bioRxiv preprint doi: https://doi.org/10.1101/2020.08.04.235804; this version posted August 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TREND-DB – A Transcriptome-wide Atlas of the Dynamic Landscape of Alternative Polyadenylation

Federico Marini[1,2,*], Denise Scherzinger[1], and Sven Danckwardt[2,3,4,5,*]

[1]Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center of the Johannes Gutenberg University Mainz, Germany
[2]Center for Thrombosis and Hemostasis (CTH), University Medical Center of the Johannes Gutenberg-University Mainz, Germany.
[3]Posttranscriptional Gene Regulation, Cancer Research and Experimental Hemostasis, University Medical Center Mainz, Germany.
[4]Institute for Clinical Chemistry and Laboratory Medicine, University Medical Center Mainz, Germany.
[5]German Center for Cardiovascular Research (DZHK), Germany.
[*]To whom correspondence should be addressed (marinif@uni-mainz.de, Sven.Danckwardt@unimedizin-mainz.de)

August 17, 2020

Abstract

Alternative polyadenylation (APA) profoundly expands the transcriptome complexity. Perturbations of APA can disrupt biological processes, ultimately resulting in devastating disorders. A major challenge in identifying mechanisms and consequences of APA (and its perturbations) lies in the complexity of RNA 3'end processing, involving poorly conserved RNA motifs and multi-component complexes consisting of far more than 50 proteins. This is further complicated in that RNA 3'end maturation is closely linked to transcription, RNA processing, and even epigenetic (histone/DNA/RNA) modifications. Here we present TREND-DB (`http://shiny.imbei.uni-mainz.de:3838/trend-db`), a resource cataloging the dynamic landscape of APA after depletion of >170 proteins involved in various facets of transcriptional, co- and posttranscriptional gene regulation, epigenetic modifications, and further processes. TREND-DB visualizes the dynamics of transcriptome 3'end diversification (TREND) in a highly interactive manner; it provides a global APA network map and allows interrogating genes affected by specific APA-regulators, and vice versa. It also permits condition-specific functional enrichment analyses of APA-affected genes, which suggest wide biological and clinical relevance across all RNAi conditions. The implementation of the UCSC Genome Browser provides additional customizable layers of gene regulation accounting for individual transcript isoforms (e.g. epigenetics, miRNA binding sites, RNA-binding proteins). TREND-DB thereby fosters disentangling the role of APA for various biological programs, including potential disease mechanisms, and helps to identify their diagnostic and therapeutic potential.

Keywords: Database; alternative polyadenylation; web application; RNAi screening; alternative splicing; interactive exploration; functional enrichment; diagnostic markers

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.04.235804; this version posted August 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

Next-generation RNA sequencing (RNA-seq) has led to the discovery of a perplexingly complex metazoan transcriptome architecture that results from the alternative use of transcription start sites, exon and introns, and polyadenylation sites (1; 2; 3). The combinatorial use and incorporation of such elements into mature transcript isoforms significantly expands genomic information and is subject to dynamic spatial and temporal modulation during development and adaptation. Recently, diversification of the transcriptome at the 3'end by alternative polyadenylation (APA) evolved as an important and evolutionarily conserved layer of gene regulation (4; 5). It results in transcript isoforms that vary at the RNA 3'end, which can have profoundly different physiological effects, by encoding proteins with distinct functions or regulatory properties, or by affecting the mRNA fate through inclusion or exclusion of regulatory elements, such as sites recognized by miRNAs or RNA-binding proteins (6).

Constitutive RNA 3'end processing relies on a complex macromolecular machinery that catalyzes endonucleolytic cleavage and polyadenylation (CPA) of pre-mRNA molecules (7). This involves the assembly of four multi-component protein complexes (CPSF, CSTF, CFI, and CFII) (8) on the pre-mRNA at dedicated, but largely poorly conserved, processing sites (9). Differential expression of individual complex components (10; 11; 12) or regulation of their binding properties (e.g. by signalling (13)) can direct the dynamic modulation of APA resulting in transcript isoforms with alternative 3'ends. In addition, other mechanisms including RNA polymerase II kinetics (14; 15) or epigenetic events (16; 17; 18; 19; 20; 21; 22) regulate APA, illustrating a previously unanticipated complex crosstalk between various cellular processes in the control of transcriptome diversity. Although difficult to detect by standard high throughput profiling techniques (23; 24; 25), dynamic changes at the transcriptome 3'end are widespread (26; 27; 28; 29). They affect more than 70% of all genes.

Global APA changes are commonly associated with differentiation and de-differentiation processes (30). This requires a well-coordinated temporal and spatial interaction between RNA motifs and the multi-component processing complex to ensure that CPA occurs timely and at the right position (31; 32). Hence 3'end processing is tightly coupled to transcription and other co- and posttranscriptional processes (7; 33), and controlled by delicate (auto-) regulatory mechanisms (34; 35). Transcripts that show dynamic regulation at the 3'end are typically encoded by phylogenetically old genes, which corresponds to the phylogenetic age of most executing APA regulators (12). Finally, individual components of the CPA machinery are epistatic during differentiation and de-differentiation; they display functional dominance over 'neighboring' factors within and across the multi-component CPA complexes in the control of APA (12).

While the underlying mechanisms of APA regulation are still being elucidated, disruption of this process clearly proved to be clinically relevant (36). For example, APA perturbations are associated with various disorders (37; 11; 38; 39; 40; 41) (and refs therein), but they can also possess direct disease eliciting activities, act as oncogenic driver, and thereby mimic genetic alterations (12; 42). Importantly, such changes are generally missed in genome profiling endeavours (43). But often they also remain undetected by standard RNA-seq technologies (44). However, even when resulting in primarily subtle changes of non-coding RNA sequence elements in the 3'UTR, APA perturbations can be functionally significant (12) and represent unexpectedly potent novel biomarkers (24; 12). Aberrant posttranscriptional expansion of the genome complexity can thus result in most devastating consequences; at the same time, it

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.04.235804; this version posted August 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

also opens novel diagnostic (24; 12) and therapeutic avenues (45; 46).

In an attempt to systematically identify drivers of APA in tumorigenic differentiation/de-differentiation processes (12), we recently applied a comprehensive RNAi screening coupled to a high-throughput sequencing approach suited to capture polyadenylated transcript 3'ends of numerous experimental conditions (TRENDseq) (44). We depleted more than 170 proteins, including all known factors involved in pre-mRNA 3'end CPA in eukaryotes (8) and selected key factors regulating transcriptional activities, splicing, RNA turnover, and other functions (47; 48; 49; 50; 51; 52; 53; 54; 55; 56; 57; 58; 59) which could directly or indirectly affecting APA (5; 60). This resulted in the identification of numerous drivers directing APA including the transcription termination factor PCF11 with a critical role in neurodevelopment and tumor formation (12).

Here, we present TREND-DB, the full roll-out of a user-friendly database, which in complementation to existing static polyA DBs (61; 62; 63; 64) (and refs therein), catalogs the dynamic nature of the APA landscape genome-wide upon targeting >170 components involved in the definition of RNA 3'ends. We provide a highly interactive APA network map displaying central hubs in APA regulation. TREND-DB allows querying effects of selected APA-regulators including condition-specific Gene Ontology (GO) (65) enrichment analyses of APA-affected genes. Including an instance of the UCSC Track Hub framework also allows the investigation of other gene regulatory mechanisms with reference to the transcriptome-wide APA data (e.g. miRNA and RNA-binding proteins, epigenetics), or alternatively the direct comparison with own data, uploaded to the genome browser. TREND-DB thereby helps to illuminate the role of APA for various biological programs and potential disease mechanisms, and allows to identify their diagnostic and therapeutic potential. The functional enrichment across all knockdown conditions already suggests wide biological and clinical relevance.

DATA COLLECTION AND PROCESSING

TRENDseq data collection and processing

Applying TRENDseq combined with an RNAi screening targeting >170 potential APA regulators (12), we previously identified more than 9.000 APA-events (out of approximately 20.000 expressed transcripts), corresponding to more than 3.600 genes significantly affected by APA. Briefly, to this end we first generated a TREND annotation assembly based on 3'end reads to identify bona fide polyadenylation events (23). This annotation was used to exclude TRENDseq reads originating from internal priming on genome encoded adenosine-rich regions. TRENDseq data originating from the RNAi screening were demultiplexed, trimmed, mapped to the human hg38 genome, and filtered from internal priming events using the TREND annotation (for further details, see (12) and Supplementary Methods).

