this concern is partially mitigated because the precise location of polyA sites can “wiggle” by up to 30 nucleotides for what is considered a single isoform, i.e. not an APA event[51], and as such researchers often group polyA sites within 30 bases into a single site[46]. Furthermore, few 100 base bins contained multiple polyA sites (Supplementary Table 3). Another potential limitation is that upstream exons are not subjected to aptardi analysis, effectively eliminating the possibility of identifying coding APA; fortunately, the vast majority of APA sites do not result in a change within the protein coding region[2]. Finally, the manually engineered DNA sequence features may not apply to taxa outside of mammals[51] and will require further research.

Transcriptome profiling is one of the most utilized approaches for investigating human diseases at the molecular level, yielding important insights into many pathologies. A prerequisite for these studies is a representative transcriptome map. Aptardi incorporates APA transcripts to produce a more accurate transcriptome map, thereby enabling future research into the role of APA transcripts – as well as other transcripts unencumbered by convoluted annotation with APA transcripts – in human health and disease.

Aptardi is implemented in Python and is freely available as open source software (https://github.com/luskry/aptardi).

Methods

Aptardi design.

The overall goal of aptardi is to accurately identify the polyA sites of expressed transcripts in a given biological sample. Specifically, aptardi analyzes the modified 3' terminal exon (see the Transcript processing section below for details on 3' terminal exon modification) of previously annotated transcripts and, using relevant RNA-Seq data and DNA sequence in a machine learning environment, identifies locations of expressed polyA sites in the region. Aptardi then annotates the 3' termini to match these locations and outputs the transcript structures to a newly assembled transcriptome (in GTF format) that can be easily incorporated into to downstream analyses. Note that aptardi does not evaluate the intron chain structure of transcripts, i.e. it only examines the modified 3' terminal exon of each transcript structure and alters the 3' terminus location(s) accordingly. Also note that aptardi outputs all original transcript structures from the original transcriptome in addition to transcripts identified through its analysis, i.e. the program only adds transcripts.

Datasets.

A total of five unique datasets, hereafter referred to as HBR, 2nd HBR, UHR, BNLx, and SHR, were subjected to aptardi's machine learning pipeline (Supplementary Table 1). In addition to RNA-Seq measurements, each dataset required DNA sequence, a transcriptome, and – since each was used to build a machine learning model – a "gold standard" data source providing locations of expressed, i.e. "true" polyA sites. HBR, 2nd HBR, and UHR are well-established RNA reference samples from the MAQC/SEQC consortium[55] (see RNA sequencing datasets for more details). BNLx and SHR represent two inbred rat strains: the congenic Brown Norway strain with

polydactyly-luxate syndrome (BN-Lx/Cub) and the spontaneous hypertensive rat strain (SHR/OlaIpcv), respectively.

DNA sequence datasets.

For BNLx and SHR, strain-specific genomes were generated from the rn6/Rnor_6.0 version of the rat genome[66] and are publicly available on the PhenoGen website (https://phenogen.org/). The human reference genome (hg38/GRCh38), accessed via the UCSC Genome Browser[67] (http://genome.ucsc.edu/), was utilized for the HBR, 2nd HBR, and UHR DNA sequence datasets.

RNA sequencing datasets.

The HBR and 2nd HBR RNA-Seq datasets were derived from the Human Brain Reference (multiple brain regions of 12 donors, Ambion, p/n AM6050), and the UHR RNA-Seq dataset was derived from the Universal Human Reference (10 pooled cancer lines, Stratagene, p/n 740000). Each of these three datasets were accessed from the Sequence Read Archive (SRA) as publicly available data (HBR[56]: Accession: PRJNA510978, SRA runs: SRR8360036-37, 2nd HBR and UHR[57]: Accession: PRJNA362835, 2nd HBR SRA runs: SRR5236425-30, UHR SRA runs: SRR5236455-60), whereas the BNLx and SHR RNA-Seq datasets[68] are available on the PhenoGen website. Briefly, all libraries were generated with the TruSeq stranded (HBR, 2nd HBR, UHR) or unstranded (BNLx and SHR) mRNA sample preparation kit (Illumina), sequenced on a HiSeq2500 Instrument (Illumina), and sequencing results processed to FASTQ files. The HBR RNA-Seq dataset originated from 1 µg RNA starting material, while 100 ng input was used for 2nd HBR and UHR RNA-Seq datasets. For more detailed descriptions on these publicly available data, see

Palomares et al.[56] (HBR) and Schuierer et al.[57] (2nd HBR and UHR). For BNLx and SHR, RNA-seq libraries prepared from the polyA+ fraction were constructed using the Illumina TruSeq RNA Sample Preparation kit from one µg of brain RNA in accordance with the manufacturer's instructions. Four µL of a 1:100 dilution of either ERCC Spike-In Mix 1 or Mix 2 (ThermoFisher Scientific) were added to each extracted RNA sample. An Agilent Technologies Bioanalyzer 2100 (Agilent Technologies) was utilized to assess sequencing library quality. RNA samples from three biological replicates per strain were processed and sequenced[68]. All reads were paired-end but differed in read length (HBR, BNLx, and SHR: 2X100, 2nd HBR and UHR: 2X75). Individual FASTQ files were assessed for quality using FastQC[69] (v.0.11.4) and, if necessary, reads were trimmed with cutadapt[70] (v.1.9.1). For the purpose of read coverage used by the aptardi algorithm, reads from technical replicates (HBR = 2, 2nd HBR = 3, UHR = 3) or biological replicates (BNLx = 3, SHR = 3) were concatenated and aligned to their respective genomes (see DNA sequence datasets section for more details) using HISAT2[71] (v.2.1.0) with the --rna-strandness (when appropriate) and --dta options specified as recommended for transcriptome assembly with StringTie[22] (see Transcriptome datasets below for more details) and otherwise default arguments (see Supplementary Table 4 for alignment results). After alignment, SAMtools[72] (v.1.9) was used to remove unmapped reads and convert the output to a sorted Binary Alignment Map (BAM) file required as input by aptardi.

True polyadenylation sites datasets.

True polyA sites (i.e. labels for machine learning) were taken from Derti et al.[46] for all datasets. Namely, total RNA from the same UHR and HBR RNA reference samples, as well as brain total

RNA from the Sprague Dawley rat (Zyagen, p/n RR-201), were subjected to PolyA-Seq analysis to identify the genomic locations of expressed polyA sites in each sample (for more information, see Derti et al.[46]). High quality filtered polyA sites from each RNA sample were accessed using the UCSC Table Browser[73] (http://genome.ucsc.edu/), and liftOver[74] (from the UCSC Genome Browser Group) was used to convert the genomic coordinates to the most recent human genome assembly (hg38/GRCh38) for the HBR and UHR samples, or rat genome assembly (rn6/Rnor_6.0) for the rat brain sample. PolyA sites identified in the HBR and UHR RNA reference samples were used for the corresponding HBR, 2nd HBR and UHR datasets, and those identified in the rat brain were used for both BNLx and SHR. Derti et al.[46] uploaded technical replicates to UCSC Table Browser for each RNA sample; however, since polyA sites within 30 bases were clustered into the single site with greatest expression, and since this was done separately for each dataset, we utilized only a single dataset for each sample, i.e. technical replicates were not combined.

