COMPUTATIONAL ANALYSES OF RNA-SEQUENCING DATA TO IDENTIFY SPLICING AND POLYADENYLATION REGULATORY ELEMENTS

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

MACIEJ BAK

Basel, 2023

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Erstbetreuerin: Prof. Dr. Mihaela Zavolan

Zweitbetreuer: Prof. Dr. Erik van Nimwegen

externer Experte: Prof Dr. Sven Bergmann

Basel, 14.12.2021

Prof. Dr. Marcel Mayor

The Dean of Faculty

Dedicated to all struggling PhD students.

Abstract

The following dissertation presents my scientific contributions to computational research in the field of RNA biology, carried out in the group of Professor Mihaela Zavolan at the Biozentrum, University of Basel, Switzerland. The projects in which I was involved focused on splicing and polyadenylation, two crucial steps in the maturation of eukaryotic mRNAs, and targets of post-transcriptional mechanisms for regulating gene expression. In the following sections I provide five concise summaries of scientific articles (published, under revision and in preparation) that I co-authored, providing my knowledge and expertise in bioinformatics. Briefly, these manuscripts cover the following aspects. Together with other colleagues from our group we developed a general purpose RNA-seq data processing workflow which utilizes best-practices in scientific software development for reproducible research. I further helped in analyzing the molecular impact of the LARP7 protein on the whole transcriptome, identifying splicing events that are deregulated upon changes in LARP7 expression level. Analogously, I worked with other members of our group in defining targets of the CFIm complex that is involved in alternative polyadenylation and in characterizing the downstream effects on CFIm-dependent polyadenylation on cellular pathways. Further, based on a previously published probabilistic framework for predicting binding sites of nucleic acid-binding proteins, I developed a user-friendly bioinformatics tool to search for binding sites of RNA-binding proteins on mRNAs. Finally, in my main PhD project I combined all the aforementioned expertise together to develop a complex computational workflow to infer the activity of RNA-binding proteins (RBPs) on alternative splicing and alternative polyadenylation from RNA-seq data. The core of my work consisted of a novel computational method to assess the impact of binding sites of RBP regulators on the inclusion of cassette exons in mature transcripts. We validated this method on various RBP knock-down datasets and uncovered proteins which are plausible drivers of differential mRNA processing in glioblastoma cancer. The workflow is implemented in a modular fashion, leaving room for expansion in potential subsequent research projects which are discussed in the final section of this dissertation. In conclusion, I present my novel scientific contribution to understanding the regulatory impact of RNA-binding proteins on mRNA processing from a computational perspective. This work is relevant not only for the results we present but also for creating an opportunity for other scientists to investigate RNA maturation in their research projects. I hope my work will aid others pursuing their scientific passion and that I indirectly contribute to many fascinating discoveries in the future.

Acknowledgments

First and foremost I wish to express my utmost gratitude to our group leader: Mihaela Zavolan, whom I profoundly respect. I feel honoured for the opportunity to work in her group. I cannot imagine having a better PhD supervisor, as she is extremely supportive and responsive but also strict and straightforward. I feel that our communication is always clear as she is open to discuss any problems that may arise along the way. I consider myself lucky to be able to learn from her advice and follow her suggestions in my scientific endeavour in computational biology. I would like to thank her a lot for her enormous patience with me and continued financial support in order to help me finish my PhD project.

I am immensely grateful for the opportunity to work with Erik van Nimwegen, from whom I learnt to appreciate the quantitative aspect of research. I recall that five years ago when I started the PhD programme I was a die-hard purist, completely blinded by the beauty of abstract conceptual frameworks in mathematics. Over time, as I have gained experience in statistical modelling, I began to understand that sometimes it is perfectly fine to use approximations in certain situations. The expertise I developed made me realise and deeply connect to the fact that "all models are wrong but some are useful".

Immediately upon joining the Zavolan group I started my PhD research in close collaboration with, at that time, a senior post-doc: Andreas Gruber. As the project grew in scope and matured we maintained a healthy professional relationship from which I picked up what analyses to perform in order to address distinct scientific questions as well as a great deal of "know-how" in the field of bioinformatics. I am extremely thankful for all the guidance I received throughout my PhD time. I feel that our teamwork helped me develop scientific rigour and vastly improved the efficiency of my work.

I would also like to thank Sven Bergman, an external advisor of my PhD committee. Our contact was not frequent yet I have kept notes after each consultation as well as every Thesis Committee Meetings and tried to follow the guidelines which we set up for me. I am also very grateful for the support and constructive help in navigating through my position as a graduate student.

I would like to express my sincere gratefulness to all the abovementioned for their time and effort they contributed towards working with me. I am very happy to admit that over the past 5 years I have been pursuing a dreamlike PhD project: one which combined development and efficient implementation of statistical models, high-quality scientific software engineering and biological discoveries in the field of RNA processing. Looking back at my time at the Biozentrum I can definitely say it was a life-changing experience. One that definitely shaped me as a scientist: sharpened my scientific method, steered my mindset towards data-driven hypotheses and increased my work culture enormously. For that, I will forever feel indebted.

I could not forget to mention all the members of the Zavolan lab. Our multiple stimulating discussions and useful suggestions I received at many occasions definitely improved my projects and helped me learn how to collaborate in research. I would especially like to communicate my appreciation towards Alexander Kanitz who, on numerous occasions, aided me in gaining competence in scientific software engineering. Additional thanks are directed towards the sciCORE team for providing a remarkably powerful computing infrastructure as well as uncommonly friendly and fast technical support. Research presented in the following dissertation would not have been possible without their backing.

Apart from all my scientific environment I am deeply thankful to all my friends, family and people who made me reach the point I am at right now. Their psychological help proved to be fundamental at times of crisis, which - of course - I experienced, as an inseparable part of the graduate studies. Samuel Abbott and Oliver Maric deverse honorable mentions here. Last but definitely not least I wish to articulate my wholehearted gratitude towards my partner. She constantly supported me and believed in me even though there were situations in which I certainly did not believe in myself.

Table of contents

Introduction

Regulation of gene expression

Eukaryotic organisms store their genetic information in the form of deoxyribonucleic acid (DNA) with the sequences encoding proteins structured into genes. The former are the most important of cell's macromolecules responsible for virtually every aspect of its life. The process of protein production based on the template of genes - gene expression - is therefore crucial for the survival of the whole organism and as such must undergo strict regulation.

Central dogma of molecular biology

The discovery of molecular structure for DNA in 1953 [1] has been marked as the cornerstone for a new branch of science, one which will later be referred to as molecular biology. While some preliminary research that highlighted differences in distinct types of nucleic acids already existed at that time [2, 3] it is that key event that led to rapid progress in the understanding of how genetic information is transferred and manifested in living organisms. Following this finding the term "central dogma of molecular biology" has been proposed, denoting that sequential information in biological systems can only be transferred from nucleic acids to proteins and not the other way [4]. Along with it the messenger RNA (mRNA) has been isolated [5, 6] and its role as an intermediary in gene expression process has been described [7]. Now, over 50 years later, when biological sciences expanded in scope and specialised into various sub-disciplines, the initial model of biological information processing still remains true to the most extent [8]. It has been proposed that due to its physico-chemical properties double-stranded DNA is more stable then single-stranded RNA [9] and thus is better suited as a medium for information storage. Consistently, DNA has emerged in evolution as the medium for genetic information storage and transmission via DNA replication [10], while various mechanisms have evolved to control gene expression, including at the level of transcription and translation [11]. With an intermediate level between DNA and proteins and the widespread presence of RNA-degrading enzymes - ribonucleases - which evolved not only to protect against RNA viruses but also to ensure the transient character of intracellular mRNAs, the kinetics of gene expression can be finely-tuned to enable nuanced responses to extracellular signals.

It is essential to note that, being critical to a cell's survival, the whole process of gene expression is closely controlled at distinct steps by various factors. Transcription initiation, apart from the formation and binding of the RNA Polymerase II holoenzyme to the promoter sequence of the transcribed gene, often requires favourable DNA topology with an open chromatin state [12] and additional binding of specific transcription factors (TFs) [13]. It has been reported that sequence-specific TFs are the most important and diverse mechanism of gene expression regulation [14], while transcriptional elongation and termination are mostly under the control of RNA polymerase II [15]. The end product of transcription of protein-coding genes in eukaryotes are the precursor mRNA molecules (pre-mRNAs), which must first undergo maturation before serving as a template for protein synthesis. The processing of pre-mRNAs and the post-transcriptional regulation is described in more detail in subsequent sections. The efficiency and speed of translation, similarly as for transcription, are controlled by various proteins which in this case directly interact with the ribosome [16]. Finally, there is a post-translational layer of regulation of gene expression with post-translational modifications (PTMs) having the most significant effect on the proteome [17]. Some proteins are tagged with so-called signal peptides that lead the cell to translocate these proteins to specific subcellular compartments or secrete them into the extracellular medium [18]. Altogether these findings show that the genetic information undergoes several controlled transformation steps before reaching its desired functional state.

mRNA processing

With the fundamental role RNA molecules have in gene expression, it is important to provide more insight into the previously mentioned process of mRNA maturation, which not only leads to gene expression but also allows the production of multiple variants - called isoforms - from a given gene, thus diversifying the functionality of eukaryotic cells [19, 20]. The first step of RNA processing is ‘capping’, which is the addition of a 7-methylguanosine nucleotide (cap) at the 5’-end of the primary transcript. The cap increases stability of the mRNA in the cytoplasm and serves as an anchor for ribosomes to initiate translation [21]. This process is tightly controlled by the C-terminal domain of RNA polymerase II. Although a few cap variants have been described [22], the 7-methylguanosine cap is generally used, leaving little room for variation at this step of RNA processing.

The RNAs generated from eukaryotic genes, especially in organisms such as mice and humans, are not contiguous copies of the respective genes. Rather, they are composed of fragments that are included in the mature RNA - exons, flanked by 3’ and 5’ splice sites - as well as fragments that are removed and degraded during the process of RNA maturation - introns. This is not to say that introns are entirely non-functional, as several studies point out various functions of intronic regions [23]. The process by which exons are excised from the pre-mRNA is called splicing. It is at this step where the first layer of transcriptome diversification takes place as alternatively splicing, the inclusion of an exon in some transcripts but not others, is a common occurrence [24]. Alternative splicing thus gives rise to distinct transcripts from the same template gene, variants that are called isoforms. The proteins translated from distinct isoforms may differ in their chemical properties like hydrophobicity or solubility but also in their biological function in cases where the structure of the proteins changes for example by the inclusion or exclusion of a protein domain. For instance, the shorter isoform of Bcl-x, a mitochondrial transmembrane protein, is pro-apoptotic whereas its long form - anti-apoptotic [25]. Splicing variants can differ in a variety of ways [26], the most prevalent variation being the skipping of ‘cassette exons’ [24]. Other variations are alternative 3’/5’ splice sites, mutually exclusive exons or retained introns. In humans over 95% of multi-exonic genes undergo alternative splicing [27] thus the effect of variations in exon composition over the whole transcriptome and the resulting proteome is vast.

The other crucial step of pre-mRNA maturation, apart from splicing, is the formation of the pre-mRNA 3’ end, which involves 3’ end cleavage and polyadenylation. The polyadenylation signal (PAS) is AAUAAA and some close variants [28, 29]. The 3’end processing complex binds PAS and cleaves the precursor mRNA sequence 10-30 nucleotides downstream after which a poly-adenosine tail of 200 nucleotides in length is added [30]. That sequence is commonly referred to as a poly(A) tail and its primary functions are to increase the stability of the mRNA by protecting it against the exonuclease-mediated degradation [31]. Most human genes contain more than one polyadenylation signal therefore a given transcript might be cleaved at different poly(A) sites [32]. This will affect the transcript sequence, as more distally-located poly(A) sites give rise to longer transcripts, which contain additional regulatory elements compared to the shorter isoforms. Frequently, in protein-coding genes, alternative polyadenylation at different poly(A) sites gives rise to transcripts that only differ in the untranslated region located beyond the protein-coding region (3’ UTR). These regions are relevant for the mRNA’s stability, subcellular localisation, or translation rate [33].

All in all, alternative splicing and alternative polyadenylation come up as two very powerful mechanisms of post-transcriptional regulation of gene expression, having a strong impact on the sequence of mature mRNAs and therefore proteins.

cis- and trans- acting regulators of transcript maturation

As both splicing and polyadenylation are context-dependent, the question arises how these processes are regulated in different cell types. In transcriptional control, the context-dependent transcription of genes is implemented by transcription factors (TFs), DNA-binding proteins that are expressed in a cell-type specific manner and bind to promoter and enhancer regions to recruit the RNA polymerase to these regions. Analogously, it has been shown that RNA-binding proteins (RBPs) with cell-type-specific pattern of expression bind to sequence elements in precursor and mature RNAs to modulate their splicing, polyadenylation, but also transport, localization and translation [34]. Recent studies highlighted multiple proteins which might be involved in more than one step during mRNA processing [35, 36, 37] although the effect is not always direct. Despite RBPs usually detecting cis-acting short sequence motifs on an unstructured single-stranded mRNA some authors report that the secondary structure of the target molecule plays a significant role in the interaction [38]. The accessibility of the binding site indeed seems to be important as well, however regulators of mRNAs maturation are commonly defined by the specific sequences they bind, suggesting that sequence-specific recognition is at the core of that regulation. Indeed, studies that explicitly investigated the role of the structural context in which the sequence motifs are located found this to be substantial only for a small number of RBPs [39].

Apart from RNA binding proteins other molecules may also detect sequence elements on both primary and processed transcripts. Three other most notable groups of regulators are microRNAs (miRNAs), small interfering RNAs (siRNAs) and long non-coding RNAs (lncRNAs), all of them very well studied and described in the literature [40, 41, 42]. These RNAs typically have regulatory functions, recognizing their RNA targets via Watson-Crick base-pairing. The miRNAs as well as the siRNAs are short, 21-23 nucleotides, binding their targets through a so-called 'seed sequence' located at the 5' end of the small RNA. In most cases their interaction sites are in the 3'UTRs of mRNAs [43] but this is not a strict rule and interactions with 5'UTRs, coding sequences and gene promoters have also been reported [44]. The most common outcome of the interaction is the degradation of the target mRNA through RNA-induced silencing complexes (RISCs). In contrast to the miRNAs, which are encoded in the genome, siRNAs are typically of exogenous origin and designed to be perfectly complementary to a particular gene [45]. Long non-coding RNAs are more than 200 nucleotides in length and are capable of interacting with DNA, mRNAs, microRNAs and proteins [41]. They are very diverse and can regulate gene expression at multiple post-transcriptional levels: interacting with splicing factors [46], behaving as miRNA sponges [47], serving as precursors for miRNA or endogenous siRNAs [48], binding to proteins that mediate mRNA decay or binding directly to mRNAs to regulate their stability [49].

As discussed, the maturation of mRNAs is a highly regulated process with multiple distinct types of regulators affecting different aspects of the transcript's fate. The vast majority of them act through their respective binding sites located on the mRNAs. It is therefore crucial to assess the overall effect of these binding sites on the two most potent mechanisms of transcriptome diversification: alternative splicing and alternative polyadenylation.

High-throughput biology

Biological sciences started as rather descriptive. In his revolutionary "Origin of Species", published in 1859, Charles Darwin compared morphological characteristics between finches living in the Galapagos Archipelago, this work having laid the foundations for the field of evolutionary biology. As considered by many, the first groundbreaking quantitative discovery in the field of genetics came in 1866 when an Augustinian friar Gregor Mendel published his work on the heredity of different traits of pea plants. He coined the terms "dominant" and "recessive" in regards to various characteristics based on their ratios in the offspring. Nowadays, over 150 years after these initial discoveries, biological research evolved into a multidisciplinary field which applies increasingly quantitatively precise methods to gather observations on cellular and molecular scale. Theoretical models are also constructed and tested against experimental measurements, similar to what is done in physics, though still at a more limited scale [50]. Some of the noteworthy high-throughput quantitative techniques in use today are: mass spectroscopy [51], ribosome profiling [52], genomic DNA sequencing [53], assay for transposase-accessible chromatin using sequencing [54], variants of cross-linking and immunoprecipitation with sequencing [55] and chromatin immunoprecipitation with sequencing [56]. All of the resulting data needs specialized computational methods for analysis and interpretation. Last but not least, novel RNA sequencing technologies, which will be elaborated upon in the next section, allow scientists to investigate gene expression at the transcriptome level.

RNA-sequencing platforms

As the gene expression profile of a cell determines the cell's repertoire of molecular functions, various approaches have been developed to quantify cellular transcriptomes. The dawn of the current century was dominated by the microarray technology, extensively described in [57] and [58]. While different versions were in use, as mentioned in [59], the key principle was that the RNA content of a population of cells was hybridized against a set of probes designed to have specific sequences, complementary to the transcripts that were annotated at that point. Due to a complicated procedure and especially to the reliance on genome annotation [60], the microarray technology was superseded by more reliable techniques that also allow the de novo discovery of RNA species present in cells. These approaches may be roughly grouped into short-read and long-read sequencing, both very well described in the literature [61, 62]. Short-read sequencing technologies further come in various flavors such as sequencing by ligation (SOLiD platform by Applied Biosystems, Complete Genomics platform) and sequencing by synthesis (platforms of: Illumina, Qiagen, 454, Ion Torrent). Both of these require prior amplification of the template material [62]. This is unlike one of the two main long-read sequencing technologies: single-molecule real-time sequencing, which can directly detect DNA without the amplification step. In contrast, synthetic long-read sequencing technologies do not produce long reads per se but instead utilise read barcodes such that downstream computational assembly of larger fragments is feasible. The leading advances in single molecule long-read sequencing came from Pacific Biosciences and Oxford Nanopore Technologies, whereas the most common methods for synthetic long-read sequencing are introduced by Illumina and 10X Genomics [62]. A separate class of technologies provides single-cell RNA sequencing data. These are described in [63] and the main challenges in the field are pointed out in [64]. However, since the following dissertation revolves around bulk RNA-Seq data analysis, I will not focus on those.

At this point it is worth highlighting that each of the previously described techniques may require a dedicated strategy for downstream data analysis. Distinct instruments can introduce technology-specific biases, which need to be accounted for in the analysis. This renders appropriate post-processing a crucial part of the whole experiment, which requires expert knowledge of both the sequencing platform as well as of bioinformatics standards for data analysis. The analysis steps following an RNA-sequencing experiment have been extensively described in [65]; they involve adapter sequence removal, read alignment against a reference genome, gene expression quantification, differential gene expression analysis and data visualisation. Each step should be accompanied by a quality control mechanism which would allow scientists to narrow down plausible sources of errors, if such arise. Subsequent steps may be project-specific and largely depend on the purpose of the experiment.

In summary, there are various RNA-Sequencing technologies available on the market and proper downstream analysis is critical for correct understanding and interpretation of the biological data.The bioinformatics software developed for these types of tasks should therefore satisfy specific standards of quality, further discussed below.

Software engineering in science

Bioinformatics is an exceedingly fast-growing branch of natural sciences and its recent progress might be attributed to the exponential increase in available data [66]. As more scientists become involved in software development, the sheer number of specialised software packages available is also increasing [67]. Technological advances in biological sciences in the 21st century resulted in multiple large-scale data-generating technologies commonly referred together as "omics". These include: genomics [68], transcriptomics [69], proteomics [70], as well as metabolomics [71]. Each of them focuses on a distinct set of molecules and can address different scientific hypotheses. However, what is universal to all "omics" approaches is that digital data generated by distinct experiments require preprocessing, analysis, integration and often additional curation over time. Society and media has coined a novel term for techniques and approaches applicable across data-rich domains: "data science". Despite this phrase being often critiqued for its vagueness and its advocates for not following a strict scientific method, the name did find its way also into the academic environment, as many scientists adopt and use it in their publications [64, 72, 73]. It might be therefore beneficial to highlight relevant technical skills often associated with positions advertised as such. Great emphasis is placed on: exploratory data analysis, data visualisation, software engineering, information integration, statistical methods, database management, data-driven inference, efficient computing, high-performance environments, machine learning, mathematical modelling as well as dealing with so-called big data. Most of these competencies are necessary (but not sufficient) to perform analyses in modern-day computational research. Naturally, this applies to bioinformatics and computational biology too, where the focus is shifted from hypothesis-driven research on small systems towards processing and inference from experimental data. This led to the development of domain-specific standards, both in terms of data and metadata formats [74, 75] as well as best coding practices [76, 77, 78, 79].

Two of the most commonly used programming languages in the broad field of bioinformatics are Python and R, both very well suited for data analysis-related tasks. Most important packages and their application in research for the former one have been extensively described in [80, 81, 82] whereas the whole scientific ecosystem of the latter is presented in [83]. A notable drawback is that both of them fall into the category of interpreted programming languages. In order to overcome the efficiency limitation for cases in which the speed of computations is a critical factor, software engineers developed dedicated interfaces between both Python & R and low-level compiled languages [84, 85]. On top of that, novel programming languages oriented around efficient data processing are being actively developed, most notably: Rust [86] and Julia [87]. Their appearance has not remained unnoticed by the scientific community [88, 89]. In addition to these efforts attention is also directed towards the development of integrated and interactive exploratory data analytic environments [90, 91] as well as platforms for collaborative work [92] and cloud computing [93]. This has been accompanied by the recent growth of strictly software engineering toolset related to: unit/integration testing with code coverage measurement, automatic documentation, static code analysis, packaging, continuous integration/continuous delivery solutions and dependencies management. All of these are considered good coding practices and facilitate the delivery of high-quality software for science. Two most notable advances are: the rise in popularity of workflow management systems and specification languages [94] as well as software encapsulation mechanisms. For the former: CWL [95], Nextflow [96], or Snakemake [97] are good representatives. However, despite their advantages, these systems alone cannot guarantee flawless execution of a computational workflow. Due to specifications of either the hardware architecture or the operating system of the host machine it is not uncommon to encounter obstacles in installation of the software required for the analysis. These problems are best addressed by container technologies like Docker [98] or Singularity [99] but also general purpose package managers and repositories: PyPI [100], conda [101] and Bioconductor [102]. Such solutions allow researchers to easily install and execute scientific software in a platform-independent manner. Together with previously mentioned workflow management systems they ensure code reusability and results reproducibility for even the most complicated data processing pipelines as individual steps are transparent and abstracted. Moreover, as the presented technologies are themselves software, they too continuously grow and expand in functionalities, according to the formerly listed software engineering techniques. That modularity allows for faster, more stable and reliable software development for research.

Currently there are numerous advancing community initiatives in order to facilitate scientific data management as well as proper metadata annotation, especially in the area of biomedical research. The most noteworthy one being the establishment of FAIR principles - a set of guidelines which emphasize four key aspects of scholarly data: Findability, Accessibility, Interoperability and Reusability [103]. Other efforts are directed, amongst others, towards deploying federated computing infrastructures [104], designing standards for metadata [105], framing policies around sensitive human-related data [106], curated software management [107, 108, 109] and integration of information across multiple research centers across different countries. To address these tasks global organizations are formed, most notable of which are: Elixir Europe and Global Alliance for Genomics & Health. Such corporations are not only highly fruitful in terms of advancing pure technical skills but also enable specialists in different aspects of research to discuss, establish partnerships, contribute their expertise and advance together towards developing a secure, strong and nowadays crucial support system for life sciences in general.

I personally believe it is essential to stress out the importance of high-quality scientific software engineering, of which I am a dedicated advocate. Due to that reason, the method which we have developed as the core part of my main PhD project, presented within this dissertation, is packaged into a fully automated computational workflow and I dedicated significant time and effort so that it meets the criteria of reproducibility. My goal is to deliver stable and open-source software for data analysis which might be later utilized by the whole scientific community in their research.

ZARP: An automated workflow for processing of RNA-seq data

RNA-seq data processing is a very common task in molecular biology projects, as gene expression analysis is key to understanding cellular function. As research in biological sciences becomes ever more interdisciplinary, it is essential for us - bioinformaticians - to share our experience and expertise with scientists outside of our specialisation, to improve the progress of the field as a whole. In our group we have frequently come across the fact that the analysis of RNA-seq data is a bottleneck in molecular biology projects, even for labs that do cover competencies in both experimental and computational biology. For this reason we have decided to develop ZARP: a general purpose computational pipeline for RNA-seq data analysis, that implements most common steps using tools and parameters that we have found optimal in our own experience. ZARP's steps are the following: sequenced reads are trimmed of adapters as well as poly(A) tails with cutadapt [110], aligned against a reference genome and transcriptome with STAR [111] and the expression levels of genes as well as transcripts are quantified with Salmon [112] and Kallisto [113]. The quality of the data is further assessed with three different approaches: the FastQC tool [114] is used to summarize various metrics regarding the quality of the sequenced reads, the ALFA tool [115] is used to functionally annotate the samples and the transcript integrity number calculation [116] is used to assess the degree of RNA degradation in the samples, which is an important factor for the accuracy of quantification of RNA levels [117]. Finally the results are presented in a user-friendly interactive report. ZARP is implemented in the snakemake workflow management system. To ensure reproducibility and reusability it makes use of best software development practices, including execution with conda or singularity technologies (alternatively). It is hosted as a public repository on GitHub; that way our work is fully transparent and the whole community of scientific software engineers can not only clone it but also interact with us in the form of suggestions or code improvements.

ZARP is a collaborative effort of many members of the Zavolan group, developed mainly during the partial lockdown due to the COVID19 pandemic; we developed it together during regular hackathons as well as individually, focusing on previously assigned tasks. My contribution to the project involved (chronologically):

- preparing tool-specific Docker containers
- preparing a separate software repository for the calculation of transcript integrity number and reporting; improvement to the original code
- adding two reporting mechanisms at the end of the workflow: snakemake report with technical information regarding the runtime and resources allocated, MultiQC [118] report with sample-wise statistics at distinct steps of the analyses; the latter included developing two MultiQC plugins for publicly available tools (TIN score calculation, ALFA)
- integrating snakemake profiles mechanism for workflow execution on a computational cluster (currently supported: SLURM Workload Manager)
- project management on GitHub as well as designing pipeline integration tests via GitHub's Action mechanism for Continuous Integration/Continuous Development.
- writing and editing the manuscript

In the end we provide a stable, reliable and transparent computational pipeline to provide first insights into RNA-seq data. The projects that I will present in the subsequent sections of this dissertation include this step of initial data processing and build on top of it in order to detect targets of alternative splicing and alternative polyadenylation and later to infer regulators of these processes.

Manuscript describing this work is currently in preparation.
Full text is included in this dissertation as Appendix B.

The Alazami Syndrome-Associated Protein LARP7 Guides U6 Small Nuclear RNA Modification and Contributes to Splicing Robustness

As mentioned in the introduction, gene expression has many layers of regulation. Recent research has identified RNA modification as a critical component of this system, governing the processing of nascent mRNA [119]. A distinct class of molecules long-known to be guiding the modification of RNAs are the small nucleolar RNAs (snoRNAs), whose importance has been highlighted in both physiological and pathological conditions [120, 121, 122]. snoRNAs generally guide the modification of ribosomal RNAs (rRNAs) and are therefore of two types, box C/D snoRNAs that guide the 2'-O-methylation of rRNAs, and box H/ACA snoRNAs that guide the pseudouridylation of rRNAs [123]. Some snoRNAs are involved in the processing of rRNAs [124], or other types of targets [125], and there are snoRNAs that are so far considered "orphan" because no target has so far been described for them [126]. In this project we were contacted by the group of Gunter Meister, to help analyze the impact of a protein known as La-related protein 7 (LARP7), which has been previously described to function in transcriptional control along with the 7SK RNA [127] and playing a role in various cancer types [128, 129, 130] as well as other diseases [131]. Despite previous research efforts, the cellular functions of LARP7 have not been fully understood. Following observations of the Meister group that LARP7 interacts with both U6 snRNA as well as U6-specific C/D box snoRNAs and that LARP7 depletion results in reduced U6 RNA modification, we were asked to help identify potential splicing targets involved with LARP7. Spliceosome assembly and canonical splicing were generally not affected by the absence of this RBP, but our analysis of RNA-seq data uncovered some changes in previously described alternative splicing events. This project provided evidence for the importance of LARP7 in RNA 2'-O-methylation of an snRNA, with consequences for the transcriptome and potentially in the context of the Alazami syndrome. I have contributed to it as an external collaborator providing:

- initial RNA-seq data analyses: preprocessing, quality control, read alignment, gene and transcript expression level quantification
- differential gene expression analyses
- quantification of alternative splicing events followed by differential splicing analyses
- Gene Ontology enrichment analyses
- short sections with results visualisations and edits to the manuscript

Within the presented study we have analysed RNA-Seq data and identified transcripts which undergo differential processing upon LARP7 depletion - targets of alternative splicing. Part of the splicing events which we have quantified and called as statistically significantly perturbed have been formerly annotated as cassette exons. Being personally intrigued by the excision of selected coding sequences from primary transcripts, I wanted to investigate how this specific process might be regulated. I aimed to develop a computational method which would infer which RNA-binding proteins could affect mRNA maturation. Pursuing this scientific question has led me to my main PhD research project, which is described in the following section of this dissertation.

This work has been published in the Molecular Cell journal.
Full text of the article is included in this dissertation as Appendix C.

CFIm-mediated alternative polyadenylation remodels cellular signaling and miRNA biogenesis

A separate mechanism of regulating gene expression at the post-transcriptional level is the alternative cleavage and polyadenylation. For the majority of human genes their transcripts undergo processing at distinct poly(A) sites thus leading to isoforms which differ in their 3'end sequences [32]. This diversification of the transcriptome contributes to various cellular processes like cell growth, proliferation, differentiation, and is especially important in many diseases as discussed in: [132] and [133]. Along with alternative promoter usage, alternative polyadenylation is the most common mechanism of transcript diversification in humans [134]. 3' end processing is carried out by the core pre-mRNA 3' end processing complex, composed of 20 proteins with notable role of four distinct subcomplexes [135]. One of these is the mammalian cleavage factor I (CFIm), responsible for the recognition of UGUA sequence motif on the primary transcript and allowing for skipping of poly(A) site upon binding to these sequences [135]. CFIm is a tetramer, composed of two copies of the CFIm25 subunit and two larger subunits of CFIm68 and/or CFIm59 [136]. The knock-down of CFIm25 or CFIm68 leads to global shortening of 3'UTRs of transcripts [137, 138, 139]. Various groups have tried to link CFIm to the 3' UTR shortening observed in cancers [140], even though the expression of CFIm seem to rather increase in proliferative cell states compared to cell states associated with reduced proliferation [141]. To shed light on this discrepancy, our group initiated a study on the downstream effects of CFIm25 and CFIm68 overexpression or knock-down in various cell lines. The first step in this effort was to construct a comprehensive list of mRNAs that respond coherently and robustly to perturbations in the expression levels of the two proteins. Further analysis of these targets pointed to the potential involvement of the ERK signalling pathway as well as of microRNAs in the downstream behavior of cells. Both of these aspects were then validated in our lab. Our systematic analysis of CFIm targets improves the understanding of CFIm's role in the integration of RNA processing with other cellular processes.

The project was mostly driven by a colleague of mine - an experimental biologist working in our group. My scientific contribution include:

- initial RNA-seq data analysis: preprocessing, quality control, read alignment, gene and transcript expression level quantification
- quantification of poly(A) sites usage, inference of CFIm-dependent targets (using the PAQR method [29], with further modifications)
- Gene Ontology enrichment analyses
- short sections with results visualisations and edits to the manuscript

My main work in this project revolved around the quantification of differential 3'UTR processing of primary transcripts. I have utilised a previously published method: PAQR in order to come up with a comprehensive set of genes which undergo coherent cleavage and polyadenylation in response to CFIm expression changes. Building on the work done in the group on PAQR as well as on KAPAC, a method for inferring the impact of short sequence motifs on 3'end processing, in my main PhD project I decided to work on a statistical model analogous to KAPAC but designed for inferring the impact of short sequence motifs on pre-mRNA splicing.

This work has been published in the Nucleic Acid Research journal.
Full text of the article is included in this dissertation as Appendix D.

Inferring binding sites of RNA-binding proteins with bindz

An essential part of any study that investigates how trans-acting regulators may influence the processing of primary transcripts is the identification of their binding sites on the pre-mRNA sequence. It is such short sequence motifs that cis-act on alternative splicing and altenative polyadenylation of a given gene. There are various tools to predict binding sites of nucleic acid binding proteins, e.g. as part of the MEME suite [142], but these are generally part of larger suites, designed for various specific problems and a simple question that we frequently come across in experimental studies, such as what binding sites can we predict occur in an RNA sequence given the current knowledge of trans-acting factors, does not have an easy-to-find answer. To address this question we set out to develop a web-based analysis tool, bindZ, that predicts binding sites for RBPs and miRNAs of interest in a transcripts sequence. We further wanted to predict the impact of specific mutations on the interactome of individual mRNAs. I have been involved in the RBP-binding site prediction component of this tool, which is completed. The miRNA-binding prediction module and the web interface remain to be finalized. At the core of the RBP-binding module we incorporated MotEvo [143] - a Bayesian probabilistic method for the prediction of binding probabilities between a selected motif (represented in the Position Weight Matrix format) and a given nucleotide sequence. We have adapted the original work such that it is suitable to infer plausible binding events for various RNA-binding proteins, for which the motifs' PWMs we obtained from the ATtRACT database [144]. As the tool's output users are presented with a plain text table which summarizes: coordinates of binding events, their energies as well as posterior probabilities. Additionally, these results are graphically represented as heatmaps so that one may quickly grasp the most important observations and draw meaningful conclusions easily. From the software engineering perspective the module is also a snakemake workflow which utilizes conda technology to ensure reproducible research. I have developed it with the help of a summer student during his internship in our group. My work may be summarised as:

- Design and development of the computational workflow encompassing the tool's functionality (from the scientific software engineering perspective).
- Development of the initial data processing steps.
- Project management and coordination with the summer student at the later stages of the project.
- Writing and editing the corresponding section of a manuscript (in preparation).