Quantification and global characteristics of APA events

The number of reads aligned to each polyadenylation site is used as a proxy for the expression of an individual 3'end transcript isoform. The expression level of each transcript isoform was examined by Fisher's exact test in comparison to the respective other 3'end isoform(s) expressed by the same gene. A contingency table included the number of reads of the tested isoform and total amount of reads of all the other isoforms of the gene (for the knockdown

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.04.235804; this version posted August 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and control samples, respectively). Obtained p-values were adjusted using the Benjamini-Hochberg method, and a threshold on the adjusted p-value ≤ 0.05 filter was applied (for further details on normalization, fold change of regulation, and calculation of the shortening index see (12)).

Interestingly, almost all depleted factors showed APA-regulation on average affecting 130 genes (Fig. 1) with a mean distance between APA-regulated sites of approximately 1.020 nucleotides (Supplementary Table 1). While the vast majority of APA events (65%) occur in the 3'UTR ('tandem polyadenylation'), 20% of the events affect internal polyadenylation (i.e. within introns or the coding region), thereby modulating the generation of truncated protein isoforms. Directly comparing the effects quantitatively and qualitatively, we observed that APA is predominantly driven by components of the CPA machinery (38% of all APA events, with key factors affecting up to 1.400 genes (e.g. NUDT21, CPSF6, and PCF11), average effect 319 genes) and that their function is mainly non-redundant (12). However, a significant proportion of APA (62%) is controlled by components involved in transcription (14%, on average affecting APA of 133 genes) and other co- and post-transcriptional events (e.g. splicing, RNA-turnover (16% and 9%; on average affecting APA of 128 and 89 genes, respectively)) or epigenetic modification. This corresponds to a quantitatively comparable extent to which individual splicing factors influence alternative splicing (66). Notably, while the depletion of the vast majority of factors results in even amounts of shortened and lengthened transcript isoforms, the depletion of the CFI and CFII complex components (CPSF6, NUDT21, and PCF11) directs APA in an almost exclusive unidirectional manner resulting in uniformly shortened (CPSF6 and NUDT21) or lengthened (PCF11) transcript isoforms (Fig. 1B). In particular, our screening also identified APA regulation to be associated with factors involved in genome surveillance or known to drive tumor-suppressive programs (e.g. TP53), as well as other processes involved in the coupling between oncogenic signals and 3'end processing, such as BARD1 (67).

DATABASE CONTENT AND USAGE

To make the dynamics of APA accessible to a wider scientific community, we constructed a user friendly, highly interactive database. In the database, we store (1) Summary files for each of the 175 knockdown experiments, where effect size (reported as fold changes) and significance (as adjusted p-values) of each isoform, together with the shortening index (SI), which describes the overall tendency of a given gene to express shortened (or lengthened, respectively) transcript isoforms (based on a proxy of the two most significantly affected APA isoforms), are reported; (2) bigWig files with the coverage for the isoform peaks, used to compute the SI values across conditions; (3) BED files with genomic positions for curated lists of poly(A) sites and known microRNA binding sites for the hg38 assembly (Fig. 2, upper section).

The definition of the software requirements has been established to cover the typical use cases, where the users can view, filter, and visualize the underlying data. An overview is followed by dedicated gene-centered and condition-centered views, offering the possibility to store the generated visual and tabular outputs (Fig. 2).

The TREND-DB application is developed in R (68) as a Shiny web application (69), leveraging the reactive programming paradigm offered by this framework. This framework

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.04.235804; this version posted August 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

has been chosen over existing other ecosystems such as BioPyramid (70) for its simplicity and efficiency of reactivity, together with the customizability of the analysis components enabled by the R programming language. Moreover, we leverage on a number of core packages and infrastructures, such as GenomicRanges (71), from the Bioconductor framework, which efficiently performs a variety of operations in omics data contexts (72). Tabular information is displayed in TREND-DB as interactive tables, generated with the DT package (73), as an R-based interface to the JavaScript DataTables library, which contains useful features, such as pagination, sorting, and filtering. Plots of genomic regions (3'UTR and gene bodies) for selected features are implemented with the Gviz package (74).

The user interface

The user interface is based on a dashboard page structure from the shinydashboard package (75), where the main functionality is divided into a series of tab-based views (represented by the blocks in Fig. 2).

The ***Data Preview*** tab serves as a starting point for exploring the provided data set (see also the dedicated guided tours in the TREND-DB web browser). It offers an overview table that contains all genes and shows their shortening index (SI) values in all involved knockdown conditions, enabling to filter for defined SI thresholds applied on absolute values. Selecting a gene in the overview table prints an additional table, listing all the knockdown conditions in which it is affected.

The ***Main View*** tab first displays an interactive view of the APA network, implemented with the visNetwork package (76). The APA network map (Fig. 2 and Fig. 3A) essentially depicts cooperative (and antagonistic) interactions between APA regulators (affecting APA of identical genes). While the radius of the nodes reflects the number of APA affected genes (per condition), the links between a pair of nodes depict synergism (green, i.e. unidirectional lengthening or shortening) or antagonism (red, i.e. reciprocal lengthening or shortening) on APA of the common pool of genes, with the edge width encoding the significance of the overrepresentation of genes being synergistically or antagonistically regulated (BH-adjusted p-values)(12). Remarkably, the network layout reflects known protein complexes involved in polyadenylation (8) and the crosstalk to other RNA processing events (7), recapitulating the underlying molecular architecture.

The remainder of the tab is separated into two sections, corresponding to the two dimensions of the database: (1) a ***Gene View*** and (2) a ***Condition View*** (Fig. 2, central section). Any gene or condition selection in the *Data Preview* tab is automatically carried over to this tab, while a (different) gene or condition can also be selected from the respective dropdown menu. Upon selecting a gene, the ***Gene View*** section displays a brief summary of the gene, including name, coordinates, strand, chromosome, and a description (Fig. 3E). Hyperlinks to additional resources, such as the NCBI RefSeq database, the GTEx Portal, the Human Protein Atlas, or the dbPTM database (77; 78; 79; 80), are provided as well to facilitate further exploration. On the other hand, once an experimental condition has been selected, the ***Condition View*** provides a list with all genes affected, their SIs, and p-values for the shorter and longer isoforms, according to the APA usage (Fig. 3D). Users may view a scatterplot for each condition, depicting the distribution of proximal and distal poly(A) site usage amongst the affected genes (Fig. 3B).

Additionally, a built-in ***GO term enrichment analysis*** tool (Fig. 3G) performs functional

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.04.235804; this version posted August 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

analysis on the set of genes affected in the selected condition against a background of all genes, as a potential starting point for more in-depth analysis - the interactive table created in TREND-DB automatically links to the corresponding entries in the AmiGO database (81), and can be represented visually as an enrichment map for better overview of the impacted biological themes (82). For example, genes affected in the AGO2 knockdown display among the top enriched results three retina-related GO terms, a tissue where 3'UTR shortening is often observed (83). However, there are also other interesting examples of GO enrichments underlining the wide biological and clinical relevance across knockdown conditions, for example suggesting either a potential bidirectional mechanistic link between alternative polyadenylation and alternative splicing, but also mechanisms in UTR biology towards immunobiology or developmental processes (Fig. 3H).

We and others earlier demonstrated that the relative abundance of transcript isoforms generated by APA has strong diagnostic (prognostic) potential, with superior performance in predicting death or high risk in cancer patients compared to commonly used clinical stratification markers (12). In addition, disentangling the functional role of (select) transcript isoforms opens ample therapeutic opportunities to interfere with polyadenylation in a selective or non-selective fashion (45; 46). This requires further in depth information on where APA occurs. We thus implemented functions that allow to visualize where APA affected genes are differentially polyadenylated, whether this affects the coding region of the gene ('internal polyadenylation') and/or non-coding UTR elements ('tandem polyadenylation', see above) and whether there are other conditions that result in a similar APA 'pattern' of the same gene.