Original transcriptome generation.

StringTie[22] (v.1.3.5) was used to reconstruct the transcriptome expressed in each dataset from their RNA-Seq data, hereafter referred to as the original transcriptome. Ensembl[58] (v.99) reference annotation from the respective species was provided to guide the StringTie reconstruction. We note that a user can simply use reference annotation directly, i.e. Ensembl annotation, in lieu of performing transcriptome assembly that takes into account expression, i.e. StringTie. If the RNA-Seq data were stranded, the read orientation was specified as an

argument to StringTie. Transcript structures from scaffold chromosomes and unstranded contigs, if present, were removed.

The data processing pipeline.

Transcript processing.

Using the original transcriptome, the 3' terminal exons of transcripts were isolated. Each transcript's 3' terminal exon was extended 10,200 bases similar to what has been done previously[43,44]. Extensions overlapping any neighboring transcripts (on the same strand) were shortened to remove the overlap. RNA-Seq coverage at single-nucleotide resolution was obtained via bedtools genomecov (BEDtools[75]; v2.29.2) and, similar to the criteria employed by Ye et al.[44] and Miura et al.[76], used to refine each of these 3' terminal exons, hereafter referred to as modified 3' terminal exons (see Supplementary Information for more details.) The refinement step either shortened the extended 3' terminal exon or kept it the same length to give the modified 3' terminal exon.

Feature extraction.

Features were engineered in 100 base increments along the modified 3' terminal exon, referred to hereafter as bins. For each bin, a total of 27 features were engineered and can be broadly classified as being derived from DNA sequence or RNA-Seq data. In both cases, information from the local environment, i.e. the 100 bases upstream and downstream the bin, as well as the bin itself, (300 bases total) was used.

DNA sequence features.

The choice of DNA sequence features was made through a combination of an exhaustive literature review[25,51,77-84] and evaluation of other algorithms that use DNA sequence to predict polyA sites[53,54,85,86]. Perhaps the most well-known indicator of polyadenylation is the polyadenylation signal (PAS), a conserved hexamer located ~10-35 nucleotides upstream the polyA site. Overrepresented sequences of DNA, or DNA sequence elements, also influence polyadenylation, and the location of these sequences are often described relative to the PAS. As such, DNA sequence features were engineered by first identifying the presence of several known PAS's. Specifically, for each bin, a six base sliding window scanned a predefined region to detect the presence or absence (binary indicator of 1/-1) of 1) the canonical PAS (AATAAA), 2) its major variant (ATTAAA), 3) a second common variant (AGTAAA), and 4) any one of nine other minor variants (AAGAAA, AAAAAG, AATACA, TATAAA, GATAAA, AATATA, CATAAA, AATAGA[25,51,77,78]) for four total PAS features. Subsequently, regions relative to the PAS (if present, otherwise predefined regions relative to the current bin) were likewise scanned using a sliding window approach to determine frequency of the following known DNA sequence elements: 1) a G-rich region downstream the PAS, 2) a downstream region near the PAS enriched in TTT, 3) a downstream region near the PAS enriched in GT/TG, and 4) a downstream region near the PAS enriched in GTGT/TGTG, 5) a T-rich region immediately downstream of the PAS, 6) a T-rich region upstream the PAS, 7) a TGTA/TATA-rich region upstream the PAS, and 8) a AT-rich region upstream and downstream the PAS[25,51,79-84] for an additional 8 features (12 DNA sequence features total). If the frequency of the given DNA sequence element was above an enrichment threshold, the feature was encoded 1, otherwise -1. (See Supplementary Information for more details.)

RNA sequencing features.

From the RNA-Seq data, coverage at single nucleotide resolution was determined using BEDtools[75] (v2.29.2). The approach for designing RNA-Seq features was to exploit localized fluctuations in RNA read coverage similar to that implemented by tools designed for APA-specific analysis from RNA-Seq data[43-45]. Intuitively, upstream but in close proximity to the end of a transcript, coverage is expected to begin to decrease gradually until its end. As a result, changes in expression were utilized when designing RNA-Seq features in two scenarios: 1) intra- and 2) inter-bin. In both cases, three regions were defined: an upstream region, a middle region, and a downstream region (Supplementary Fig. 5). Changes in expression between these regions were quantified using various mathematical combinations of coverage values in each region to generate 14 unique features (see Supplementary Information for more details). To account for local variability in RNA-Seq coverage, median coverage values in each region were used.

A final RNA-Seq feature was derived from the original transcriptome. If the 3' base of any annotated transcript from the original reconstruction was located within a bin, this feature was encoded 1, otherwise -1. Supplementary Fig. 4 summarizes the data processing pipeline prior to machine learning.

Building aptardi.

The machine learning task is two class classification (polyA site or no polyA site) of each 100 base bin. Supervised learning was used where labels for training were provided from the polyA

sites datasets (see PolyA sites datasets for more details). A bidirectional long short term memory recurrent neural network[87-90] was implemented using the Keras framework (http://keras.io). This machine learning paradigm was chosen because of its design to analyze sequential data, i.e. it takes into account all the 100 base bins of a given transcript when learning model parameters for each individual bin. Each direction of the bidirectional long short term memory recurrent neural network consisted of 20 nodes (40 total), and this layer was followed by a fully connected dense layer with a sigmoid activation function that outputs a probability value.

Training aptardi.

To prevent duplicate bins, overlapping modified 3' terminal exons were merged prior to training. Additionally, all merged modified 3' terminal exons were masked to a length of 300 bins (30,000 bases total) to generate equal lengths, which is required for the sequential model. A total of 778,166 bins were present, of which 42,977 possessed a polyA site. Merged modified 3' terminal exons were split into 60/20/20 training, validation, testing sets, respectively. Quantitative measures were standardized using the training set, and the training set was used to build the model in 25 epochs. Model weights were optimized using a binary cross entropy loss function and Adam[91] optimizer. Precision and recall metrics on the training and validation sets were monitored during training to prevent overfitting, and the model that produced the minimum loss was kept. For evaluation purposes (see Results), individual prediction models were generated from each of the five datasets.

Evaluating aptardi.