Similar to ZARP, bindZ is being developed having the whole community of molecular as well as computational biologists in mind. Its primary use case is to aid researchers in the design of experiments with sequence variants that may differ in their ability to bind RBPs and microRNAs, due to specific point mutations that might create/destroy binding sites. More advanced users might automatise the analyses in such a way which would allow for global screening of selected RPBs over a wide range of pre-mRNA sequences, thus enabling more general studies on the regulation of gene expression at the post-transcriptional level. It is the second approach that I will utilise in my main research project, presented in the next section of this dissertation. Given Position Weight Matrices which represent binding motifs of distinct RNA-binding proteins I will infer probabilities of binding of these regulators in the proximity of 3'/5' splice sites of cassette exons as well as proximal and distal poly(A) sites. This information will serve as input data for our statistical models which assess the regulatory impact of RBPs on the mRNA maturation process.

An application note describing this tool is currently in preparation.
Full text is included in this dissertation as Appendix E.

MAPP unravels frequent co-regulation of splicing and polyadenylation by RBPs and their dysregulation in cancer

My main PhD project focused on the inference of sequence motifs and RBPs that regulate splicing and polyadenylation from RNA-seq data and ties together themes I have previously presented in this dissertation: RNA-Seq data processing, scientific software engineering, statistical data analysis, quantification of alternative splicing events, quantification of differential poly(A) sites usage and the inference of binding sites for RNA-binding proteins on the sequences of transcripts. With the knowledge and expertise I had acquired during my other scientific contributions I aimed to combine them into a high-quality bioinformatics tool which would uncover novel aspects of mRNA processing. In pursuing this scientific journey we have developed a complex computational pipeline which, given raw sequencing data, is able to infer the most plausible sequence motifs driving both differential inclusion of cassette exons and the choice of distinct poly(A) sites. Our results confirm a dual-action of both HNRNPC and PTBP1 proteins on both processes. We further uncover a position-dependent effect of RBFOX1 on alternative splicing for which binding in exonic vs. intronic sequences seem to have an opposite effect on the inclusion of an exon. We present a list of multiple regulatory RBPs together with their impact profiles around 3'/5' splice sites and poly(A) sites. Finally we investigate the patterns of mRNA processing in glioblastoma, a very aggressive cancer. We conclude that RNA-binding proteins which we previously studied: PTBP1 and RBFOX1 act in concert and are the two main regulators responsible for the global excision of cassette exons in this disease. We analyse downstream targets of differential splicing and discover that multiple proteins which have been previously reported to be associated with the cancer phenotype are indeed targets of these RBPs. In conclusion, we developed a complex but high-quality bioinformatics tool to analyze RNA-Seq data and infer potential regulators of mRNA maturation. We presented the validity of our method on several datasets providing insight into molecular mechanisms of RBPs action on distinct conditions. Our pipeline is developed under open source license and once published in a scientific journal will be publicly available for other scientists to utilise in their projects. My work during this project included:

- Investigation into available tools related to quantification of alternative splicing events; improvement of an existing strategy in order to obtain reliable estimates of cassette exon inclusion from RNA-Seq data.
- Design and implementation of a Bayesian statistical model to explain differential cassette exon inclusion across RNA-Seq samples with the activity of short sequence motifs.
- Design and development of most of the computational workflow (from the software engineering perspective); modular development of sub-pipelines dedicated to distinct functionalities.
- Method validation and parametrization on selected RBPs RNA-Seq knock-down experiments publicly available from NCBI servers. Global screening of almost 500 RNA-Seq knock-down experiments of various RNA binding proteins from the ENCODE project; post-processing and downstream analyses.
- Analysis and post-processing of glioblastoma RNA-Seq data sets from The Cancer Genome Atlas project.
- Writing and editing the manuscript.

Manuscript describing this work is currently under revision in the Nature Communications journal.
Article preprint has been uploaded to the bioRxiv server.
Full text is included in this dissertation as Appendix F.

Conclusions and future prospects

The research detailed within this dissertation is dedicated to crafting computational methodologies aimed at unraveling the intricate stages of RNA processing pivotal in deciphering gene expression levels. Commencing from the processing of raw sequencing reads as well as quality control procedures, through the accurate quantification of gene and isoform expressions with the most highlight on the modeling of the intricate regulatory influence wielded by RNA-binding proteins upon primary transcripts. Finally, I explored the downstream targets encompassing differential exon inclusion and poly(A)-site usage. This comprehensive approach aimed not only to illuminate specific molecular mechanisms related to splicing and plyadenylation of pre-mRNA but also to provide an understanding of their global impact on gene expression. Throughout my doctoral journey, I presented the practicality and efficacy of these methods in addressing paramount queries within the field of computational biology. A focal point of my endeavor was the development of user-friendly and reproducible tools. Notably, the methodologies, especially the MAPP method, stand as open-source computational workflows nwhich I maintain and enhance from the software engineering point of view up to this day. Moreover, my horizon in biological sciences, particularly in the domain of RNA biology, has been considerably broadened. Crucially, I've acquired the proficiency to approach and address complex scientific inquiries with computational approaches and therefore I stand equipped with a comprehensive skill set required for driving computational research independently, competently and at the highest standard of research reproducibility.

In my primary research during my PhD, the primary goal was to pinpoint the RNA-binding proteins responsible for orchestrating two pivotal facets of gene expression: the inclusion of cassette exons and the choice of distinct poly(A) sites at the stage of 3'end processing. These steps, as I outlined earlier, hold fundamental significance in how proteins are produced within eukariotic cells. In 2017 it has been reported that cancer is the second leading cause of death worldwide [145]. Previous studies already extensively detailed the pivotal roles that alternative splicing and polyadenylation play in the development of tumors [146] and [147]. These aberrant modifications in mRNA processing trigger a cascade of effects, impacting crucial cellular processes: sustaining proliferative signaling pathways via ERK and MAPK, circumventing natural cell growth inhibitors through PI3K-AKT signaling, activating specific proto-oncogenes while suppressing tumor-inhibitors, resisting programmed cell death, augmenting cell replication through the Wnt pathway, fostering the growth of new blood vessels via the vascular endothelial growth factor (VEGF), facilitating the metastasis of cancer cells, altering cellular metabolism, and evading immune system detection. Another noteworthy aspect is the profound influence of alternative splicing within the microenvironment surrounding tumors, succinctly summarised in [148]. These modifications ripple through the functionalities of T and B cells, altering their functions, along with affecting proteins present in the extracellular matrix (ECM). Moreover, the hypoxic conditions often characteristic of the tumor environment also trigger a specific program governing splicing of primary transcripts. Collectively, the intricate process of eukaryotic mRNA processing emerges as a pivotal mechanism significantly influencing the progression and growth of cancer. Beyond the realm of cancer, various diseases, such as spinal muscular atrophy, retinitis pigmentosa, cystic fibrosis, type 2 diabetes, beta-Thalassemia, and several neurological and urogenital disorders, stand as testament to the ramifications of incorrect mRNA processing [132, 149]. These are not usually linked to a global deregulation but rather some well-described mRNA processing changes on very specific genes. Nonetheless, understanding the nuances governing these processes isn't just a matter of academic curiosity; it opens doors for designing novel and effective treatment strategies and paving the way for innovative therapies.

A major group of regulators in cellular mechanisms comprises RNA-binding proteins, whose intricate involvement in human genetic diseases, both Mendelian and somatic, has been extensively detailed in recent literature [150]. The author underscores the adverse impacts of mutations occurring in genes that encode RBPs, shedding light on their multifaceted repercussions. These mutations might wield a diverse array of effects on RNA-binding proteins. For instance, they could potentially alter the expression levels of these proteins, leading to shifts in the relative proportions of alternative isoforms displaying distinct functionalities. Alternatively, mutations might truncate proteins or modify their amino acid sequences, severely impacting their interactions with cofactors or RNA targets. For those RBPs that serve dual roles as enzymes some mutations might even alter their enzymatic properties. Another plausible effect of such mutations could involve the mislocalization and aggregation of RNA-binding proteins. Additionally, products derived from mutated RBP genes might undergo incorrect post-translational modifications. Such mutations could even alter the physiochemical properties of the mutant proteins, affecting their solubility and consequently impacting subsequent binding events downstream [150]. In all these instances, the molecular functionality of the defective protein stands at risk of impairment, potentially leading to improper cellular pre-mRNA processing. This underscores the critical necessity of accurately identifying distinct RNA-binding proteins that oversee the processing of specific mRNA molecules across diverse conditions. From a clinical standpoint, RBPs are surfacing as promising targets for novel therapeutic interventions. Over recent years, several successful strategies have emerged for RBP-based treatments [151, 152, 153]. These strategies encompass a wide range of approaches, such as administering small natural or synthesized molecules that modulate the spliceosome machinery, developing inhibitory molecules targeting protein kinases involved in splicing, and indirectly influencing splicing regulation through transcriptional elongation - a process kinetically linked to alternative splicing. Additionally, artificially synthesized antisense oligonucleotides have gained substantial interest within the scientific community. One may distinghuish two separate cetegories of particualr interest: splice-switching oligonucleotides (SSOs) and RNA decoy oligonucleotides. The former, short sequences spanning 15-30 nucleotides, competitively bind to splicing factors in pre-mRNA, aiming to modulate splicing by interfering with spliceosome recognition of splice sites. Conversely, decoy oligonucleotides contain repeated sequence motifs recognized by specific RBP regulators, acting as protein sponges, effectively hindering the biological activities of these regulatory proteins [151]. Despite these advancements, there lacks a clinically approved strategy utilizing antisense oligonucleotides to interfere with the 3' end cleavage and polyadenylation of transcripts. Nonetheless, decoy oligonucleotides present a promising therapeutic approach, particularly given the involvement of multiple RNA-binding proteins through the same binding sites, as detailed in our manuscript. Hence, it becomes imperative to critically investigate the specific impact of RNA-binding proteins on both splicing and polyadenylation processes, delving into the modes of action at a fine-grained level of details in order to identify the primary mRNA processing regulators active under various conditions. This exploration isn't solely confined to the realms of basic research; it holds profound implications for understanding the intricate machinery governing gene expression at the cellular level. Moreover, these novel biological insights could potentially lay the groundwork for the development of new drugs and therapies — an overarching objective driving the design and development of MAPP.

While my PhD research has been instrumental in unearthing novel regulatory interactions, the inherent nature of scientific exploration dictates that each answer attained leads to a myriad of new questions. Work that culminated in this dissertation is no different in that sense. The profound revelations gleaned from this research serve not as conclusions but as catalysts propelling future scientific inquiries and I am particularly drawn to the following domains for further exploration and scholarly pursuit:

- **Insight into mRNA processing profiles of distinct tissues.** It is well established that alternative splicing constributes to the acquisition of tissue function and identity [154]. Key changes which drive the establishment of cell types arise early in the embryonic development stage and persist for distinct tissues throughout the adulthood. Splicing changes in numerous genes are usually coordianted by global regulators - mainly RNA-binding proteins. Thus it seems especially interesting to investigate the patterns of differential RBP activities between various types of cells. Our prelimiary and unused results of analyzing samples from over 20 distinct tissues publicly available from the Human Protein Atlas project uncover a striking pattern for U-rich motifs acting on inclusion of cassette exons in skeletal muscles and heart as well as C/T-motifs with strong activity in cerebral cortex, skeletal muscles and heart. These results suggest that the differences in gene expression profiles across various tissues are not only due to regulation at the transcriptional level but also because of tissue-specific regulation of mRNA maturation via RNA-binding proteins. As these are complex physiological systems with dynamic interation of all splicing regulators at once it seems reasonable to expect a different repertoir of multiple proteins affecting the processing of mRNA in each tissue. With MAPP we could aim to construct "impact profiles" which reflect spllicing regulatory network and may point towards similarities and differences in not only cassette exon inclusion but also alternative 3' end processing between specific cell types.

- **A broader analysis of differential mRNA processing in various cancer types.** It is well-known that both alternative splicing as well as alternative polyadenylation play a significant role in tumorigenesis [146, 155]. So far I have applied MAPP to a curated set of RNA-Seq samples obtained from glioblastoma patients, where we confirmed the pivotal role of PTBP1 and RBFOX1 in mRNA processing deregulation [29]. However, The Cancer Genome Atlas contains substantial amount of data coming from multiple distinct cancer types as well as different stages of cancers. It would be quite fascinating to construct an "Atlas of mRNA processing profiles in cancer" in order to uncover main RBPs driving the inclusion of skipped exons and poly(A) site choice in cancerous cells originating from various tissues. Additionally, a comprehensive exploration could involve delving into the intricacies of tumor heterogeneity, specifically examining the distinct patterns of mRNA processing that manifest within separate subpopulations of cancer cells. This investigation, however, necessitates the availability of RNA-sequencing samples derived from meticulously obtained tumor biopsies.

- **Investigation of the activity of RNA-binding proteins on intron retention.** Within the vast landscape of splicing events documented on the human genome, a notably significant category comprises the retained introns [156]. These particular introns play a pivotal role in the nuclear retention of host transcripts and are intricately linked to cellular responses to various signaling mechanisms [157]. The likelihood-based model for exon inclusion we developed during my PhD time is general enough to be easily adapted to encompass these splicing events as well. Conceptually speaking, both processes involve excision of a specific subsequence from primary mRNA and, provided quantified intron inclusion values, the downstream inference of motifs activities may follow the same principle. That is not to say that the same biological factors need to control and affect both processes. A slight modification to our existing method might provide a wider overview of the regulation of mRNA processing by orchestrated RNA-binding proteins. This expanded approach holds the promise of broadening the scope of our analytical pipeline significantly as integrating a different kind of splicing events into the analysis offers a more comprehensive panorama of splicing regulation at the stage of RNA processing.

- **Modelling differential mRNA processing with RNA-binding proteins' activity on single-cell level.** The advent of single-cell RNA-sequencing technologies has revolutionized our ability to distinguish distinct cell subpopulations based on their unique expression profiles. We currently observe a rapid growth in the field of scRNA-seq analysis [158]. However, despite this rapid advancement, the extent to which RNA processing diverges among individual cells within a population remains largely unexplored. While studies have shed light on the influence of transcription factors' expression on discrete cell subgroups [159], the potential impact of RNA-binding proteins (RBPs) on defining these distinct cellular subsets remains to be explored. Considering the pivotal role of RBPs in governing mRNA maturation, it is not unreasonable to hypothesise that distinct subpopulations of cells could be defined by differential RBPs activities. Although one could adapt our workflow to process single-cell sequencing data, obtaining reliable estimates of cassette exon inclusion and poly(A) site usage from single cells is very challenging. Unlike for bulk RNA-sequencing technologies, current sequencing depth for single cell methodologies is not sufficient to accurately estimate expression levels of specific gene isoforms. Despite that, recent advancements have introduced methodologies that offer promising avenues for quantifying alternative splicing events and poly(A) site expressions from single cells [160, 161]. These emerging techniques signify a potential breakthrough in the analysis of differential mRNA processing at the single-cell level. As such, investigating the nuances of mRNA processing regulation within cell populations holds tremendous promise as an exciting and forward-thinking research direction in this field.

- **Exploring proximal RBP interactions surrounding specific sites.** In a strict sense, our novel method associates quantified measures of mRNA processing — such as cassette exon inclusion or poly(A) site usage — with with the count of binding sites allocated for distinct regulators situated along the primary transcript. As such, these are cis-acting elements and may only serve as an indirect proof of RNA-binding protein's action. Complementary to the inferred motif activities one should ensure that the trans-acting regulators indeed bind to their respective binding sites. In our manuscript we utilised publicly available data from enhanced cross-linking and immunoprecipitation experiments (eCLIP) [162]. This extensive dataset empowered us to scrutinize whether endogenous pre-mRNAs exhibiting the most substantial alterations in mRNA processing are indeed targeted by their presumed regulators and that this binding is more pronounced then for a contol group of primary transcripts. This facet of our analysis not only solidifies our findings but also lays the groundwork for generalizing this approach to quantify and visually represent the effects of RBP binding on any set of foreground and background mRNA sequences. As we move forward, the intention is to encapsulate this aspect of our research into a separate bioinformatics tool tailored specifically for CLIP data post-processing. This tool, designed to unravel and interpret the intricate effects of RBP binding, holds the potential to stand as an independent entity, serving the broader scientific community in their exploration of RNA regulation mechanisms.

The aforementioned points offer intriguing opportunities stemming from my current findings, showcasing promising directions for future exploration. I believe I would like to pursue an academic career path and thus would be very eager to follow up on these fascinating ideas. I find myself increasingly drawn towards the methodological facets of computational research. Upon the successful defense of my doctoral thesis, my intent is to actively seek out a postdoctoral position. In the next phase in my journey I aim to immense myself in the realms of statistical modeling and machine learning within the field of bioinformatics, thereby enriching my expertise in these areas.

Bibliography

[1] J. D. Watson and F. H. Crick, “Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid,” *Nature*, vol. 171, pp. 737–738, Apr. 1953.

[2] F. W. Allen, “The biochemistry of the nucleic acids, purines, and pyrimidines,” *Annu. Rev. Biochem.*, vol. 10, pp. 221–244, June 1941.

[3] J. Hämmerling, “Nucleo-cytoplasmic relationships in the development of acetabularia,” in *International Review of Cytology* (G. H. Bourne and J. F. Danielli, eds.), vol. 2, pp. 475–498, Academic Press, Jan. 1953.

[4] F. Crick, “Central dogma of molecular biology,” *Nature*, vol. 227, pp. 561–563, Aug. 1970.

[5] S. Brenner, F. Jacob, and M. Meselson, “An unstable intermediate carrying information from genes to ribosomes for protein synthesis,” *Nature*, vol. 190, pp. 576–581, May 1961.

[6] F. Gros, H. Hiatt, W. Gilbert, C. G. Kurland, R. W. Risebrough, and J. D. Watson, “Unstable ribonucleic acid revealed by pulse labelling of escherichia coli,” *Nature*, vol. 190, pp. 581–585, May 1961.

[7] F. Jacob and J. Monod, “Genetic regulatory mechanisms in the synthesis of proteins,” *J. Mol. Biol.*, vol. 3, pp. 318–356, June 1961.

[8] E. V. Koonin, “Does the central dogma still stand?,” *Biol. Direct*, vol. 7, p. 27, Aug. 2012.

[9] A. Lazcano, R. Guerrero, L. Margulis, and J. Oró, “The evolutionary transition from RNA to DNA in early cells,” *J. Mol. Evol.*, vol. 27, no. 4, pp. 283–290, 1988.

[10] I. Frouin, A. Montecucco, S. Spadari, and G. Maga, “DNA replication: a complex matter,” *EMBO Rep.*, vol. 4, pp. 666–670, July 2003.

[11] S. Ramírez-Clavijo and G. Montoya-Ortíz, *Gene expression and regulation.* El Rosario University Press, July 2013.

[12] F. Kouzine, D. Levens, and L. Baranello, “DNA topology and transcription,” *Nucleus*, vol. 5, pp. 195–202, May 2014.

[13] K. M. Lelli, M. Slattery, and R. S. Mann, “Disentangling the many layers of eukaryotic transcriptional regulation,” *Annu. Rev. Genet.*, vol. 46, pp. 43–68, Aug. 2012.

[14] B. Pulverer, “Getting specific,” *Nat. Rev. Mol. Cell Biol.*, vol. 6, pp. S12–S12, Dec. 2005.

[15] D. Reines, R. C. Conaway, and J. W. Conaway, “Mechanism and regulation of transcriptional elongation by RNA polymerase II,” *Curr. Opin. Cell Biol.*, vol. 11, pp. 342–346, June 1999.

[16] K. Van Der Kelen, R. Beyaert, D. Inzé, and L. De Veylder, “Translational control of eukaryotic gene expression,” *Crit. Rev. Biochem. Mol. Biol.*, vol. 44, pp. 143–168, July 2009.

[17] G. A. Khoury, R. C. Baliban, and C. A. Floudas, “Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database,” *Sci. Rep.*, vol. 1, Sept. 2011.

[18] H. Owji, N. Nezafat, M. Negahdaripour, A. Hajiebrahimi, and Y. Ghasemi, “A comprehensive review of signal peptides: Structure, roles, and applications,” *Eur. J. Cell Biol.*, vol. 97, pp. 422–441, Aug. 2018.

[19] B. Modrek and C. Lee, “A genomic view of alternative splicing,” *Nat. Genet.*, vol. 30, pp. 13–19, Jan. 2002.

[20] M. Zavolan, S. Kondo, C. Schonbach, J. Adachi, D. A. Hume, Y. Hayashizaki, T. Gaasterland, RIKEN GER Group, and GSL Members, “Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome,” *Genome Res.*, vol. 13, pp. 1290–1300, June 2003.

[21] E. Decroly, F. Ferron, J. Lescar, and B. Canard, “Conventional and unconventional mechanisms for capping viral mRNA,” *Nat. Rev. Microbiol.*, vol. 10, pp. 51–65, Dec. 2011.

[22] A. Galloway, A. Atrih, R. Grzela, E. Darzynkiewicz, M. A. J. Ferguson, and V. H. Cowling, “CAP-MAP: cap analysis protocol with minimal analyte processing, a rapid and sensitive approach to analysing mRNA cap structures,” *Open Biol.*, vol. 10, p. 190306, Feb. 2020.

[23] M. Chorev and L. Carmel, "The function of introns," *Front. Genet.*, vol. 3, p. 55, Apr. 2012.

[24] Y. Cui, M. Cai, and H. E. Stanley, "Comparative analysis and classification of cassette exons and constitutive exons," *Biomed Res. Int.*, vol. 2017, p. 7323508, Dec. 2017.

[25] M. Stevens and S. Oltean, "Modulation of the apoptosis gene bcl-x function through alternative splicing," *Front. Genet.*, vol. 10, p. 804, Sept. 2019.

[26] Y. Wang, J. Liu, B. O. Huang, Y.-M. Xu, J. Li, L.-F. Huang, J. Lin, J. Zhang, Q.-H. Min, W.-M. Yang, and X.-Z. Wang, "Mechanism of alternative splicing and its regulation," *Biomed Rep*, vol. 3, pp. 152–158, Mar. 2015.

[27] W. Jiang and L. Chen, "Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 183–195, 2021.

[28] C. S. Lutz, "Alternative polyadenylation: a twist on mRNA 3' end formation," *ACS Chem. Biol.*, vol. 3, pp. 609–617, Oct. 2008.

[29] A. J. Gruber, R. Schmidt, S. Ghosh, G. Martin, A. R. Gruber, E. van Nimwegen, and M. Zavolan, "Discovery of physiological and cancer-related regulators of 3' UTR processing with KAPAC," *Genome Biol.*, vol. 19, p. 44, Mar. 2018.

[30] M. D. Sheets and M. Wickens, "Two phases in the addition of a poly(a) tail," *Genes Dev.*, vol. 3, pp. 1401–1412, Sept. 1989.

[31] C. J. Wilusz, M. Wormington, and S. W. Peltz, "The cap-to-tail guide to mRNA turnover," *Nat. Rev. Mol. Cell Biol.*, vol. 2, pp. 237–246, Apr. 2001.

[32] A. J. Gruber and M. Zavolan, "Alternative cleavage and polyadenylation in health and disease," *Nat. Rev. Genet.*, vol. 20, pp. 599–614, Oct. 2019.

[33] A. J. Gruber, R. Schmidt, A. R. Gruber, G. Martin, S. Ghosh, M. Belmadani, W. Keller, and M. Zavolan, "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 Res.*, vol. 26, pp. 1145–1159, Aug. 2016.

[34] S. M. García-Mauriño, F. Rivero-Rodríguez, A. Velázquez-Cruz, M. Hernández-Vellisca, A. Díaz-Quintana, M. A. De la Rosa, and I. Díaz-Moreno, "RNA binding protein regulation and Cross-Talk in the control of AU-rich mRNA fate," *Front Mol Biosci*, vol. 4, p. 71, Oct. 2017.

[35] M. J. Mallory, S. P. McClory, R. Chatrikhi, M. R. Gazzara, R. J. Ontiveros, and K. W. Lynch, "Reciprocal regulation of hnRNP C and CELF2 through translation and transcription tunes splicing activity in T cells," *Nucleic Acids Res.*, vol. 48, pp. 5710–5719, Apr. 2020.

[36] M. Nazim, A. Masuda, M. A. Rahman, F. Nasrin, J.-I. Takeda, K. Ohe, B. Ohkawara, M. Ito, and K. Ohno, "Competitive regulation of alternative splicing and alternative polyadenylation by hnRNP H and CstF64 determines acetylcholinesterase isoforms," *Nucleic Acids Res.*, vol. 45, pp. 1455–1468, Feb. 2017.

[37] M. Movassat, T. L. Crabb, A. Busch, C. Yao, D. J. Reynolds, Y. Shi, and K. J. Hertel, "Coupling between alternative polyadenylation and alternative splicing is limited to terminal introns," *RNA Biol.*, vol. 13, pp. 646–655, July 2016.

[38] X. Li, G. Quon, H. D. Lipshitz, and Q. Morris, "Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure," *RNA*, vol. 16, pp. 1096–1107, June 2010.

[39] D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A. Smith, C. A. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. F. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, and T. R. Hughes, "A compendium of RNA-binding motifs for decoding gene regulation," *Nature*, vol. 499, pp. 172–177, July 2013.

[40] J. O'Brien, H. Hayder, Y. Zayed, and C. Peng, "Overview of MicroRNA biogenesis, mechanisms of actions, and circulation," *Front. Endocrinol.*, vol. 9, p. 402, Aug. 2018.

[41] X. Zhang, W. Wang, W. Zhu, J. Dong, Y. Cheng, Z. Yin, and F. Shen, "Mechanisms and functions of long Non-Coding RNAs at multiple regulatory levels," *Int. J. Mol. Sci.*, vol. 20, Nov. 2019.

[42] H. Dana, G. M. Chalbatani, H. Mahmoodzadeh, R. Karimloo, O. Rezaiean, A. Moradzadeh, N. Mehmandoost, F. Moazzen, A. Mazraeh, V. Marmari, M. Ebrahimi, M. M. Rashno, S. J. Abadi, and E. Gharagouzlo, "Molecular mechanisms and biological functions of siRNA," *Int. J. Biomed. Sci.*, vol. 13, pp. 48–57, June 2017.

[43] M. Ha and V. N. Kim, "Regulation of microRNA biogenesis," *Nat. Rev. Mol. Cell Biol.*, vol. 15, pp. 509–524, Aug. 2014.

[44] J. P. Broughton, M. T. Lovci, J. L. Huang, G. W. Yeo, and A. E. Pasquinelli, "Pairing beyond the seed supports MicroRNA targeting specificity," *Mol. Cell*, vol. 64, pp. 320–333, Oct. 2016.

[45] J. K. W. Lam, M. Y. T. Chow, Y. Zhang, and S. W. S. Leung, "siRNA versus miRNA as therapeutics for gene silencing," *Mol. Ther. Nucleic Acids*, vol. 4, p. e252, Sept. 2015.

[46] N. Romero-Barrios, M. F. Legascue, M. Benhamed, F. Ariel, and M. Crespi, "Splicing regulation by long noncoding RNAs," *Nucleic Acids Res.*, vol. 46, pp. 2169–2184, Mar. 2018.

[47] M. D. Paraskevopoulou and A. G. Hatzigeorgiou, "Analyzing MiRNA-LncRNA interactions," *Methods Mol. Biol.*, vol. 1402, pp. 271–286, 2016.

[48] I. Ulitsky, "Interactions between short and long noncoding RNAs," *FEBS Lett.*, vol. 592, pp. 2874–2883, Sept. 2018.

[49] M. Sebastian-delaCruz, I. Gonzalez-Moro, A. Olazagoitia-Garmendia, A. Castellanos-Rubio, and I. Santin, "The role of lncRNAs in gene expression regulation through mRNA stabilization," *Noncoding RNA*, vol. 7, Jan. 2021.

[50] S. Shekhar, L. Zhu, L. Mazutis, A. E. Sgro, T. G. Fai, and M. Podolski, "Quantitative biology: where modern biology meets physical sciences," *Mol. Biol. Cell*, vol. 25, pp. 3482–3485, Nov. 2014.

[51] G. L. Glish and R. W. Vachet, "The basics of mass spectrometry in the twenty-first century," *Nat. Rev. Drug Discov.*, vol. 2, pp. 140–150, Feb. 2003.

[52] G. A. Brar and J. S. Weissman, "Ribosome profiling reveals the what, when, where and how of protein synthesis," *Nat. Rev. Mol. Cell Biol.*, vol. 16, pp. 651–664, Nov. 2015.

[53] G. A. Logsdon, M. R. Vollger, and E. E. Eichler, "Long-read human genome sequencing and its applications," *Nat. Rev. Genet.*, vol. 21, pp. 597–614, Oct. 2020.

[54] F. Yan, D. R. Powell, D. J. Curtis, and N. C. Wong, "From reads to insight: a hitchhiker's guide to ATAC-seq data analysis," *Genome Biol.*, vol. 21, p. 22, Feb. 2020.

[55] M. Hafner, M. Katsantoni, T. Köster, J. Marks, J. Mukherjee, D. Staiger, J. Ule, and M. Zavolan, "CLIP and complementary methods," *Nature Reviews Methods Primers*, vol. 1, pp. 1–23, Mar. 2021.

[56] T. S. Furey, "ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions," *Nat. Rev. Genet.*, vol. 13, pp. 840–852, Dec. 2012.

[57] J. D. Hoheisel, "Microarray technology: beyond transcript profiling and genotype analysis," *Nat. Rev. Genet.*, vol. 7, pp. 200–210, Mar. 2006.

[58] R. Bumgarner, "Overview of DNA microarrays: types, applications, and their future," *Curr. Protoc. Mol. Biol.*, vol. Chapter 22, p. Unit 22.1., Jan. 2013.

[59] P. Jaluria, K. Konstantopoulos, M. Betenbaugh, and J. Shiloach, "A perspective on microarrays: current applications, pitfalls, and potential uses," *Microb. Cell Fact.*, vol. 6, p. 4, Jan. 2007.

[60] R. Jaksik, M. Iwanaszko, J. Rzeszowska-Wolny, and M. Kimmel, "Microarray experiments and factors which affect their reliability," *Biol. Direct*, vol. 10, p. 46, Sept. 2015.

[61] M. L. Metzker, "Sequencing technologies - the next generation," *Nat. Rev. Genet.*, vol. 11, pp. 31–46, Jan. 2010.

[62] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nat. Rev. Genet.*, vol. 17, pp. 333–351, May 2016.

[63] B. Hwang, J. H. Lee, and D. Bang, "Single-cell RNA sequencing technologies and bioinformatics pipelines," *Exp. Mol. Med.*, vol. 50, pp. 1–14, Aug. 2018.

[64] D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz, L. Pinello, P. Skums, A. Stamatakis, C. S.-O. Attolini, S. Aparicio, J. Baaijens, M. Balvert, B. d. Barbanson, A. Cappuccio, G. Corleone, B. E. Dutilh, M. Florescu, V. Guryev, R. Holmer, K. Jahn, T. J. Lobo, E. M. Keizer, I. Khatri, S. M. Kielbasa, J. O. Korbel, A. M. Kozlov, T.-H. Kuo, B. P. F. Lelieveldt, I. I. Mandoiu, J. C. Marioni, T. Marschall, F. Mölder, A. Niknejad, L. Raczkowski, M. Reinders, J. d. Ridder, A.-E. Saliba, A. Somarakis, O. Stegle, F. J. Theis, H. Yang, A. Zelikovsky, A. C. McHardy, B. J. Raphael, S. P. Shah, and A. Schönhuth, "Eleven grand challenges in single-cell data science," *Genome Biol.*, vol. 21, p. 31, Feb. 2020.

[65] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi, "A survey of best practices for RNA-seq data analysis," *Genome Biol.*, vol. 17, p. 13, Jan. 2016.

[66] X. Gao, J. Y. Chen, and M. J. Zaki, "Multiscale and multimodal analysis for computational biology," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 15, pp. 1951–1952, Nov. 2018.

[67] L. Clément, D. Emeric, B. J. Gonzalez, M. Laurent, L. David, H. Eivind, and V. Kristian, "A data-supported history of bioinformatics tools," July 2018.

[68] L. Del Giacco and C. Cattaneo, "Introduction to genomics," *Methods Mol. Biol.*, vol. 823, pp. 79–88, 2012.