The ***Gene Plot*** tab creates overview plots of a gene and its associated data (Fig. 3F). These plots contain a gene region reference track, coverage tracks for the selected condition (otherwise defaulting to the condition where the maximum SI is observed) as well as the control, and a genomic track for known APA sites. By default, overview plots are centered around the 3'UTR region of the gene; however, one can display the whole gene body region in the plot. Generated plots can be downloaded using a download button. Similarly affected genes, suggested in the interface component below the plot, can also be selected and displayed.

To enable a seamless integration with the wealth of genomic information provided by other sources, such as the UCSC Genome Browser (84), we set up an instance of the UCSC Track Hub (85), where the data is stored on a secure FTP server and the required configuration files (hub.txt, genomes.txt, and trackDb.txt) are specified in the TREND-DB GitHub repository, specifying the properties of the hub data tracks (genome assemblies, name, labels, descriptions, and display parameters). This content is integrated in the ***Genome Browser*** tab of TREND-DB via an iframe HTML element (for displaying web pages within other web pages). A custom URL is generated, accounting for all current selections made by the user, and passes it to the iframe, where the Genome Browser mirrors the behavior of the TREND-DB application.

While the ***Gene Plot*** view has a more specific aim on visualizing TRENDseq data (e.g. whether APA primarily affects the 3'UTR or the coding region), the ***Genome Browser*** view constitutes a bridge to a more flexible environment, where users can upload additional resources, such as existing tracks for their own genomic data, and facilitate the integration with the dataset of TREND-DB. As a proof-of-concept we implemented miRNA tracks; however, this can also be expanded to other resources e.g. CLIPseq (86), global interactome data (87; 88), or DNA/RNA modifications (19; 21; 22).

Altogether, these functions enable accessible and extendable exploration for the TREND-

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.04.235804; this version posted August 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DB data sets, for unbiased and targeted analysis of downstream functions, for assessment of the diagnostic potential of APA signatures, and for the development and design of potential therapeutic strategies (89). TREND-DB thereby facilitates access and reuse of such complex data by a broader scientific community complying with the FAIR data principles (90).

Simplifying the exploration for users

A common issue when exploring a series of data with a new application is how to efficiently guide the users across the different pieces of functionality offered. While text-based tutorials and walk-through documents (also in video format) can work efficiently for a specific snapshot of the application, it is not a convenient approach for maintenance once new features are added. Moreover, an approach which encourages the users to try out specific actions tightly linked to defined elements in the user interface has proven useful for other applications such as ideal (91) and iSEE (92), offering a learning-by-doing paradigm for showcasing the application features in a more cohesive, workflow-like way.

Each tab of TREND-DB has a dedicated question mark action button, which triggers bespoke step-by-step tours of the functionality offered in each view, via the rintrojs package (93). We guide the audience in the original documentation by displaying use cases that refer to a few selected findings reported earlier (12).

Moreover, as TRENDseq data is not as common as e.g. transcriptome datasets, we included a Glossary section in the main dropdown menu, precisely defining the terms commonly used throughout the app, and specifically referring to the data type under inspection.

Availability of database, code, and data

The TREND-DB application has been deployed on the Shiny Server at the Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), and can be accessed at `http://shiny.imbei.uni-mainz.de:3838/trend-db`.

The data (sequencing raw and processed) displayed in TREND-DB is also available on GEO, under the accession GSE95057 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE95057`).

TREND-DB is freely available as a Shiny web application, with source code and processing information available under the MIT license at `https://github.com/federicomarini/trend-db`.

SUMMARY AND FUTURE DIRECTIONS

TREND-DB has been conceived as an application to enable easy access to the main datasets from a recently published massive RNAi screening approach on the mechanisms and consequences underlying APA regulation (12). The deployment of such a portal can provide immediate usability of the resources showcased in publications, and thus increase their impact by releasing a more user-friendly version of the data, thus reaching a wider number of interested researchers.

The R/Shiny framework is highly flexible in regard to data manipulation and analysis, therefore we expect straightforward extensions and adaptations to the scenario we proposed

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.04.235804; this version posted August 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in this work, also to other experimental settings. As computational reproducibility gains more attention (94; 95; 96) in the scientific community, it will be desirable to support the creation of a report-like document, e.g. implemented with knitr/R Markdown, tracking the performed analyses, similar to the approach followed by the pcaExplorer package (97).

Finally, we anticipate TREND-DB to become a rich resource to explore and validate the impact of genetic and epigenetic variants on APA, and thereby also help to expand our understanding of the genetic origins of disease (98).

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to current and former members of the Danckwardt lab.

SUPPLEMENTARY DATA

Supplementary Data (Supplementary Table 1, RNAi targets and effects on alternative polyadenylation (APA)) and Supplementary Methods are available online.

FUNDING

Work in the laboratory of S.D. is supported by the DFG (DA 1189/2-1), the GRK 1591, the DFG Priority Program SPP 1935 (Deciphering the mRNP code: RNA-bound Determinants of Post-transcriptional Gene Regulation), by the Federal Ministry of Education and Research (BMBF 01EO1003), by the Hella Bühler Award for Cancer Research, and by the German Society of Clinical and Laboratory Medicine (DGKL). The work of F.M. is supported by the German Federal Ministry of Education and Research (BMBF 01EO1003).

Conflict of interest statement.

None declared.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.04.235804; this version posted August 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bibliography

[1] Carninci, P., Klein, S. A., Seidel, D. J., Pearson, B. D., Singer, S. F., Michaels, P. J., Cox, D. I., Seidel, D. J., Eskridge, R. E., Peterson, T. C., Vose, R. S., Miao, B. Q., Krishnaiah, P. R., Rao, C. R., Wu, F., Free, M., and Wang, J. (sep, 2005) The Transcriptional Landscape of the Mammalian Genome. *Science*, **309**(5740), 1559–1563.

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

[3] Reyes, A. and Huber, W. (jan, 2018) Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Research*, **46**(2), 582–592.

[4] Elkon, R., Ugalde, A. P., and Agami, R. (jul, 2013) Alternative cleavage and polyadenylation: extent, regulation and function. *Nature Reviews Genetics*, **14**(7), 496–506.

[5] Tian, B. and Manley, J. L. (jan, 2017) Alternative polyadenylation of mRNA precursors. *Nature Reviews Molecular Cell Biology*, **18**(1), 18–30.

[6] Mayr, C. (nov, 2017) Regulation by 3′-Untranslated Regions. *Annual Review of Genetics*, **51**(1), 171–194.

[7] Danckwardt, S., Hentze, M. W., and Kulozik, A. E. (2008) 3′ end mRNA processing: Molecular mechanisms and implications for health and disease. *EMBO Journal*, **27**(3), 482–498.

[8] Shi, Y., Di Giammartino, D. C., Taylor, D., Sarkeshik, A., Rice, W. J., Yates, J. R., Frank, J., and Manley, J. L. (2009) Molecular Architecture of the Human Pre-mRNA 3′ Processing Complex. *Molecular Cell*, **33**(3), 365–376.

[9] Gruber, A. J., Schmidt, R., Gruber, A. R., Martin, G., Ghosh, S., Belmadani, M., Keller, W., and Zavolan, M. (2016) A comprehensive analysis of 3′ end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Research*, **26**(8), 1145–1159.

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

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

[12] Ogorodnikov, A., Levin, M., Tattikota, S., Tokalov, S., Hoque, M., Scherzinger, D., Marini, F., Poetsch, A., Binder, H., Macher-Göppinger, S., Probst, H. C., Tian, B., Schaefer, M., Lackner, K. J., Westermann, F., and Danckwardt, S. (dec, 2018) Transcriptome 3′end organization by PCF11 links alternative polyadenylation to formation and neuronal differentiation of neuroblastoma. *Nature Communications*, **9**(1), 5331.

[13] Danckwardt, S., Gantzert, A. S., Macher-Goeppinger, S., Probst, H. C., Gentzel, M., Wilm, M., Gröne, H. J., Schirmacher, P., Hentze, M. W., and Kulozik, A. E. (2011) P38 MAPK Controls Prothrombin Expression by Regulated RNA 3′ End Processing. *Molecular Cell*, **41**(3), 298–310.