Precision and recall at the default probability threshold (0.5) were used to evaluate model performance defined as follows:

$$P = \frac{T_p}{T_p + F_p} \quad (1)$$

$$R = \frac{T_p}{T_p + F_n} \quad (2)$$

where $T_p = true\ positive$, and $F_p = false\ positive$, $F_n = false\ negative$, $P = precision$, and $R = recall$

To generalize model performance over the range of probability thresholds, average precision was used in place of the receiver operating curve due to the highly imbalanced nature of the data[92] (far fewer bins with polyA sites than bins without a polyA site):

$$AP = \sum_n (R_n - R_{n-1}) P_n \quad (3)$$

where $AP = average\ precision$ and R_n and P_n are the precision and recall at the nth threshold, respectively.

Integrating aptardi results with the original transcriptome.

For 100 base bins where a polyA site is predicted, new transcript structures are annotated to the 3' most base position unless either 1) the input transcript's stop site is already in the region or 2) the 3' most base position is within 100 bases of the input transcript's stop site. Any new transcript structures were added to the original transcriptome, and this aptardi modified transcriptome was outputted as a GTF file.

Software.

A user has the option of using the pre-existing aptardi prediction model or building a new prediction model (if a reliable polyA sites dataset is available). The pre-built model, i.e. the aptardi prediction model, provided on GitHub (https://github.com/luskry/aptardi) was generated from the HBR dataset[56]. Several other algorithm options are available.

TAPAS analysis

TAPAS (Tool for Alternative Polyadenylation site AnalysiS) identifies the locations of polyA sites from RNA-Seq and reference annotation (genome or transcriptome)[45]. Its performance was evaluated on HBR, specifically using the HBR RNA-Seq data and StringTie assembled transcriptome. Since TAPAS requires noncoding sequence coordinates for analysis, the 3' terminal exon of each transcript was provided as the noncoding region, and default arguments were used.

CFIm25 knockdown analysis.

RNA-Seq from HeLa cells and RNA-Seq after RNA interference on HeLa cells was used to generate the control RNA-Seq dataset and the treatment CFIm25 knockdown RNA-Seq dataset that induces APA switching, respectively. The RNA-Seq datasets were accessed from SRA as publicly available data (Accession: PRJNA182153, control SRA: SRR1238549, CFIm25 knockdown SRA: SRR1238551). The RNA-Seq library preparation was unstranded, and 100 base paired end reads were sequenced on an Illumina HiSeq 2000 instrument (see Masamha et al.[59] for more details). Reads were processed in a manner identical to all other datasets to produce a sorted

BAM file (see Supplementary Table 5 for alignment results). An original transcriptome using StringTie/Ensembl (without the read orientation argument) and an aptardi modified transcriptome were generated for each RNA-Seq dataset (four total). The aptardi prediction model produced using the HBR dataset was used to generate the aptardi modified transcriptomes.

Rat differential expression analysis.

To generate a single transcriptome representing both rat strains, their genome-aligned RNA-Seq data were merged using SAMtools[72] (v.1.9) followed by the production of an original transcriptome using the merged RNA-Seq dataset and StringTie/Ensembl (without the read orientation argument). The aptardi modified transcriptome was produced using the original transcriptome as input, along with the merged RNA-Seq data, the rn6/Rnor_6 DNA sequence (accessed via the UCSC Genome Browser[67]; http://genome.ucsc.edu/), and the aptardi prediction model built from HBR. RSEM[93] (v.1.2.31) was used to estimate the abundances of the isoforms identified within each transcriptome (the original transcriptome and aptardi modified transcriptome). Prior to quantitation, transcripts from scaffold chromosomes and unstranded contigs were removed from both transcriptomes. Isoform level expression estimates were determined for each biological sample (BNLx = 3, SHR = 3). Isoforms without at least 50 counts in two of the three biological replicates for at least one strain were removed, and differential expression between the two strains (with BNLx as reference) were evaluated using DESeq2[94] (v.1.28.0) for the remaining set of isoforms in each transcriptome (Supplementary Fig. 5 summarizes these analysis steps). A significance threshold of 0.001 was applied to the

unadjusted p-values to allow for comparisons across the two datasets (original transcriptome and aptardi modified transcriptome) that differ in the number of transcripts tested.

Acknowledgements

We would like to thank Drs. Richard A. Radcliffe, Paula L. Hoffman, and Peter L. Anderson for their helpful discussions and Spencer Mahaffey for his help managing the data. This research was supported by the following US National Institutes of Health (NIH) grants: NIAAA F31AA027430, NIAAA R24AA013162, and NIDA P30DA044223.

Contributions

Conceptualization: R.L and L.S.; methodology: R.L., L.S., B.T., and K.K.; machine learning: L.S., R.L., E.S., and F.B.; data curation: R.L., L.S., and B.T.; implementation: R.L.; analysis: R.L., L.S., B.T., and K.K.; visualization: R.L. and L.S.; software: R.L. and E.S.; funding acquisition: R.L., L.S., and B.T.; supervision: L.S.; writing: R.L. and L.S. with input from all authors

Ethics declarations

Competing interests

The authors declare no competing interests.

Data availability

All data were accessed through free, publicly available means:

UCSC Genome Browser (http://genome.ucsc.edu/):

1. Human reference genome (hg38/GRCh38)
2. Rat reference genome (rn6/Rnor_6.0)

NCBI Sequence Read Archive (SRA)

1. Human brain reference RNA sequencing (Accession: PRJNA510978, SRA runs: SRR5236425-30)
2. 2nd human brain reference and universal human reference RNA sequencing: 2nd HBR SRA runs: Accession: PRJNA362835, SRR5236425-30, UHR SRA runs: SRR5236455-60)
3. Control vs CFIm25 knockdown RNA sequencing (Accession: PRJNA182153, control SRA: SRR1238549, CFIm25 knockdown SRA: SRR1238551)

PhenoGen (https://phenogen.org/):

1. BN-Lx/Cub strain-specific genome
2. SHR/OlaIpcv strain-specific genome
3. BN-Lx/Cub and SHR/OlaIpcv RNA sequencing

Code availability

The software aptardi is maintained on GitHub repository (https://github.com/luskry/aptardi).

References

1 Di Giammartino, D. C., Nishida, K. & Manley, J. L. Mechanisms and consequences of alternative polyadenylation. *Molecular cell* **43**, 853-866, doi:10.1016/j.molcel.2011.08.017 (2011).

2 Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nature reviews. Molecular cell biology* **18**, 18-30, doi:10.1038/nrm.2016.116 (2017).

3 Park, J. Y. *et al.* Comparative analysis of mRNA isoform expression in cardiac hypertrophy and development reveals multiple post-transcriptional regulatory modules. *PLoS ONE* **6**, e22391, doi:10.1371/journal.pone.0022391 (2011).