[69] D. C. Chambers, A. M. Carew, S. W. Lukowski, and J. E. Powell, "Transcriptomics and single-cell RNA-sequencing," *Respirology*, vol. 24, pp. 29–36, Jan. 2019.

[70] F. Lottspeich, "Introduction to proteomics," *Methods Mol. Biol.*, vol. 564, pp. 3–10, 2009.

[71] A. Amberg, B. Riefke, G. Schlotterbeck, A. Ross, H. Senn, F. Dieterle, and M. Keck, "NMR and MS methods for metabolomics," *Methods Mol. Biol.*, vol. 1641, pp. 229–258, 2017.

[72] S. A. Shetty and L. Lahti, "Microbiome data science," *J. Biosci.*, vol. 44, Oct. 2019.

[73] W. S. Pittard, C. k. Villaveces, and S. Li, "A bioinformatics primer to data science, with examples for metabolomics," in *Computational Methods and Data Analysis for Metabolomics* (S. Li, ed.), pp. 245–263, New York, NY: Springer US, 2020.

[74] L. Mills, "Common file formats," *Curr. Protoc. Bioinformatics*, vol. 45, pp. A.1B.1–18, Mar. 2014.

[75] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants," *Nucleic Acids Res.*, vol. 38, pp. 1767–1771, Dec. 2009.

[76] F. d. V. Leprevost, V. C. Barbosa, E. L. Francisco, Y. Perez-Riverol, and P. C. Carvalho, "On best practices in the development of bioinformatics software," *Front. Genet.*, vol. 5, p. 199, July 2014.

[77] G. Wilson, D. A. Aruliah, C. T. Brown, N. P. Chue Hong, M. Davis, R. T. Guy, S. H. D. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, B. Waugh, E. P. White, and P. Wilson, "Best practices for scientific computing," *PLoS Biol.*, vol. 12, p. e1001745, Jan. 2014.

[78] G. Wilson, J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, and T. K. Teal, "Good enough practices in scientific computing," *PLoS Comput. Biol.*, vol. 13, p. e1005510, June 2017.

[79] R. C. Jiménez, M. Kuzak, M. Alhamdoosh, M. Barker, B. Batut, M. Borg, S. Capella-Gutierrez, N. Chue Hong, M. Cook, M. Corpas, M. Flannery, L. Garcia, J. L. Gelpí, S. Gladman, C. Goble, M. González Ferreiro, A. Gonzalez-Beltran, P. C. Griffin, B. Grüning, J. Hagberg, P. Holub, R. Hooft, J. Ison, D. S. Katz, B. Leskošek, F. López Gómez, L. J. Oliveira, D. Mellor, R. Mosbergen, N. Mulder, Y. Perez-Riverol, R. Pergl, H. Pichler, B. Pope, F. Sanz, M. V. Schneider, V. Stodden, R. Suchecki, R. Svobodová Vařeková, H.-A. Talvik, I. Todorov, A. Treloar, S. Tyagi, M. van Gompel, D. Vaughan, A. Via, X. Wang, N. S. Watson-Haigh, and S. Crouch, "Four simple recommendations to encourage best practices in research software," *F1000Res.*, vol. 6, June 2017.

[80] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. Del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, Sept. 2020.

[81] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: fundamental algorithms for scientific computing in python," *Nat. Methods*, vol. 17, pp. 261–272, Mar. 2020.

[82] The pandas development team, "pandas-dev/pandas: Pandas 1.3.4," 2021.

[83] H. Wickham, M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani, "Welcome to the tidyverse," *J. Open Source Softw.*, vol. 4, p. 1686, Nov. 2019.

[84] W. Jakob, J. Rhinelander, and D. Moldovan, "pybind11 – seamless operability between c++11 and python," 2017.

[85] D. Eddelbuettel and R. François, "Rcpp: Seamless R and C++Integration," *J. Stat. Softw.*, vol. 40, no. 8, 2011.

[86] N. D. Matsakis and F. S. Klock, II, "The rust language," in *Proceedings of the 2014 ACM SIGAda annual conference on High integrity language technology - HILT '14*, (New York, New York, USA), ACM Press, 2014.

[87] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," *SIAM Rev.*, vol. 59, pp. 65–98, Jan. 2017.

[88] J. M. Perkel, "Why scientists are turning to rust," *Nature*, vol. 588, pp. 185–186, Dec. 2020.

[89] J. M. Perkel, "Julia: come for the syntax, stay for the speed," *Nature*, vol. 572, pp. 141–142, Aug. 2019.

[90] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, and C. Willing, "Jupyter notebooks – a publishing format for reproducible computational workflows," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (F. Loizides and B. Schmidt, eds.), pp. 87–90, 2016.

[91] RStudio Team, "RStudio: Integrated development environment for R," 2020.

[92] E. Bisong, "Google colaboratory," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners* (E. Bisong, ed.), pp. 59–64, Berkeley, CA: Apress, 2019.

[93] S. Koppad, A. B, G. V. Gkoutos, and A. Acharjee, "Cloud computing enabled big Multi-Omics data analytics," *Bioinform. Biol. Insights*, vol. 15, p. 11779322211035921, July 2021.

[94] L. Wratten, A. Wilm, and J. Göke, "Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers," *Nat. Methods*, pp. 1–8, Sept. 2021.

[95] M. R. Crusoe, S. Abeln, A. Iosup, P. Amstutz, J. Chilton, N. Tijanić, H. Ménager, S. Soiland-Reyes, B. Gavrilovic, and C. Goble, "Methods included: Standardizing computational reuse and portability with the common workflow language," May 2021.

[96] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, "Nextflow enables reproducible computational workflows," *Nat. Biotechnol.*, vol. 35, pp. 316–319, Apr. 2017.

[97] F. Mölder, K. P. Jablonski, B. Letcher, M. B. Hall, C. H. Tomkins-Tinch, V. Sochat, J. Forster, S. Lee, S. O. Twardziok, A. Kanitz, A. Wilm, M. Holtgrewe, S. Rahmann, S. Nahnsen, and J. Köster, "Sustainable data analysis with snakemake," *F1000Res.*, vol. 10, p. 33, Jan. 2021.

[98] D. Merkel and Others, "Docker: lightweight linux containers for consistent development and deployment," *Linux J.*, vol. 2014, no. 239, p. 2, 2014.

[99] G. M. Kurtzer, V. Sochat, and M. W. Bauer, "Singularity: Scientific containers for mobility of compute," *PLoS One*, vol. 12, p. e0177459, May 2017.

[100] "Python package index - PyPI." `https://pypi.org/`.

[101] "Anaconda documentation — anaconda documentation." https://docs.anaconda.com. Accessed: 2021-8-23.

[102] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang, "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biol.*, vol. 5, p. R80, Sept. 2004.

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

[104] E. Afgan, V. Jalili, N. Goonasekera, J. Taylor, and J. Goecks, "Federated galaxy: Biomedical computing at the frontier," *IEEE Int Conf Cloud Comput*, vol. 2018, July 2018.

[105] S. Soiland-Reyes, P. Sefton, M. Crosas, L. J. Castro, F. Coppens, J. M. Fernández, D. Garijo, B. Grüning, M. La Rosa, S. Leo, E. Ó. Carragáin, M. Portier, A. Trisovic, RO-Crate Community, P. Groth, and C. Goble, "Packaging research artefacts with RO-Crate," Aug. 2021.

[106] A. M. Arellano, W. Dai, S. Wang, X. Jiang, and L. Ohno-Machado, "Privacy policy and technology in biomedical data science," *Annu Rev Biomed Data Sci*, vol. 1, pp. 115–129, July 2018.

[107] P. A. Ewels, A. Peltzer, S. Fillinger, H. Patel, J. Alneberg, A. Wilm, M. U. Garcia, P. Di Tommaso, and S. Nahnsen, "The nf-core framework for community-curated bioinformatics pipelines," *Nat. Biotechnol.*, vol. 38, pp. 276–278, Mar. 2020.

[108] "The Snakemake-Workflows project," 2017.

[109] "WorkflowHub," 2020.

[110] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, pp. 10–12, May 2011.

[111] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, pp. 15–21, Jan. 2013.

[112] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression," *Nat. Methods*, vol. 14, pp. 417–419, Apr. 2017.

[113] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification," *Nat. Biotechnol.*, vol. 34, pp. 525–527, May 2016.

[114] S. Andrew, "FastQC: a quality control tool for high throughput sequence data." http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, 2010.

[115] M. Bahin, B. F. Noël, V. Murigneux, C. Bernard, L. Bastianelli, H. Le Hir, A. Lebreton, and A. Genovesio, "ALFA: annotation landscape for aligned reads," *BMC Genomics*, vol. 20, p. 250, Mar. 2019.

[116] L. Wang, J. Nie, H. Sicotte, Y. Li, J. E. Eckel-Passow, S. Dasari, P. T. Vedell, P. Barman, L. Wang, R. Weinshiboum, J. Jen, H. Huang, M. Kohli, and J.-P. A. Kocher, "Measure transcript integrity using RNA-seq data," *BMC Bioinformatics*, vol. 17, p. 58, Feb. 2016.

[117] I. Gallego Romero, A. A. Pai, J. Tung, and Y. Gilad, "RNA-seq: impact of RNA degradation on transcript quantification," *BMC Biol.*, vol. 12, p. 42, May 2014.

[118] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, pp. 3047–3048, Oct. 2016.

[119] M. Frye, B. T. Harada, M. Behm, and C. He, "RNA modifications modulate gene expression during development," *Science*, vol. 361, pp. 1346–1349, Sept. 2018.

[120] D. Bergeron, É. Fafard-Couture, and M. S. Scott, "Small nucleolar RNAs: continuing identification of novel members and increasing diversity of their molecular mechanisms of action," *Biochem. Soc. Trans.*, vol. 48, pp. 645–656, Apr. 2020.

[121] J. Liang, J. Wen, Z. Huang, X.-P. Chen, B.-X. Zhang, and L. Chu, "Small nucleolar RNAs: Insight into their function in cancer," *Front. Oncol.*, vol. 9, p. 587, July 2019.

[122] I. Barbieri and T. Kouzarides, "Role of RNA modifications in cancer," *Nat. Rev. Cancer*, vol. 20, pp. 303–322, June 2020.

[123] T. Bratkovič, J. Božič, and B. Rogelj, "Functional diversity of small nucleolar RNAs," *Nucleic Acids Res.*, vol. 48, pp. 1627–1651, Feb. 2020.

[124] S. Ojha, S. Malla, and S. M. Lyons, "snoRNPs: Functions in ribosome biogenesis," *Biomolecules*, vol. 10, May 2020.

[125] C. Huang, J. Shi, Y. Guo, W. Huang, S. Huang, S. Ming, X. Wu, R. Zhang, J. Ding, W. Zhao, J. Jia, X. Huang, A. P. Xiang, Y. Shi, and C. Yao, "A snoRNA modulates mRNA 3' end processing and regulates the expression of a subset of mRNAs," *Nucleic Acids Res.*, vol. 45, pp. 8647–8660, Sept. 2017.

[126] H. Jorjani, S. Kehr, D. J. Jedlinski, R. Gumienny, J. Hertel, P. F. Stadler, M. Zavolan, and A. R. Gruber, "An updated human snoRNAome," *Nucleic Acids Res.*, vol. 44, pp. 5068–5082, June 2016.

[127] A. Markert, M. Grimm, J. Martinez, J. Wiesner, A. Meyerhans, O. Meyuhas, A. Sickmann, and U. Fischer, "The la-related protein LARP7 is a component of the 7SK ribonucleoprotein and affects transcription of cellular and viral polymerase II genes," *EMBO Rep.*, vol. 9, pp. 569–575, June 2008.

[128] Y. Cheng, Z. Jin, R. Agarwal, K. Ma, J. Yang, S. Ibrahim, A. V. Olaru, S. David, H. Ashktorab, D. T. Smoot, M. D. Duncan, D. F. Hutcheon, J. M. Abraham, S. J. Meltzer, and Y. Mori, "LARP7 is a potential tumor suppressor gene in gastric cancer," *Lab. Invest.*, vol. 92, pp. 1013–1019, July 2012.

[129] N. He, N. S. Jahchan, E. Hong, Q. Li, M. A. Bayfield, R. J. Maraia, K. Luo, and Q. Zhou, "A la-related protein modulates 7SK snRNP integrity to suppress P-TEFb-dependent transcriptional elongation and tumorigenesis," *Mol. Cell*, vol. 29, pp. 588–599, Mar. 2008.

[130] X. Ji, H. Lu, Q. Zhou, and K. Luo, "LARP7 suppresses P-TEFb activity to inhibit breast cancer progression and metastasis," *Elife*, vol. 3, p. e02907, July 2014.

[131] A. M. Alazami, M. Al-Owain, F. Alzahrani, T. Shuaib, H. Al-Shamrani, Y. H. Al-Falki, S. M. Al-Qahtani, T. Alsheddi, D. Colak, and F. S. Alkuraya, "Loss of function mutation in LARP7, chaperone of 7SK ncRNA, causes a syndrome of facial dysmorphism, intellectual disability, and primordial dwarfism," *Hum. Mutat.*, vol. 33, pp. 1429–1434, Oct. 2012.

[132] J. W. Chang, H. S. Yeh, and J. Yong, "Alternative polyadenylation in human diseases," *Endocrinol Metab (Seoul)*, vol. 32, pp. 413–421, Dec. 2017.

[133] A. Curinha, S. Oliveira Braz, I. Pereira-Castro, A. Cruz, and A. Moreira, "Implications of polyadenylation in health and disease," *Nucleus*, vol. 5, pp. 508–519, Oct. 2014.

[134] A. Reyes and W. Huber, "Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues," *Nucleic Acids Res.*, vol. 46, pp. 582–592, Jan. 2018.

[135] B. Tian and J. L. Manley, "Alternative polyadenylation of mRNA precursors," *Nat. Rev. Mol. Cell Biol.*, vol. 18, pp. 18–30, Jan. 2017.

[136] Y. Zhu, X. Wang, E. Forouzmand, J. Jeong, F. Qiao, G. A. Sowd, A. N. Engelman, X. Xie, K. J. Hertel, and Y. Shi, "Molecular mechanisms for CFIm-Mediated regulation of mRNA alternative polyadenylation," *Mol. Cell*, vol. 69, pp. 62–74.e4, Jan. 2018.

[137] W. Li, B. You, M. Hoque, D. Zheng, W. Luo, Z. Ji, J. Y. Park, S. I. Gunderson, A. Kalsotra, J. L. Manley, and B. Tian, "Systematic profiling of poly(a)+ transcripts modulated by core 3' end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation," *PLoS Genet.*, vol. 11, p. e1005166, Apr. 2015.

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

[139] A. R. Gruber, G. Martin, W. Keller, and M. Zavolan, "Cleavage factor im is a key regulator of 3' UTR length," *RNA Biol.*, vol. 9, pp. 1405–1412, Dec. 2012.

[140] C. P. Masamha, Z. Xia, J. Yang, T. R. Albrecht, M. Li, A.-B. Shyu, W. Li, and E. J. Wagner, "CFIm25 links alternative polyadenylation to glioblastoma tumour suppression," *Nature*, vol. 510, pp. 412–416, June 2014.

[141] A. R. Gruber, G. Martin, W. Keller, and M. Zavolan, "Means to an end: mechanisms of alternative polyadenylation of messenger RNA precursors," *Wiley Interdiscip. Rev. RNA*, Nov. 2013.

[142] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, "MEME SUITE: tools for motif discovery and searching," *Nucleic Acids Res.*, vol. 37, pp. W202–8, July 2009.

[143] P. Arnold, I. Erb, M. Pachkov, N. Molina, and E. van Nimwegen, "MotEvo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences," *Bioinformatics*, vol. 28, pp. 487–494, Feb. 2012.

[144] G. Giudice, F. Sánchez-Cabo, C. Torroja, and E. Lara-Pezzi, "ATtRACT-a database of RNA-binding proteins and associated motifs," *Database*, vol. 2016, Apr. 2016.

[145] H. Nagai and Y. H. Kim, "Cancer prevention from the perspective of global cancer burden patterns," *J. Thorac. Dis.*, vol. 9, pp. 448–451, Mar. 2017.

[146] F. Qi, Y. Li, X. Yang, Y.-P. Wu, L.-J. Lin, and X.-M. Liu, "Significance of alternative splicing in cancer cells," *Chin. Med. J.*, vol. 133, pp. 221–228, Jan. 2020.

[147] B. P. Jain, "The role of alternative polyadenylation in cancer progression," *Gene Reports*, vol. 12, pp. 1–8, Sept. 2018.

[148] X. Song, X. Li, Y. Ge, J. Song, Q. Wei, M. He, M. Wei, Y. Zhang, T. Chen, and L. Zhao, "Alternative splicing events and function in the tumor microenvironment: New opportunities and challenges," *Int. Immunopharmacol.*, vol. 123, p. 110718, Oct. 2023.

[149] N. A. Faustino and T. A. Cooper, "Pre-mRNA splicing and human disease," *Genes Dev.*, vol. 17, pp. 419–437, Feb. 2003.

[150] F. Gebauer, T. Schwarzl, J. Valcárcel, and M. W. Hentze, "RNA-binding proteins in human genetic disease," *Nat. Rev. Genet.*, vol. 22, pp. 185–198, Mar. 2021.

[151] J. Desterro, P. Bak-Gordon, and M. Carmo-Fonseca, "Targeting mRNA processing as an anticancer strategy," *Nat. Rev. Drug Discov.*, vol. 19, pp. 112–129, Feb. 2020.

[152] T. C. Roberts, R. Langer, and M. J. A. Wood, "Advances in oligonucleotide drug delivery," *Nat. Rev. Drug Discov.*, vol. 19, pp. 673–694, Oct. 2020.

[153] C. F. Bennett, "Therapeutic antisense oligonucleotides are coming of age," *Annu. Rev. Med.*, vol. 70, pp. 307–321, Jan. 2019.

[154] F. E. Baralle and J. Giudice, "Alternative splicing as a regulator of development and tissue identity," *Nat. Rev. Mol. Cell Biol.*, vol. 18, pp. 437–451, July 2017.

[155] F. Yuan, W. Hankey, E. J. Wagner, W. Li, and Q. Wang, "Alternative polyadenylation of mRNA and its role in cancer," *Genes Dis*, vol. 8, pp. 61–72, Jan. 2021.

[156] J.-T. Zheng, C.-X. Lin, Z.-Y. Fang, and H.-D. Li, "Intron retention as a mode for RNA-Seq data analysis," *Front. Genet.*, vol. 11, p. 586, July 2020.

[157] O. Mauger, F. Lemoine, and P. Scheiffele, "Targeted intron retention and excision for rapid gene regulation in response to neuronal activity," *Neuron*, vol. 92, pp. 1266–1278, Dec. 2016.

[158] L. Zappia and F. J. Theis, "Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape," *Genome Biol.*, vol. 22, p. 301, Oct. 2021.

[159] W. E. Heavner, S. Ji, J. H. Notwell, E. S. Dyer, A. M. Tseng, J. Birgmeier, B. Yoo, G. Bejerano, and S. K. McConnell, "Transcription factor expression defines subclasses of developing projection neurons highly similar to single-cell RNA-seq subtypes," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, pp. 25074–25084, Oct. 2020.

[160] S. Liu, B. Zhou, L. Wu, Y. Sun, J. Chen, and S. Liu, “Single-cell differential splicing analysis reveals high heterogeneity of liver tumor-infiltrating T cells,” *Sci. Rep.*, vol. 11, p. 5325, Mar. 2021.

[161] R. Patrick, D. T. Humphreys, V. Janbandhu, A. Oshlack, J. W. K. Ho, R. P. Harvey, and K. K. Lo, “Sierra: discovery of differential transcript usage from polya-captured single-cell RNA-seq data,” *Genome Biol.*, vol. 21, p. 167, July 2020.

[162] M. Hafner, M. Katsantoni, T. Köster, J. Marks, J. Mukherjee, D. Staiger, J. Ule, and M. Zavolan, “CLIP and complementary methods,” *Nature Reviews Methods Primers*, vol. 1, pp. 1–23, Mar. 2021.

Appendices

Appendix A: Complete list of scientific articles

1. Bak M., van Nimwegen E., Kouzel I.U., Schmidt R., Zavolan M., Gruber A.J. **MAPP unravels frequent co-regulation of splicing and polyadenylation by RBPs and their dysregulation in cancer**. (in revision; Nature Communications)

2. Bak M., Agarwal K., Katsantoni M., Zavolan M. **Inferring binding sites of RNA-binding proteins with bindz**. (in preparation)

3. Katsantoni M., Gypas F., Herrmann C.J., Iborra de Toledo P., Burri D., Bak M., Börsch A., Agarwal K., Kandhari R., Ataman M., Zavolan M., Kanitz A. (2021). **ZARP: An automated workflow for processing of RNA-seq data**. (in preparation)

4. Ghosh S., Ataman M., Bak M., Borsch A., Schmidt A., Buczak K., Martin G., Dimitriades B., Kanitz A., Zavolan M. (2021). **CFIm-mediated alternative polyadenylation remodels cellular signaling and miRNA biogenesis**. Nucleic Acids Res. 50(6):3096-3114 (published)

5. Hasler D., Rajyalakshmi M., Bak M., Lehmann G., Heizinger L., Wang X., Li., Z., Sement F. M., Bruckmann A., Dock-Bregeon A., Merkel R., Kalb R., Grauer E., Kunstmann E., Zavolan M., Liu M., Fisher U., Meister G. (2020). **The Alazami Syndrome-associated protein LARP7 guides U6 small nuclear RNA modification and contributes to splicing robustness**. Mol Cell 77(5):1014-1031.e13 (published)

Appendix B: ZARP: An automated workflow for processing of RNA-seq data (Manuscript)

ZARP: An automated workflow for processing of RNA-seq data

Maria Katsantoni [1,2], Foivos Gypas [3], Christina J. Herrmann [1,2], Paula Iborra de Toledo [1,2], Dominik Burri [1,2], Maciej Bak [1,2], Anastasiya Börsch [1,2], Krish Agarwal [1], Meric Ataman [1,2], Mihaela Zavolan [1,2,*] & Alexander Kanitz [1,2,*]

1. Biozentrum, University of Basel, Basel, 4056, Switzerland
2. Swiss Institute of Bioinformatics, Switzerland
3. Friedrich Miescher Institute for Biomedical Research, Basel, 4058, Switzerland.

* Corresponding author

Abstract

Bioinformatics is a rapidly expanding field, with a plethora of new open source software tools developed to address specific biological questions. As RNA sequencing is a basic component of many scientific studies, multiple models and packages have been developed for processing and analysis of such data. Still, identification of appropriate tools remains a time consuming process that requires an in-depth understanding of the data, as well as of the principles and parameters of each tool. In addition, packages designed for individual tasks are developed in different programming languages and have dependencies of various degrees of complexity, which renders their installation and execution challenging for users with limited computational expertise. The recent emergence of workflow languages and execution engines have enormously facilitated these tasks. Computational workflows can be reliably shared with the scientific community, enhancing reusability while improving the reproducibility of results, as individual analysis steps are more transparent. In the following work we present ZARP, a general purpose RNA-seq analysis workflow which builds on state of the art software in the field to facilitate the analysis of RNA-seq data sets. ZARP is developed in the snakemake workflow language using best software development practices. It can run locally or in a cluster environment, generating extensive reports not only of the data but also of the options utilized. It is built using modern technologies with the ultimate goal to reduce the hands-on time for bioinformaticians and non-expert users.

Contact: mihaela.zavolan@unibas.ch, alexander.kanitz@unibas.ch

Keywords

Computational workflow, pipeline, RNA-seq, high-throughput, reproducible research, FAIR, transcriptomics, bioinformatics

Main body

Introduction

Recent years have seen an exponential growth in bioinformatics tools [1], a large proportion of which are dedicated to High Throughput Sequencing (HTS) data analysis. For example, for transcript-level analyses there are tools to quantify the expression level of transcripts and genes from RNA-seq data [2], identify RNA-binding protein (RBP) binding sites from crosslinking and immunoprecipitation (CLIP) data [3,4], improve transcript annotation with the help of RNA 3'end-sequencing data [5,6], estimate gene expression at the single cell level [7] or improve the annotation of transcripts and quantification of splicing events based on long read sequencing (e.g. on the Oxford Nanopore platform) to [8,9]. Such tools are written in different programming languages (e.g. Python, R, C, Rust) and have distinct library requirements and dependencies. In most cases, the tools expect the input to be in one of the widely accepted file formats (e.g. FASTQ [10], BAM [11]), but custom formats are also frequently used. In addition, the variations in protocols or instruments across experiments may make it necessary to use different parameterization for every sample, rendering a joint analysis of samples from multiple studies challenging. Combining tools into an analysis protocol is a time consuming and error prone process. As these tasks have become so common, and as the data sets and analyses continue to increase in size and complexity, there is an urgent need for expertly curated, well-tested, maintained and easy-to-use computational workflows.

Workflow management systems and specification languages [12,13] like CWL [14], snakemake [15,16], nextflow [17] are now available, making it possible for such workflows to be developed, tested and shared. This leads to reusable code and reproducible results, while fostering scientific collaborations along the way.

Despite their apparent advantages, workflow execution languages cannot guarantee the flawless execution of a workflow. Differences in the hardware architecture or in the host operating system may lead to (sometimes impossible to resolve) difficulties in installation, or a lack of reproducibility during execution of the workflows. Software containers like docker [18] or singularity [19] or general-purpose package and environment managers like conda [20] in combination with scientific channels like bioconda [21] allow users to easily install and run scientific tools or general purpose software packages.

The execution of a workflow generates metadata along with the expected results. These can be useful for re-analyses of the data but may also provide insights into the results, facilitating their interpretation. There are agreed-upon principles on how these metadata should be organised that are followed by the scientific community [22].

Methods/Results

ZARP: a general purpose RNA-seq workflow

ZARP (**Z**avolan-Lab **A**utomated **R**NA-Seq **P**ipeline) is a generic RNA-seq analysis workflow that allows users to carry out the most general steps in the analysis of Illumina short-read sequencing libraries with minimum effort. The workflow is developed in snakemake [15,16], a widely used workflow management system [12]. It relies on publicly available bioinformatics tools that follow best practices in the field [23], and handles bulk, stranded RNA-seq data, single or paired-end.

Workflow input

ZARP uses three distinct input files, two of which are mandatory. The first is a tab-delimited file with sample-specific information, such as paths to the sequencing data (FASTQ format), reference genome sequence (FASTA format), transcriptome annotation (GTF format) and additional experiment protocol and library preparation specifications like adapter sequences or fragment size. The second input is a configuration file in YAML format, containing workflow-related parameters (e.g. results and log directories location, user-related information etc). Advanced users can take advantage of ZARP's flexibility to provide rule-specific configuration parameters in an optional input file, thus adjusting the behaviour of some of the tools used in the workflow. More information on this "sample table" and the rest of the input files can be found in ZARP's extensive documentation (https://github.com/zavolanlab/zarp/blob/dev/pipeline_documentation.md#preparatory).

Analysis steps

A general schema of the workflow in its current version, 0.3.0, is presented in Figure 1.

In a first step, the workflow generates the indices required by the alignment tools, STAR [24], Salmon [25], kallisto [26] and ALFA [27], with the aid of gffread [28]. These rules are applied once for every provided genome.

After calculating per-sample quality statistics by applying fastqc [29] to the FASTQ files, adapters are trimmed with cutadapt [30]; 5' or/and 3' adapters, but also poly(A/T) stretches are removed, as indicated by the user.

The trimmed reads are then aligned to the genome with STAR. The resulting bam-formatted files are sorted (based on coordinates) and an index is created using samtools [11]. This output is later reformatted into bedgraph (STAR) and BigWig (BedGraphtoBigWig from ucsc tools [31]), which allow for library normalisation and are therefore useful when visualising coverages for multiple samples with a viewer like IGV [32]. The functional annotation in terms of gene segments (e.g. CDS, 3’UTR, intergenic, etc.) and biotypes (e.g. protein coding genes, rRNA, etc.) ZARP is computed with ALFA [27]. The aligned reads are also used to calculate per-transcript Transcript Integrity Numbers (TIN scores) [33], which can be used to assess the degree of RNA degradation in the sample. This is done with a customized tool, tin-score-calculation [34], which is based on a script originally included in the RSeQC package [35]. The tin-score-calculation tool processes a BED12-formatted list of transcripts that is generated with the aid of zgtf [36].

The reads are also used by the Salmon and kallisto tools along with a transcript annotation to infer transcript and gene expression estimates. The outputs of these tools, in Transcripts Per Million (TPM) [37] as well as raw counts are collected by ZARP with the aid of Salmon and merge_kallisto [38] to generate summary tables for all analyzed samples. These are then used for a principal component analysis (PCA) with zpca [39]. The workflow produces two user-friendly reports: one with a summary of samples-related information (by MultiQC [40]) and the other with estimates of utilized computational resources (by snakemake).

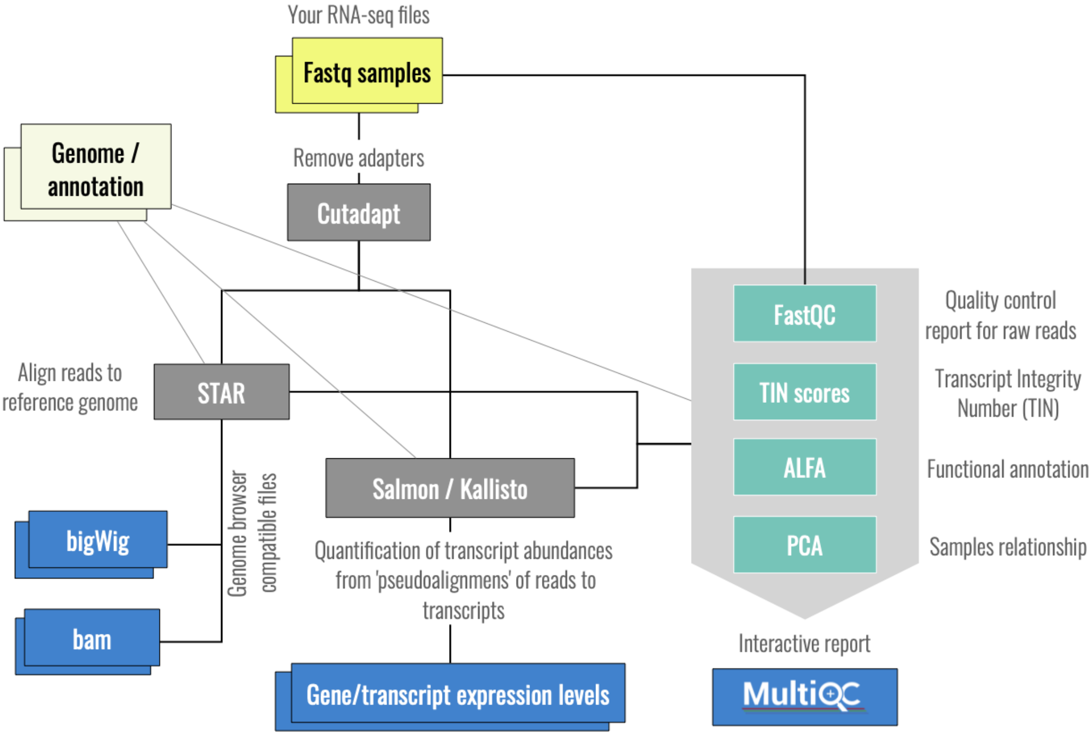

Your RNA-seq files
Fastq samples
Genome / annotation
Remove adapters
Cutadapt
Align reads to reference genome
STAR
Genome browser compatible files
Salmon / Kallisto
bigWig
bam
Quantification of transcript abundances from 'pseudoalignmens' of reads to transcripts
Gene/transcript expression levels
FastQC
Quality control report for raw reads
TIN scores
Transcript Integrity Number (TIN)
ALFA
Functional annotation
PCA
Samples relationship
Interactive report
MultiQC

Figure 1: General overview of the ZARP workflow.

Reproducibility and reusability

To enforce reproducibility of results and enhance reusability of the workflow, each step (rule) of the workflow relies either on conda environments or on singularity images (originally developed in docker) [19] that are built and hosted in the bioconda channel [21]. Users can choose between the two technologies by selecting an appropriate profile before the workflow execution step. The conda environments are built on the fly, while the docker images are pulled from external servers and converted to singularity images before the execution is initiated.

Output and documentation

In addition to the transcript/gene expression tables, ZARP collects log files and metadata for downstream analyses. Intermediate files can be optionally cleaned up by ZARP to minimize disk space usage. As the workflow is hosted in its own GitHub repository, each ZARP version released is accompanied by an up-to-date workflow-oriented description.

Continuous Integration and Testing

To enable continuous integration and community development, the built-in GitHub Actions mechanism for CI/CD is implemented. Each modification to the remote repository triggers a variety of integration tests to guarantee ZARP's correct execution throughout the development cycle as the source code is refactored and new features are added.