[14] Pinto, P. A. B., Henriques, T., Freitas, M. O., Martins, T., Domingues, R. G., Wyrzykowska, P. S., Coelho, P. A., Carmo, A. M., Sunkel, C. E., Proudfoot, N. J., and Moreira, A. (jun, 2011) RNA polymerase II kinetics in polo polyadenylation signal selection. *The EMBO Journal*, **30**(12), 2431–2444.

[15] Neve, J., Burger, K., Li, W., Hoque, M., Patel, R., Tian, B., Gullerova, M., and Furger, A. (jan, 2016) Subcellular RNA profiling links splicing and nuclear DICER1 to alternative cleavage and polyadenylation. *Genome Research*, **26**(1), 24–35.

[16] Wood, A. J., Schulz, R., Woodfine, K., Koltowska, K., Beechey, C. V., Peters, J., Bourc'his, D., and Oakey, R. J. (may, 2008) Regulation of alternative polyadenylation by genomic imprinting. *Genes & Development*, **22**(9), 1141–1146.

[17] Spies, N., Nielsen, C. B., Padgett, R. A., and Burge, C. B. (oct, 2009) Biased Chromatin Signatures around Polyadenylation Sites and Exons. *Molecular Cell*, **36**(2), 245–254.

[18] Li, W., Park, J. Y., Zheng, D., Hoque, M., Yehia, G., and Tian, B. (dec, 2016) Alternative cleavage and polyadenylation in spermatogenesis connects chromatin regulation with post-transcriptional control. *BMC Biology*, **14**(1), 6.

[19] Ke, S., Alemu, E. A., Mertens, C., Gantman, E. C., Fak, J. J., Mele, A., Haripal, B., Zucker-Scharff, I., Moore, M. J., Park, C. Y., Vågbø, C. B., Kusśnierczyk, A., Klungland, A., Darnell, J. E., and Darnell, R. B. (oct, 2015) A majority of m 6 A residues are in the last exons, allowing the potential for 3′ UTR regulation. *Genes & Development*, **29**(19), 2037–2053.

[20] Brumbaugh, J., Di Stefano, B., Wang, X., Borkent, M., Forouzmand, E., Clowers, K. J., Ji, F., Schwarz, B. A., Kalocsay, M., Elledge, S. J., Chen, Y., Sadreyev, R. I., Gygi, S. P., Hu, G., Shi, Y., and Hochedlinger, K. (2018) Nudt21 Controls Cell Fate by Connecting Alternative Polyadenylation to Chromatin Signaling. *Cell*, **172**(1-2), 106–120.e21.

[21] Yue, Y., Liu, J., Cui, X., Cao, J., Luo, G., Zhang, Z., Cheng, T., Gao, M., Shu, X., Ma, H., Wang, F., Wang, X., Shen, B., Wang, Y., Feng, X., He, C., and Liu, J. (dec, 2018) VIRMA mediates preferential m6A mRNA methylation in 3′UTR and near stop codon and associates with alternative polyadenylation. *Cell Discovery*, **4**(1), 10.

[22] Parker, M. T., Knop, K., Sherwood, A. V., Schurch, N. J., Mackinnon, K., Gould, P. D., Hall, A. J., Barton, G. J., and Simpson, G. G. (jan, 2020) Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. *eLife*, **9**, 1–35.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.04.235804; this version posted August 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[23] Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., Park, J. Y., Yehia, G., and Tian, B. (2013) Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nature Methods*, **10**(2), 133–139.

[24] Xia, Z., Donehower, L. A., Cooper, T. A., Neilson, J. R., Wheeler, D. A., Wagner, E. J., and Li, W. (dec, 2014) Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. *Nature Communications*, **5**(1), 5274.

[25] Ha, K. C. H., Blencowe, B. J., and Morris, Q. (dec, 2018) QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biology*, **19**(1), 45.

[26] Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A., and Burge, C. B. (jun, 2008) Proliferating Cells Express mRNAs with Shortened 3′ Untranslated Regions and Fewer MicroRNA Target Sites. *Science*, **320**(5883), 1643–1647.

[27] Mayr, C. and Bartel, D. P. (2009) Widespread Shortening of 3′UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell*, **138**(4), 673–684.

[28] Ozsolak, F., Kapranov, P., Foissac, S., Kim, S. W., Fishilevich, E., Monaghan, A. P., John, B., and Milos, P. M. (2010) Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell*, **143**(6), 1018–1029.

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

[30] Turner, R. E., Pattison, A. D., and Beilharz, T. H. (mar, 2018) Alternative polyadenylation in the regulation and dysregulation of gene expression. *Seminars in Cell & Developmental Biology*, **75**, 61–69.

[31] Lutz, C. S. and Moreira, A. (2011) Alternative mRNA polyadenylation in eukaryotes: An effective regulator of gene expression. *Wiley Interdisciplinary Reviews: RNA*, **2**(1), 22–31.

[32] Gruber, A. J. and Zavolan, M. (oct, 2019) Alternative cleavage and polyadenylation in health and disease. *Nature Reviews Genetics*, **20**(10), 599–614.

[33] Millevoi, S. and Vagner, S. (2009) Molecular mechanisms of eukaryotic pre-mRNA 3′ end processing regulation. *Nucleic Acids Research*, **38**(9), 2757–2774.

[34] Kamieniarz-Gdula, K., Gdula, M. R., Panser, K., Nojima, T., Monks, J., Wiśniewski, J. R., Riepsaame, J., Brockdorff, N., Pauli, A., and Proudfoot, N. J. (2019) Selective Roles of Vertebrate PCF11 in Premature and Full-Length Transcript Termination. *Molecular Cell*, **74**(1), 158–172.e9.

[35] Wang, R., Zheng, D., Wei, L., Ding, Q., and Tian, B. (2019) Regulation of Intronic Polyadenylation by PCF11 Impacts mRNA Expression of Long Genes. *Cell Reports*, **26**(10), 2766–2778.e6.

[36] Nourse, J., Spada, S., and Danckwardt, S. (2020) Emerging Roles of RNA 3′-end Cleavage and Polyadenylation in Pathogenesis, Diagnosis and Therapy of Human Disorders. *Biomolecules*, **10**(6), 8–12.

[37] Lin, Y., Li, Z., Ozsolak, F., Kim, S. W., Arango-Argoty, G., Liu, T. T., Tenenbaum, S. A., Bailey, T., Monaghan, A. P., Milos, P. M., and John, B. (2012) An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Research*, **40**(17), 8460–8471.

[38] Weng, T., Ko, J., Masamha, C. P., Xia, Z., Xiang, Y., Chen, N.-y., Molina, J. G., Collum, S., Mertens, T. C., Luo, F., Philip, K., Davies, J., Huang, J., Wilson, C., Thandavarayan, R. A., Bruckner, B. A., Jyothula, S. S., Volcik, K. A., Li, L., Han, L., Li, W., Assassi, S., Karmouty-Quintana, H., Wagner, E. J., and Blackburn, M. R. (apr, 2019) Cleavage factor 25 deregulation contributes to pulmonary fibrosis through alternative polyadenylation. *Journal of Clinical Investigation*, **129**(5), 1984–1999.

[39] Patel, R., Brophy, C., Hickling, M., Neve, J., and Furger, A. (2019) Alternative cleavage and polyadenylation of genes associated with protein turnover and mitochondrial function are deregulated in Parkinson's, Alzheimer's and ALS disease. *BMC Medical Genomics*, **12**(1), 1–14.

[40] Ye, C., Zhou, Q., Hong, Y., and Li, Q. Q. (2019) Role of alternative polyadenylation dynamics in acute myeloid leukaemia at single-cell resolution. *RNA Biology*, **16**(6), 785–797.

[41] Soetanto, R., Hynes, C., Patel, H., Humphreys, D., Evers, M., Duan, G., Parker, B., Archer, S., Clancy, J., Graham, R., Beilharz, T., Smith, N., and Preiss, T. (may, 2016) Role of miRNAs and alternative mRNA 3′-end cleavage and polyadenylation of their mRNA targets in cardiomyocyte hypertrophy. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, **1859**(5), 744–756.

[42] Lee, S. H., Singh, I., Tisdale, S., Abdel-Wahab, O., Leslie, C. S., and Mayr, C. (2018) Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. *Nature*, **561**(7721), 127–131.