4 de Klerk, E. *et al.* Poly(A) binding protein nuclear 1 levels affect alternative polyadenylation. *Nucleic Acids Research* **40**, 9089-9101, doi:10.1093/nar/gks655 (2012).
5 Jenal, M. *et al.* The Poly(A)-Binding Protein Nuclear 1 Suppresses Alternative Cleavage and Polyadenylation Sites. *Cell* **149**, 538-553, doi:10.1016/j.cell.2012.03.022 (2012).
6 Lembo, A., Di Cunto, F. & Provero, P. Shortening of 3'UTRs correlates with poor prognosis in breast and lung cancer. *PLoS One* **7**, e31129, doi:10.1371/journal.pone.0031129 (2012).
7 Bishop, D. F., Kornreich, R. & Desnick, R. J. Structural organization of the human alpha-galactosidase A gene: further evidence for the absence of a 3' untranslated region. *Proceedings of the National Academy of Sciences* **85**, 3903-3907 (1988).
8 Lin, C. L. *et al.* Aberrant RNA processing in a neurodegenerative disease: the cause for absent EAAT2, a glutamate transporter, in amyotrophic lateral sclerosis. *Neuron* **20**, 589-602 (1998).
9 Gieselmann, V., Polten, A., Kreysing, J. & von Figura, K. Arylsulfatase A pseudodeficiency: loss of a polyadenylylation signal and N-glycosylation site. *Proceedings of the National Academy of Sciences* **86**, 9436-9440 (1989).
10 Lemmers, R. J. L. F. *et al.* A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science* **329**, 1650-1653, doi:10.1126/science.1189044 (2010).
11 Yoon, O. K., Hsu, T. Y., Im, J. H. & Brem, R. B. Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells. *PLoS Genetics* **8**, e1002882, doi:10.1371/journal.pgen.1002882 (2012).
12 Manning, K. S. & Cooper, T. A. The roles of RNA processing in translating genotype to phenotype. *Nature reviews. Molecular cell biology* **18**, 102-114, doi:10.1038/nrm.2016.139 (2016).
13 Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42**, D1001-D1006, doi:10.1093/nar/gkt1229 (2013).
14 Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American journal of human genetics* **94**, 559-573, doi:10.1016/j.ajhg.2014.03.004 (2014).
15 Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600-604, doi:10.1126/science.aad9417 (2016).
16 Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186, doi:10.1016/j.cell.2017.05.038 (2017).
17 Shi, Y. Alternative polyadenylation: new insights from global analyses. *RNA (New York, N.Y.)* **18**, 2105-2117, doi:10.1261/rna.035899.112 (2012).
18 Yong, H.-S. Y. a. J. Alternative Polyadenylation of mRNAs: 3’-Untranslated Region Matters in Gene Expression. *Molecules and cells* **39**, 281-285, doi:10.14348/molcells.2016.0035 (2016).
19 Zhang, H., Lee, J. Y. & Tian, B. Biased alternative polyadenylation in human tissues. *Genome Biol* **6**, R100, doi:10.1186/gb-2005-6-12-r100 (2005).
20 Beaudoing, E. & Gautheret, D. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res* **11**, 1520-1526, doi:10.1101/gr.190501 (2001).

21 Sanfilippo, P., Wen, J. & Lai, E. C. Landscape and evolution of tissue-specific alternative polyadenylation across Drosophila species. *Genome biology* **18**, 229, doi:10.1186/s13059-017-1358-0 (2017).
22 Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**, 290-295, doi:10.1038/nbt.3122 (2015).
23 Schurch, N. J. *et al.* Improved Annotation of 3' Untranslated Regions and Complex Loci by Combination of Strand-Specific Direct RNA Sequencing, RNA-Seq and ESTs. *PLOS ONE* **9**, e94270, doi:10.1371/journal.pone.0094270 (2014).
24 Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* **13**, 233-245, doi:10.1038/nrg3163 (2012).
25 Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research* **33**, 201-212, doi:10.1093/nar/gki158 (2005).
26 Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res* **23**, 169-180, doi:10.1101/gr.139618.112 (2013).
27 Shepard, P. J. *et al.* Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA (New York, N.Y.)* **17**, 761-772, doi:10.1261/rna.2581711 (2011).
28 Shenker, S., Miura, P., Sanfilippo, P. & Lai, E. C. IsoSCM: improved and alternative 3' UTR annotation using multiple change-point inference. *RNA* **21**, 14-27, doi:10.1261/rna.046037.114 (2015).
29 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**, 511-515, doi:10.1038/nbt.1621 (2010).
30 Shao, M., Ma, J. & Wang, S. DeepBound: accurate identification of transcript boundaries via deep convolutional neural fields. *Bioinformatics* **33**, i267-i273, doi:10.1093/bioinformatics/btx267 (2017).
31 Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods* **8**, 469-477, doi:10.1038/nmeth.1613 (2011).
32 Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**, 503-510, doi:10.1038/nbt.1633 (2010).
33 Huber, W., Toedling, J. & Steinmetz, L. M. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* **22**, 1963-1970, doi:10.1093/bioinformatics/btl289 (2006).
34 Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods* **10**, 1177-1184, doi:10.1038/nmeth.2714 (2013).
35 Chen, M. *et al.* A survey on identification and quantification of alternative polyadenylation sites from RNA-seq data. *Brief Bioinform*, doi:10.1093/bib/bbz068 (2019).

36 Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**, 1009-1015, doi:10.1038/nmeth.1528 (2010).
37 Ha, K. C. H., Blencowe, B. J. & Morris, Q. QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biol* **19**, 45, doi:10.1186/s13059-018-1414-4 (2018).
38 Gruber, A. J. *et al.* Discovery of physiological and cancer-related regulators of 3' UTR processing with KAPAC. *Genome Biol* **19**, 44, doi:10.1186/s13059-018-1415-3 (2018).
39 Birol, I. *et al.* Kleat: cleavage site analysis of transcriptomes. *Pac Symp Biocomput*, 347-358 (2015).
40 Bonfert, T. & Friedel, C. C. Prediction of Poly(A) Sites by Poly(A) Read Mapping. *PLoS One* **12**, e0170914, doi:10.1371/journal.pone.0170914 (2017).
41 Szkop, K. J. & Nobeli, I. Untranslated Parts of Genes Interpreted: Making Heads or Tails of High-Throughput Transcriptomic Data via Computational Methods: Computational methods to discover and quantify isoforms with alternative untranslated regions. *Bioessays* **39**, doi:10.1002/bies.201700090 (2017).
42 Bayerlova, M. *et al.* Newly Constructed Network Models of Different WNT Signaling Cascades Applied to Breast Cancer Expression Data. *PLoS One* **10**, e0144014, doi:10.1371/journal.pone.0144014 (2015).
43 Xia, Z. *et al.* Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nature Communications* **5**, 5274, doi:10.1038/ncomms6274 (2014).
44 Ye, C., Long, Y., Ji, G., Li, Q. Q. & Wu, X. APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics* **34**, 1841-1849, doi:10.1093/bioinformatics/bty029 (2018).
45 Arefeen, A., Liu, J., Xiao, X. & Jiang, T. TAPAS: tool for alternative polyadenylation site analysis. *Bioinformatics* **34**, 2521-2529, doi:10.1093/bioinformatics/bty110 (2018).
46 Derti, A. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Research* **22**, 1173-1183, doi:10.1101/gr.132563.111 (2012).
47 Hoque, M. *et al.* Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nature Methods* **10**, 133-139, doi:10.1038/nmeth.2288 (2013).
48 Shepard, P. J. *et al.* Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**, 761-772, doi:10.1261/rna.2581711 (2011).
49 Elkon, R., Ugalde, A. P. & Agami, R. Alternative cleavage and polyadenylation: extent, regulation and function. *Nature Reviews Genetics* **14**, 496-506, doi:10.1038/nrg3482 (2013).
50 Ji, G., Guan, J., Zeng, Y., Li, Q. Q. & Wu, X. Genome-wide identification and predictive modeling of polyadenylation sites in eukaryotes. *Brief Bioinform* **16**, 304-313, doi:10.1093/bib/bbu011 (2015).
51 Tian, B. & Graber, J. H. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdisciplinary Reviews: RNA* **3**, 385-396, doi:10.1002/wrna.116 (2012).
52 Arefeen, A., Xiao, X. & Jiang, T. DeepPASTA: deep neural network based polyadenylation site analysis. *Bioinformatics* **35**, 4577-4585, doi:10.1093/bioinformatics/btz283 (2019).