Use Cases

ZARP was tested on an RNA-seq dataset obtained by Ham et al. [41] (GEO [42] accession number GSE139213), while analyzing the role of mTORC1 signalling in the age-related loss of muscle mass and function in mice. The dataset consists of 20 samples corresponding to 4 cohorts of 3-months old mice: wild-type, rapamycin-treated, tuberous sclerosis complex 1 (TSC1) knockout and rapamycin-treated TSC1 knockout. Each cohort contains 5 biological replicates, and the libraries are single-end. The samples were mapped against ENSEMBL's [43] GRCm38 genome primary assembly and the gene annotation for standard chromosomes was used. Other parameters for populating ZARP's samples table were obtained from the GEO accession entries of the respective samples.

As shown in Figure 2 the samples are of high quality, with metrics such as GC content (Figure 2A), not showing any bias across the samples. Adapters constitute only a few bases out of each sequence (~7 nucleotides) and the large majority of reads successfully pass the filters after trimming, which indicates that there is no adapter contamination (Figure 2B). The statistics of STAR-based read alignments to the mouse genome are consistent across samples, with more than 75% uniquely mapped and less than 3% unmapped reads in a library (Figure 2E).

Transcript integrity is high across the whole transcriptome for all 20 samples, with the highest density of transcripts at TIN scores of 75 to 85 (Figure 2C). As expected, ALFA analysis of transcript categories shows that uniquely mapped reads overwhelmingly originated from protein coding genes (over 86% for all samples) (Figure 2D).

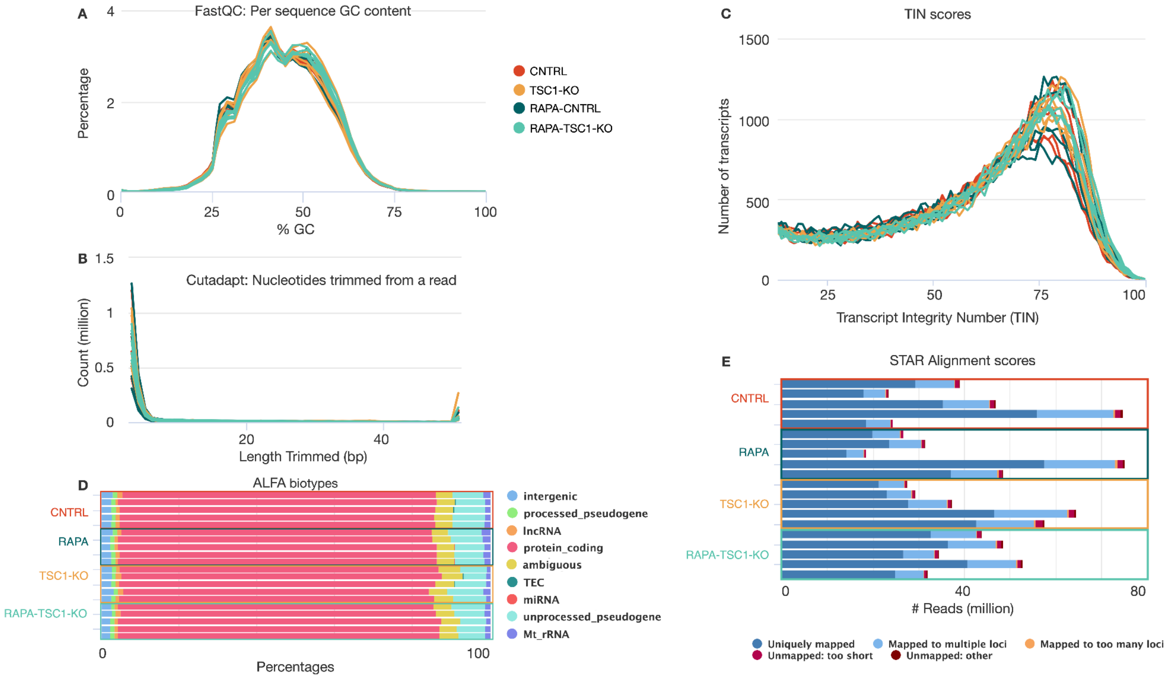


Figure 2: Zarp metrics (A) GC content, (B) Nucleotides trimmed from a read, (C) Transcript Integrity Number (TIN) score, (D) ALFA biotypes, (E) STAR alignment scores

The distribution of samples in the space of the first two principal components shows a clustering by condition, with a clear separation between knockout and wild type, as well as between the untreated and rapamycin-treated TSC1 knockout mice (Figure 3). This separation is more pronounced at the gene expression level (Figure 3A), but is also present at the transcript level (Figure 3B).

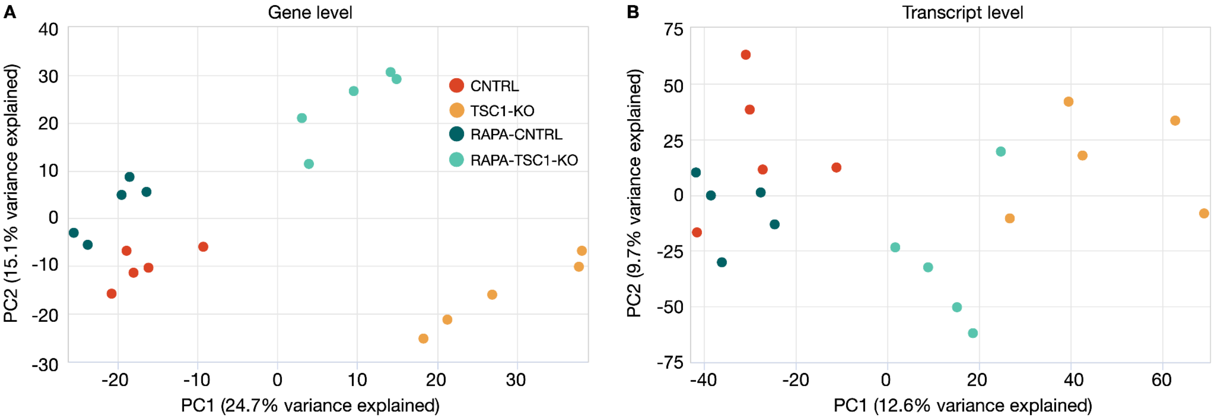


Figure 3: PCA at the (A) gene and (B) transcript level

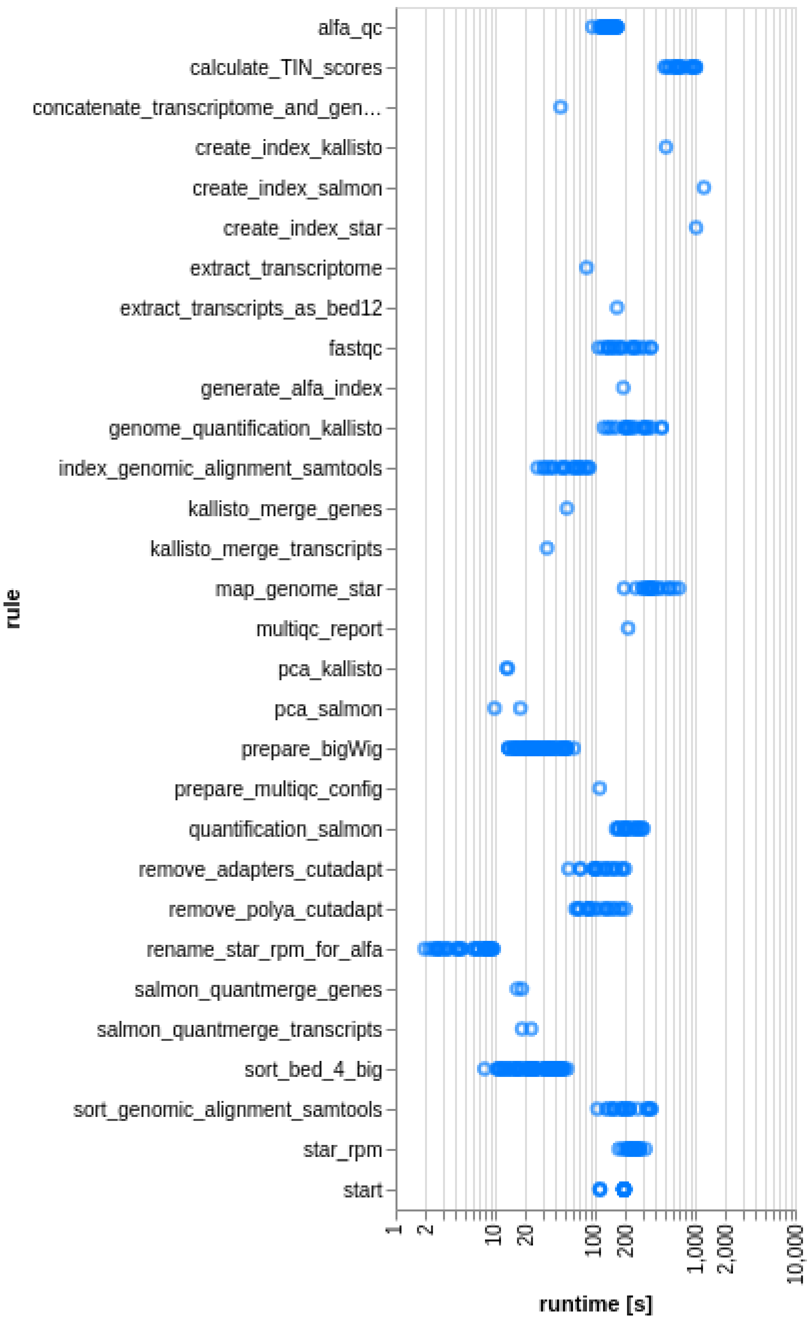


Figure 4: Run time (in seconds) of the different steps (rules) of the workflow. The workflow was executed in an HPCcluster (SLURM), so the execution runtimes include the time that the jobs were in the queue. The machines that are part of the cluster also have slightly different specifications, so the execution times might show some additional variation based on which machine was used for their execution.

Discussion/Conclusions

ZARP is a general purpose, easy-to-use, stable and efficient RNA-seq processing workflow that can be used by molecular biologists with minimal programming experience. Scientists with access to a UNIX-based computer (ideally a linux machine with enough memory to align sequencing reads) or a computing cluster can run the workflow to get an initial view of their data on a relatively short time scale (Figure 4). The advantage of using ZARP is that it has been fine tuned to process bulk RNA-seq datasets, allowing users to run it out of the box with default parameters. At the same time ZARP allows users to customize different options of the tools (e.g. via the rule config) making it a helpful tool for handling special cases. The files that ZARP provides can serve as entry points for other project-specific analyses such as differential gene and transcript isoform expression. As ZARP is publicly available and open source (Apache License, Version 2.0), contributions from the bioinformatics community are very welcome and will likely further enhance the functionality of the code. Please address all development-related inquiries as issues at the official GitHub repository: https://github.com/zavolanlab/zarp.

Data and Software Availability

Data

Raw data analysed in the section: Use cases are publicly available for anyone to download from the NCBI:GEO server, accession number GSE139213.

Software

ZARP lives on GitHub, the official repository is located at: https://github.com/zavolanlab/zarp under Apache License, Version 2.0. Version 0.3.0 of the workflow described in this manuscript has been additionally uploaded to zenodo platform where it will be stored permanently, doi: XXX. Both services are public and allow anyone to download the software without prior registration.

Results

Analysis results presented in the section: Use cases are publicly available for anyone to download from the zenodo platform, doi: 10.5281/zenodo.5683525.

Author Contributions

MK, FG, MZ, AK conceived the project. MK, FG, CJH, PI, DB, MB, AK developed the method. MK, FG, CJH, PI, DB, MB, KA, MZ, AK developed custom tools used in the study. MA, AB tested the method with real datasets. MK, FG, CJH, PI, DB, MB, MA, MZ, AK wrote the manuscript. MK, MZ, AK supervised the study. MK, MB, AK managed the software repository. All of the authors approved the manuscript.

Competing Interests

None declared.

Grant Information

M.K., M.B., D.B. were supported by Biozentrum Basel International Ph.D. Program Fellowships for Excellence.

Acknowledgements

This work would not have been possible without the sciCORE Team at the University of Basel. We are thankful for their support regarding the computational infrastructure as well as dedicated time and effort to aid us in this project. We would also like to express our deepest gratitude towards all members of the Zavolan Lab who contributed to this work with numerous pieces of advice, during the initial development as well as by testing the workflow in later stages.

References

1. Clément L, Emeric D, J GB, Laurent M, David L, Eivind H, et al. A data-supported history of bioinformatics tools [Internet]. arXiv [cs.DL]. 2018. Available from: http://arxiv.org/abs/1807.06808

2. Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biol. 2015;16:150.

3. Khorshid M, Rodak C, Zavolan M. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. Nucleic Acids Res. 2011;39:D245–52.

4. Hafner M, Katsantoni M, Köster T, Marks J, Mukherjee J, Staiger D, et al. CLIP and complementary methods. Nature Reviews Methods Primers. Nature Publishing Group;

2021;1:1–23.

5. Gruber AJ, Gypas F, Riba A, Schmidt R, Zavolan M. Terminal exon characterization with TECtool reveals an abundance of cell-specific isoforms. Nat Methods [Internet]. 2018; Available from: http://dx.doi.org/10.1038/s41592-018-0114-z

6. Herrmann CJ, Schmidt R, Kanitz A, Artimo P, Gruber AJ, Zavolan M. PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. Nucleic Acids Res. Oxford Academic; 2019;48:D174–9.

7. Breda J, Zavolan M, van Nimwegen E. Bayesian inference of gene expression states from single-cell RNA-seq data. Nat Biotechnol [Internet]. 2021; Available from: http://dx.doi.org/10.1038/s41587-021-00875-x

8. Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. Nat Commun. 2019;10:3359.

9. Karousis ED, Gypas F, Zavolan M, Mühlemann O. Nanopore sequencing reveals endogenous NMD-targeted isoforms in human cells. bioRxiv [Internet]. biorxiv.org; 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.04.30.442116v1.abstract

10. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2009;38:1767–71.

11. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

12. Perkel JM. Workflow systems turn raw data into scientific knowledge. Nature. 2019;573:149–50.

13. Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat Methods. Nature Publishing Group; 2021;1–8.

14. Amstutz P, Crusoe MR, Tijanić N, Chapman B, Chilton J, Heuer M, et al. Common Workflow Language, v1.0 [Internet]. Figshare; 2016. Available from: http://dx.doi.org/10.6084/M9.FIGSHARE.3115156.V2

15. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics [Internet]. academic.oup.com; 2012; Available from: https://academic.oup.com/bioinformatics/article-abstract/28/19/2520/290322

16. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Res. 2021;10:33.

17. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316–9.

18. Merkel D, Others. Docker: lightweight linux containers for consistent development and

deployment. Linux J. 2014;2014:2.

19. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS One. 2017;12:e0177459.

20. Anaconda Documentation — Anaconda documentation [Internet]. [cited 2021 Aug 23]. Available from: https://docs.anaconda.com

21. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018;15:475–6.

22. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.

23. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.

24. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

25. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods [Internet]. 2017; Available from: http://dx.doi.org/10.1038/nmeth.4197

26. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.

27. Bahin M, Noël BF, Murigneux V, Bernard C, Bastianelli L, Le Hir H, et al. ALFA: annotation landscape for aligned reads. BMC Genomics. 2019;20:250.

28. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Res. 2020;9:304.

29. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010.

30. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10–2.

31. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinform. 2013;14:144–61.

32. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer [Internet]. Nature Biotechnology. 2011. p. 24–6. Available from: http://dx.doi.org/10.1038/nbt.1754

33. Wang L, Nie J, Sicotte H, Li Y, Eckel-Passow JE, Dasari S, et al. Measure transcript integrity using RNA-seq data. BMC Bioinformatics. 2016;17:58.

34. tin-score-calculation: Given a set of BAM files and a gene annotation BED file, calculates the Transcript Integrity Number (TIN) for each transcript [Internet]. Github; [cited 2021 Aug 23].

Available from: https://github.com/zavolanlab/tin-score-calculation

35. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012;28:2184–5.

36. zgtf: gtf conversion utlity [Internet]. Github; [cited 2021 Aug 23]. Available from: https://github.com/zavolanlab/zgtf

37. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 2012;131:281–5.

38. merge_kallisto: Merge kallisto results from multiple runs [Internet]. Github; [cited 2021 Aug 23]. Available from: https://github.com/zavolanlab/merge_kallisto

39. zpca: PCA analysis [Internet]. Github; [cited 2021 Aug 23]. Available from: https://github.com/zavolanlab/zpca

40. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32:3047–8.

41. Ham DJ, Börsch A, Lin S, Thürkauf M, Weihrauch M, Reinhard JR, et al. The neuromuscular junction is a focal point of mTORC1 signaling in sarcopenia. Nat Commun. 2020;11:4510.

42. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41:D991–5.

43. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic Acids Res. 2021;49:D884–91.

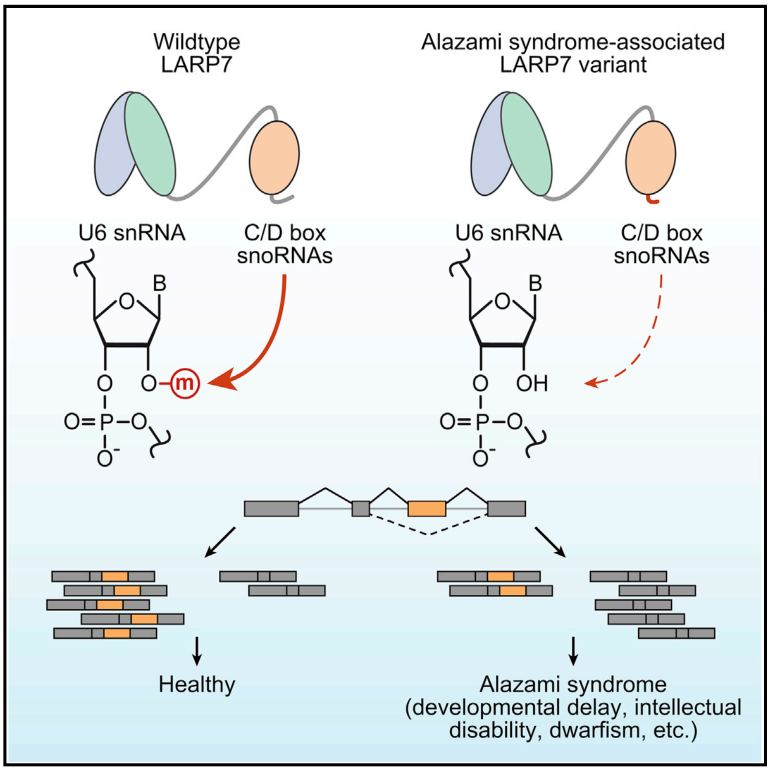
Appendix C: The Alazami Syndrome-Associated Protein LARP7 Guides U6 Small Nuclear RNA Modification and Contributes to Splicing Robustness (Published)

Article

Molecular Cell

The Alazami Syndrome-Associated Protein LARP7 Guides U6 Small Nuclear RNA Modification and Contributes to Splicing Robustness

Graphical Abstract


Wildtype
LARP7
Alazami syndrome-associated
LARP7 variant
U6 snRNA
C/D box
snoRNAs
B
O
m
OH
O=P–O
O-
Healthy
Alazami syndrome
(developmental delay, intellectual
disability, dwarfism, etc.)

Highlights

- LARP7 interacts simultaneously with the U6 snRNA and U6-specific C/D box snoRNAs
- Depletion of LARP7 results in reduced 2′-O-methylation of the U6 snRNA
- Changes in alternative splicing are observed in the absence of LARP7
- A *LARP7* mutation causes splicing alterations in Alazami syndrome patients

Authors

Daniele Hasler, Rajyalakshmi Meduri, Maciej Bąk, ..., Mo-Fang Liu, Utz Fischer, Gunter Meister

Correspondence

utz.fischer@biozentrum.uni-wuerzburg.de (U.F.), gunter.meister@ur.de (G.M.)

In Brief

Mutations in the *LARP7* gene are associated with the Alazami syndrome, a severe developmental disorder. Hasler et al. report that the RNA-binding protein LARP7 is required for efficient 2′-O-methylation of the spliceosomal U6 snRNA. Perturbation of this function results in splicing alterations, which contribute to the etiology of the disease.

Hasler et al., 2020, Molecular Cell *77*, 1014–1031
March 5, 2020 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.molcel.2020.01.001

The Alazami Syndrome-Associated Protein LARP7 Guides U6 Small Nuclear RNA Modification and Contributes to Splicing Robustness

Daniele Hasler,[1] Rajyalakshmi Meduri,[2] Maciej Bąk,[3] Gerhard Lehmann,[1] Leonhard Heizinger,[4] Xin Wang,[5] Zhi-Tong Li,[5] François M. Sement,[6] Astrid Bruckmann,[1] Anne-Catherine Dock-Bregeon,[6] Rainer Merkl,[2] Reinhard Kalb,[7] Eva Grauer,[7] Erdmute Kunstmann,[7] Mihaela Zavolan,[3] Mo-Fang Liu,[5] Utz Fischer,[2,*] and Gunter Meister[1,8,*]

[1]Biochemistry Center Regensburg (BZR), Laboratory for RNA Biology, University of Regensburg, 93053 Regensburg, Germany
[2]Department of Biochemistry, Theodor Boveri Institute, University of Würzburg, 97074 Würzburg, Germany
[3]Computational and Systems Biology, Biozentrum, University of Basel, 4056 Basel, Switzerland
[4]Biochemistry Center Regensburg (BZR), Institute of Biophysics and Physical Biochemistry, University of Regensburg, 93053 Regensburg, Germany
[5]State Key Laboratory of Molecular Biology, Shanghai Key Laboratory of Molecular Andrology, CAS Center for Excellence in Molecular Cell Science, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences-University of Chinese Academy of Sciences, Shanghai 200031, China
[6]Sorbonne Université, Centre National de la Recherche Scientifique (CNRS), Laboratory of Integrative Biology of Marine Models (UMR8227), Station Biologique de Roscoff, 29680 Roscoff, France
[7]Department of Human Genetics, Biozentrum, University of Würzburg, 97074 Würzburg, Germany
[8]Lead Contact
*Correspondence: utz.fischer@biozentrum.uni-wuerzburg.de (U.F.), gunter.meister@ur.de (G.M.)
https://doi.org/10.1016/j.molcel.2020.01.001

SUMMARY

The La-related protein 7 (LARP7) forms a complex with the nuclear 7SK RNA to regulate RNA polymerase II transcription. It has been implicated in cancer and the Alazami syndrome, a severe developmental disorder. Here, we report a so far unknown role of this protein in RNA modification. We show that LARP7 physically connects the spliceosomal U6 small nuclear RNA (snRNA) with a distinct subset of box C/D small nucleolar RNAs (snoRNAs) guiding U6 2′-O-methylation. Consistently, these modifications are severely compromised in the absence of LARP7. Although general splicing remains largely unaffected, transcriptome-wide analysis revealed perturbations in alternative splicing in LARP7-depleted cells. Importantly, we identified defects in 2′-O-methylation of the U6 snRNA in Alazami syndrome siblings carrying a *LARP7* mutation. Our data identify LARP7 as a bridging factor for snoRNA-guided modification of the U6 snRNA and suggest that alterations in splicing fidelity contribute to the etiology of the Alazami syndrome.

INTRODUCTION

Small nucleolar RNAs (snoRNAs) are conserved non-coding RNAs that guide post-transcriptional modification of various RNAs. Based on common sequence motifs, two main families of snoRNAs have been defined. The H/ACA box family guides the conversion of uridine to pseudouridine, whereas C/D box snoRNAs guide the methylation of the 2′ hydroxyl of the ribose (2′-O-methylation). Members of either snoRNA family assemble with distinct sets of proteins to form snoRNA protein complexes (snoRNPs), in which the snoRNAs serve as guides for enzymes catalyzing the chemical reactions (Matera et al., 2007). The catalytic subunit of H/ACA snoRNPs is dyskerin, whereas the enzymatic activity of C/D box snoRNPs is provided by fibrillarin (FBL) (Lui and Lowe, 2013). To function as guides, snoRNAs hybridize to complementary sequences within their substrate RNAs. The best characterized targets of H/ACA and C/D box snoRNP are ribosomal RNAs (rRNAs) but also other non-coding RNAs, such as small nuclear RNA (snRNAs), are frequent substrates (Bohnsack and Sloan, 2018). In the current model, RNA-RNA interactions are thought to be sufficient for a stable contact between snoRNPs and substrate RNAs, thus allowing for efficient RNA modification.

The large family of RNA-binding proteins (RBPs) is characterized by its ability to directly contact RNA molecules via RNA binding domains (RBDs) (Gerstberger et al., 2014; Hentze et al., 2018). The function of RBPs in RNA biology can be highly diverse, ranging from scaffolding and catalysis to regulation. Indeed, it is now well established that RBPs pervade virtually every layer of gene expression, including RNA processing, trafficking, stability, localization, and translation (Gehring et al., 2017).

A prominent and well-characterized RBP is the lupus autoantigen La that functions as RNA chaperone, assisting RNA polymerase III (Pol III) transcripts in adopting their correctly folded state (Hasler et al., 2016; Kucera et al., 2011; Maraia et al., 2017; Naeeni et al., 2012; Pannone et al., 1998). La is the founding member of the La-related proteins (LARPs) family (Maraia et al., 2017), which are characterized by an N-terminal La module

© 2020 Elsevier Inc.

composed of the La motif (LaM) followed by an RNA recognition motif (RRM1). Both motifs are organized as tandem arrangements to synergistically bind RNA substrates (Alfano et al., 2004). In higher eukaryotes, some LARP family members contain a second RRM (referred to as RRM2), located at their C-terminal ends. Biophysical studies revealed that RRM2 of La in isolation possesses only weak RNA binding activity on its own (Brown et al., 2016). Interestingly, addition of a short sequence stretch downstream of the RRM2 increases the affinity to RNA *in vitro*, suggesting a more complex and so far only poorly understood RNA binding mode. It is conceivable that, under *in vivo* conditions, RRM2 and its short extension synergistically function with the La module to allow various, presumably transient interactions with RNA substrates important for RNA chaperone activity (Martino et al., 2012).

LARP7 is most similar to La in terms of structural architecture and sequence homology. In contrast to La, which binds transiently to the 3′ termini of all Pol III transcripts, LARP7 forms stable RNP structures with distinct RNAs in various species. In *Tetrahymena*, the LARP7 homolog p65 stably interacts with the telomerase RNA (TER), which is transcribed by Pol III in this organism (Jiang et al., 2013). P65 is an integral component of the telomerase RNP and is important for proper telomerase function. Intriguingly, p65 functions in a RNA-chaperone-like manner during telomerase assembly (Stone et al., 2007). Pof8, a La-related protein in fission yeast, has also been identified as a constitutive component of the telomerase complex, underscoring evolutionary conservation of this interaction (Collopy et al., 2018; Mennie et al., 2018; Páez-Moscoso et al., 2018).

The best-characterized human LARP7 target is the 7SK RNA, which sequesters the positive transcription elongation factor b (P-TEFb) (Markert et al., 2008; Yang et al., 2001). Sequestration prevents the P-TEFb-dependent phosphorylation of Ser2 in the C-terminal domain (CTD) of RNA Pol II and hence impairs transcription elongation (Peterlin et al., 2012). The 7SK RNA is transcribed by Pol III and initially bound by La. During 7SK RNP assembly, La is replaced by LARP7, which remains as an integral component of the 7SK RNP (He et al., 2008; Muniz et al., 2013). Loss of LARP7 leads to a strong reduction of 7SK RNA levels and activation of RNA Pol II transcription, highlighting the importance of LARP7 for the integrity and function of the 7SK RNP. 7SK is structurally organized in four distinct hairpins connected by single-stranded sequence elements. LARP7 uses its La module to interact with the terminal UUU-3′OH and its RRM2 to contact the apical loop of the 3′ terminal hairpin (Eichhorn et al., 2016, 2018; Muniz et al., 2013; Uchikawa et al., 2015).

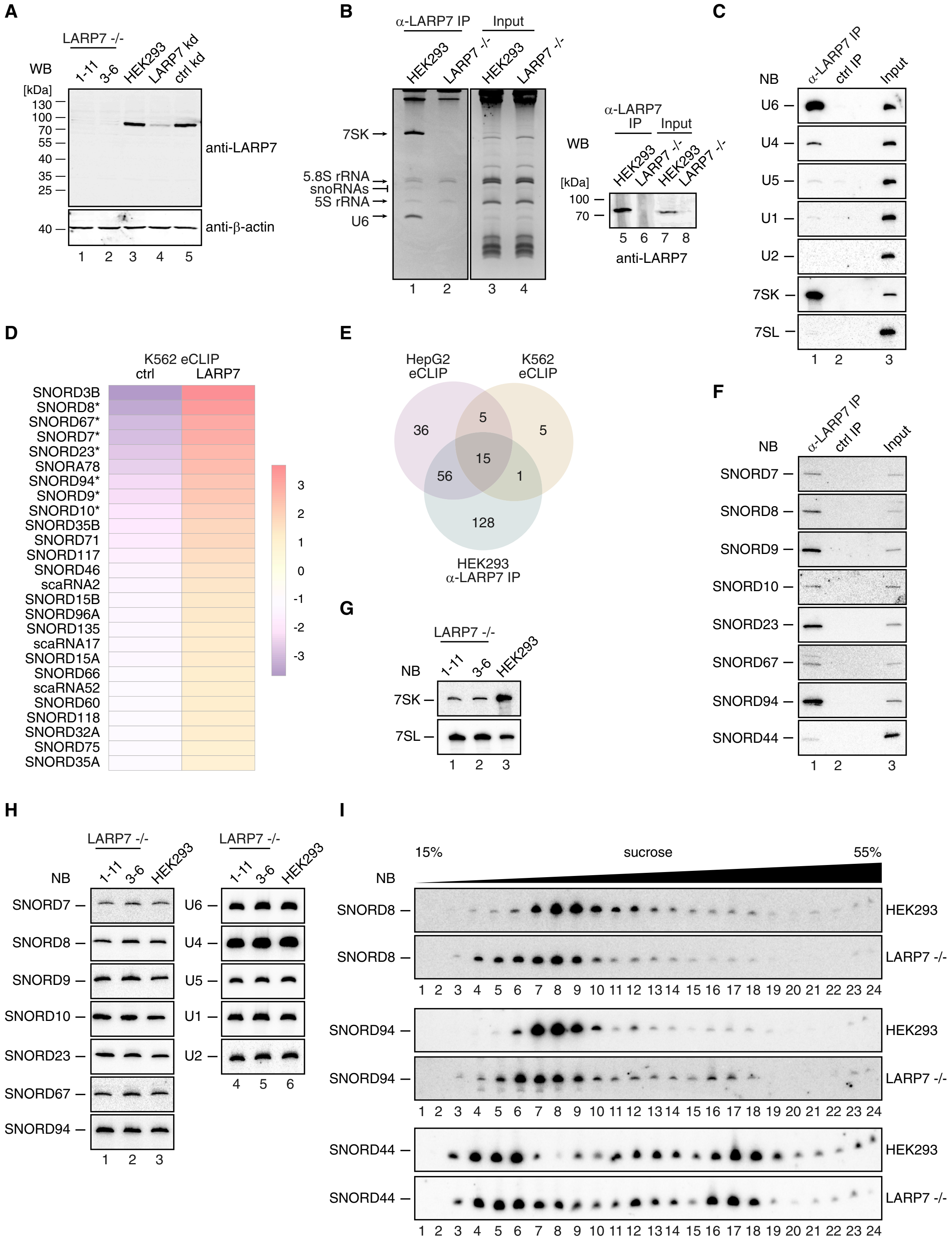
Reduced LARP7 levels have been associated with several cancers, and it has been suggested that this is due to its impaired function in transcriptional regulation through P-TEFb sequestration (Cheng et al., 2012; He et al., 2008; Ji et al., 2014). However, LARP7 loss-of-function mutations in humans are not associated with hyper-proliferation and cancer-like phenotypes. Instead, affected individuals suffer from primordial dwarfism, mental disabilities, and facial dysmorphism, collectively referred to as the Alazami syndrome (Alazami et al., 2012). These observations and the finding that loss of LARP7 reduces proliferation of embryonic stem cells (ESCs) lead to the assumption that the Alazami syndrome phenotype is independent of the LARP7-7SK axis (Dai et al., 2014). Closer examination of several Alazami syndrome patients revealed a reduced telomerase activity, and thus, it has been suggested that LARP7, alike p65 in *Tetrahymena*, affects telomerase function (Holohan et al., 2016). However, the underlying molecular mechanisms of this disease remain elusive.