[43] Stacey, S. N., Sulem, P., Jonasdottir, A., Masson, G., Gudmundsson, J., Gudbjartsson, D. F., Magnusson, O. T., Gudjonsson, S. A., Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K. R., Nexø, B. A., Tjønneland, A., Overvad, K., Rudnai, P., Gurzau, E., Koppova, K., Hemminki, K., Corredera, C., Fuentelsaz, V., Grasa, P., Navarrete, S., Fuertes, F., García-Prats, M. D., Sanambrosio, E., Panadero, A., De Juan, A., Garcia, A., Rivera, F., Planelles, D., Soriano, V., Requena, C., Aben, K. K., Van Rossum, M. M., Cremers, R. G., Van Oort, I. M., Van Spronsen, D. J., Schalken, J. A., Peters, W. H., Helfand, B. T., Donovan, J. L., Hamdy, F. C., Badescu, D., Codreanu, O., Jinga, M., Csiki, I. E., Constantinescu, V., Badea, P., Mates, I. N., Dinu, D. E., Constantin, A., Mates, D., Kristjansdottir, S., Agnarsson, B. A., Jonsson, E., Barkardottir, R. B., Einarsson, G. V., Sigurdsson, F., Moller, P. H., Stefansson, T., Valdimarsson, T., Johannsson, O. T., Sigurdsson, H., Jonsson, T., Jonasson, J. G., Tryggvadottir, L., Rice, T., Hansen, H. M., Xiao, Y., Lachance, D. H., O'Neill, B. P., Kosel, M. L., Decker, P. A., Thorleifsson, G., Johannsdottir, H., Helgadottir, H. T., Sigurdsson, A., Steinthorsdottir, V., Lindblom, A., Sandler, R. S., Keku, T. O., Banasik, K., Jørgensen, T., Witte, D. R., Hansen, T., Pedersen, O., Jinga, V., Neal, D. E., Catalona, W. J., Wrensch, M., Wiencke, J., Jenkins, R. B., Nagore, E., Vogel, U., Kiemeney, L. A., Kumar, R., Mayordomo, J. I., Olafsson, J. H., Kong, A., Thorsteinsdottir, U., Rafnar, T., and Stefansson, K. (2011) A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nature Genetics*, **43**(11), 1098–1103.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.04.235804; this version posted August 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[44] Ogorodnikov, A., Kargapolova, Y., and Danckwardt, S. (jun, 2016) Processing and transcriptome expansion at the mRNA 3′ end in health and disease: finding the right end. *Pflügers Archiv - European Journal of Physiology*, **468**(6), 993–1012.

[45] Vorlová, S., Rocco, G., LeFave, C. V., Jodelka, F. M., Hess, K., Hastings, M. L., Henke, E., and Cartegni, L. (2011) Induction of Antagonistic Soluble Decoy Receptor Tyrosine Kinases by Intronic PolyA Activation. *Molecular Cell*, **43**(6), 927–939.

[46] Araki, S., Nakayama, Y., Sano, O., Nakao, S., Shimizu-Ogasawara, M., Toyoshiba, H., Nakanishi, A., and Aparicio, S. (dec, 2018) Decoding Transcriptome Dynamics of Genome-Encoded Polyadenylation and Autoregulation with Small-Molecule Modulators of Alternative Polyadenylation. *Cell Chemical Biology*, **25**(12), 1470–1484.e5.

[47] Bava, F.-A., Eliscovich, C., Ferreira, P. G., Miñana, B., Ben-Dov, C., Guigó, R., Valcárcel, J., and Méndez, R. (mar, 2013) CPEB1 coordinates alternative 3′-UTR formation with translational regulation. *Nature*, **495**(7439), 121–125.

[48] Fusby, B., Kim, S., Erickson, B., Kim, H., Peterson, M. L., and Bentley, D. L. (2015) Coordination of RNA Polymerase II Pausing and 3′ end processing factor recruitment with alternative polyadenylation. *Molecular and Cellular Biology*, **36**(2), MCB.00898–15.

[49] Dermody, J. L. and Buratowski, S. (2010) Leo1 subunit of the yeast Paf1 complex binds RNA and contributes to complex recruitment. *Journal of Biological Chemistry*, **285**(44), 33671–33679.

[50] Batra, R., Charizanis, K., Manchanda, M., Mohan, A., Li, M., Finn, D. J., Goodwin, M., Zhang, C., Sobczak, K., Thornton, C. A., and Swanson, M. S. (oct, 2014) Loss of MBNL Leads to Disruption of Developmentally Regulated Alternative Polyadenylation in RNA-Mediated Disease. *Molecular Cell*, **56**(2), 311–322.

[51] Neve, J., Burger, K., Li, W., Hoque, M., Patel, R., Tian, B., Gullerova, M., and Furger, A. (2016) Subcellular RNA profiling links splicing and nuclear DICER1 to alternative cleavage and polyadenylation. *Genome Research*, **26**(1), 24–35.

[52] Kaida, D., Berg, M. G., Younis, I., Kasim, M., Singh, L. N., Wan, L., and Dreyfuss, G. (2010) U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*, **468**(7324), 664–668.

[53] Katahira, J., Okuzaki, D., Inoue, H., Yoneda, Y., Maehara, K., and Ohkawa, Y. (2013) Human TREX component Thoc5 affects alternative polyadenylation site choice by recruiting mammalian cleavage factor I. *Nucleic Acids Research*, **41**(14), 7060–7072.

[54] Di Giammartino, D. C., Li, W., Ogami, K., Yashinskie, J. J., Hoque, M., Tian, B., and Manley, J. L. (2014) RBBP6 isoforms regulate the human polyadenylation machinery and modulate expression of mRNAs with AU-rich 39 UTRs. *Genes and Development*, **28**(20), 2248–2260.

[55] Dutertre, M., Chakrama, F. Z., Combe, E., Desmet, F.-O., Mortada, H., Polay Espinoza, M., Gratadou, L., and Auboeuf, D. (may, 2014) A recently evolved class of alternative 3′-terminal exons involved in cell cycle regulation by topoisomerase inhibitors. *Nature Communications*, **5**(1), 3395.

[56] Huang, Y., Li, W., Yao, X., jiang Lin, Q., wen Yin, J., Liang, Y., Heiner, M., Tian, B., Hui, J., and Wang, G. (2012) Mediator Complex Regulates Alternative mRNA Processing via the MED23 Subunit. *Molecular Cell*, **45**(4), 459–469.

[57] Hallais, M., Pontvianne, F., Andersen, P. R., Clerici, M., Lener, D., Benbahouche, N. E. H., Gostan, T., Vandermoere, F., Robert, M. C., Cusack, S., Verheggen, C., Jensen, T. H., and Bertrand, E. (2013) CBC-ARS2 stimulates 3′-end maturation of multiple RNA families and favors cap-proximal processing. *Nature Structural and Molecular Biology*, **20**(12), 1358–1366.

[58] Müller-McNicoll, M., Botti, V., de Jesus Domingues, A. M., Brandl, H., Schwich, O. D., Steiner, M. C., Curk, T., Poser, I., Zarnack, K., and Neugebauer, K. M. (2016) SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. *Genes and Development*, **30**(5), 553–566.

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

[60] Proudfoot, N. J. (jun, 2016) Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science*, **352**(6291), aad9926–aad9926.

[61] Muller, S., Rycak, L., Afonso-Grunz, F., Winter, P., Zawada, A. M., Damrath, E., Scheider, J., Schmah, J., Koch, I., Kahl, G., and Rotter, B. (jul, 2014) APADB: a database for alternative polyadenylation and microRNA regulation events. *Database*, **2014**, bau076–bau076.

[62] Wang, R., Nambiar, R., Zheng, D., and Tian, B. (2018) PolyA-DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Research*, **46**(D1), D315–D319.

[63] Hong, W., Ruan, H., Zhang, Z., Ye, Y., Liu, Y., Li, S., Jing, Y., Zhang, H., Diao, L., Liang, H., and Han, L. (jan, 2020) APAatlas: decoding alternative polyadenylation across human tissues. *Nucleic Acids Research*, **48**(D1), D34–D39.