53 Magana-Mora, A., Kalkatawi, M. & Bajic, V. B. Omni-PolyA: a method and tool for accurate recognition of Poly(A) signals in human genomic DNA. *BMC Genomics* **18**, 620, doi:10.1186/s12864-017-4033-7 (2017).
54 Leung, M. K. K., Delong, A. & Frey, B. J. Inference of the human polyadenylation code. *Bioinformatics* **34**, 2889-2898, doi:10.1093/bioinformatics/bty211 (2018).
55 Consortium, M. *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology* **24**, 1151-1161, doi:10.1038/nbt1239 (2006).
56 Palomares, M. A. *et al.* Systematic analysis of TruSeq, SMARTer and SMARTer Ultra-Low RNA-seq kits for standard, low and ultra-low quantity samples. *Sci Rep* **9**, 7550, doi:10.1038/s41598-019-43983-0 (2019).
57 Schuierer, S. *et al.* A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples. *BMC Genomics* **18**, 442, doi:10.1186/s12864-017-3827-y (2017).
58 Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res* **48**, D682-D688, doi:10.1093/nar/gkz966 (2020).
59 Masamha, C. P. *et al.* CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature* **510**, 412-416, doi:10.1038/nature13261 (2014).
60 Mayr, C. & Bartel, D. P. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**, 673-684, doi:10.1016/j.cell.2009.06.016 (2009).
61 Kubo, T., Wada, T., Yamaguchi, Y., Shimizu, A. & Handa, H. Knock-down of 25 kDa subunit of cleavage factor Im in Hela cells alters alternative polyadenylation within 3'-UTRs. *Nucleic Acids Research* **34**, 6264-6271, doi:10.1093/nar/gkl794 (2006).
62 Pamplona, F. A., Vendruscolo, L. F. & Takahashi, R. N. Increased sensitivity to cocaine-induced analgesia in Spontaneously Hypertensive Rats (SHR). *Behav Brain Funct* **3**, 9, doi:10.1186/1744-9081-3-9 (2007).
63 Vendruscolo, L. F., Izidio, G. S. & Takahashi, R. N. Drug reinforcement in a rat model of attention deficit/hyperactivity disorder--the Spontaneously Hypertensive Rat (SHR). *Curr Drug Abuse Rev* **2**, 177-183, doi:10.2174/1874473710902020177 (2009).
64 Papadimitriou, E. *et al.* Pleiotrophin and its receptor protein tyrosine phosphatase beta/zeta as regulators of angiogenesis and cancer. *Biochim Biophys Acta* **1866**, 252-265, doi:10.1016/j.bbcan.2016.09.007 (2016).
65 Le Grevès, P. Pleiotrophin gene transcription in the rat nucleus accumbens is stimulated by an acute dose of amphetamine. *Brain Research Bulletin* **65**, 529-532, doi:https://doi.org/10.1016/j.brainresbull.2005.03.010 (2005).
66 Hoffman, P. L., Saba, L. M., Vanderlinden, L. A. & Tabakoff, B. Voluntary exposure to a toxin: the genetic influence on ethanol consumption. *Mamm Genome* **29**, 128-140, doi:10.1007/s00335-017-9726-3 (2018).
67 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).
68 Saba, L. M. *et al.* The sequenced rat brain transcriptome – its use in identifying networks predisposing alcohol consumption. *The FEBS Journal* **282**, 3556-3578, doi:10.1111/febs.13358 (2015).

69 Andrews, S. *FASTQC. A quality control tool for high throughput sequence data\ \ }.* (2010).
70 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10, doi:10.14806/ej.17.1.200 (2011).
71 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915, doi:10.1038/s41587-019-0201-4 (2019).
72 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
73 Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493-496, doi:10.1093/nar/gkh103 (2004).
74 Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Research* **12**, 996-1006, doi:10.1101/gr.229102 (2002).
75 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
76 Miura, P., Sanfilippo, P., Shenker, S. & Lai, E. C. Alternative polyadenylation in the nervous system: to what lengths will 3' UTR extensions take us? *BioEssays : news and reviews in molecular, cellular and developmental biology* **36**, 766-777, doi:10.1002/bies.201300174 (2014).
77 Neve, J., Patel, R., Wang, Z., Louey, A. & Furger, A. M. Cleavage and polyadenylation: Ending the message expands gene regulation. *RNA Biol* **14**, 865-890, doi:10.1080/15476286.2017.1306171 (2017).
78 Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M. & Gautheret, D. Patterns of variant polyadenylation signal usage in human genes. *Genome Research* **10**, 1001-1010 (2000).
79 Hu, J., Lutz, C. S., Wilusz, J. & Tian, B. Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA* **11**, 1485-1493, doi:10.1261/rna.2107305 (2005).
80 Salisbury, J., Hutchison, K. W. & Graber, J. H. A multispecies comparison of the metazoan 3'-processing downstream elements and the CstF-64 RNA recognition motif. *BMC Genomics* **7**, 55, doi:10.1186/1471-2164-7-55 (2006).
81 Hutchins, L. N., Murphy, S. M., Singh, P. & Graber, J. H. Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics* **24**, 2684-2690, doi:10.1093/bioinformatics/btn526 (2008).
82 Legendre, M. & Gautheret, D. Sequence determinants in human polyadenylation site selection. *BMC Genomics* **4**, 7, doi:10.1186/1471-2164-4-7 (2003).
83 McDevitt, M. A., Hart, R. P., Wong, W. W. & Nevins, J. R. Sequences capable of restoring poly(A) site function define two distinct downstream elements. *EMBO J* **5**, 2907-2913 (1986).
84 Gil, A. & Proudfoot, N. J. Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit beta-globin mRNA 3' end formation. *Cell* **49**, 399-406, doi:10.1016/0092-8674(87)90292-3 (1987).
85 Kalkatawi, M. *et al.* Dragon PolyA Spotter: predictor of poly(A) motifs within human genomic DNA sequences. *Bioinformatics* **29**, 1484, doi:10.1093/bioinformatics/btt161 (2013).