To further elucidate the cellular functions of LARP7 and their potential link to Alazami syndrome, we biochemically characterized LARP7 in human cells. We find that LARP7 not only interacts with the 7SK RNA but also with other non-coding RNAs, including the U6 snRNA and a subset of C/D box snoRNAs. Strikingly, this highly specific set of snoRNAs guides the 2′-O-methylation of the U6 snRNA. We demonstrate that LARP7 functions as an adaptor protein that connects these two RNA species. Importantly, this function is critical for 2′-O-methylation as U6 snRNA modification is severely impaired in the absence of LARP7. Although U6 snRNA modification does not detectably affect spliceosome assembly and canonical splicing, RNA sequencing (RNA-seq) data reveal alterations in the usage of alternative splice sites upon LARP7 depletion. Splicing alterations, however, are moderate under optimal cell culture conditions, but several distinct splicing events are more affected under temperature stress in LARP7 knockout cells. This suggested that 2′-O-methylation of the U6 snRNA contributes to splicing robustness under changing environmental conditions. In agreement with this model, two Alazami syndrome patients display reduced 2′-O-methylation on their U6 snRNA, which is accompanied with strong alterations in alternative splicing. Strikingly, a very high ratio of the affected genes is associated with human diseases manifesting clinical symptoms similar to the Alazami phenotype. Our study unravels a so far unknown function of LARP7 as auxiliary factor in U6 2′-O-methylation and suggests that perturbation of this process contributes to the etiology of the Alazami syndrome.

RESULTS

LARP7 Associates with Distinct RNPs

Many symptoms of the Alazami syndrome are apparently incompatible with the well-established role of LARP7 in 7SK RNA-mediated transcriptional regulation. Therefore, we hypothesized that LARP7 might have additional 7SK-independent functions with relevance for this disease. To identify and characterize such putative functions, we generated two CRISPR-mediated LARP7 knockout cell lines and in addition established effective knockdown conditions as confirmed by western blot analysis (Figure 1A). For functional investigation, we generated polyclonal anti-LARP7 antibodies, which do not cross-react with the homologous La protein (Figure S1A) and performed immunoprecipitations from wild-type (WT) and LARP7 knockout cells (Figure 1B). Interestingly, we observed co-immunoprecipitation of a specific set of RNAs in WT, but not in knockout, cells. Analysis of these RNAs by sequencing and northern blotting identified the well-known target 7SK RNA but also so far unknown interactors, such as snoRNAs or the spliceosomal U6 snRNA (Figures 1B and S1B). Other RNAs with similar abundance in immunoprecipitates from WT and LARP7 knockout lysates were considered non-specific.



(legend on next page)

The U6 snRNA is present predominantly as U4/U6 and U4/U6.U5 particles, both being major building blocks of the spliceosome (Fica and Nagai, 2017; Wahl et al., 2009). It was therefore surprising that only the U6 snRNA but no other spliceosomal snRNAs were enriched in the anti-LARP7 immunoprecipitates (Figure 1B). Indeed, northern blotting revealed that only U6 became enriched to a comparable extent as 7SK RNA, whereas all other snRNAs and the 7SL RNA, which served as negative control, were detected only at background levels (Figure 1C).

Interestingly, we also found a subset of C/D box snoRNAs that specifically associated with LARP7 (Figures 1B and S1B). To further corroborate this finding, we analyzed the publicly available LARP7 crosslinking and immunoprecipitation (CLIP) data provided by the Encode project (Sloan et al., 2016) and found that snoRNAs were specifically associated with LARP7 in these studies as well (Figures 1D and 1E; Table S1). Indeed, a common set of 15 C/D box snoRNAs were identified in all Encode datasets and in our immunoprecipitation (Figure 1E; Table S1). Strikingly, these include all seven C/D box snoRNAs that guide 2′-O-methylation of the human U6 snRNA (Gumienny et al., 2017; Lestrade and Weber, 2006). To validate this specific interaction pattern, we immunoprecipitated endogenous LARP7 and performed northern blot analyses. U6-modifying snoRNAs were efficiently enriched, in contrast to SNORD44, which modifies rRNA and served as negative control (Figure 1F).

Next, we investigated whether LARP7 associates with free RNAs or assembled RNPs. Lysates from HEK293 cells stably expressing FH-LARP7 were incubated with anti-FLAG antibodies, and co-immunoprecipitated proteins were analyzed by SDS-PAGE (Figure S1C) and by mass spectrometry (Figure S1D; Table S2). The top gene ontology (GO) terms of the co-isolated proteins are “snoRNA binding,” “7SK RNA binding,” “snRNA binding,” “mRNA/5′ UTR binding,” “rRNA binding,” and “nucleocytoplasmic carrier activity.” Of note, well-known U6 snRNP components, such as the LSm proteins, were not found in our analysis. These data suggest that LARP7 specifically interacts with premature U6 snRNPs but fully assembled box C/D snoRNPs (Figure S1D).

LARP7 Does Not Affect snoRNA/U6 snRNA Levels or Their Assembly into RNPs

As LARP7 stabilizes the 7SK RNA, it is conceivable that it may act in a similar fashion on the bound snoRNAs and/or U6 snRNA. To test this possibility, we analyzed RNA levels in LARP7 WT and knockout cell lines by northern blotting. Consistent with earlier findings, 7SK levels were strongly reduced in LARP7 knockout cell lines, although 7SL, which is not bound by LARP7, remained unaffected (Figure 1G). In contrast to 7SK, neither the levels of the LARP7-associated snoRNAs (Figure 1H, left panel) nor the U6 snRNA (Figure 1H, right panel) were affected in the knockout cell lines. It has also been reported that LARP7 influences snRNA transcription (Egloff et al., 2017), which is not observed under our conditions. Based on these results, we reasoned that LARP7 exerts functions beyond the stabilization of associated RNAs.

To explore these putative functions, we asked whether the assembly of snoRNPs and/or the U6 snRNP is affected upon LARP7 deprivation. We first tested the U6 association with the LSm complex, which is a hallmark of functionally assembled U6 snRNP (Figure S1E) and SART3, a marker for U6 and U4/U6 di-snRNPs (Figure S1F). In both cases, no major differences in the association of these proteins with the U6 snRNA in LARP7-depleted cells were evident. To evaluate snoRNP assembly, we performed immunoprecipitations from LARP7 knockout as well as WT HEK293 cell line lysates using antibodies directed against FBL, the catalytic subunit of C/D box snoRNPs. In both cases, U6-specific SNORD8 and SNORD94 were co-immunoprecipitated as determined by northern blotting (Figure S1G), suggesting efficient incorporation of FBL into snoRNPs in the absence of LARP7. To further test whether LARP7 is a stable component of U6-modifying C/D box snoRNPs, we fractionated lysates from WT or LARP7 knockout cells on sucrose density gradients and analyzed snoRNP sedimentation by northern blotting (Figure 1I). The main peak of the U6-specific SNORD8 and SNORD94 shifted toward lower molecular weights in the absence of LARP7, although rRNA-specific SNORD44 remained unaffected. Our data thus suggest that LARP7 is stably associated with assembled and presumably functional U6-specific snoRNPs.

LARP7 Contacts RNA Substrates with Distinct Domains

LARP7 contains a La module and a C-terminal RRM2, which are both utilized for 7SK RNA binding (Figure 2A). To identify the domains mediating the interaction with the U6 snRNA and C/D box snoRNAs, we generated a panel of different LARP7 mutants (Figure 2B) and tested their RNA binding activity (Figures 2C, S2A, and S2B). LARP7 knockout cells were stably transfected with FH-LARP7 variants and used for immunoprecipitation experiments. Co-isolated RNAs were subsequently analyzed by northern blotting (Figure 2C). As expected from previous studies

Figure 1. LARP7 Associates with the U6 snRNA and with U6-Specific C/D Box snoRNAs

(A) Lysates from WT, LARP7 knockout HEK293 cells (1–11 and 3–6) and cells transfected with small interfering RNAs (siRNAs) against LARP7 or control siRNAs were analyzed by western blotting (WB) against LARP7. β-actin served as loading control.

(B) LARP7 was immunoprecipitated from WT or knockout cell lines, and associated RNAs were analyzed by RNA PAGE followed by ethidium bromide staining. The panel to the right shows a western blot against LARP7.

(C) Samples were obtained as described in (B) and used for northern blotting (NB) against the indicated RNAs.

(D) Analysis of LARP7 snoRNA association in K562 cells from publicly available enhanced CLIP (eCLIP) datasets.

(E) Overlap of LARP7 eCLIP datasets from HepG2, K562 cells and our sequencing data obtained from LARP7 isolations from HEK293 cells.

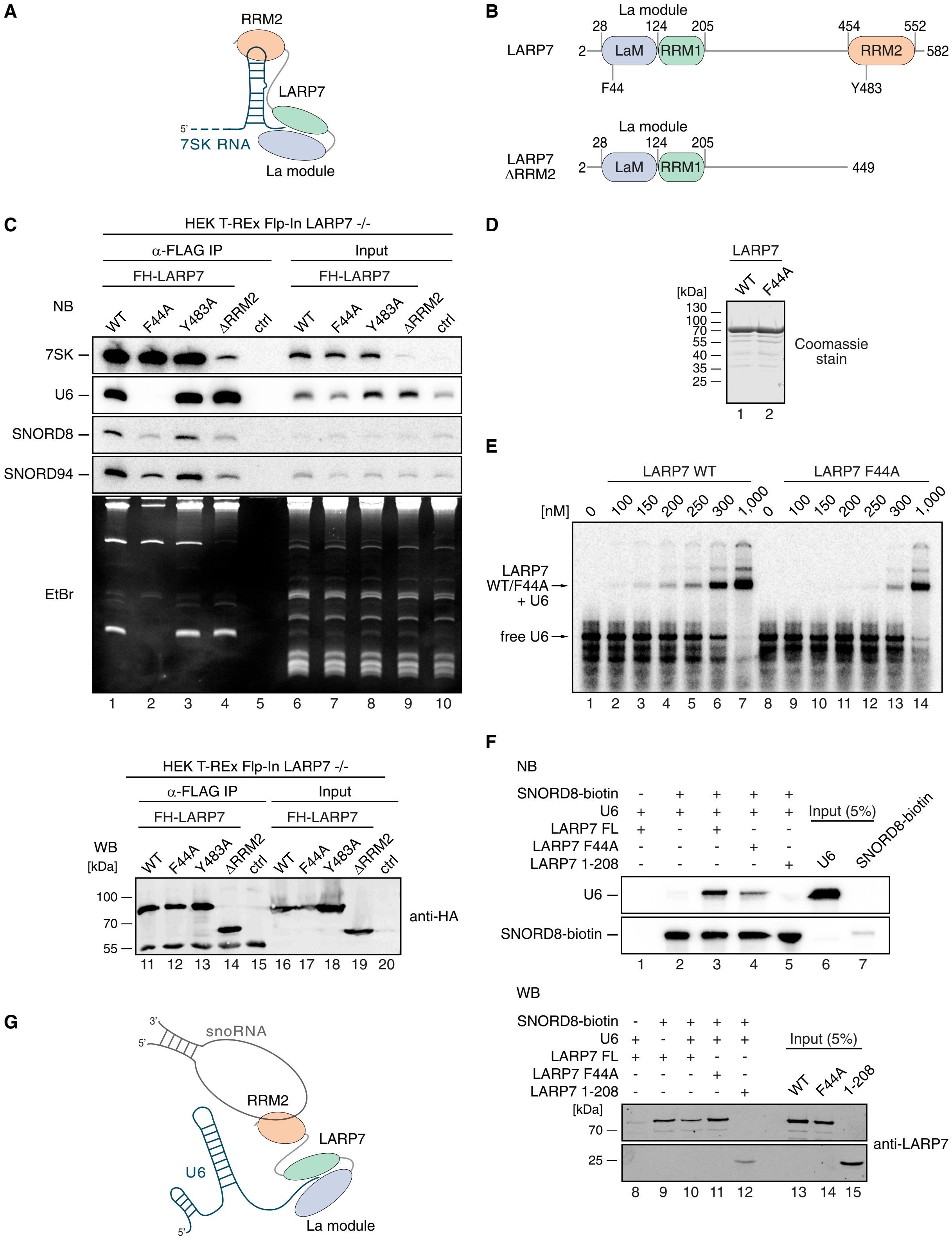
(F) Samples were obtained as described in (B) and used for northern blotting against the indicated snoRNAs.

(G) Total RNA from WT and LARP7 knockout HEK293 cells (1–11 and 3–6) were analyzed by northern blotting using probes against 7SK and 7SL RNAs.

(H) Samples used in (G) were analyzed by northern blotting using probes against the indicated C/D box snoRNAs (left) and the spliceosomal U snRNAs (right).

(I) Nuclear extracts from WT and LARP7 knockout HEK293 cells were fractionated by sucrose gradient centrifugation. RNAs extracted from each fraction were assayed by northern blotting against the indicated snoRNAs.

See also Figure S1 and Tables S1 and S2.



(legend on next page)

(Eichhorn et al., 2018), FH-LARP7 lacking RRM2 (LARP7 ΔRRM2) bound only weakly to 7SK (Figure 2C, lane 4). In contrast, this truncation still bound efficiently to U6 snRNA. Furthermore, a point mutation within RRM2 (Y483A) did not affect U6 or 7SK RNA binding (lane 3). Strikingly, the substitution F44A located in the La motif abolished U6 binding without affecting 7SK interaction (lane 2), indicating that U6 is contacted only via the La module. Of note, LARP7 mutants that bind to 7SK also rescued the loss of 7SK RNA in LARP7-deficient cells, suggesting that the generated mutants are fully functional (Figure S2C). We next used the same set of LARP7 mutants to test for SNORD8 and SNORD94 binding. Both RNAs associated only weakly with FH-LARP7 ΔRRM2 (Figure 2C, lane 4), although the Y483A substitution was not affected (lane 3). The F44A mutation, which fails to interact with U6, also reduced snoRNA binding (lane 2). We tested a number of additional mutants and consistently confirmed these observations (Figures S2A and S2B). Finally, we investigated whether the observed interaction of LARP7 with the U6 snRNA is direct or mediated by a putative bridging factor. For this, bacterially expressed recombinant LARP7 or its F44A mutant (Figure 2D) was incubated with radiolabeled U6 snRNA and complex formation was tested by electromobility shift assays (EMSAs) (Figure 2E). WT LARP7 directly bound to U6 with an estimated K_d in the range of 250 nM, whereas the binding affinity of the mutant form of LARP7 was strongly reduced. The observation that LARP7 binds to U6 and snoRNAs with different domains led us to investigate whether this results in the formation of a trimeric complex that can be assembled *in vitro* (Figure 2F). Pull-down experiments using biotinylated SNORD8 revealed that efficient association of U6 with SNORD8 was only observed in the presence of recombinant LARP7 (lane 3). However, LARP7 variants that are either deficient in U6 binding (F44A, lane 4) or encompassing only the La module, which is not sufficient for snoRNA binding (1–208, lane 5), failed to promote efficient formation of the trimeric complex.

Taken together, our results indicate that LARP7 uses specific RNA binding surfaces to directly bind to the U6 or 7SK RNAs, most likely forming functionally distinct complexes. LARP7 also binds to C/D box snoRNAs directly, as evident from UV crosslinking in CLIP experiments (Figures 1D and 1E), but molecular interactions appear to be more complex and require multiple RNA-protein contacts. Although LARP7 uses its two RNA binding domains to contact different sites of the 7SK RNA intramolecularly (Figure 2A), the same RNA binding domains are used to target U6 and snoRNA intermolecularly (Figure 2G).

U6 snRNA-Targeting C/D Box snoRNAs Contain a Conserved Sequence Motif

LARP7 binds to a distinct set of snoRNAs, and we wondered whether this interaction is facilitated by a specific RNA motif. Comparing the sequences of C/D box snoRNAs targeting the U6 snRNA in a number of different species and to all other, non-U6-specific C/D box snoRNAs, we found that 73 out of 118 (62%) U6-modifying C/D box snoRNAs contain a CAGGG sequence motif. In contrast, this motif is only present in 92 out of 4,759 (1.9%) C/D box snoRNAs that target other RNAs (Figures 3A and S3A). To assess the relevance of this motif for LARP7 binding, we mutated a short sequence stretch in SNORD8 (SNORD8 motif mutant; Figure 3B), transfected it into HEK293 SNORD8/9 double knockout cells (Figure S3B), and examined interaction with endogenous LARP7 in anti-LARP7 co-immunoprecipitations (Figure 3C). In addition, we also generated a SNORD8 variant with mutated U6 targeting sequence (SNORD8 target mutant; Figures 3B and 3D). Indeed, LARP7 binding to the SNORD8 motif mutant was strongly decreased. Similarly, the target motif mutant also bound much weaker compared to WT, suggesting that both sequence elements are important for this interaction. As LARP7 interacts with U6, which in turn base pairs with SNORD8, we generated a double mutant to assess the contribution of snoRNA-U6 interactions and thus indirect LARP7 binding. In agreement, the double mutant completely lost its LARP7 binding activity, confirming that the residual binding of the SNORD8 motif mutant is likely due to RNA-RNA contacts. To exclude that the mutations introduced into our SNORD8 constructs interfered with snoRNP assembly and function, we immunoprecipitated FBL and analyzed the bound snoRNAs by northern blotting (Figure 3E). All SNORD8 mutants were readily detected in complex with FBL, suggesting that they are integrated into presumably functional snoRNPs. Based on these results, we conclude that U6-modifying C/D box snoRNAs contain a LARP7 binding motif, which we name LARP7 binding box or LAB box. This sequence motif contributes, together with snoRNA-U6 snRNA interaction, to the formation of a trimeric complex composed of LARP7, U6 snRNA, and a corresponding snoRNP (Figure 3D).

U6 snRNA Lacks 2′-O-Methylation in LARP7-Deficient Cells

The data above raised the possibility that LARP7 is required for snoRNA-guided 2′-O-methylation of U6 snRNA. To directly test this hypothesis, we isolated U6 snRNA from WT or LARP7

Figure 2. Mapping of LARP7 Interactions with 7SK, U6, and C/D Box snoRNAs

(A) Current model of LARP7-7SK RNA interactions.

(B) Schematic representation of the used LARP7 mutations.

(C) FH-LARP7 WT, F44A, Y483A, or ΔRRM2 were stably transfected into LARP7 knockout cells and were immunoprecipitated using anti-FLAG antibodies. Co-immunoprecipitated RNAs were analyzed by northern blotting and ethidium bromide (EtBr) staining. Input samples are presented in lanes 6–10. The lower panel shows an anti-hemagglutinin (HA) western blot of the immunoprecipitated LARP7 variants as well as input LARP7 expression levels.

(D) Coomassie staining of recombinant LARP7 proteins.

(E) Electromobility shift assay (EMSA) of LARP7 WT and the F44A mutant. Indicated concentrations of the recombinant proteins were incubated with constant amounts of radiolabeled U6 snRNA. Samples were subsequently analyzed by native gel electrophoresis.

(F) The formation of the trimeric complex consisting of the biotinylated U6-specific SNORD8 snoRNA, the U6 snRNA, and LARP7 was investigated by *in vitro* pull-down assays followed by northern blotting (upper panels) and western blotting (lower panels).

(G) Model of the contacts formed by LARP7 to the U6 snRNA and to U6-specific snoRNAs.

See also Figure S2.

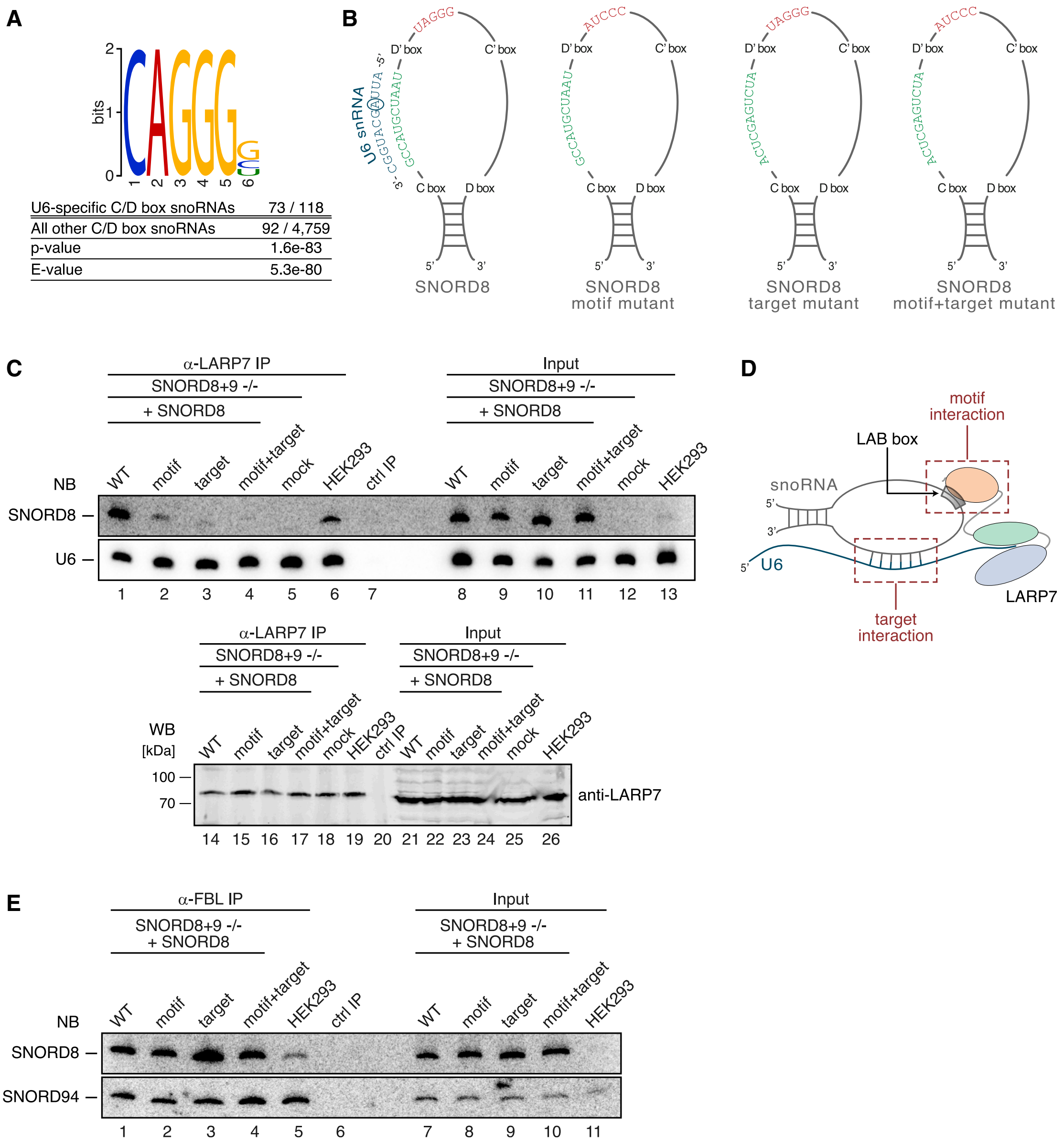


Figure 3. A Sequence Motif Enriched in U6-Modifying C/D Box snoRNAs Interacts with LARP7

(A) The sequences of U6-specific snoRNAs from different species were analyzed for the enrichment of a specific motif compared to all other C/D box snoRNAs. A Fisher's exact test classifies the enrichment of this motif as highly significant.

(B) Schematic representation of the different SNORD8 variants that were used.

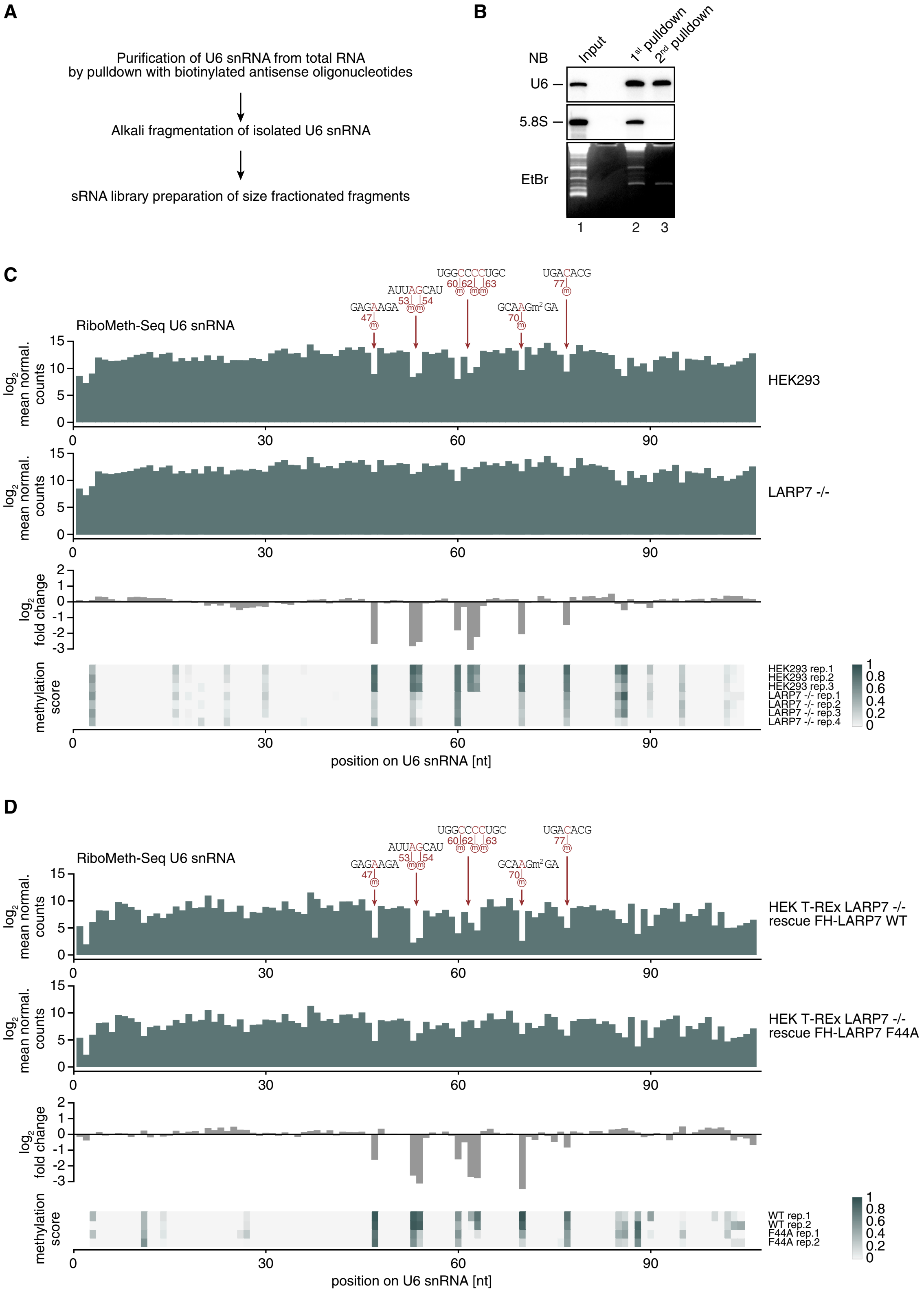
(C) SNORD8 and SNORD9 double knockout HEK293 cells (SNORD8+9−/−) were transfected with the indicated SNORD8 constructs and endogenous LARP7 was immunoprecipitated from the lysates. Associated SNORD8 (upper panel) or U6 (lower panel) were analyzed by northern blotting. Lane 5 shows mock transfected and lane 6 WT HEK293 cells. Input samples are shown to the right and an anti-LARP7 western blot in the lower part.

(D) Model of snoRNA-U6-LARP7 interactions.

(E) SNORD8+9−/− cells were transfected as described in (C), and endogenous FBL was immunoprecipitated. Co-isolated SNORD8 variants as well as SNORD94 were analyzed by northern blotting. In lane 5, WT HEK293 cells were used. Lanes 7–11 show input samples.

See also Figure S3.

knockout HEK293 cells using biotinylated oligonucleotides (Figure 4A). After two consecutive purification steps, U6 snRNA was highly enriched (Figure 4B) and then used for 2′-O-methylation mapping applying RiboMeth-seq (Birkedal et al., 2015). U6 snRNA was treated with alkaline conditions to hydrolyze the phosphodiester backbone of the RNA. Because this reaction requires a 2′ hydroxyl group of the ribose, 2′-O-methylated nucleotides are protected from alkaline hydrolysis, thus allowing the identification of modified sites by RNA-seq and bioinformatic analysis. 2′-O-methylation sites are uncovered by mapping the



(legend on next page)

start and the end of the sequenced reads and each modified position result in a drop of coverage at the respective nucleotide (Figure 4C). In our experiments, all known 2′-O-methylation sites on U6 snRNA are readily detectable (Figure 4C, upper panel). Interestingly, these modifications are strongly reduced in LARP7-deficient cells (Figure 4C, lower panel), indicating that LARP7 is indeed required for efficient U6 2′-O-methylation. As control, we performed similar experiments with the U2 snRNA, which also contains a defined set of site-specific 2′-O-methylated ribose. None of these sites was affected in LARP7 knockout cells, suggesting that LARP7 specifically acts on U6 snRNA (Figure S4). To corroborate these findings, we rescued the LARP7 knockout cells by integrating either WT LARP7 or the F44A mutant stably into the genome (Figure 4D). In agreement with our model, WT LARP7 fully rescued U6 2′-O-methylation, but the F44A mutant, which cannot bind to U6, had a much weaker effect. Taken together, our data reveal an essential role of LARP7 in snoRNP-mediated 2′-O-methylation of the U6 snRNA.

Effects of U6 2′-O-Methylation Deficiency on Splicing Fidelity

Although the general functions of 2′-O-methylation of the U6 snRNA are largely elusive, it is conceivable that the lack of 2′-O-methylation could influence splicing. Hence, we tested whether general pre-mRNA splicing is affected and investigated spliceosome assembly as well as pre-mRNA splicing *in vitro*. We found that, under these experimental conditions, neither spliceosome formation nor splicing of a model pre-mRNA substrate is dependent on LARP7 (Figures S5A and S5B). In order to unravel consequences of LARP7 depletion in a physiological context, we performed total RNA-seq from WT as well as LARP7-deficient cells. Comparing overall gene expression levels, we observed only modest changes (Figure 5A). Dissociation of the P-TEFb kinase complex from the 7SK snRNP results in enhanced RNA Pol II Ser2 phosphorylation (Ser2-P), which in turn increases elongation rates. To assess potential P-TEFb-linked effects in our gene expression data, we investigated Ser2-P using phospho-specific antibodies (Figure 5B). Unexpectedly, we did not observe detectable differences between WT and LARP7 knockout cells. Of note, the western blot signal is phospho-specific because it disappeared upon phosphatase treatment (right panel). Furthermore, it has been demonstrated that CDK9 levels, i.e., the kinase subunit of P-TEFb acting on Ser2, are reduced to compensate for RNA Pol II hyperactivation upon LARP7 knockdown (Dai et al., 2014). Indeed, we observed a mild reduction of CDK9 levels in our LARP7 knockout cells, suggesting adaptation to the 7SK loss (Figure S5C). Thus, global transcriptional effects that may have resulted from the destabilization of the 7SK RNA accompanied by increased RNA Pol II transcription were not evident in this analysis.

We next assessed our RNA-seq data for alternative splicing events affected in LARP7 knockout cells (Figure 5C; Table S3) and observed moderate but significant changes at distinct splice sites. We therefore conclude that splicing is not generally affected and hypothesize that U6 2′-O-methylation could contribute to a general splicing robustness or fidelity by noise reduction. To test this, we examined alterations in distinct splicing patterns in our RNA-seq data and found indeed a number of events that were sensitive to LARP7 loss. These events, however, were not of a particular mode of alternative splicing, which is generally consistent with broader effects on overall splicing fidelity (Figure 5C).

We selected a number of splicing events and validated them directly. We first confirmed that the observed effects are found in two independent LARP7 knockout cell lines. We then performed semiquantitative RT-PCR (Figure 5D) as well as radioactive RT-PCR experiments (Figures 5E and S6A). Strikingly, quantification of the radioactive signals revealed that the candidates PARP6, KMT2D (also known as MLL2), and SETMAR show modest but highly reproducible changes in alternative splicing patterns. Of note, some of the investigated effects could not be confirmed in a second knockout cell line and were considered unspecific (Figure S6B).

The data above suggest that LARP7-supported U6 modification has only a minor impact on splicing under normal conditions. However, splicing robustness might become more important under stress conditions, such as higher temperatures, that all cells experience when organisms react to infections with fever, for example. We hence shifted WT or LARP7 knockout cells to 40°C and performed RNA-seq (Figures S6C and S6D). Consistent with this idea, under high-temperature conditions, several distinct splice sites are much stronger affected in LARP7 knockout cells compared to the control cell line.

Thus, our findings are consistent with the model that LARP7-guided 2′-O-methylation of the U6 snRNA is not necessary for its function in splicing per se but rather contributes to splicing robustness. These effects might be relevant under conditions of cellular stress, such as high temperature, where RNA-RNA interactions are stabilized to reduce splicing noise.