[64] Herrmann, C. J., Schmidt, R., Kanitz, A., Artimo, P., Gruber, A. J., and Zavolan, M. (oct, 2019) PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Research*, **48**(D1), D174–D179.

[65] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (may, 2000) Gene ontology: Tool for the unification of biology. *Nature Genetics*, **25**(1), 25–29.

[66] Brooks, A. N., Duff, M. O., May, G., Yang, L., Bolisetty, M., Landolin, J., Wan, K., Sandler, J., Booth, B. W., Celniker, S. E., Graveley, B. R., and Brenner, S. E. (2015) Regulation of alternative splicing in Drosophila by 56 RNA binding proteins. *Genome Research*, **25**(11), 1771–1780.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.04.235804; this version posted August 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[67] Kleiman, F. E. and Manley, J. L. (2001) The BARD1-CstF-50 interaction links mRNA 3' end formation to DNA damage and tumor suppression. *Cell*, **104**(5), 743–753.

[68] R Core Team R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing Vienna, Austria (2019).

[69] Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. shiny: Web Application Framework for R (2020) R package version 1.4.0.2.

[70] Stephenson, L., Wakeham, Y., Seidenman, N., and Choi, J. (sep, 2018) Building online genomics applications using BioPyramid. *Bioinformatics*, **34**(17), 3055–3057.

[71] Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (aug, 2013) Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, **9**(8), e1003118.

[72] Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. a., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., Pagès, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., and Morgan, M. (jan, 2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, **12**(2), 115–121.

[73] Xie, Y., Cheng, J., and Tan, X. DT: A Wrapper of the JavaScript Library 'DataTables' (2020) R package version 0.13.

[74] Hahne, F. and Ivanek, R. (2016) Visualizing Genomic Data Using Gviz and Bioconductor. In Mathé, E. and Davis, S., (eds.), *Statistical Genomics*, Vol. 1418 of Methods in Molecular Biology, pp. 335–351 Springer New York New York, NY.

[75] Chang, W. and Borges Ribeiro, B. shinydashboard: Create Dashboards with 'Shiny' (2018) R package version 0.7.1.

[76] Almende B.V., Thieurmel, B., and Robert, T. visNetwork: Network Visualization using 'vis.js' Library (2019) R package version 2.0.9.

[77] O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, **44**(D1), D733–D745.

[78] Gamazon, E. R., Segrè, A. V., van de Bunt, M., Wen, X., Xi, H. S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E. M., Aguet, F., Quan, J., Nicolae, D. L., Eskin, E., Kellis, M., Getz, G., McCarthy, M. I., Dermitzakis, E. T., Cox, N. J., and Ardlie, K. G. (jul, 2018) Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nature Genetics*, **50**(7), 956–967.

[79] Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., Sanli, K., von Feilitzen, K., Oksvold, P., Lundberg, E., Hober, S., Nilsson, P., Mattsson, J., Schwenk, J. M., Brunnström, H., Glimelius, B., Sjöblom, T., Edqvist, P.-H., Djureinovic, D., Micke, P., Lindskog, C., Mardinoglu, A., and Ponten, F. (2017) A pathology atlas of the human cancer transcriptome. *Science*, **357**(6352), eaan2507.

[80] Huang, K.-Y., Lee, T.-Y., Kao, H.-J., Ma, C.-T., Lee, C.-C., Lin, T.-H., Chang, W.-C., and Huang, H.-D. (jan, 2019) dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications. *Nucleic Acids Research*, **47**(D1), D298–D308.

[81] Carbon, S., Douglass, E., Dunn, N., Good, B., Harris, N. L., Lewis, S. E., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L. P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan, A., Huang, X., Poudel, S., Mushayahama, T., Hu, J. C., LaBonte, S. A., Siegele, D. A., Antonazzo, G., Attrill, H., Brown, N. H., Fexova, S., Garapati, P., Jones, T. E., Marygold, S. J., Millburn, G. H., Rey, A. J., Trovisco, V., Dos Santos, G., Emmert, D. B., Falls, K., Zhou, P., Goodman, J. L., Strelets, V. B., Thurmond, J., Courtot, M., Osumi, D. S., Parkinson, H., Roncaglia, P., Acencio, M. L., Kuiper, M., Lreid, A., Logie, C., Lovering, R. C., Huntley, R. P., Denny, P., Campbell, N. H., Kramarz, B., Acquaah, V., Ahmad, S. H., Chen, H., Rawson, J. H., Chibucos, M. C., Giglio, M., Nadendla, S., Tauber, R., Duesbury, M. J., Del, N. T., Meldal, B. H., Perfetto, L., Porras, P., Orchard, S., Shrivastava, A., Xie, Z., Chang, H. Y., Finn, R. D., Mitchell, A. L., Rawlings, N. D., Richardson, L., Sangrador-Vegas, A., Blake, J. A., Christie, K. R., Dolan, M. E., Drabkin, H. J., Hill, D. P., Ni, L., Sitnikov, D., Harris, M. A., Oliver, S. G., Rutherford, K., Wood, V., Hayles, J., Bahler, J., Lock, A., Bolton, E. R., De Pons, J., Dwinell, M., Hayman, G. T., Laulederkind, S. J., Shimoyama, M., Tutaj, M., Wang, S. J., D'Eustachio, P., Matthews, L., Balhoff, J. P., Aleksander, S. A., Binkley, G., Dunn, B. L., Cherry, J. M., Engel, S. R., Gondwe, F., Karra, K., MacPherson, K. A., Miyasato, S. R., Nash, R. S., Ng, P. C., Sheppard, T. K., Shrivatsav Vp, A., Simison, M., Skrzypek, M. S., Weng, S., Wong, E. D., Feuermann, M., Gaudet, P., Bakker, E., Berardini, T. Z., Reiser, L., Subramaniam, S., Huala, E., Arighi, C., Auchincloss, A., Axelsen, K., Argoud, G. P., Bateman, A., Bely, B., Blatter, M. C., Boutet, E., Breuza, L., Bridge, A., Britto, R., Bye-A-Jee, H., Casals-Casas, C., Coudert, E., Estreicher, A., Famiglietti, L., Garmiri, P., Georghiou, G., Gos, A., Gruaz-Gumowski, N., Hatton-Ellis, E., Hinz, U., Hulo, C., Ignatchenko, A., Jungo, F., Keller, G., Laiho, K., Lemercier, P., Lieberherr, D., Lussi, Y., Mac-Dougall, A., Magrane, M., Martin, M. J., Masson, P., Natale, D. A., Hyka, N. N., Pedruzzi, I., Pichler, K., Poux, S., Rivoire, C., Rodriguez-Lopez, M., Sawford, T., Speretta, E., Shypitsyna, A., Stutz, A., Sundaram, S., Tognolli, M., Tyagi, N., Warner, K., Zaru, R., Wu, C., Chan, J., Cho, J., Gao, S., Grove, C., Harrison, M. C., Howe, K., Lee, R., Mendel, J., Muller, H. M., Raciti, D., Van Auken, K., Berriman, M., Stein, L., Sternberg, P. W., Howe, D., Toro, S., and Westerfield, M. (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, **47**(D1), D330–D338.

[82] Merico, D., Isserlin, R., Stueker, O., Emili, A., and Bader, G. D. (nov, 2010) Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *PLoS ONE*, **5**(11), e13984.

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

[84] Haeussler, M., Zweig, A. S., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., Lee, C. M., Lee, B. T., Hinrichs, A. S., Gonzalez, J. N., Gibson, D., Diekhans, M., Clawson, H., Casper, J., Barber, G. P., Haussler, D., Kuhn, R. M., and Kent, W. J. (2019) The UCSC Genome Browser database: 2019 update. *Nucleic Acids Research*, **47**(D1), D853–D858.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.04.235804; this version posted August 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[85] Raney, B. J., Dreszer, T. R., Barber, G. P., Clawson, H., Fujita, P. A., Wang, T., Nguyen, N., Paten, B., Zweig, A. S., Karolchik, D., and Kent, W. J. (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**(7), 1003–1005.

[86] Kargapolova, Y., Levin, M., Lackner, K., and Danckwardt, S. (jun, 2017) sCLIP—an integrated platform to study RNA–protein interactomes in biomedical research: identification of CSTF2tau in alternative processing of small nuclear RNAs. *Nucleic Acids Research*, **45**(10), 6074–6086.