86 Akhtar, M. N., Bukhari, S. A., Fazal, Z., Qamar, R. & Shahmuradov, I. A. POLYAR, a new computer program for prediction of poly(A) sites in human sequences. *BMC Genomics* **11**, 646, doi:10.1186/1471-2164-11-646 (2010).

87 Gers, F. A., Schmidhuber, J. & Cummins, F. Learning to forget: continual prediction with LSTM. *Neural Comput* **12**, 2451-2471, doi:10.1162/089976600300015015 (2000).

88 Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput* **9**, 1735-1780, doi:10.1162/neco.1997.9.8.1735 (1997).

89 Baldi, P., Brunak, S., Frasconi, P., Soda, G. & Pollastri, G. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* **15**, 937-946, doi:10.1093/bioinformatics/15.11.937 (1999).

90 Schuster, M. & Paliwal, K. K. Bidirectional recurrent neural networks. *Ieee T Signal Proces* **45**, 2673-2681, doi:Doi 10.1109/78.650093 (1997).

91 Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, arXiv:1412.6980 (2014). <https://ui.adsabs.harvard.edu/abs/2014arXiv1412.6980K>.

92 Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, e0118432, doi:10.1371/journal.pone.0118432 (2015).

93 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323, doi:10.1186/1471-2105-12-323 (2011).

94 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. doi:10.1101/002832 (2014).

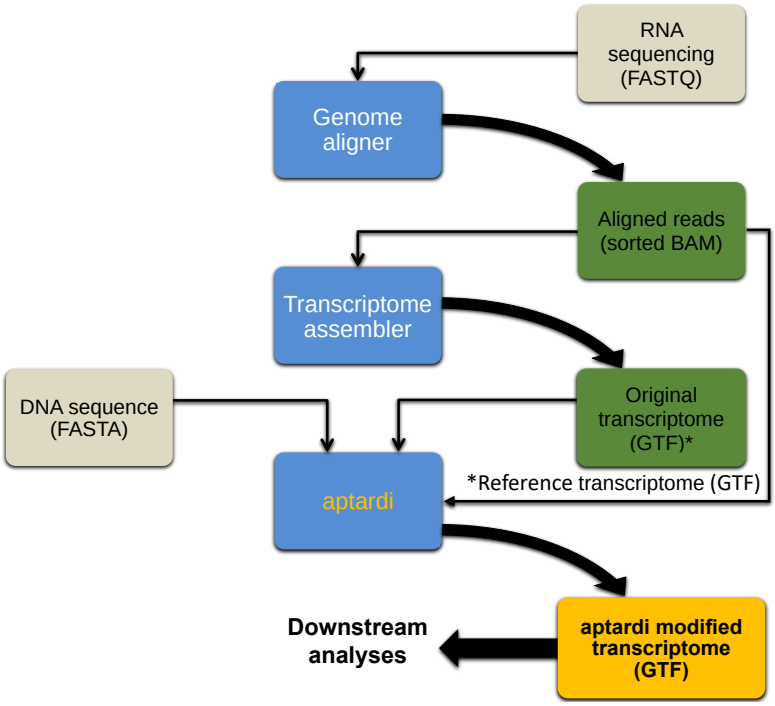

Fig. 1: Overview for using aptardi. Aptardi requires three files as input (tan boxes): 1) FASTA file of DNA sequence with headers by chromosome, 2) sorted Binary Alignment Map (BAM) file of reads aligned to the genome, and 3) General Feature Format (GTF) file of transcript structures. Blue and green boxes represent software and intermediate files, respectively. Yellow writing/boxes indicate aptardi incorporation. Note transcript structures can be derived from a reference transcriptome (i.e. Ensembl annotation) in lieu of the original transcriptome generated from a transcriptome assembler.

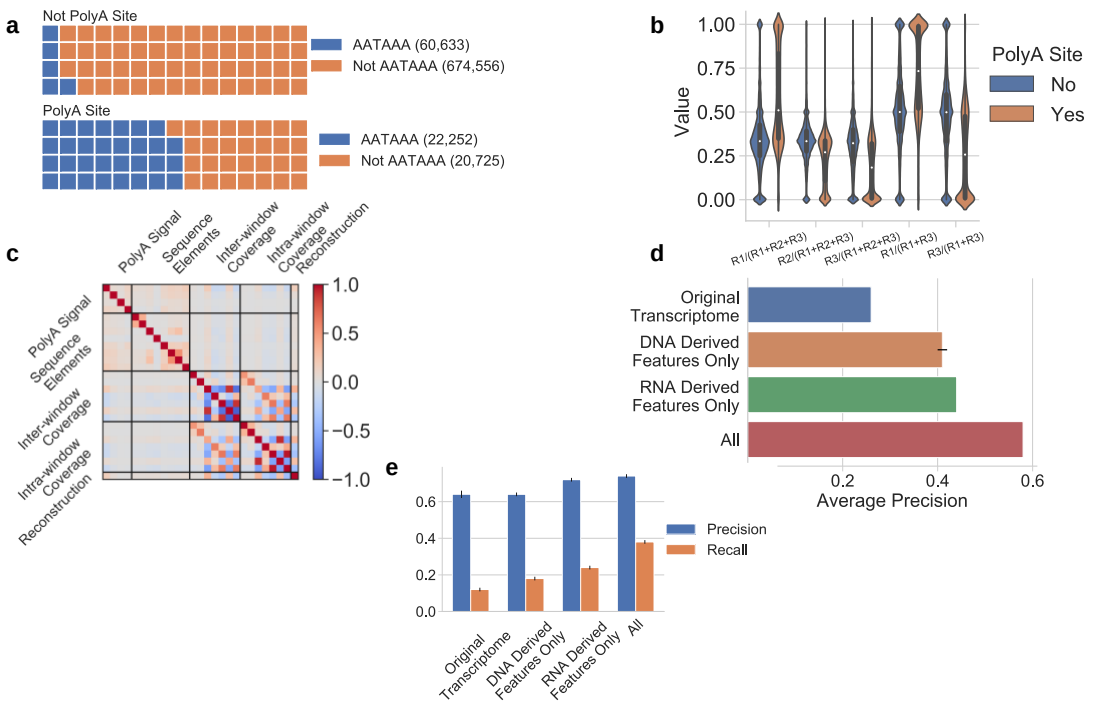

Fig. 2: DNA sequence and RNA sequencing (RNA-Seq) features are individually associated with polyadenylation (polyA) sites. a, Blue and orange boxes denote 100 base bins containing or not containing the canonical polyA signal DNA sequence feature, respectively, stratified by the bin not containing (top) or containing (bottom) a polyA site. **b,** the inter-bin RNA sequencing features are associated with polyA sites (values were standardized using the training set). **c**, RNA-Seq features and DNA sequence features display little correlation across omics type. Pearson correlation was used. The combination of RNA-Seq information and DNA sequence information improves **d**, average precision and **e**, precision and recall at a specific prediction threshold (probability > 0.50) over each separately. Values were averaged on the test set for five unique train-validate-test splits and the standard deviation is shown to the nearest hundredths. For several bars, the standard deviation is not visible at this resolution. Data shown are from the Human Brain Reference dataset.