A LARP7 Mutation Affects U6 2′-O-Methylation in a Family with Alazami Syndrome

Because mutations in the *LARP7* gene are linked to the Alazami syndrome and this phenotype cannot be readily explained by the role of LARP7 in transcription, we asked whether the newly discovered function of LARP7 in RNA modification might

Figure 4. LARP7 Knockout Affects 2′-O-Methylation of the U6 snRNA

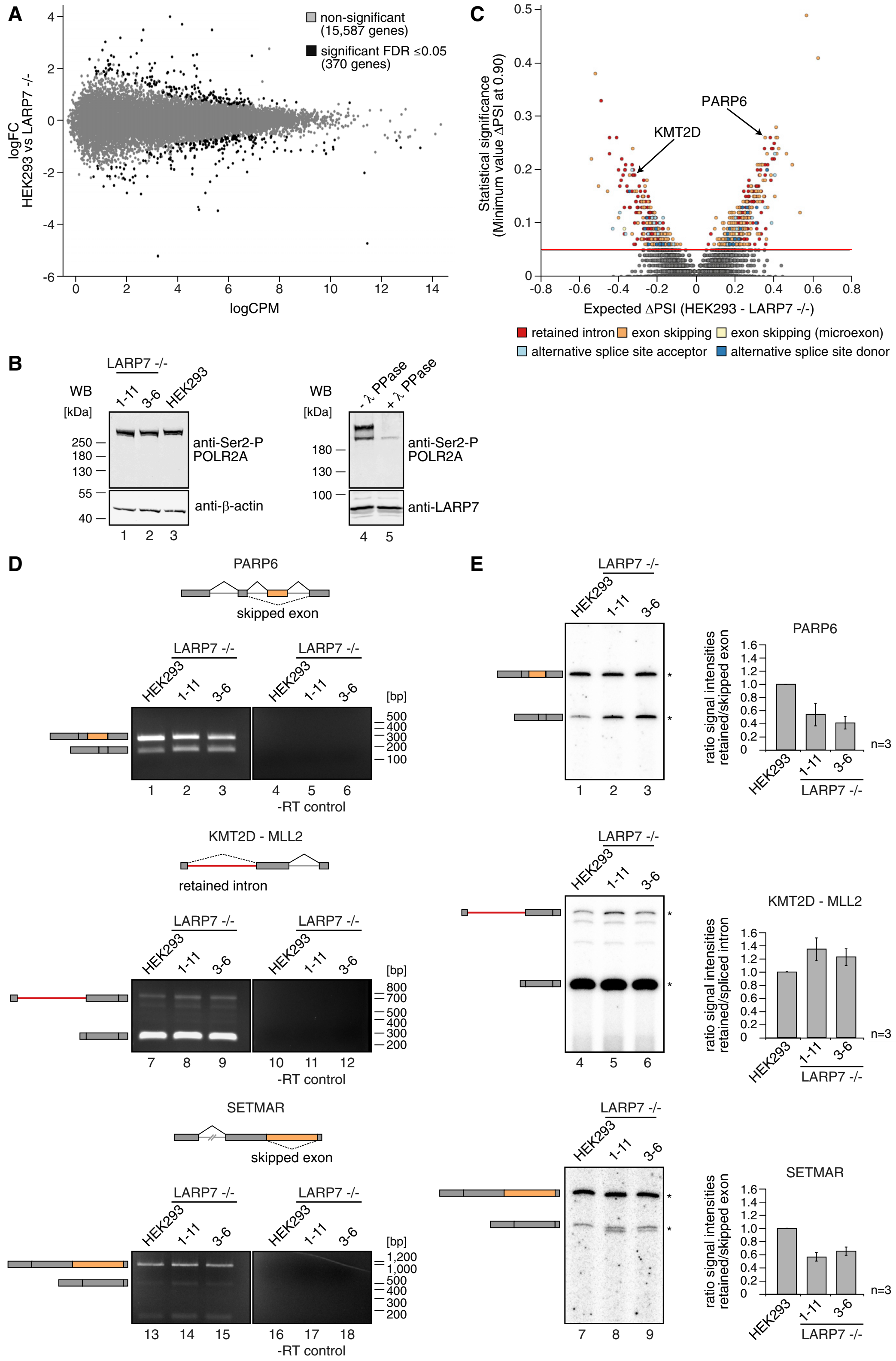
(A) Workflow of U6 snRNA enrichment using biotinylated antisense probes and generation of RiboMeth-seq libraries.

(B) The isolated RNA was analyzed by EtBr staining and northern blotting using probes against the U6 snRNA and the 5.8S rRNA.

(C) The isolated U6 snRNA was fragmented, cloned, and sequenced (RiboMeth-seq). Reads are plotted as $\log_2$ mean-normalized counts for WT HEK293 cells and LARP7 knockout cells (first and second panels) or as $\log_2$ fold change ($\log_2$FC) between the two cell lines (third panel). Panel 4 shows the “methylation score” calculated as described in Birkedal et al. (2015).

(D) LARP7 knockout cells were rescued by stably expressing WT LARP7 or the LARP7 F44A mutants, and RiboMeth-seq experiments as well as data analysis were performed as described in (C).

See also Figure S4.



(legend on next page)

contribute to this disease. To test this hypothesis, we analyzed the genomic DNA as well as RNA from blood samples of a family with two siblings diagnosed for Alazami syndrome (Figures 6A–6D). Exome analysis revealed that both siblings are homozygous for the deletion c.1669-1_1671del, which destroys the 3′ splice site of intron14/exon15 of *LARP7* (Figure 6B).

Analysis of the LARP7 mRNA revealed that the identified deletion leads to splicing of exon 14 to a cryptic splice acceptor located further downstream in exon 15 (Figure 6B), which results in a 13-nt deletion. This would generate a frameshift resulting in a premature stop codon (Figure 6C; Würzburg variant). We hence examined whether the mutated mRNA is removed from cells by the nonsense-mediated decay (NMD) pathway. qPCR analysis of the heterozygous parents and the homozygous siblings excluded this possibility, as LARP7 RNA levels are unaffected (Figure 6D). We reasoned that the patients express a mutant LARP7 protein containing an altered C terminus. To test this, we established B-lymphoblastoid cell lines (B-LCLs) from donated blood samples. Western blot analysis revealed that LARP7 protein levels were comparable between parents and their affected children. As the mutated LARP7 is predicted to differ only in eleven amino acids from the WT protein, the two protein forms could not be distinguished by protein gel electrophoresis (Figure 6E). We therefore tested by northern blotting of blood samples whether LARP7 stabilizes the 7SK RNA in the Alazami patients (Figure 6F). Although the mutated LARP7 variant was present in normal amounts, the level of 7SK RNA was strongly reduced in both children compared to their heterozygous parents. Control RNAs, such as U2, but also U6-specific SNORD23 and SNORD94, were unaffected. For further analysis of the associated RNA species, we immunoprecipitated the mutated LARP7 variant from patient-derived B-LCLs. As expected, binding to 7SK was severely compromised in these cells. Although the interaction to the U6 snRNA was comparable between the parent- and patient-derived cell lines, we observed a markedly reduced interaction of the LARP7 mutant with the U6-specific C/D box snoRNAs (Figure 6G). Finally, we expressed the C-terminal domain of WT LARP7 and of the corresponding Würzburg variant (Figure 6H) and assessed direct binding to SNORD8 in EMSA experiments (Figure 6I). Also in these direct interaction assays, binding of the Alazami variant was markedly reduced.

U6 2′-O-Methylation Defects and Alterations in Splice Site Usage in Alazami Patients

Based on our finding that the 2′-O-methylation of U6 requires LARP7, we performed RiboMeth-seq experiments with the U6 snRNA isolated from the generated B-LCLs (Figure 7A). All known 2′-O-methylation sites can be readily detected in the cell line derived from a parent. Interestingly, most 2′-O-methylation sites are affected in the patient-derived cell line. Of note, RiboMeth-seq of U6 from blood samples shows strong effects on two distinct sites, although others were less affected (Figure S7A), which might be due to very low RNA amounts that were available for our experiments. The observed modification defects may affect alternative splicing, as we already observed in HEK293 cells (Figure 5C). To test this directly, we performed RNA-seq experiments using the parent- and patient-derived B-LCLs (Figures 7B and S7B). Interestingly, splice site selection is strongly altered in the patient with mutated LARP7. Consistently with our observations in HEK293 cells, RNA Pol II Ser2-P is unchanged and CDK9 levels are mildly reduced in the patient-derived cells, suggesting that loss of 7SK might play a minor role in gene expression changes in the Alazami patients (Figures S7C and S7D). Strikingly, when closely investigating the affected genes, more than 16% of them are associated with diseases sharing similar clinical phenotypes as Alazami patients (Figure 7C). To test the contribution of the Alazami variant on splicing fidelity more rigorously, we selected the PARP6 exon skipping event and performed rescue experiments in LARP7-deficient HEK293 cells (Figure 7D). Strikingly, WT LARP7, but not the Würzburg Alazami variant, rescued the observed splicing change. Thus, our data uncover a so far unknown biochemical defect in Alazami syndrome patients, which likely contributes to the multifaceted disease phenotype.

DISCUSSION

LARP7 serves as an integral component of the 7SK RNP (He et al., 2008), where it replaces the canonical La protein from the 3′ end during RNP assembly and stabilizes the 7SK RNA. Using a biochemical approach, we have identified another so far unknown 7SK-independent function of LARP7 in human cells. We show that LARP7 acts as an adaptor protein connecting the U6 snRNA with C/D box snoRNAs for 2′-O-methylation. In

Figure 5. LARP7 Knockout Cells Exhibit Changes in Splice Site Usage

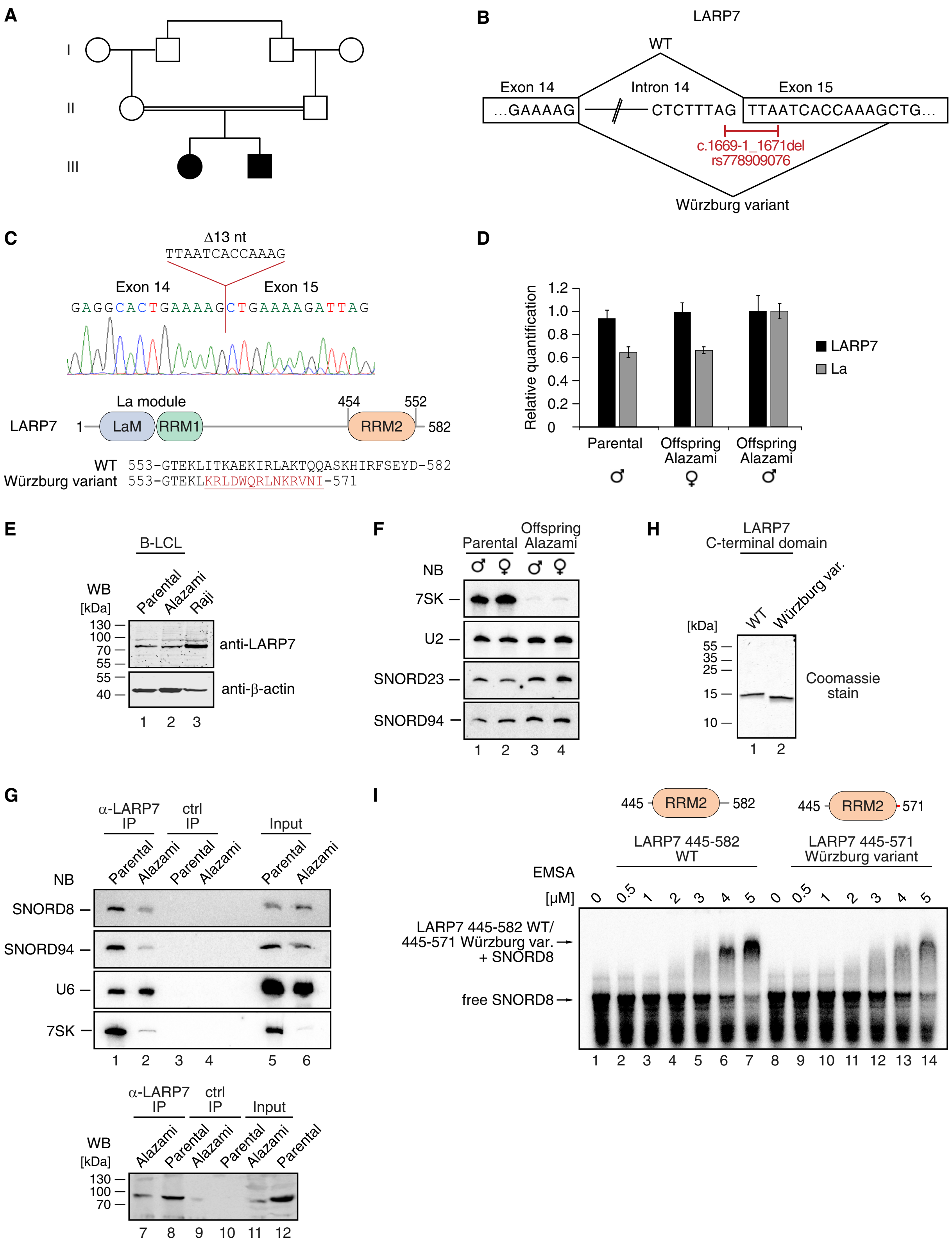
(A) Total RNA from LARP7 knockout and WT HEK293 cells was used for RNA-seq experiments. Gene expression was quantified, and the natural logarithm of the fold change (logFC) between WT and LARP7 knockout is plotted (y axis) against the logarithm of the counts per million (logCPM) value (x axis). Differentially expressed genes with a false discovery rate (FDR) $\leq$ 0.05 are shown in black. The genes that do not show significant changes in expression are shown in gray.

(B) Ser2-P levels of the RNA Pol II subunit POLR2A were assayed by western blotting using lysates from LARP7 knockout (lanes 1 and 2) and WT (lane 3) HEK293 cell lines (panels to the left). In the right panel, antibody specificity for phosphorylated POLR2A was tested by lambda protein phosphatase treatment of HEK293 cell lysate (lane 5).

(C) The RNA-seq data shown in (A) were analyzed for changes in alternative splicing between WT and LARP7 knockout cells. The Volcano plot shows differentially regulated splicing events, color coded as indicated for the different event types. The statistical significance, quantified as the minimum absolute difference (MV) of percent spliced in (ΔPSI) for a given event that is supported at a 0.90 probability, is plotted on the y axis. The expected value for ΔPSI (HEK293 WT $-$ LARP7 knockout) is plotted on the x axis. The red line indicates the threshold for statistical significance of a minimal absolute ΔPSI value above 0.05. Two of the events chosen for validation experiments are highlighted.

(D and E) Validation of alternative splicing events between LARP7 knockout and WT HEK293 cells by conventional RT-PCR (D) or radioactive RT-PCR (E). A schematic representation of the affected splice patterns is shown. Representative gels are shown in (E), and the bands used for quantification of the signals are indicated with asterisks. The ratio of the signals from three biological replicates is shown to the right with error bars depicting ± SD.

See also Figures S5 and S6 and Table S3.



(legend on next page)

LARP7 knockout cells, 2′-O-methylation of U6 is strongly reduced, and our data suggest LARP7-mediated fine-tuning effects on splicing. This is particularly important in tissues development and cell differentiation, where specific alternative splicing events are crucial (see accompanying study by Wang et al., 2020 [this issue of *Molecular Cell*]). We further report that 2′-O-methylation of the U6 snRNA is impaired in Alazami syndrome patients, suggesting a contribution of splicing alterations to the disease.

LARP7 uses two separate RNA binding domains to contact different RNA species. The La module anchors LARP7 on the U6 snRNA, and the C-terminal RRM2 interacts with a specific subset of C/D box snoRNAs. We have identified a conserved sequence motif required for LARP7 binding, which we refer to as LAB box. This motif is specifically enriched in C/D box snoRNAs targeting U6. Based on our data, we propose a model in which the bivalent RBP LARP7 functions as a scaffold for two different RNAs, which hybridize to each other through complementary sequences. Although a major contact between U6 and the C/D box snoRNAs is mediated by base pairing, U6 remains almost unmethylated in LARP7-deficient cells highlighting the importance of LARP7 as an auxiliary factor for U6-specific 2′-O-methylation. A number of different modes of action could be envisioned for LARP7 in this process. First, LARP7 might use its two RBDs that are connected by a presumably flexible linker region to anchor U6 and the C/D box snoRNAs and thus accelerating the generation of RNA-RNA contacts. Second, base pairing between the U6 snRNA and its modifying snoRNAs might not be stable enough to allow catalysis and thus requires LARP7-mediated clamping to further stabilize the interaction, allowing for efficient 2′-O-methylation. Third, LARP7 might play a more active role in catalysis by stimulating FBL activity or positioning. It would therefore be interesting to study whether LARP7 engages in direct protein-protein interactions with FBL. Fourth, it has been suggested that LARP7 can function as RNA chaperone (Hussain et al., 2013). It is hence conceivable that LARP7 induces a specific structural conformation of U6 that allows for efficient C/D box snoRNA binding. For example, LARP7 might rearrange or prevent secondary structures prior to 2′-O-methylation. Structural investigations of LARP7 simultaneously bound to a C/D box snoRNA and U6 snRNA will provide valuable insights into this U6 biogenesis intermediate complex.

The mode of action of LARP7 in snoRNP-mediated RNA modification serves as a paradigm for a double-sided RBP as auxiliary factor for 2′-O-methylation. It also raises the possibility that other modular RBPs exert similar functions. For example, the U6 snRNA is also modified by pseudouridylation, and this process requires H/ACA box snoRNAs. These snoRNAs are structurally different from C/D box snoRNAs and thus might find their specific targets in a different way. However, it is also possible that so far unknown RBPs fulfill LARP7-like functions in U6 pseudouridylation. Moreover, other spliceosomal U snRNAs, such as the U2 snRNA, are also heavily modified by 2′-O-methylation and pseudouridylation (Bohnsack and Sloan, 2018), and specific RBPs may support the modification of other snRNAs. Finally, most human snoRNAs target rRNAs during ribosome biogenesis. A vast number of RBPs are essential for ribosome biogenesis, and a direct role in 2′-O-methylation or pseudouridylation has not been investigated. Interestingly, recent studies associated different RBPs with 2′-O-methylation of rRNAs (D'Souza et al., 2018; Nachmani et al., 2019). A mechanism for these observations has not been unraveled. Together with our findings, these data suggest that effects of RBPs on snoRNA-guided RNA modification might be a widespread phenomenon. In situations of weak base pairing, the requirement of LARP7-like auxiliary factors might be critical although other RNAs may stably pair without the help of such factors.

Although all spliceosomal snRNAs carry modifications, the precise functions of individual modified nucleotides are poorly understood. 2′-O-methylation enhances RNA duplex stability by stabilizing the A-form RNA-RNA helix and affects global folding by preventing hydrogen bond formation with the 2′ hydroxyl group (Ayadi et al., 2019; Prusiner et al., 1974). During the splice cycle, U6 base pairs with U4, U2, and also the 5′ splice site of the pre-mRNA at later stages of catalysis (Bohnsack and Sloan, 2018). Consistently, a U2-U6 complex is stabilized by RNA modification, but this contribution is rather mild and a general requirement for splicing is unlikely (Karunatilaka and Rueda, 2014). Instead, U6 2′-O-methylation may influence RNA structure formation and thus splicing fidelity under specific conditions, such as cellular or environmental stress, or in specific tissues. Strikingly, an accompanying paper (Wang et al., 2020) shows mouse male germline-specific alternative splicing changes that can be attributed to LARP7 deficiency and U6

Figure 6. A Novel *LARP7* Mutation in Alazami Syndrome Patients

(A) Family tree of the two Alazami syndrome patients.

(B) Schematic representation of the *LARP7* micro-deletion found in the two patients shown in (A).

(C) Sequencing of the LARP7 cDNA from blood samples revealed the usage of a cryptic splice site leading to an alternative C terminus of LARP7.

(D) RNA was isolated from blood samples donated by the parents or the two Alazami syndrome siblings, and LARP7 mRNA expression levels were determined by qPCR. La expression levels were measured as control. The error bars represent ± SD from three technical replicates.

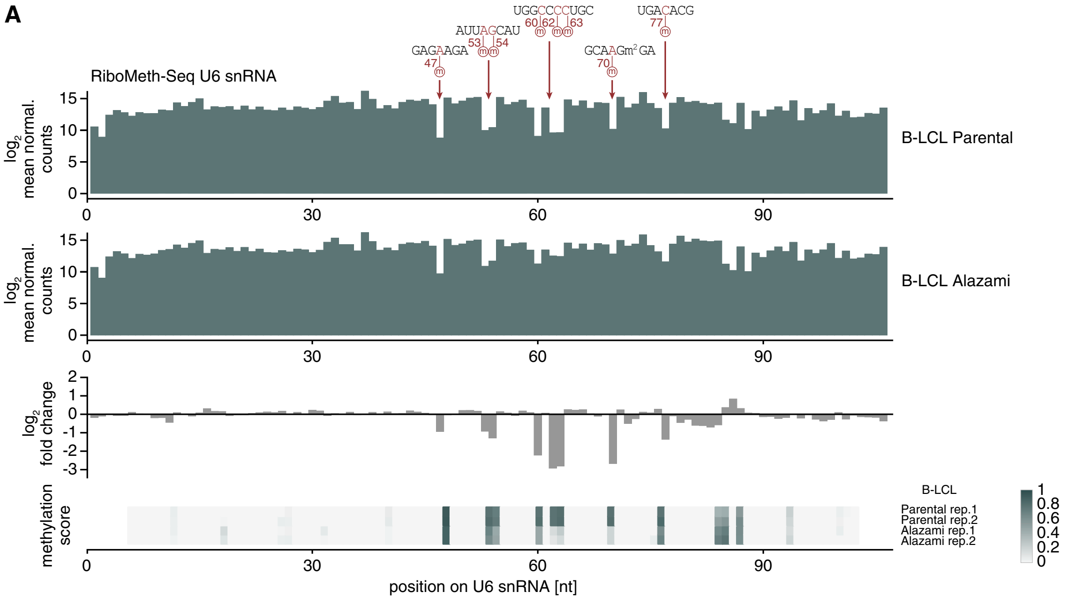
(E) LARP7 protein expression levels in B-LCLs derived from a healthy parent (lane 1) or from an Alazami syndrome child (lane 2), as well as the B-cell-derived Raji cell line (lane 3) were determined by western blotting. β-actin served as loading control.

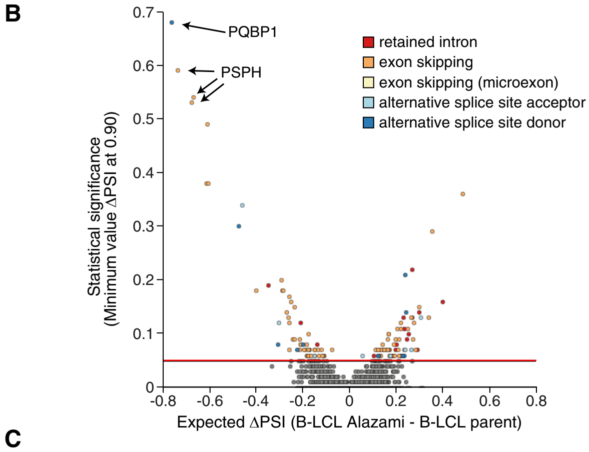
(F) RNA from blood samples donated by healthy parents (lanes 1 and 2) or Alazami patients (lanes 3 and 4) was analyzed by northern blotting for the indicated RNAs.

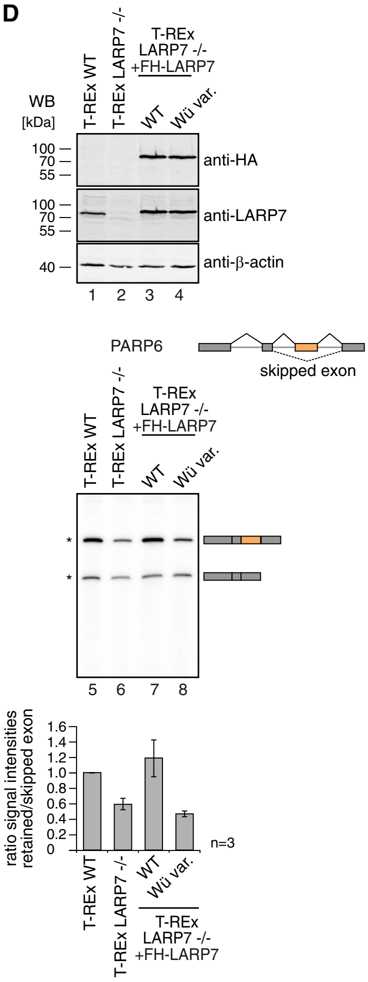
(G) RNA co-purified in anti-LARP7 immunoprecipitations from B-LCLs shown in (E) were analyzed by northern blotting. Input samples are shown in lanes 5 and 6 (upper panels). The corresponding western blot analysis is shown in the lower panel.

(H) Recombinant expression of the C-terminal domain of the LARP7 WT (lane 1) or of the Würzburg variant (lane 2) followed by SDS-PAGE and Coomassie staining.

(I) Binding of the LARP7 protein variants shown in (H) to *in vitro* transcribed and radiolabeled SNORD8 RNA was assayed in EMSA experiments. The concentrations of the recombinant proteins used for each condition are indicated on top of the lanes.







C

Gene	Disease (OMIM No.)	Symptoms shared with Alazami syndrome
PQPB1	Renpenning syndrome (309500)	🟧 🟩 🟦 🟪 ⬜ 🟫
PSPH	Phosphoserine phosphatase deficiency (614023)	🟧 🟩 🟦 🟪 ⬜ ▫
ACADSB	2-methylbutyrylglycinuria (610006)	🟧 ⬜ ▫
BICD2	Spinal muscular atrophy, lower extremity-predominant, 2B (618291)	🟧 ▫
STRADA	Polyhydramnios, megalencephaly, and symptomatic epilepsy (611087)	🟧 🟪 ⬜ ▫ 🟫
ATG5	Spinocerebellar ataxia, autosomal recessive 25 (617584)	🟧
CHKB	Muscular dystrophy, congenital, megaconial type (602541)	🟧 🟦 🟫
EHMT1	Kleefstra syndrome 1 (610253)	🟧 🟦 ⬜ ▫
TINF2	Dyskeratosis congenita, autosomal dominant 3 (613990) Revesz syndrome (268130)	🟧 🟩 🟦
GMNN	Meier-Gorlin syndrome 6 (616835)	🟧 🟩 🟦 🟪
BIN1	Centronuclear myopathy 2 (255200)	🟧 ▫
CC2D1A	Mental retardation, autosomal recessive 3 (608443)	🟧
SLC25A26	Combined oxidative phosphorylation deficiency 28 (616794)	🟧 ▫
PHF21A	Potocki-Shaffer syndrome (601224)	🟧 🟪 ⬜
CNTNAP1	Hypomyelinating neuropathy, congenital, 3 (618186)	🟧 🟩 🟦 🟪 ⬜ ▫
TRIO	Mental retardation, autosomal dominant 44 (617061)	🟧 🟦
SIK3	Spondyloepimetaphyseal dysplasia, Krakow type (618162)	🟧 🟩 ⬜ 🟫
RFWD3	Fanconi anemia, complementation group W (617784)	🟩 🟦
NT5C2	Spastic paraplegia 45, autosomal recessive (613162)	🟧 🟩 🟦

🟧 intellectual disability / developmental delay
🟩 short stature / growth retardation
🟦 microcephaly
🟪 dysmorphic facial features (e.g., triangular face, short philtrum, broad nose, cleft palate)
⬜ seizures
▫ hypotonia
🟫 atrial septal defects

(legend on next page)

hypo-methylation. These data support our model that U6 2′-O-methylation becomes essential under the physiological conditions in a living organism. Nonetheless, even under optimal tissue culture conditions, splicing patterns of several mRNAs are moderately changed. Most importantly, when we analyze alternative splicing under temperature stress, LARP7 knockout cells produce different splicing patterns compared to WT cells. Under fever conditions, cells are exposed to high temperature, but accurate splicing needs nonetheless be maintained. Hence, the establishment of fever strategies as defense line against infections might be the positive selection pressure for 2′-O-methylation of spliceosomal snRNAs during evolution. This might even be true for various other RNA species that carry such modifications.

The Alazami syndrome is caused by loss-of-function mutations within the LARP7 gene (Alazami et al., 2012). We have analyzed a pair of siblings with a short deletion spanning the last exon-intron boundary. This results in the use of an alternative splice acceptor and a frameshift leading to a marginally shorter protein carrying a different C terminus. Using immortalized blood cells from the patients, we find that the mutated mRNA escapes NMD and produces a stable protein variant. Interestingly, although the mutation is outside of the RRM2, a clear reduction of 7SK is observed. This is consistent with the findings that RRM2 is immediately followed by an α helix, which appears to be important for RNA interaction in addition to RRM2. Because we found that this region is also involved in snoRNA binding, effects on U6 modification efficiency could be predicted. Indeed, RiboMeth-seq revealed strongly reduced U6 2′-O-methylation in the Alazami patients, identifying a so far unknown facet of this disease.

The Alazami syndrome has been associated with loss of 7SK function as well as reduced telomerase activity. Because the loss of 7SK leads to hyper-proliferation and cancer, it has been questioned whether 7SK alone is the cause of the Alazami syndrome, which is characterized by primordial dwarfism. Moreover, the involvement of telomerase is also unclear, because we did not observe changes in telomerase RNA levels in the patient samples that we investigated (data not shown). Based on our study, we propose that loss of U6 2′-O-methylation and thus changes in splicing fidelity or robustness contribute to the polymorphic phenotype of the Alazami syndrome. Interestingly, it was reported that mutations in the minor spliceosome can also lead to primordial dwarfism (Verma et al., 2018), a phenotype that is characteristic for Alazami patients. This model is further underscored by our observation that many genes that are associated with the clinical symptoms of the Alazami disease are also alternatively spliced in Alazami patients (Figure 7C). Taken together, these findings are in agreement with the model that the Alazami syndrome belongs to the growing group of splicing diseases.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Patient recruitment and clinical report
  - ○ Generation of B-lymphoblastoid cell lines
  - ○ Generation of knockout cell lines
  - ○ Generation of stable FH-LARP7 cell lines
  - ○ Other cell lines and cell culture conditions
- METHOD DETAILS
  - ○ Sequencing from human samples
  - ○ RNA sequencing and analysis
  - ○ Small RNA sequencing
  - ○ Pulldown of RNAs for RiboMeth-Seq analysis
  - ○ RiboMeth-sequencing and data analysis
  - ○ Databases for annotation of sequencing data
  - ○ Analysis of ENCODE CLIP datasets
  - ○ Analysis of small RNA sequencing datasets
  - ○ Sequence analysis of U6-specific snoRNAs
  - ○ Plasmids
  - ○ Transfections
  - ○ Immunoprecipitations
  - ○ Synthesis of cDNA and quantitative PCR
  - ○ Protein expression and purification
  - ○ Generation of polyclonal antibodies
  - ○ *In vitro* transcription of RNA
  - ○ ^{32}P-labeling of oligonucleotides
  - ○ Northern blot
  - ○ Electromobility shift assay

Figure 7. Reduced U6 snRNA 2′-O-Methylation and Changes in Alternative Splicing in Alazami Syndrome Patients

(A) RiboMeth-seq analysis of U6 snRNA isolated from the B-LCLs derived from a healthy parent or a child homozygous for the Alazami-syndrome-associated *LARP7* c.1669-1_1671del variant (upper two plots). The coverage of the termini of the sequenced fragments is expressed in $\log_2$ of mean normalized counts. The $\log_2$FC for each position and the resulting methylation scores are depicted in the lower graph and heatmaps. Known 2′-O-methylated positions are indicated on top.

(B) Poly(A) RNA from the B-LCLs derived from an Alazami syndrome child and the healthy father were used for RNA-seq experiments, and differences in splicing patterns were determined according to Figure 5C.

(C) Genes affected by differential splice site usage as determined in (B) were analyzed for their association with human diseases. Genes and diseases, which share similar clinical symptoms with the Alazami phenotype, are summarized in the table.

(D) Effects on the alternative splicing of PARP6 upon stable expression of FH-LARP7 WT or FH-LARP7 Würzburg variant in LARP7 knockout cell lines. Protein levels were assayed by western blotting with the indicated antibodies. Cell lysates prepared from HEK T-REx Flp-In WT (lane 1), HEK T-REx Flp-In LARP7 knockout (lane 2), and from the two FH-LARP7 overexpression cell lines (lanes 3 and 4) were used (upper panels). Radioactive RT-PCRs were performed as in Figure 5E with RNA obtained from the corresponding samples (lanes 5–8). The autoradiogram of a representative gel is shown in the central panel with a schematic representation of the affected splice pattern. The signals indicated with asterisks were quantified from three biological replicates, and the ratio between retained or skipped exon events was determined with error bars depicting ± SD (lower panel).

See also Figure S7.

- ○ Radioactive RT-PCR
- ○ Detection of radioactive signals
- ○ *In vitro* pull down of biotinylated snoRNA-U6 snRNA-LARP7 complexes
- ○ Spliceosome assembly/*in vitro* splicing assay
- ○ Fractionation of nuclear extracts by sucrose gradient centrifugation
- ○ SDS-PAGE and western blotting
- ○ Mass spectrometric analysis
- ○ GO term enrichment analysis
- ● QUANTIFICATION AND STATISTICAL ANALYSIS
- ● DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.molcel.2020.01.001.