[87] Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B. M., Strein, C., Davey, N. E., Humphreys, D. T., Preiss, T., Steinmetz, L. M., Krijgsveld, J., and Hentze, M. W. (2012) Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell*, **149**(6), 1393–1406.

[88] Baltz, A. G., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., Wyler, E., Bonneau, R., Selbach, M., Dieterich, C., and Landthaler, M. (2012) The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Molecular Cell*, **46**(5), 674–690.

[89] Yang, F., Wang, W., Cetinbas, M., Sadreyev, R. I., and Blower, M. D. (mar, 2020) Genome-wide analysis identifies cis -acting elements regulating mRNA polyadenylation and translation during vertebrate oocyte maturation. *RNA*, **26**(3), 324–344.

[90] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (mar, 2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**, 160018.

[91] Marini, F., Linke, J., and Binder, H. (2020) ideal : an R / Bioconductor package for Interactive Differential Expression Analysis. *bioRxiv*,.

[92] Rue-Albrecht, K., Marini, F., Soneson, C., and Lun, A. T. (jun, 2018) iSEE: Interactive SummarizedExperiment Explorer. *F1000Research*, **7**(0), 741.

[93] Ganz, C. (2016) rintrojs: A Wrapper for the Intro.js Library. *The Journal of Open Source Software*, **1**(6), 2016.

[94] Peng, R. D. (dec, 2011) Reproducible Research in Computational Science. *Science*, **334**(6060), 1226–1227.

[95] Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E. L., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E. J., Wilson, R., and Yarkoni, T. (jun, 2015) Promoting an open research culture. *Science*, **348**(6242), 1422–1425.

[96] Collins, F. S. and Tabak, L. A. (2014) NIH plans to enhance reproducibility. *Nature*, **505**(7485), 612–613.

[97] Marini, F. and Binder, H. (dec, 2019) pcaExplorer: an R/Bioconductor package for interacting with RNA-seq principal components. *BMC Bioinformatics*, **20**(1), 331.

[98] Bogard, N., Linder, J., Rosenberg, A. B., and Seelig, G. (jun, 2019) A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. *Cell*, **178**(1), 91–106.e23.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.04.235804; this version posted August 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

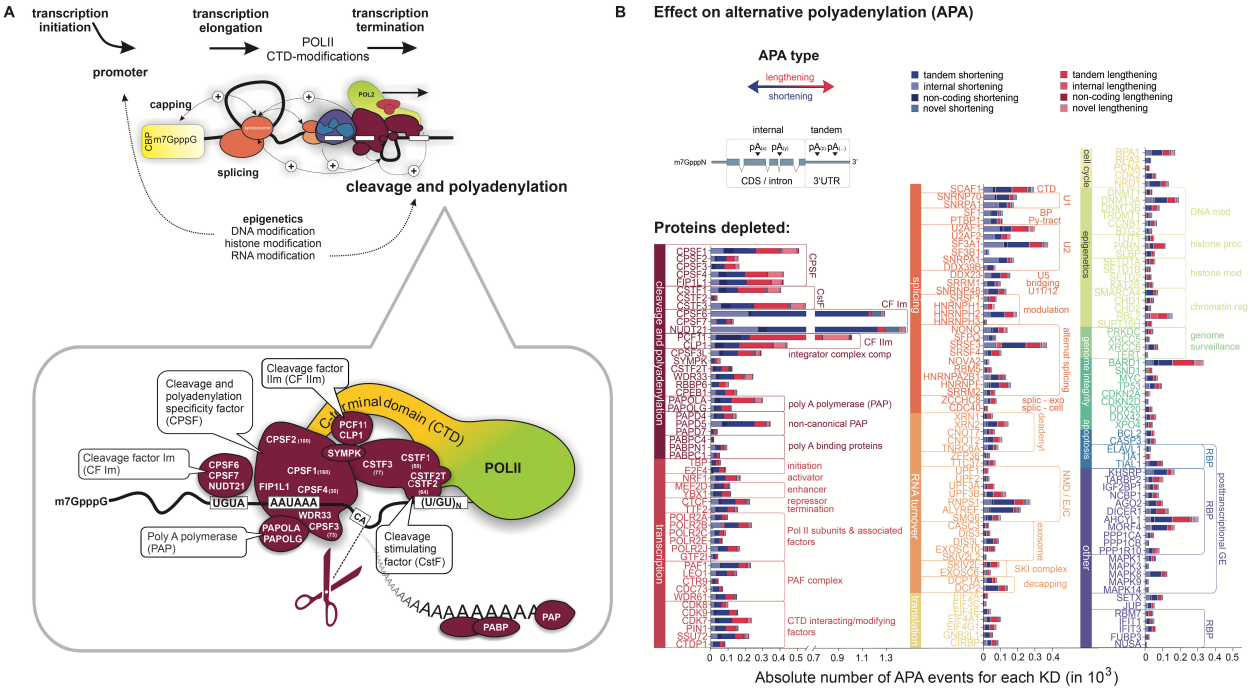

Figure 1: Data preview of TREND-DB cataloging the dynamic landscape of alternative polyadenylation **(A)** 3'end processing of polyadenylated RNAs relies on a complex macromolecular machinery that catalyzes endonucleolytic cleavage and polyadenylation (CPA) of the nascent pre-mRNA molecule. CPA is carried out after assembly of four multi-component protein complexes (CPSF, CSTF, CFI, and CFII) interacting with multiple RNA motifs. This is tightly coupled to transcription, other steps of RNA processing, and even epigenetic modifications. Differential expression of individual complex components or regulation of their binding properties can redirect CPA to alternative processing sites. In addition, further mechanisms including RNA polymerase II kinetics can regulate alternative polyadenylation. **(B)** Global overview of alternative polyadenylation (obtained by transcriptome-wide TRENDseq) after depletion >170 components involved in various facets of transcriptional, co- and posttranscriptional gene regulation, epigenetic modifications, and further processes with a potential role in the definition of RNA 3'ends. APA shortening and lengthening refers to CPA at an upstream or downstream positioned poly(A) site (PAS). This can modulate CPA at alternative PASs located in the 3'UTR ("tandem" APA) or in the coding sequence (CDS) and/or intronic region ("internal" APA), resulting in transcript isoforms with distinct regulatory properties or encoding c-terminally modified proteins. TREND-DB is designed to provide interactive and accessible exploration of this data set.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.04.235804; this version posted August 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