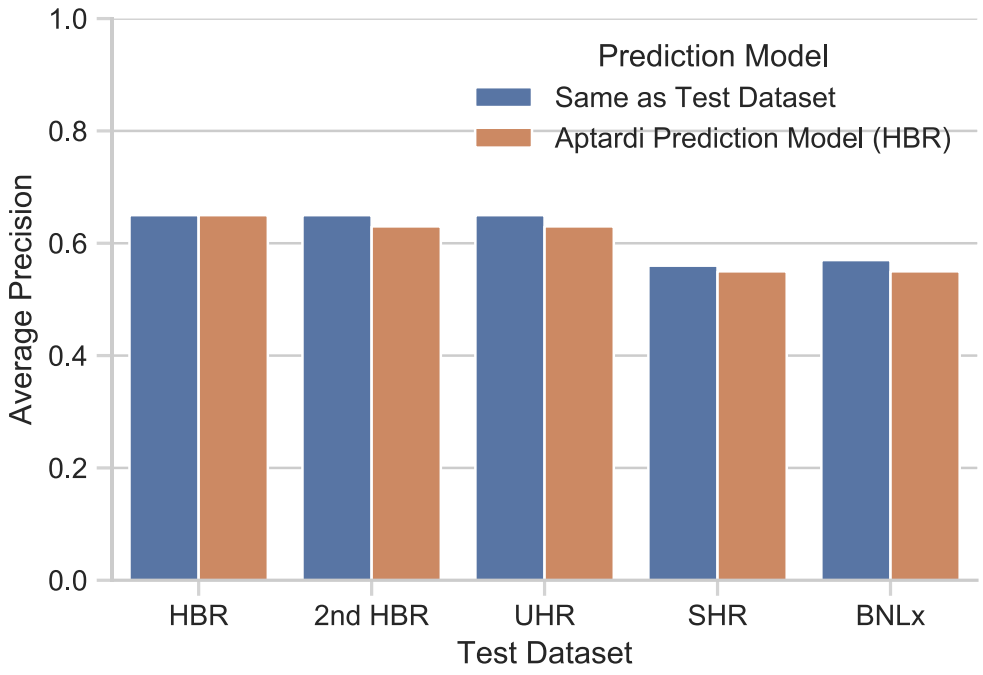

Fig. 3: The machine learning pipeline used to build aptardi is robust to different datasets and the aptardi prediction model generated from the Human Brain Reference dataset is applicable across diverse datasets. Blue bars indicate the performance of the dataset-specific prediction model on its own dataset, i.e. the model was built and evaluated on a single dataset. Orange bars represent the performance of the aptardi prediction model – built from the Human Brain Reference dataset – on the given dataset (x-axis).

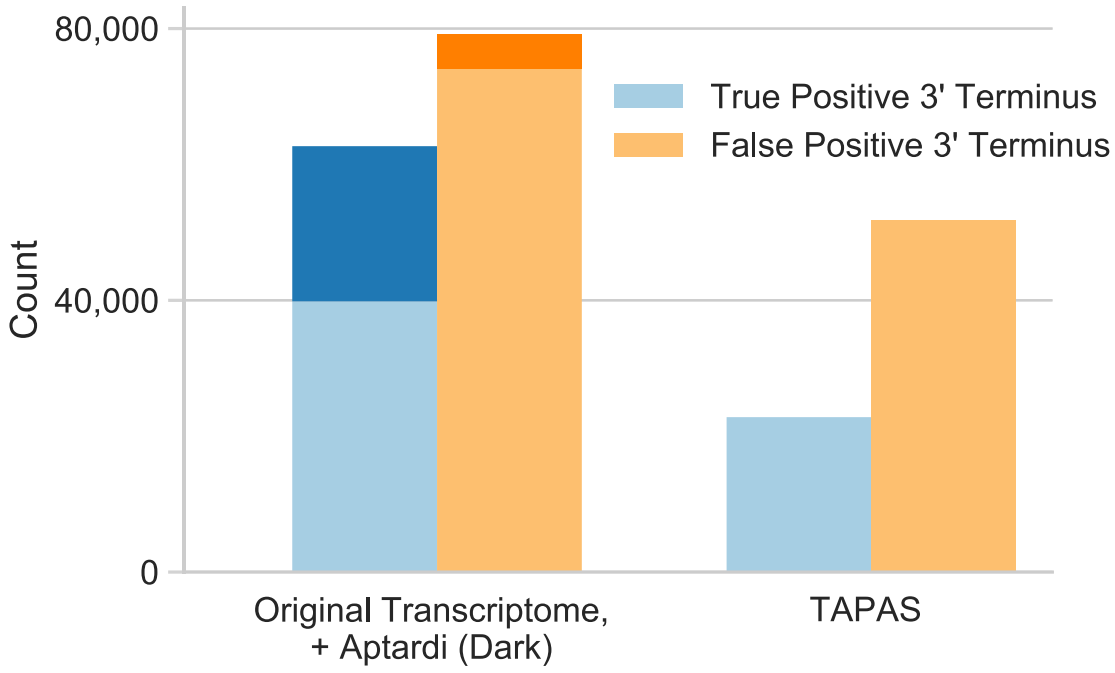

Fig. 4: Incorporating aptardi transcripts into the original transcriptome improves the ratio of true positive to false positive 3' termini compared to the original transcriptome and compared to the Tool for Alternative Polyadenylation site AnalysiS (TAPAS) analysis on the original transcriptome. Results from novel transcripts added by aptardi to the original transcriptome are shaded in dark. Transcripts whose 3' terminus was plus or minus 100 bases of a true polyadenylation site from PolyA-Seq data were considered a true positive and otherwise counted as a false positive. Data shown are from the Human Brain Reference dataset.