ACKNOWLEDGMENTS

We thank S. Ammon, C. Friederich, E. Hochmuth, N. Pluta, A. Prusty, N. Eichner, and R. Hett for technical support and N. Treiber for VP5- and pcDNA5-LARP7 WT plasmids. Our work is supported by the Deutsche Forschungsgemeinschaft (Fi573/15-1/20-1, Me2064/9-1, SPP 1935, and SFB 960), the European Union (ERC grant 682291 ‘‘moreRNAs’’), and the Bavarian Ministry for Education and Science (BioSysNet). F.M.S. was funded by a post-doctoral grant provided by Région Bretagne SAD2-2017-RNA2init (17009) and R.M. by the Postdoc-Plus program of the GSLS. M.-F.L. was supported by the National Key R&D Program of China (2017YFA0504400) and the National Natural Science Foundation of China (31961133022 and 31830109). We are grateful to the family participating to our study.

AUTHOR CONTRIBUTIONS

D.H. performed most experiments. R.M., X.W., and Z.-T.L. contributed to experiments. M.B., G.L., and M.Z. designed and contributed bioinformatic analysis. L.H. and R.M. performed motif enrichment. F.M.S. and A.-C.D.-B. provided recombinant LARP7 protein. E.G., R.K., and E.K. recruited patients and established cell lines. A.B. performed mass spectrometry. D.H., M.-F.L., U.F., and G.M. designed and discussed experiments. D.H., U.F., and G.M. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 15, 2019
Revised: November 25, 2019
Accepted: December 26, 2019
Published: February 3, 2020

REFERENCES

Alazami, A.M., Al-Owain, M., Alzahrani, F., Shuaib, T., Al-Shamrani, H., Al-Falki, Y.H., Al-Qahtani, S.M., Alsheddi, T., Colak, D., and Alkuraya, F.S. (2012). Loss of function mutation in LARP7, chaperone of 7SK ncRNA, causes a syndrome of facial dysmorphism, intellectual disability, and primordial dwarfism. Hum. Mutat. *33*, 1429–1434.

Alfano, C., Sanfelice, D., Babon, J., Kelly, G., Jacks, A., Curry, S., and Conte, M.R. (2004). Structural analysis of cooperative RNA binding by the La motif and central RRM domain of human La protein. Nat. Struct. Mol. Biol. *11*, 323–329.

Ayadi, L., Galvanin, A., Pichot, F., Marchand, V., and Motorin, Y. (2019). RNA ribose methylation (2′-O-methylation): occurrence, biosynthesis and biological functions. Biochim. Biophys. Acta Gene Regul. Mech. *1862*, 253–269.

Bailey, T.L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics *27*, 1653–1659.

Birkedal, U., Christensen-Dalsgaard, M., Krogh, N., Sabarinathan, R., Gorodkin, J., and Nielsen, H. (2015). Profiling of ribose methylations in RNA by high-throughput sequencing. Angew. Chem. Int. Ed. Engl. *54*, 451–455.

Bohnsack, M.T., and Sloan, K.E. (2018). Modifications in small nuclear RNAs and their roles in spliceosome assembly and function. Biol. Chem. *399*, 1265–1276.

Breuza, L., Poux, S., Estreicher, A., Famiglietti, M.L., Magrane, M., Tognolli, M., Bridge, A., Baratin, D., and Redaschi, N.; UniProt Consortium (2016). The UniProtKB guide to the human proteome. Database (Oxford) *2016*, bav120.

Brown, K.A., Sharifi, S., Hussain, R., Donaldson, L., Bayfield, M.A., and Wilson, D.J. (2016). Distinct dynamic modes enable the engagement of dissimilar ligands in a promiscuous atypical RNA recognition motif. Biochemistry *55*, 7141–7150.

Cheng, Y., Jin, Z., Agarwal, R., Ma, K., Yang, J., Ibrahim, S., Olaru, A.V., David, S., Ashktorab, H., Smoot, D.T., et al. (2012). LARP7 is a potential tumor suppressor gene in gastric cancer. Lab. Invest. *92*, 1013–1019.

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

Collopy, L.C., Ware, T.L., Goncalves, T., Í Kongsstovu, S., Yang, Q., Amelina, H., Pinder, C., Alenazi, A., Moiseeva, V., Pearson, S.R., et al. (2018). LARP7 family proteins have conserved function in telomerase assembly. Nat. Commun. *9*, 557.

D'Souza, M.N., Gowda, N.K.C., Tiwari, V., Babu, R.O., Anand, P., Dastidar, S.G., Singh, R., James, O.G., Selvaraj, B., Pal, R., et al. (2018). FMRP interacts with C/D box snoRNA in the nucleus and regulates ribosomal RNA methylation. iScience *9*, 399–411.

Dai, Q., Luan, G., Deng, L., Lei, T., Kang, H., Song, X., Zhang, Y., Xiao, Z.X., and Li, Q. (2014). Primordial dwarfism gene maintains Lin28 expression to safeguard embryonic stem cells from premature differentiation. Cell Rep. *7*, 735–746.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46* (D1), D794–D801.

Dignam, J.D., Lebovitz, R.M., and Roeder, R.G. (1983). Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. Nucleic Acids Res. *11*, 1475–1489.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Egloff, S., Vitali, P., Tellier, M., Raffel, R., Murphy, S., and Kiss, T. (2017). The 7SK snRNP associates with the little elongation complex to promote snRNA gene expression. EMBO J. *36*, 934–948.

Eichhorn, C.D., Chug, R., and Feigon, J. (2016). hLARP7 C-terminal domain contains an xRRM that binds the 3′ hairpin of 7SK RNA. Nucleic Acids Res. *44*, 9977–9989.

Eichhorn, C.D., Yang, Y., Repeta, L., and Feigon, J. (2018). Structural basis for recognition of human 7SK long noncoding RNA by the La-related protein Larp7. Proc. Natl. Acad. Sci. USA *115*, E6457–E6466.

Fica, S.M., and Nagai, K. (2017). Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine. Nat. Struct. Mol. Biol. *24*, 791–799.

Gehring, N.H., Wahle, E., and Fischer, U. (2017). Deciphering the mRNP code: RNA-bound determinants of post-transcriptional gene regulation. Trends Biochem. Sci. *42*, 369–382.

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. Nat. Rev. Genet. *15*, 829–845.

Gumienny, R., Jedlinski, D.J., Schmidt, A., Gypas, F., Martin, G., Vina-Vilaseca, A., and Zavolan, M. (2017). High-throughput identification of C/D box snoRNA targets with CLIP and RiboMeth-seq. Nucleic Acids Res. *45*, 2341–2353.

Han, H., Braunschweig, U., Gonatopoulos-Pournatzis, T., Weatheritt, R.J., Hirsch, C.L., Ha, K.C.H., Radovani, E., Nabeel-Shah, S., Sterne-Weiler, T., Wang, J., et al. (2017). Multilayered control of alternative splicing regulatory networks by transcription factors. Mol. Cell *65*, 539–553.e7.

Hasler, D., Lehmann, G., Murakawa, Y., Klironomos, F., Jakob, L., Grässer, F.A., Rajewsky, N., Landthaler, M., and Meister, G. (2016). The lupus autoantigen La prevents mis-channeling of tRNA fragments into the human microRNA pathway. Mol. Cell *63*, 110–124.

He, N., Jahchan, N.S., Hong, E., Li, Q., Bayfield, M.A., Maraia, R.J., Luo, K., and Zhou, Q. (2008). A La-related protein modulates 7SK snRNP integrity to suppress P-TEFb-dependent transcriptional elongation and tumorigenesis. Mol. Cell *29*, 588–599.

Hentze, M.W., Castello, A., Schwarzl, T., and Preiss, T. (2018). A brave new world of RNA-binding proteins. Nat. Rev. Mol. Cell Biol. *19*, 327–341.

Holohan, B., Kim, W., Lai, T.P., Hoshiyama, H., Zhang, N., Alazami, A.M., Wright, W.E., Meyn, M.S., Alkuraya, F.S., and Shay, J.W. (2016). Impaired telomere maintenance in Alazami syndrome patients with LARP7 deficiency. BMC Genomics *17* (*Suppl 9*), 749.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods *12*, 115–121.

Hussain, R.H., Zawawi, M., and Bayfield, M.A. (2013). Conservation of RNA chaperone activity of the human La-related proteins 4, 6 and 7. Nucleic Acids Res. *41*, 8715–8725.

Ji, X., Lu, H., Zhou, Q., and Luo, K. (2014). LARP7 suppresses P-TEFb activity to inhibit breast cancer progression and metastasis. eLife *3*, e02907.

Jiang, J., Miracco, E.J., Hong, K., Eckert, B., Chan, H., Cash, D.D., Min, B., Zhou, Z.H., Collins, K., and Feigon, J. (2013). The architecture of Tetrahymena telomerase holoenzyme. Nature *496*, 187–192.

Jorjani, H., Kehr, S., Jedlinski, D.J., Gumienny, R., Hertel, J., Stadler, P.F., Zavolan, M., and Gruber, A.R. (2016). An updated human snoRNAome. Nucleic Acids Res. *44*, 5068–5082.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D., and Petrov, A.I. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. *46* (D1), D335–D342.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al.; University of California Santa Cruz (2003). The UCSC Genome Browser Database. Nucleic Acids Res. *31*, 51–54.

Karunatilaka, K.S., and Rueda, D. (2014). Post-transcriptional modifications modulate conformational dynamics in human U2-U6 snRNA complex. RNA *20*, 16–23.

Kucera, N.J., Hodsdon, M.E., and Wolin, S.L. (2011). An intrinsically disordered C terminus allows the La protein to assist the biogenesis of diverse noncoding RNA precursors. Proc. Natl. Acad. Sci. USA *108*, 1308–1313.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lestrade, L., and Weber, M.J. (2006). snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res. *34*, D158–D162.

Love, M.I., Soneson, C., and Patro, R. (2018). Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. F1000Res. *7*, 952.

Lui, L., and Lowe, T. (2013). Small nucleolar RNAs and RNA-guided post-transcriptional modification. Essays Biochem. *54*, 53–77.

Maraia, R.J., Mattijssen, S., Cruz-Gallardo, I., and Conte, M.R. (2017). The La and related RNA-binding proteins (LARPs): structures, functions, and evolving perspectives. Wiley Interdiscip. Rev. RNA *8*, https://doi.org/10.1002/wrna.1430.

Markert, A., Grimm, M., Martinez, J., Wiesner, J., Meyerhans, A., Meyuhas, O., Sickmann, A., and Fischer, U. (2008). The La-related protein LARP7 is a component of the 7SK ribonucleoprotein and affects transcription of cellular and viral polymerase II genes. EMBO Rep. *9*, 569–575.

Martino, L., Pennell, S., Kelly, G., Bui, T.T., Kotik-Kogan, O., Smerdon, S.J., Drake, A.F., Curry, S., and Conte, M.R. (2012). Analysis of the interaction with the hepatitis C virus mRNA reveals an alternative mode of RNA recognition by the human La protein. Nucleic Acids Res. *40*, 1381–1394.

Matera, A.G., Terns, R.M., and Terns, M.P. (2007). Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nat. Rev. Mol. Cell Biol. *8*, 209–220.

Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., and Tuschl, T. (2004). Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. Mol. Cell *15*, 185–197.

Mellacheruvu, D., Wright, Z., Couzens, A.L., Lambert, J.P., St-Denis, N.A., Li, T., Miteva, Y.V., Hauri, S., Sardiu, M.E., Low, T.Y., et al. (2013). The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. Nat. Methods *10*, 730–736.

Mennie, A.K., Moser, B.A., and Nakamura, T.M. (2018). LARP7-like protein Pof8 regulates telomerase assembly and poly(A)+TERRA expression in fission yeast. Nat. Commun. *9*, 586.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P.D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. *45* (D1), D183–D189.

Muniz, L., Egloff, S., and Kiss, T. (2013). RNA elements directing in vivo assembly of the 7SK/MePCE/Larp7 transcriptional regulatory snRNP. Nucleic Acids Res. *41*, 4686–4698.

Nachmani, D., Bothmer, A.H., Grisendi, S., Mele, A., Bothmer, D., Lee, J.D., Monteleone, E., Cheng, K., Zhang, Y., Bester, A.C., et al. (2019). Germline NPM1 mutations lead to altered rRNA 2′-O-methylation and cause dyskeratosis congenita. Nat. Genet. *51*, 1518–1529.

Naeeni, A.R., Conte, M.R., and Bayfield, M.A. (2012). RNA chaperone activity of human La protein is mediated by variant RNA recognition motif. J. Biol. Chem. *287*, 5472–5482.

Neitzel, H. (1986). A routine method for the establishment of permanent growing lymphoblastoid cell lines. Hum. Genet. *73*, 320–326.

Páez-Moscoso, D.J., Pan, L., Sigauke, R.F., Schroeder, M.R., Tang, W., and Baumann, P. (2018). Pof8 is a La-related protein and a constitutive component of telomerase in fission yeast. Nat. Commun. *9*, 587.

Pannone, B.K., Xue, D., and Wolin, S.L. (1998). A role for the yeast La protein in U6 snRNP assembly: evidence that the La protein is a molecular chaperone for RNA polymerase III transcripts. EMBO J. *17*, 7442–7453.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419.

Peterlin, B.M., Brogie, J.E., and Price, D.H. (2012). 7SK snRNA: a noncoding RNA that plays a major role in regulating eukaryotic transcription. Wiley Interdiscip. Rev. RNA *3*, 92–103.

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L.I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. *45*, D833–D839.

Prusiner, P., Yathindra, N., and Sundaralingam, M. (1974). Effect of ribose O(2′)-methylation on the conformation of nucleosides and nucleotides. Biochim. Biophys. Acta *366*, 115–123.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. *8*, 2281–2308.

Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T., et al. (2016). ENCODE data at the ENCODE portal. Nucleic Acids Res. *44* (D1), D726–D732.

Stone, M.D., Mihalusova, M., O'connor, C.M., Prathapam, R., Collins, K., and Zhuang, X. (2007). Stepwise protein-mediated RNA folding directs assembly of telomerase ribonucleoprotein. Nature *446*, 458–461.

Tapial, J., Ha, K.C.H., Sterne-Weiler, T., Gohr, A., Braunschweig, U., Hermoso-Pulido, A., Quesnel-Vallières, M., Permanyer, J., Sodaei, R., Marquez, Y., et al. (2017). An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res. *27*, 1759–1768.

Uchikawa, E., Natchiar, K.S., Han, X., Proux, F., Roblin, P., Zhang, E., Durand, A., Klaholz, B.P., and Dock-Bregeon, A.C. (2015). Structural insight into the mechanism of stabilization of the 7SK small nuclear RNA by LARP7. Nucleic Acids Res. *43*, 3373–3388.

Verma, B., Akinyi, M.V., Norppa, A.J., and Frilander, M.J. (2018). Minor spliceosome and disease. Semin. Cell Dev. Biol. *79*, 103–112.

Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. Cell *136*, 701–718.

Wang, X., Li, Z.-T., Yan, Y., Lin, P., Tang, W., Hasler, D., Meduri, R., Li, Y., Hua, M.-M., Qi, H.-T., et al. (2020). LARP7-mediated U6 snRNA modification ensures splicing fidelity and spermatogenesis in mice. Mol. Cell *77*. Published online February 3, 2020. https://doi.org/10.1016/j.molcel.2020.01.002.

Yang, Z., Zhu, Q., Luo, K., and Zhou, Q. (2001). The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. Nature *414*, 317–322.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al. (2018). Ensembl 2018. Nucleic Acids Res. *46* (D1), D754–D761.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit polyclonal anti-LARP7 antibody	This paper	SY7862
Rabbit polyclonal anti-LARP7 antibody	This paper	SY7863
Rabbit polyclonal anti-Fibrillarin antibody	Bethyl Laboratories	Cat#A303-891A; RRID: AB_2620241
Rabbit polyclonal anti-LSM4 antibody	Antibody Genie	Cat#CAB5891
Rat monoclonal anti-RNA polymerase II subunit B1 (phospho CTD Ser-2) antibody, clone 3E10	Merck Millipore	Cat#04-1571; RRID: AB_11212363
Rabbit polyclonal anti-CDK9 antibody	Bethyl Laboratories	Cat#A303-493A-M, RRID: AB_10949230
Rabbit polyclonal anti-SART3 antibody	This paper	1631
Mouse monoclonal anti-beta actin antibody (clone AC-15)	GeneTex	Cat#GTX26276; RRID: AB_367161
Mouse monoclonal anti-influenza hemagglutinin (HA) antibody HA.11 (clone 16B12)	Covance Research Products	Cat#MMS-101P; RRID: AB_2314672
Goat polyclonal anti-Rabbit IgG, IRDye 800CW conjugated antibody	LI-COR Biosciences	Cat#926-32211; RRID: AB_621843
Goat polyclonal anti-Mouse IgG, IRDye 800CW conjugated antibody	LI-COR Biosciences	Cat#925-32210; RRID: AB_2687825
Goat polyclonal anti-Rat IgG, IRDye 800CW conjugated antibody	LI-COR Biosciences	Cat#926-32219; RRID: AB_1850025
Goat anti-Mouse IgG, IRDye 680RD conjugated antibody	LI-COR Biosciences	Cat#926-68070; RRID: AB_10956588
Bacterial and Virus Strains		
Escherichia coli Rosetta (DE3)	Our laboratory	N/A
Biological Samples		
Human blood samples Alazami patients and parents	This paper	N/A
B-lymphoblastoid cell lines (B-LCLs) derived from healthy parent and Alazami patient	This paper	AASD72, AAAS14
Chemicals, Peptides, and Recombinant Proteins		
6xHis-LARP7 (human)	Markert et al., 2008	N/A
LARP7 full-length wildtype (human)	Uchikawa et al., 2015	N/A
LARP7 F44A (human)	This paper	N/A
LARP7 1-208 (human)	Uchikawa et al., 2015	N/A
LARP7 C-terminal domain 445-582 (human)	This paper	N/A
LARP7 C-terminal domain 445-571 Würzburg variant (human)	This paper	N/A
T7 RNA polymerase	Our laboratory	N/A
Thermostable inorganic pyrophosphatase	New England Biolabs	Cat#M0296
RiboLock RNase inhibitor	Thermo Scientific	Cat#EO0384
RNase-Free DNase Set	QIAGEN	Cat#79254
T4 RNA ligase 1	New England Biolabs	Cat#M0204
Truncated T4 RNA ligase 2	Our laboratory	N/A
Calf intestinal alkaline phosphatase	New England Biolabs	Cat#M0290
T4 polynucleotide kinase	Thermo Scientific	Cat#EK0031
Tobacco acid pyrophosphatase	Epicenter	Cat#T19100 (discontinued)
Lambda protein phosphatase	New England Biolabs	Cat#P0753S
Rotiphorese sequencing gel system	Carl Roth	Cat#A431.1

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
TRIzol reagent	Invitrogen	Cat#15596018
TRIzol LS reagent	Invitrogen	Cat#10296010
Dynabeads M-270 streptavidin	Invitrogen	Cat#65306
ANTI-FLAG M2 affinity gel	Sigma-Aldrich	Cat#A2220
nProtein A Sepharose Fast Flow antibody purification resin	GE Healthcare	Cat#17528003
Illustra MicroSpin G-25 columns	GE Healthcare	Cat#27532501
pCp-Biotin	Jena Bioscience	Cat#NU-1706-BIO
SYBR gold nucleic acid gel stain	Invitrogen	Cat#S11494
N-(3-dimethylaminopropyl)-N-ethylcarbodiimide hydrochloride (EDC)	Sigma-Aldrich	Cat#E7750; CAS: 25952-53-8
1-Methylimidazole	Sigma-Aldrich	Cat#M50834; CAS: 616-47-7
Critical Commercial Assays		
Ovation SoLo RNA-seq systems	NuGEN Technologies	Cat#0407
Universal Plus mRNA-Seq	NuGEN Technologies	Cat#0508
SuperScriptIII first strand synthesis super mix	Invitrogen	Cat#18080400
First-strand cDNA synthesis kit	Thermo Scientific	Cat#K1612
SsoFast EvaGreen supermix	Bio-Rad Laboratories	Cat#1725205
RNeasy Mini Kit	QIAGEN	Cat#74104
QIAshredder	QIAGEN	Cat# 79654
NuPAGE 4-12% Bis-Tris protein gels	Invitrogen	Cat#NP0321PK2
Deposited Data		
Raw and analyzed deep-sequencing data	This paper	GEO: GSE126911
LARP7-enriched snoRNAs	This paper	Table S1
Mass spectrometric analysis of FH-LARP7 associated proteins	This paper	Table S2
Alternative splicing events in LARP7 knockout cells	This paper	Table S3
LARP7 mutation associated with Alazami syndrome	This paper	ClinVar: RCV000678485.1
ENCODE CLIP dataset K562 LARP7	Davis et al., 2018	ENCSR456KXI
ENCODE CLIP dataset K562 mock	Davis et al., 2018	ENCSR863ZGZ
ENCODE CLIP dataset HepG2 LARP7	Davis et al., 2018	ENCSR961OKA
ENCODE CLIP dataset HepG2 mock	Davis et al., 2018	ENCSR095SIV
Annotations for the human reference genome (assembly GRCh38.p12) Ensembl release 92	Zerbino et al., 2018	http://apr2018.archive.ensembl.org/index.html; RRID: SCR_006773
Annotations for RepeatMasker, tRNAs and snaRs from UCSC table browser hg38 (February 2018)	Karolchik et al., 2003	https://genome.ucsc.edu/cgi-bin/hgTables; RRID: SCR_005780
Annotation for human snoRNA from snoRNA atlas	Jorjani et al., 2016	http://snoatlas.bioinf.uni-leipzig.de/
Annotation for snoRNA of different species from rfam 14.1	Kalvari et al., 2018	http://rfam.xfam.org/; RRID: SCR_007891
Online Mendelian Inheritance in Man (OMIM), catalog of Human Genes and Genetic Disorders (June 2019)	McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University	https://www.omim.org/; RRID: SCR_006437
DisGeNET, collections of genes and variants associated to human diseases	Piñero et al., 2017	https://www.disgenet.org/; RRID: SCR_006178
UniProtKB/Swiss-Prot *Homo sapiens* database	Breuza et al., 2016	https://www.uniprot.org/; RRID: SCR_004426
Experimental Models: Cell Lines		
Human: HEK293T	Our laboratory	RRID: CVCL_0063
Human: HEK293T LARP7 −/− clone 1-11	This paper	N/A
Human: HEK293T LARP7 −/− clone 3-6	This paper	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Human: HEK293T SNORD8 + SNORD9 −/−	This paper	N/A
Human: Flp-In T-REx 293 cell line	Invitrogen	Cat#R78007; RRID: CVCL_U427
Human: HEK293 T-REx FH-LARP7	This paper	N/A
Human: HEK293 T-REx Flp/In LARP7 −/−	This paper	N/A
Human: HEK293 T-REx LARP7 −/− + FH-LARP7 WT	This paper	N/A
Human: HEK293 T-REx LARP7 −/− + FH-LARP7 F44A	This paper	N/A
Human: HEK293 T-REx LARP7 −/− + FH-LARP7 Y483A	This paper	N/A
Human: HEK293 T-REx LARP7 −/− + FH-LARP7 ΔRRM2	This paper	N/A
Human: HEK293 T-REx LARP7 −/− + FH-LARP7 Würzburg variant	This paper	N/A
Human: Raji	Laboratory of Friedrich A. Grässer	RRID: CVCL_0511
Oligonucleotides		
LARP7 siRNA pool (siPool)	siTOOLs Biotech	N/A
Control siRNA pool (siPool)	siTOOLs Biotech	N/A
Biotinylated DNA oligonucleotide sequences for RNA pulldowns	metabion international	Table S4
Northern blot probe sequences	metabion international	Table S4
DNA oligonucleotide sequences for cloning, mutagenesis and DNA amplifications in general	metabion international Eurofins Genomics	Table S4
DNA oligonucleotide sequences for RT-PCRs and qRT-PCRs	metabion international Eurofins Genomics	Table S4
Recombinant DNA		
pET28b + 6xHis-LARP7	Markert et al., 2008	N/A
pnEA + 6xHis-LARP7 WT	Uchikawa et al., 2015	N/A
pnEA + 6xHis-LARP7 F44A	This paper	N/A
pnEA + 6xHis-LARP7 1-208	Uchikawa et al., 2015	N/A
pnEA + 6xHis-LARP7 C-terminal domain 445-582	This paper	N/A
pnEA + 6xHis-LARP7 C-terminal domain 445-571 Würzburg variant	This paper	N/A
pSpCas9(BB)-2A-Puro (PX459) V2.0	Ran et al., 2013	Addgene plasmid #62988
PX459 V2.0 + LARP7-1	This paper	N/A
PX459 V2.0 + LARP7-2	This paper	N/A
PX459 V2.0 + SNORD8 up + down	This paper	N/A
PX459 V2.0 + SNORD9 up + down	This paper	N/A
pOG44	Invitrogen	Cat#V600520
pcDNA5/FRT/TO modified with N-terminal FH-tag	Invitrogen	Cat#V652020
pcDNA5/FRT/TO + FH-LARP7 WT	This paper	N/A
pcDNA5/FRT/TO + FH-LARP7 F44A	This paper	N/A
pcDNA5/FRT/TO + FH-LARP7 Y483A	This paper	N/A
pcDNA5/FRT/TO + FH-LARP7 ΔRRM2	This paper	N/A
pcDNA5/FRT/TO + FH-LARP7 Würzburg variant	This paper	N/A
pGEM-T-Easy	Promega	Cat#A1360
pIRES-VP5 (VP5) modified	Meister et al., 2004	N/A
VP5 + FH-La	Hasler et al., 2016	N/A
VP5 + FH-LARP7 WT	This paper	N/A
VP5 + FH-LARP7 F44A	This paper	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
VP5 + FH-LARP7 D54A F56A	This paper	N/A
VP5 + FH-LARP7 Y128A E130A	This paper	N/A
VP5 + FH-LARP7 Y483A	This paper	N/A
VP5 + FH-LARP7 ΔLa module	This paper	N/A
VP5 + FH-LARP7 2-561	This paper	N/A
VP5 + FH-LARP7 ΔRRM2	This paper	N/A
pSUPER modified	OligoEngine	Cat#VEC-PBS-0002
pSUPER + U6 snRNA	This paper	N/A
pcDNA3.1 (+) modified multiple cloning site	Laboratory of Jan Medenbach	N/A
pcDNA3.1 + SNORD8 WT	This paper	N/A
pcDNA3.1 + SNORD8 motif mutant	This paper	N/A
pcDNA3.1 + SNORD8 target mutant	This paper	N/A
pcDNA3.1 + SNORD8 motif+target mutant	This paper	N/A
Software and Algorithms		
Discriminative regular expression motif elicitation (DREME)	Bailey, 2011	http://meme-suite.org/tools/dreme; RRID: SCR_016860
GO term enrichment analysis (PANTHER)	Mi et al., 2017	http://amigo.geneontology.org/rte; RRID: SCR_004869
CRAPome	Mellacheruvu et al., 2013	http://www.crapome.org/
STAR	Dobin et al., 2013	https://github.com/alexdobin/STAR; RRID: SCR_015899
Bowtie2	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml; RRID: SCR_016368
Salmon	Patro et al., 2017	https://github.com/COMBINE-lab/Salmon
Bioconductor	Huber et al., 2015	http://www.bioconductor.org/
R Project	R Core Team, 2018	https://www.r-project.org/; RRID: SCR_001905
Vast-tools	Tapial et al., 2017	https://github.com/vastgroup/vast-tools;
Mascot 2.5.1	Matrix Science	http://www.matrixscience.com/server.html; RRID: SCR_014322
Biopython	Cock et al., 2009	https://biopython.org; RRID: SCR_007173

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Gunter Meister (gunter.meister@vkl.uni-regensburg.de). All unique/stable reagents generated in this study are available from the Lead Contact with a completed Materials Transfer Agreement. Transfer of the B-LCLs underlies the additional approval from the Ethics Committee of the University of Würzburg.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Patient recruitment and clinical report

The nine year old daughter and the two year old son of healthy Arabian consanguineous parents were referred for genetic counseling because both children show psychomotor retardation. The mother reported that the girl was born after an uncomplicated pregnancy via Caesarean section in Yemen. Several neonatal problems occurred like hypotonia and impaired suction whereas deglutition seemed to be unaffected. At the age of three, the girl's family immigrated to Germany and a submucous cleft palate was diagnosed and was surgically corrected. At the age of seven the girl presented with short stature (107 cm, −3.13 standard deviation), low weight (16.5 kg, −2.72 standard deviation) and a head circumference of 49.5 cm. Moreover, she showed developmental and speech delay, scoliosis and hypertelorism. Her parents noticed episodes of rigidity and anxiety not due to external stimuli. An EEG revealed no pathological features. A cranial MRT has not been performed so far. A chromosome analysis as well as an array comparative genomic hybridization (CGH) analysis were performed and both exhibited no pathological findings.

The brother was delivered at 36+3 weeks of gestation and his birth weight was 1,740 g. He was seen at the age of 2 years and presented the following clinical features: hypertelorism, strabism, *dextroversio cordis*, *pes adductus et supinatus* and a leftsided multicystic dysplastic degeneration of the kidney. In addition, psychomotor retardation was evident. No cytogenetic analyses were performed on this patient.

Blood samples and cell lines were obtained with informed consent of the parents who allowed chromosome analyses, exome sequencing and Sanger sequencing procedures in the interests of the affected patients, who were minors. The study was approved by the Ethics Committee of the University of Würzburg.

Generation of B-lymphoblastoid cell lines

Lymphoblastoid cell lines (LCLs) were established from Ficoll-isolated blood lymphocytes via transformation by Epstein-Barr virus as described previously (Neitzel, 1986) and maintained in RPMI-1640 with L-glutamine (Sigma-Aldrich) supplemented with 15% fetal bovine serum (Sigma-Aldrich). The cell line AASD72 was established from the father (at the age of 46 years) and the Alazami-patient derived cell line AAAS14 from the son (at the age of 4 years). The genotypes and the resulting sequences for the affected splice site within the LARP7 mRNA were confirmed by PCR-amplification from genomic DNA and cDNA respectively using the primers indicated in Table S4. The PCR products were subcloned into the pGEM-T-Easy vector (Promega) according to the manufacturer's instructions and were subsequently analyzed by Sanger sequencing.

Generation of knockout cell lines

LARP7 knockout cell lines were generated from HEK293 and Flp-In T-REx 293 (Invitrogen) parental cell lines by CRISPR/Cas9-mediated genome editing. The HEK293 LARP7 −/− clones 1-11 and 3-6 were generated using two independent guide RNAs (contained in PX459 V2.0 + LARP7-1 and PX459 V2.0 + LARP7-2 respectively). The HEK293 T-REx Flp/In LARP7 −/− cell line was generated using the PX459 V2.0 + LARP7-1 construct. Two consecutive rounds of transfections with the indicated plasmids and puromycin selection were performed before singularizing cells in 96-well plates. The single clones were allowed to recover and were expanded into 6-well plates until confluency was reached. For each clonal line, half of the well was used to generate cryopreserved stocks, while to other half was used to screen for LARP7 knockout clones by western blotting.

For the generation of the HEK293 SNORD8 + SNORD9 double knockout cell line, single SNORD8 knockout cell lines were generated first using PX459 V2.0 + SNORD8 up + down. One of the positive clonal cell lines was later used to delete SNORD9 using PX459 V2.0 + SNORD9 up + down. Successful deletion of the snoRNAs was first screened by PCR amplification of the targeted genomic region (the primers used are listed in Table S4) and was subsequently confirmed by Northern blot analyses.

Generation of stable FH-LARP7 cell lines

Stable inducible cell lines were generated using the Flp-In T-REx 293 system (Invitrogen) according to the manufacturer's instructions. In short, either the Flp-In T-REx 293 cell line or the LARP7 knockout cell line derived from it, were seeded into 12-well plates and were then co-transfected with the plasmids pOG44 and pcDNA5-FRT/TO in a 9:1 ratio. The latter plasmid contained the FH-LARP7 construct intended to be integrated into the genome. After 48 h, cells were transferred to a cell culture dish (100 mm diameter) and selection of stable clones was achieved by addition of 200 μg/mL hygromycin B (GIBCO) to the medium in addition to 5 μg/mL blasticidin (GIBCO). Single colonies were picked approximately two weeks later. Following expansion, single clones were tested for the expression of the FH-LARP7 construct upon induction with tetracycline (1 μg/mL).

Other cell lines and cell culture conditions

The HEK293T cell line (female, embryonic kidney) and its derived LARP7 knockout cell lines (clones 1-11 and 3-6) were cultivated in Dulbecco's modified Eagle's medium (DMEM; Sigma). The Raji cell line (male, 11 years old, Epstein-Barr virus-related Burkitt lymphoma) was cultivated in RPMI-1640 medium (Sigma). Both media were supplemented with 10% fetal bovine serum (Sigma-Aldrich), and 100 U/mL penicillin, and 100 mg/mL streptomycin (Sigma). The DMEM medium used for the maintenance of the Flp-In T-REx 293 cell line (female, embryonic kidney) (Invitrogen) and its derived HEK293 T-REx Flp/In LARP7 knockout cell line was further supplemented with 100 μg/mL zeocin and 5 μg/mL blasticidin (GIBCO). The HEK293 T-REx FH-LARP7 cell line and the HEK293 T-REx Flp/In LARP7 knockout cell lines stably expressing the different FH-LARP7 variants were maintained DMEM supplemented with 5 μg/mL blasticidin (GIBCO) and 100 μg/mL hygromycin B (GIBCO). All cells were grown in high-humidity incubators at 5% (v/v) CO_2 and 37°C. Heat shocks were performed in 6-well plates sealed with parafilm and incubated in a pre-warmed water bath for 2 hours at 40°C.