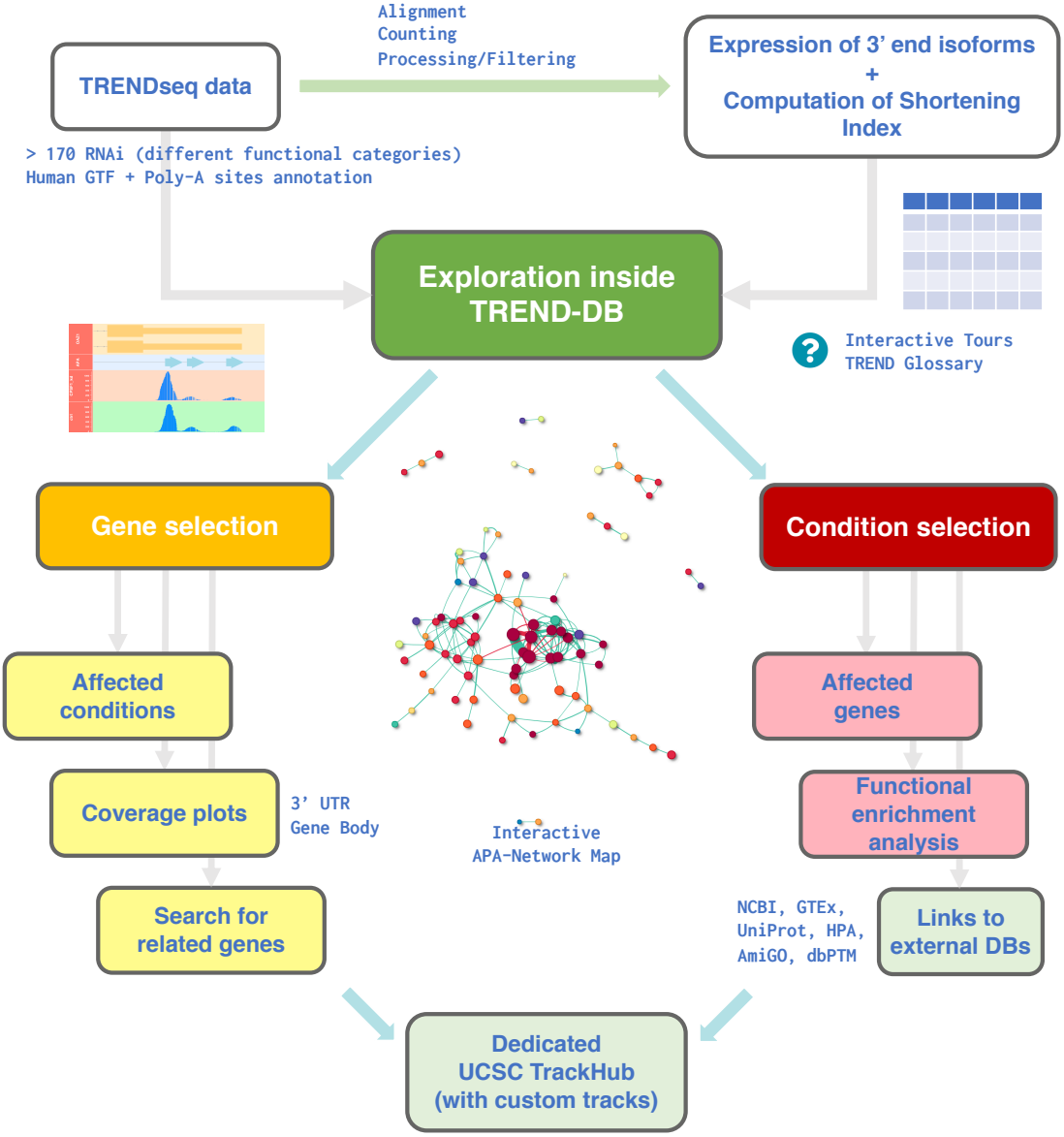

Figure 2: Overview of the exploration workflow with TREND-DB. TRENDseq data were aligned, quantified, and processed (12) to obtain the expression values of APA transcript isoforms, and the corresponding Shortening Index (SI) across all siRNA-conditions (displayed in the ***Data Preview*** tab). The ***Main View***, where an interactive APA-network map is displayed, enables users to inspect both dimensions of the underlying resource. Upon selection of a gene (left, ***Gene View***), all siRNA conditions affecting APA of this gene are displayed, accompanied by the TRENDseq coverage plot centered on the desired region (3'UTR or gene body); genes showing a similar APA phenotype (based on the SI profiles) can also be retrieved. When selecting a siRNA-condition (right, ***Condition View***), all APA-affected genes are listed, and users can perform functional enrichment analysis on the fly. External databases can be reached via dynamically generated action buttons, reducing the time to retrieve additional information. A dedicated instance of the UCSC Browser is provided, containing all the genomic tracks of the TRENDseq dataset, with the possibility to integrate custom tracks.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.04.235804; this version posted August 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

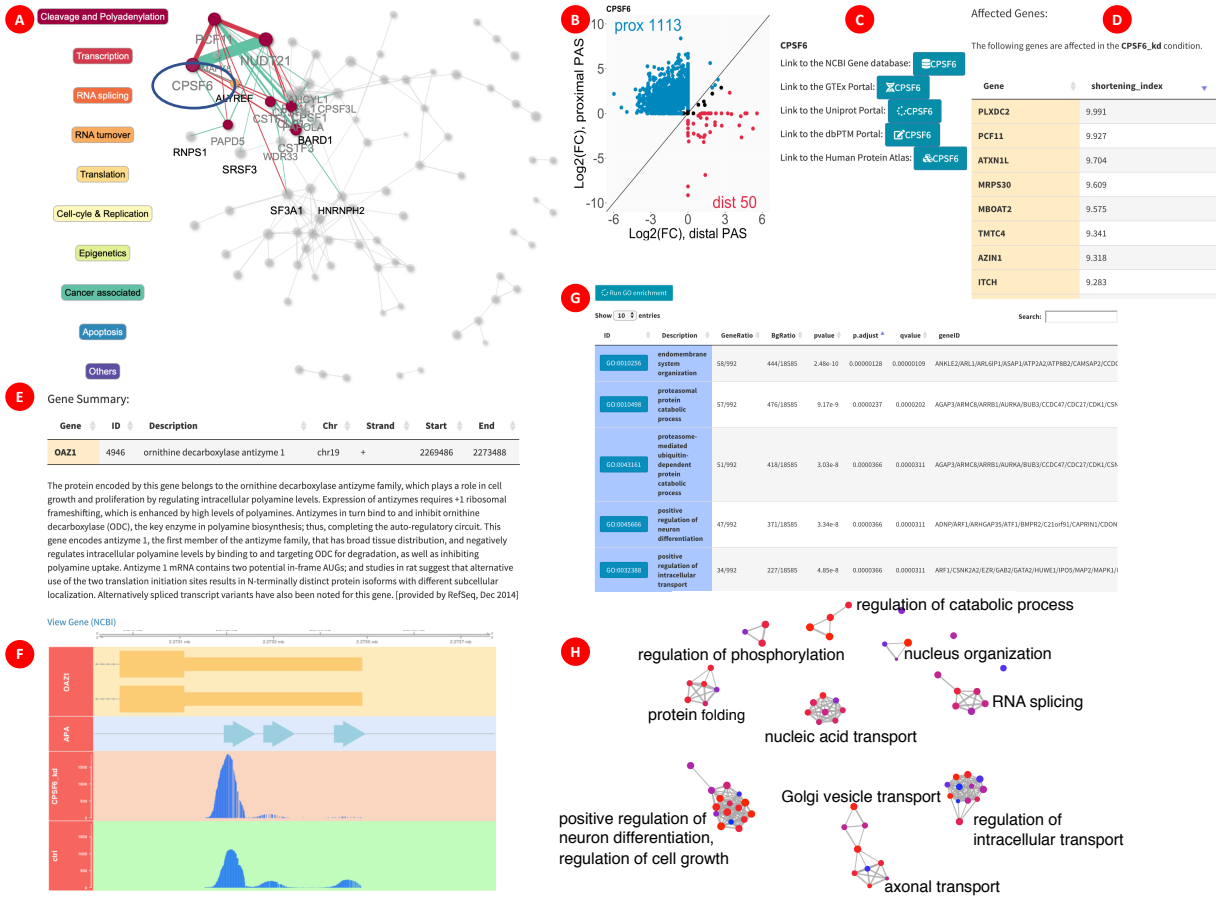

Figure 3: Selected screenshots to illustrate the exploration of the APA landscape with TREND-DB. From the ***Main View*** tab, users can select any condition among those included in the RNAi screening, either from a dropdown menu or from the interactive APA network **(A)**. For example, upon selection of CPSF6 (component of the CFIm protein complex, highlighted in the circle) 'neighboring' components/complexes that also regulate identical APA-affected target genes are highlighted. In this case CPSF6 was found to regulate APA with a global shortening bias, as displayed by the scatterplot in **(B)** summarizing the PAS usage. Links to external databases **(C)** are automatically displayed for simplifying further exploration steps beyond TREND-DB. Moreover, a table containing the affected genes (sorted by the Shortening Index in **(D)**) is displayed, and can be used to further explore the APA phenotype of a particular gene - e.g. OAZ1, both in the ***Main View*** **(E)**, where general information is shown, and in the ***Gene Plot*** tab, where users can inspect the coverage tracks for a locus of interest, such as the 3'UTR, where the effects of isoform shortening or lengthening is shown **(F)**. Using the whole set of affected genes in a particular screening condition (i.e. knockdown), TREND-DB enables the direct calculation of functional enrichment of APA-affected target genes. The functional enrichment is displayed in an interactive table **(G)**, or graphically as an enrichment map **(H)** for easier interpretation (for improving the readability, the single node labels have been replaced with a representative biological process for the clustered subset).