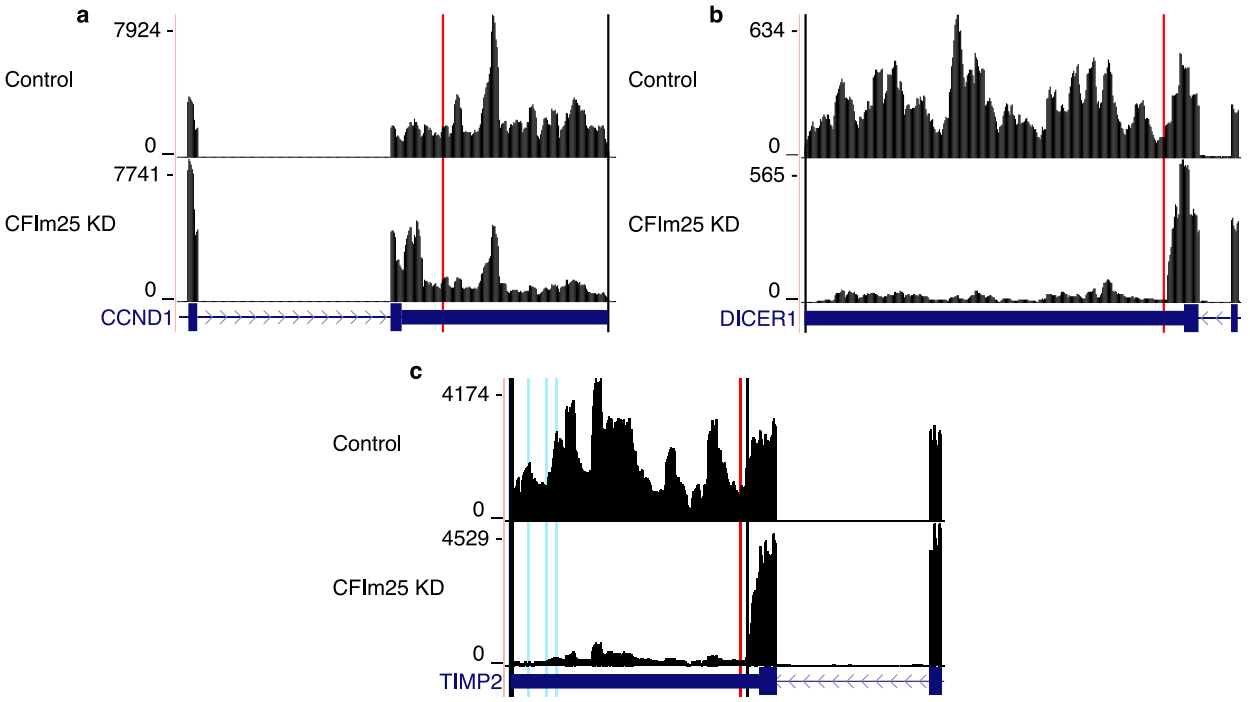

Fig. 5: Aptardi displays sample specific sensitivity when annotating transcription stop sites. RNA sequencing (RNA-Seq) read densities for **a,** *CCND1*, **b,** *DICER1*, and **c,** *TIMP2* after control (Control) siRNA treatment and CFIm25 knockdown (KD) in HeLa cells. Numbers on y-axis indicate RNA-Seq read coverage. After knockdown, each gene preferentially expresses a proximal alternative polyadenylation (APA) site compared to under control conditions. Transcript structures shown are from RefSeq annotation (dark blue), where boxes and lines indicate exons and introns, respectively. Black vertical lines indicate transcript stop sites identified in the original transcriptome, red vertical lines indicate transcript stop sites only identified in the aptardi modified transcriptome and that match the original study's findings, and blue vertical lines indicate transcript stop sites only identified in the aptardi modified transcriptome that are not described in the original study. Graphics were generating using the UCSC Genome Browser (https://genome.ucsc.edu/) using the hg38 human genome assembly.

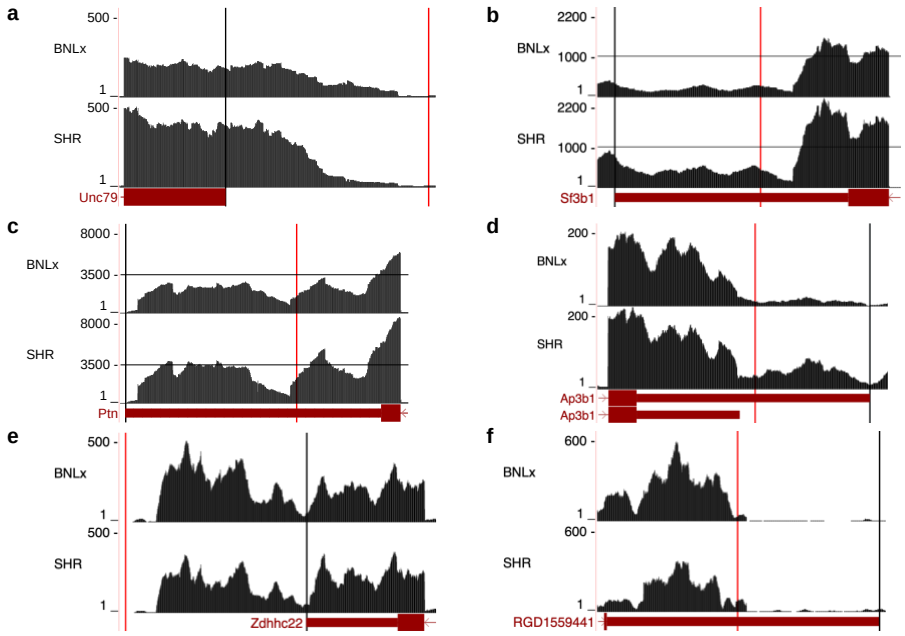

Fig. 6: Incorporating aptardi into differential expression analyses may uncover new insights. RNA sequencing (RNA-Seq) read densities for six genes in BNLx and SHR inbred rat strains. Numbers on y-axis indicate RNA-Seq read coverage. Read coverage represents the aggregate of three biological samples for each strain. Transcript structures shown are from Ensembl annotation (dark red), where boxes and lines indicate exons and introns, respectively. Black vertical lines denote transcript stop sites identified in the original transcriptome derived using StringTie, and red vertical lines indicate transcript stop sites identified in the aptardi modified transcriptome only. No transcripts were identified as differentially expressed between strains in the original transcriptome ($p > 0.001$), but at least one differentially expressed transcript for each gene was identified in the aptardi modified transcriptome ($p <= 0.001$). For **a,** *Unc79* **b,** *Sf3b1* **c,** *Ptn* and **d** *Ap3b1* the original transcript isoform (black line) was differentially expressed in the aptardi modified transcriptome, and for **e** *Zdhhc22* and **f** *RGD1559441* the novel aptardi transcript was differentially expressed (red line). Graphics were generating using the UCSC Genome Browser (https://genome.ucsc.edu/) using the rn6 rat genome assembly.