METHOD DETAILS

Sequencing from human samples

Genomic DNA was extracted from peripheral blood leukocytes of both index patients and of the parents using standard protocols. Whole exome sequencing was performed only in the girl. Her genomic DNA was amplified with the Nextera Library Prep Kit (Illumina) and Nextera xGen Exome Research Panel (IDT). Sequencing was done on a NextSeq desktop sequencer (Illumina). Data were analyzed with GensearchNGS (PhenoSystems SA) and Alamut Visual (Interactive Biosoftware) by using a phenotype-based approach for variant filtering. A homozygous variant in the LARP7 gene (NM_016648.3) was identified: c.1669-1_1671del. This variant is reported in the gnomAD population databases (https://gnomad.broadinstitute.org) in five heterozygotes (MAF < 0.01%), but not in homozygous state. The same variant is deposited in dbSNP (https://www.ncbi.nlm.nih.gov/SNP/) under the accession rs778909076. Sanger sequencing of the brother and of the parents confirmed autosomal recessive inheritance and segregation of the identified variant.

RNA sequencing and analysis

For the generation of RNA-seq libraries the quality of the starting material was assayed on a TapeStation 4200 (Agilent Technologies) device. For none of the samples the RIN^e value was below 9.0. Two biological replicates were used per condition.

For the analyses shown in Figures 5A and 5C and Figure S6A, total RNA from HEK293 wild-type and LARP7 knockout (clone 1-11) cells was extracted using the TRIzol reagent (Invitrogen) and libraries were prepared with the Ovation SoLo RNA-seq system (NuGEN Technologies) according to the manufacturer's instructions. Libraries were sequenced at the Biomedical Sequencing Facility of the CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences (Vienna, Austria) using the Ovation SoLo Custom R1 primer included in the Ovation SoLo RNA-seq kit.

For the analyses shown in Figures S6C and S6D, RNA was extracted from heat shocked LARP7 knockout (clones 1-11 and 3-6) and WT HEK293 cells with the RNeasy Mini Kit (QIAGEN) including the shredding step via QIAshredder columns (QIAGEN) and the DNase digestion step.

RNA from the B-LCL cell lines AAAS14 and AASD72 was isolated by the same procedure. In these cases, libraries of polyA$^+$-selected RNA were prepared with the Universal Plus mRNA-Seq Kit (NuGEN Technologies) according to the manufacturer's instructions. Sequencing was performed according to standard procedures at the Biomedical Sequencing Facility of the CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences (Vienna, Austria).

RNA-Seq data was mapped to ENSEMBL human genome ver. hg38 with STAR aligner (Dobin et al., 2013). Genes and transcript expression levels were quantified from these pre-computed alignments with the salmon tool (Patro et al., 2017). Differential gene expression analysis was performed with edgeR Bioconductor package using a well-described workflow (Love et al., 2018). Genes with FDR < 0.05 were considered significantly differentially expressed. The vast-tools analysis workflow (Tapial et al., 2017) was used to assess alternative splicing events by computing the Percent Spliced In (PSI) scores, and their statistical significance was evaluated with the additional 'diff' module (Han et al., 2017). Events that passed a minimum read coverage (defined in https://github.com/vastgroup/vast-tools#combine-output-format) and for which the probability of a PSI difference of at least 0.05 (MV value) between conditions was 0.9 or higher were called as differentially spliced. To determine whether specific types of splicing events preferentially occurred in the data, we calculated p values for the enrichment from hypergeometric distributions, where the background set of events was composed of all events that passed the minimum quality score, irrespective of their statistical significance, and the foreground was the set of events that were considered significant (defined above).

Genes identified to be differentially spliced in the B-LCL RNA-seq data were further analyzed by searching the Online Mendelian Inheritance in Man (OMIM) database for associated diseases. Information regarding the clinical symptoms of the resulting diseases was obtained from the OMIM and the DisGeNET databases and was manually compared to the phenotypes reported for patients diagnosed for the Alazami syndrome.

Small RNA sequencing

Small RNA libraries were generated from RNAs co-purifying with endogenous LARP7 by immunoprecipitation as well as from total RNA isolated with TRIzol reagent (Invitrogen) from HEK293 cell lysate. Truncated T4 RNA Ligase 2 was used to ligate the RNAs to an adenylated 3′ adaptor [5′-App-TGGAATTCTCGGGTGCCAAGG-(C7-amino)-3′]. The ligation of the 5′ RNA adaptor (5′-GUUCAGA GUUCUACAGUCCGACGAUC-3′) was performed with T4 RNA Ligase 1 (New England Biolabs). The resulting ligation products were reverse-transcribed using the SuperScriptIII first strand synthesis super mix (Invitrogen), followed by a PCR amplification, wherein index sequences and other Illumina-specific sequences were added. The samples were resolved on a 6% polyacrylamide (acrylamid/bisacrylamid 19:1) urea gel (Carl Roth) and the bands corresponding to PCR amplification products containing inserts were cut out and eluted overnight in 300 mM NaCl and 2 mM EDTA. The supernatants containing the libraries were collected using Costar Spin-X filter tubes (Corning), precipitated with ethanol overnight at −20°C, pelleted and dissolved in water. Bioinformatic analysis of the sequenced reads was conducted similarly to the analysis of the ENCODE CLIP datasets.

Sequencing of fractionated RNAs of different length was achieved by resolving the sample on a polyacrylamide (acrylamid/bisacrylamid 19:1) urea gel (Carl Roth). RNAs were visualized under UV-light with the SYBR gold nucleic acid gel stain (Invitrogen). The RNA contained in excised gel material was eluted overnight at 4°C in 300 mM NaCl and 2 mM EDTA and was subsequently precipitated. The generation of libraries for sequencing occurred as described above, with the exception that the RNAs were first treated with the tobacco acid pyrophosphatase (Epicenter) to remove RNA cap structures and allow the ligation of the 5′ RNA adaptor.

Pulldown of RNAs for RiboMeth-Seq analysis

The purification of spliceosomal snRNAs for RiboMeth-Seq analysis occurred by pulldowns with complementary biotinylated oligonucleotides and magnetic streptavidin beads (Dynabeads M-270, Invitrogen). Therefore, total RNA was extracted with the TRIzol reagent (Invitrogen) and 40-75 μg RNA - in the case of RNA isolated from human blood sample 3-5 μg - were diluted in 10 mM Tris pH 7.5, 0.5 M LiCl, 1 mM EDTA. The complementary oligonucleotide mixture (Table S4) was added to the samples, which were then denatured at 95°C for 45 s before allowing annealing of the oligonucleotides to the target RNAs by incubation at 37°C for 20 min. Meanwhile, the magnetic streptavidin beads were washed with wash buffer (20 mM Tris pH 7.5, 1 M LiCl, 2 mM EDTA and 0.1% Tween-20) and were then resuspended in binding buffer (20 mM Tris pH 7.5, 1 M LiCl, 2 mM EDTA) in an equal volume as the annealing reaction. The samples were mixed with the resuspended magnetic beads and binding of the biotinylated oligonucleotides to the streptavidin beads was allowed to occur at 20°C under constant shaking. Following this incubation step, the beads were collected with a magnetic rack and were washed three times with wash buffer. The elution of the bound RNA was performed by incubation of the beads for 2 min at 80°C in pre-warmed buffer consisting of 10 mM Tris pH 7.5 and 1 mM EDTA. The whole pulldown procedure consisting of denaturation, annealing of biotinylated oligonucleotides and pulldown with fresh magnetic streptavidin beads was repeated for a second time with the eluate of the first pulldown. This resulted in higher purity of the purified target RNAs. Finally, TRIzol reagent (Invitrogen) was added to the last eluate and the RNA was extracted.

RiboMeth-sequencing and data analysis

RiboMeth-sequencing experiments were performed adapting the protocol described in Birkedal et al. (2015) to the small RNA sequencing protocol described above. In brief, RNA isolated from total RNA samples by antisense oligonucleotide pulldowns was fragmented by alkaline hydrolysis in 50 mM Na_2CO_3 pH 9.0 for 50-55 min at 90°C. RNA fragments were resolved on polyacrylamide (acrylamid/bisacrylamid 19:1) urea gels (Carl Roth) together with synthetic RNA size markers of 21 nt and 35 nt length loaded on different lanes. RNAs were visualized under UV-light by SYBR gold (Invitrogen) staining and gel slices containing RNA fragments in the range between 21 nt and 35 nt were excised. Upon elution and precipitation of the fragments, the RNA was treated with 0.5 U/mL calf intestinal alkaline phosphatase (New England Biolabs) for 1 h at 37°C. Dephosphorylated RNAs were purified by phenol/chloroform/isoamyl alcohol extraction followed by precipitation. Phosphorylation of the 5′ ends occurred with the T4 polynucleotide kinase (Thermo Scientific), followed by heat-inactivation of the enzyme, removal of unincorporated ATPs via Illustra MicroSpin G-25 gel filtration columns (GE Healthcare), phenol/chloroform/isoamyl alcohol extraction of the RNA fragments and precipitation. The generation of small RNA sequencing libraries from the products obtained by this strategy was conducted as described in a previous section with the exception that the fragments ligated to the adenylated 3′ adaptor were gel purified prior to ligation to the 5′ RNA adaptor.

For data analysis, adaptor trimming was performed using cutadapt with the parameters–overlap = 8–minimum-length = 15–discard-untrimmed. Mapping with Bowtie2 (Langmead and Salzberg, 2012) and further analysis was carried out according to Birkedal et al. (2015) using R (R Core Team, 2018, https://www.R-project.org) and the Bioconductor (Huber et al., 2015) packages ShortRead and ROCR. For plotting the Bioconductor package Gviz was used. A Matthews correlation analysis was carried out with known 2′-O-methylation sites of the U6 snRNA, giving Score A as the score with the highest area under curve (data not shown). Fold change was calculated using mean normalized counts according to the DESeq2 normalization.

Databases for annotation of sequencing data

The following annotation databases were used and were processed as described below. Biotypes were adopted from Ensembl hsa GRCh38.92 (Zerbino et al., 2018), RepeatMasker, tRNA and snaR annotations were retrieved from the UCSC table browser (Karolchik et al., 2003) hg38 (status February 2018).

RRNA were combined from Ensembl and RepeatMasker annotations and additional rRNA annotations were retrieved from NCBI GRCh38 (status July 2018).

SnoRNA annotations originated from snoRNA atlas (status July 2018) (Jorjani et al., 2016). The category "Pol III" was created by combining the following annotations: 5S-rRNA from Ensembl and RepeatMasker, tRNA from RepeatMasker and UCSC, U6, 7SL, 7SK and scRNA from Ensembl and RepeatMasker, vaultRNA, RMRP and RNaseP_nuc (RPPH1) from Ensembl, and SNAR from UCSC. Overlapping or multiple annotations were merged together. SnoRNA annotations from snoRNA atlas were converted from h19 to hg38 coordinates and were combined together with Ensembl snoRNAs. Overlapping and multiple annotations from these two databases were merged together.

Several subbiotypes from the Ensembl annotations were grouped into the following categories: "Protein coding," "Pseudogene," "Long noncoding" and "RepeatMasker." To reduce multiple counting the annotations for small RNAs (i.e., miRNAs, Pol III transcripts, snoRNAs and Pol I rRNAs) were cut out from the remaining annotations with an additional spacer of ten bases.

Analysis of ENCODE CLIP datasets

CLIP datasets generated by the group of Gene Yeo (UCSD, USA) were obtained from the ENCODE portal (Davis et al., 2018). The mapped files with the following identifiers were downloaded: ENCSR456KXI and ENCSR961OKA (LARP7 CLIP performed respectively in K562 and HepG2 cells) and ENCSR863ZGZ and ENCSR095SIV (corresponding mock control datasets).

Following steps were carried out with R (R Core Team, 2018, https://www.R-project.org) and the Bioconductor (Huber et al., 2015) packages ShortRead, Biostrings and rtracklayer. The databases described in the previous section were used for annotation. Overlaps between annotation and sequenced reads were counted on exon level with at least one base overlap. For plotting, the R packages ggplot2, cowplot and VennDiagram were used. Differential analysis was carried out with the Bioconductor package DESeq2 (significant targets: fold change $\geq$ 4, p-adjust $\leq$ 0.05). Heatmaps were plotted with the R package pheatmap.

Analysis of small RNA sequencing datasets

Adaptor trimming was performed using cutadapt with the parameters–overlap = 5–minimum-length = 16. Mapping was carried out using Bowtie2 (Langmead and Salzberg, 2012) with standard parameters. Annotation and counting was performed as described in the previous section. The counts were normalized according to DESeq2 normalization. Due to the lack of replicates, targets were called regulated with a fold change of normalized expression $\geq$ 4 and a normalized expression value of 10 in at least one condition.

Sequence analysis of U6-specific snoRNAs

The database Rfam 14.1 (Kalvari et al., 2018) was used to compile sequence sets for the detection of a conserved motif in U6-specific C/D box snoRNAs. The file *family.txt* was downloaded from the Rfam FTP-server and was filtered for the search term "CD-box." This resulted in the list L_{CD} consisting of the names of 475 C/D box snoRNA families. The list L_{U6} was assembled manually and contained the names of six U6-specific snoRNA families. The list $L_{CD \setminus U6}$ contained the names of all C/D box snoRNA families except the U6-specific snoRNAs. Based on L_{U6} and $L_{CD \setminus U6}$, the two sequence sets S_{U6} (comprising 118 sequences of U6-specific C/D box snoRNAs) and $S_{CD \setminus U6}$ (comprising 4,759 sequences of non-U6-specific C/D box snoRNAs) were compiled. The lists consisted of the seed alignment sequences representing the selected Rfam families. Scripts were programmed in Python using the Biopython package (Cock et al., 2009).

The detection of discriminative motifs was performed using the Discriminative Regular Expression Motif Elicitation (DREME) algorithm (Bailey, 2011). This tool detects short, ungapped motifs, which are relatively enriched in one sequence set, i.e., S_{U6}, compared to a second set of control sequences, i.e., $S_{CD \setminus U6}$. For each of the computed motifs, statistical significance is determined by means of a Fisher's exact test. The most significant motif of six nucleotides length resulting from this analysis is shown in Figure 3A.

Plasmids

Expression of FH-tagged protein was achieved from modified pIRES-VP5 (VP5) plasmids (Meister et al., 2004). The open reading frame (ORF) of human LARP7 was PCR-amplified from cDNA using oligonucleotides LARP7-NotI-F and LARP7-BamHI-R. The PCR product was digested with NotI and BamHI and was inserted into VP5 generating VP5 + FH-LARP7 WT. The VP5 constructs encoding the FH-LARP7 mutants F44A, D54A F56A, Y128A E130A and Y483A were created by subcloning LARP7 WT into the pGEM-T-Easy vector (Promega) using LARP7-NotI-F and LARP7-BamHI-R. Subsequently, mutagenesis PCRs were performed with the primer pairs LARP7-F44A-F/R, LARP7-D54A-F56A-F/R, LARP7-Y128A-E130A-F/R and LARP7-Y483A-F/R. A positive clone was used for PCR-amplifications of the mutated sequences using LARP7-NotI-F and LARP7-BamHI-R and subsequent cloning into VP5 as described above. The primer pairs LARP7-NotI-F and LARP7-dRRM2-BamHI-R, LARP7-NotI-F and LARP7-tr-561-BamHI-R and LARP7-dLAM-Not-F and LARP7-BamHI-R were used to amplify and clone the LARP7 truncations ΔRRM2, 2-561 and ΔLa module into VP5 via NotI and BamHI restriction digest. The VP5 + FH-La plasmid has been described in Hasler et al. (2016).

The pcDNA5/FRT/TO + FH-LARP7 WT/ F44A / Y483A / ΔRRM2 plasmids were generated by PCR-amplification of the LARP7 construct from the corresponding VP5 plasmids and ligation into a modified pcDNA5/FRT/TO vector already containing the N-terminal FH-tag and a modified multiple cloning site. The same primers and restriction sites were used as for their VP5 counterparts. For the generation of the pcDNA5/FRT/TO + FH-LARP7 Würzburg variant plasmid, the ORF of the LARP7 Würzburg variant was PCR-amplified from the cDNA obtained from the blood sample of an Alazami patient using the oligonucleotides LARP7-NotI-F and LARP7-patient-Bam-R.

The plasmid pnEA + 6xHis-LARP7 F44A used for the purification of the recombinant LARP7 F44A protein was generated by PCR amplification from VP5 + FH-LARP7 F44A with the primers pnEA-LARP7-NdeI-F and pnEA-LARP7-BamHI-R and cloning into the pnEA vector via NdeI and BamHI. The C-terminal domain of the LARP7 WT or of the LARP7 Würzburg variant were PCR-amplified using the primer pairs pnEA-LARP7-445-NdeI-F and LARP7-BamHI-R or pnEA-LARP7-445-NdeI-F and LARP7-patient-Bam-R, respectively. The obtained PCR products were cloned via NdeI and BamHI restriction digest into a modified pnEA version carrying a N-terminal 6xHis-tag. The pSUPER + U6 snRNA plasmid was obtained by amplification of the U6 snRNA from HEK293 genomic DNA with the primers U6-BglII-F and U6-HindIII-R. The resulting PCR product was then cloned into a modified pSUPER vector via BglII and HindIII restriction digest.

SNORD8 was expressed from a modified pcDNA3.1 (+) vector containing approximately 600 bp of the SNORD8-encompassing intronic region of the CHD8 host gene. This was amplified from HEK293 genomic DNA with the primer pair SNORD8-genotyping-F/R. The resulting PCR product was phosphorylated and was inserted by blunt end ligation into the pcDNA3.1 (+) vector linearized by EcoRV restriction digest. The pcDNA3.1 + SNORD8 motif and target mutants were obtained by mutagenesis PCR with the primer pairs D8-UAGGG-mut-F/R and D8-target-mut-F/R respectively. The plasmid pcDNA3.1 + SNORD8 motif+target mutant was obtained by mutagenesis PCR of pcDNA3.1 + SNORD8 target mutant with the primer pair D8-UAGGG-mut-F/R.

Genome editing was achieved with the CRISPR/Cas9 system. To this end, guide sequences against the LARP7 gene were designed and ligated into the pSpCas9(BB)-2A-Puro (PX459) V2.0 vector (Ran et al., 2013) created by the Zhang group. This occurred by annealing and phosphorylation of the complementary oligonucleotide pairs LARP7-S1/AS1 and LARP7-S2/AS2. The annealed products possessed single-stranded overhangs, which were compatible with the ends of the BbsI-digested PX459 vector. The guide sequences SNORD8-up-S/AS, SNORD8-down-S/AS, SNORD9-up-S/AS, SNORD9-down-S/AS were cloned similarly into PX459 V2.0. However, snoRNAs were deleted from the genome by directing the Cas9 endonuclease shortly upstream and downstream of the mature snoRNA sequence. To avoid co-transfection of distinct plasmids, the two independent single-guide RNAs targeting the intronic region of the snoRNA host gene were expressed from single plasmids. This was obtained as follows: the whole single-guide expression unit, including the U6 promoter, was amplified with XbaI-U6-Fw and U6-Rv-KpnI from PX459 V2.0 + SNORD8 up or from PX459 V2.0 + SNORD9 down. The resulting PCR products were inserted via XbaI and KpnI restriction into PX459 V2.0 + SNORD8 down and PX459 V2.0 + SNORD9 up, respectively. By that, PX459 V2.0 + SNORD8 up + down and PX459 V2.0 + SNORD9 up + down were obtained.

For the sequences of the oligonucleotides used for cloning, please refer to Table S4.

Transfections

For immunoprecipitations of overexpressed proteins, HEK293 cells were plated and transfected 3-4 h later with calcium phosphate. Therefore, 10 μg of plasmid DNA for each cell culture dish with a diameter of 15 cm were used. Cells were harvested after 48 h or 72 h. Transfections for the generation of knockout cell lines or stable HEK293 T-REx cell lines were performed with Lipofectamine 2000 (Invitrogen) in a 24-well or 12-well format according to the manufacturer's instructions.

Immunoprecipitations

Binding of rabbit antibodies to 40 μL nProtein A Sepharose bead slurry (GE Healthcare) occurred in 1 mL phosphate buffered saline (PBS) rotating overnight at 4°C. For anti-FBL immunoprecipitations 6-8 μg antibody (Bethyl Laboratories) for each coupling reaction and for anti-LSM4 immunoprecipitations 6 μL antibody (Antibody Genie) for each coupling reaction were used. In case of anti-LARP7 (SY7862) or anti-SART3 (1631) immunoprecipitations, beads were incubated with 20 μL serum each. FH-tagged proteins were precipitated with 30 μL of ANTI-FLAG M2 affinity gel (Sigma-Aldrich).

For each immunoprecipitation reaction, adherent cells were harvested from one to three cell culture dishes (150 mm diameter) and suspension cell lines were collected from three to four T75 flasks. Stable induction of FH-LARP7 WT, F44A, Y483A or ΔRRM2 expression occurred for 48 h by addition to the growth medium of 1 μg/mL doxycycline. The washed and pelleted cells were lysed on ice in 1.5 mL IP lysis buffer composed of 25 mM Tris pH 7.5, 150 mM KCl, 2 mM EDTA, 1 mM NaF, 0.5% (v/v) NP-40 alternative, 1 mM dithiothreitol (DTT) and 0.5 mM AEBSF. Clarified lysates were obtained by full-speed centrifugation at 4°C and the protein concentration was determined by a Bradford assay to adjust the volumes of cell lysates used in the downstream applications. Aliquots were taken, which served as input samples for subsequent western blot and Northern blot analyses. The remaining lysate was transferred to a fresh tube containing the antibody-coupled beads. Control reactions were performed by incubating lysates with beads only. The mixtures were incubated under constant rotation for 2-3 h at 4°C. The beads were then transferred to a fresh reaction tube and were washed four to five times with ice-cold wash buffer [50 mM Tris pH 7.5, 350 mM KCl, 1 mM $MgCl_2$, 0.5% (v/v) NP-40 alternative], and once with ice-cold PBS. The beads were resuspended in 100 μL PBS and a 20 μL aliquot was taken for western blot analysis. The co-precipitated RNAs were purified from the remaining beads by performing a digestion with proteinase K (Thermo Scientific) and a phenol/chloroform extraction. Total RNA was isolated directly from an aliquot of the lysates using the TRIzol reagent (Invitrogen).

The immunoprecipitation of FH-LARP7 used for MS analyses was performed as follows: cells were grown to 90% confluency and were harvested following tetracycline (1 μg/mL) treatment for 16 h. The cells were lysed on ice for 10 min in buffer containing 50 mM HEPES pH 7.5, 150 mM NaCl, 2.5 mM $MgCl_2$, 1% NP-40, protease inhibitors and RNase inhibitors. The lysate was passed 6 times through a 26G needle followed by water bath sonication. After sonication, the lysates were centrifuged for 20 min at 14,000 rpm and 4°C. Equal amounts of total protein lysate from control HEK293 cells and FH-LARP7 overexpressing cells were taken for immunoprecipitations. The pre-equilibrated ANTI-FLAG M2 affinity gel (Sigma-Aldrich) was incubated with the lysates for 3 h on a head-over-tail rotor at 4°C. After incubation, beads were collected and were washed three times with washing buffer containing 50 mM HEPES pH 7.5, 300 mM NaCl and 2.5 mM $MgCl_2$, followed by the last wash using 1 x PBS. FH-LARP7 and the interacting proteins were eluted using 200 μg/mL 3x FLAG peptide (Sigma-Aldrich) in PBS.

Synthesis of cDNA and quantitative PCR

cDNA synthesis was carried out using the first strand cDNA synthesis kit (Thermo Scientific), 1 μg of total RNA and random hexamer primers (for RT-PCRs used for the validation of alternative splicing) or oligo(dT)$_{18}$ primers (for other applications). For qPCRs, cDNA was diluted 1:10 and thereof 2 μL per sample were mixed with 10 pmol each of forward and reverse primer (Table S4) and 10 μL of SsoFast EvaGreen supermix (Bio-Rad) in a total volume of 20 μL. Measurements were performed on a CFX96 Real-Time System (Bio-Rad). The error bars display ± standard deviations of the normalized signals from three technical replicates.

Protein expression and purification

For immunizations, N-terminally 6xHis-tagged human LARP7 was expressed in bacterial cells from the pET28b vector described in Markert et al. (2008). Induction occurred with 0.5 mM IPTG and cells were incubated at 30°C overnight. Bacteria were then harvested, resuspended in buffer containing 50 mM HEPES pH 7.5, 20 mM NaCl, 25 mM imidazole and 5 mM 2-mercaptoethanol and cocktail of protease inhibitors. The resuspended cells were sonicated and centrifuged at 45,000 rpm at 4°C. The clarified lysate was incubated over pre-equilibrated Ni-NTA resin. The incubated beads were collected, washed and the bound protein was eluted in buffer containing 50 mM HEPES pH 7.5, 200 mM NaCl, 250 mM imidazole and 5 mM 2-mercaptoethanol. The eluted 6xHis-LARPP7 protein was dialyzed in PBS.

The expression of recombinant LARP7 wild-type, F44A mutant and LARP7 1-208 proteins as well as the C-terminal domain constructs used for EMSA experiments occurred in *E. coli* Rosetta (DE3) from the pnEA vector, which is derived from pET15. Full-length proteins had a 6xHis-tag at their C terminus, while the LARP7 1-208 and the LARP7 C-terminal domain constructs were tagged at the N terminus. The purifications were performed as described in Uchikawa et al. (2015). In short, upon induction with 1 mM IPTG (for 4 hours at 28°C for the domains or overnight at 20°C for the full-length versions), cells were lysed by sonication in buffer containing 50 mM Tris pH 7.6, 500 mM NaCl, 5 mM $MgCl_2$ and 1 mM DTT. The recombinant protein variants were purified by nickel affinity and were eluted with lysis buffer containing 300 mM imidazole. Removal of the tag by TEV protease occurred during dialysis in 20 mM HEPES pH 7.2, 500 mM NaCl and 1 mM DTT. In the case of the LARP7 truncations, the excised tag was removed by a second nickel affinity purification. The proteins were then subjected to chromatography on heparin Hiload to reach a high degree of homogeneity. Fractions containing LARP7 variants were pooled, and concentrated (in the 300 μM range), supplemented with 10% glycerol and flash-frozen in liquid nitrogen and stored at −80°C. For the full-length versions, proteins were directly loaded on the heparin column after cleavage with the TEV protease without the second nickel affinity purification. The collected fractions were concentrated (in the 50 μM range) and dialyzed overnight in 20 mM HEPES pH 7.2, 500 mM NaCl 2 mM DTT and 10% glycerol before flash-freezing.

Generation of polyclonal antibodies

Immunization of two rabbits with N-terminally 6xHis-tagged human LARP7 protein was performed by Eurogentec. The anti-LARP7 serum SY7862 was used for immunoprecipitations and the serum SY7863 was used for the detection of LARP7 by western blotting.

In vitro transcription of RNA

The templates for the *in vitro* transcription of the U6 snRNA or of the SNORD8 snoRNA were PCR amplified respectively from pSUPER + U6 snRNA with the primer pair T7-U6-F and U6-ivt-R and from pcDNA3.1 + SNORD8 WT with the primer pairs T7-SNORD8-F and SNORD8-ivt-R (for sequences refer to Table S4). By that, the T7 promoter sequence was added upstream of the sequences intended to be transcribed. *In vitro* transcriptions were carried out using 0.1 mg/mL T7 RNA polymerase in 30 mM Tris pH 8.0, 25 mM MgCl2, 10 mM each NTP, 2 mM spermidine, 1 mM DTT, 0.01% Triton X-100 and 2 U/mL thermostable inorganic pyrophosphatase (New England Biolabs) for 4 h at 37°C. The *in vitro* transcribed RNA was purified on a 6% polyacrylamide (acrylamid/bisacrylamid 19:1) urea gel (Carl Roth), eluted in water and precipitated.

^{32}P-labeling of oligonucleotides

DNA oligonucleotides used as probes for Northern blot assays (Table S4) were labeled by incubating 20 pmol of oligonucleotides with 20 μCi of γ-^{32}P-ATP (Hartmann Analytic) and 0.5 U/μl T4 Polynucleotide Kinase (PNK) in 1x PNK buffer A (Thermo Scientific) at 37°C for 30-60 min. The reaction was stopped by adding EDTA, pH 8.0, and the labeled oligonucleotides were purified with a G-25 column (GE Healthcare).

Labeling of 50 pmol primers used for radioactive PCRs was performed accordingly, except that the T4 PNK was heat-inactivated prior loading of the reaction onto the G-25 column. The flow-through was then precipitated and the labeled oligonucleotide was dissolved in water.

For EMSA experiments, *in vitro* transcribed U6 or SNORD8 RNAs were dephosphorylated prior ^{32}P-labeling by incubating 30 pmol RNA with 0.1 U/mL FastAP (Thermo Scientific) in 1x PNK buffer A supplemented with 2 U/μL RiboLock RNase inhibitor (Thermo Scientific). The reactions were carried out for 30 min at 37°C and the enzyme was heat-inactivated for 20 min at 75°C. The ^{32}P-labeling reactions were performed as described above with the exception that the T4 PNK was heat-inactivated for 10 min at 75°C without addition of EDTA prior gel filtration.

Northern blot

Northern blots were carried out with 10-20 μg of total RNA or RNA isolated from immunoprecipitations. RNAs were separated in 1x TBE on 6% polyacrylamide (acrylamid/bisacrylamid 19:1) urea gels (Carl Roth). After electrophoresis, the RNA was stained with ethidium bromide to ensure equal loading of the lanes and to determine the RNA quality. The RNA was then blotted for 45 min at 20 V onto an Amersham Hybond-N membrane (GE Healthcare) and crosslinked to the membrane for 1 h at 50°C using an EDC solution. An additional UV-crosslinking (254 nm, 120 mJoules/cm^2) was performed in a UV Stratalinker (Stratagene). The membrane was incubated overnight at 50°C in hybridization solution (5x SSC, 7% SDS, 20 mM sodium phosphate buffer pH 7.2, 1x Denhardt's solution) with a ^{32}P-labeled oligonucleotide antisense to the RNA to detect (Table S4). The membrane was washed twice with 5x SSC, 1% SDS, once with 1x SSC, 1% SDS before being wrapped in saran and exposed to a storage phosphor screen. Before re-probing a membrane, the hybridized oligonucleotides were removed by incubating the membrane twice with a boiling 0.1% SDS solution for at least 10 min.

Electromobility shift assay

EMSA experiments were performed according to Uchikawa et al. (2015). In short, 500 pM ^{32}P-labeled RNA were incubated with various amounts of recombinant proteins in 25 mM HEPES pH 7.2, 250 mM NaCl, 5 mM $MgCl_2$, 2 mM DTT, 0.05 mg/mL bovine serum albumin, 0.005% NP-40 alternative, 10% glycerol and 5 μM yeast tRNAs. Complexes were allowed to form for 15 min on ice and were then resolved on 6% native polyacrylamide gels (acrylamid/bisacrylamid 37.5:1) containing 5% glycerol and 0.5x Tris borate (TB) buffer (45 mM Tris [pH 8.0] and 45 mM borate). Electrophoresis was carried out for 2.5 h at 4°C and 230 V in 0.5x TB buffer. The gels were dried prior exposure for signal detection.

Radioactive RT-PCR

The RNA used for radioactive RT-PCRs was extracted with the Nucleospin RNA kit (Macherey-Nagel) and an additional digestion with DNase I (Thermo Scientific) was performed prior cDNA synthesis.

First, conventional RT-PCRs with the primers indicated in Table S4 were performed, either using the *Thermus aquaticus* DNA polymerase or the Phusion DNA polymerase (Thermo Scientific) to test the amplification condition and verify the presence of the expected products.

Radioactive RT-PCRs were performed with one of the two primers end-labeled with ^{32}P as described in a previous section. The reactions were carried out in a volume of 12.5 µL and contained approximately 125 pmol primers and 0.5 µL cDNA. The splice patterns of PARP6 and SETMAR were analyzed using the *Taq* DNA polymerase for 20 and 25 amplification cycles respectively. For KMT2D (MLL2), the Phusion DNA polymerase (Thermo Scientific) was used for 22 amplification cycles. The samples were then diluted with an equal volume of deionized formamide containing bromophenol blue and xylencyanol, denatured for 3 min at 95°C and loaded on 5%–6% polyacrylamide (acrylamid/bisacrylamid 19:1) urea gels (Carl Roth). Electrophoresis was performed in 1x TBE until the bromophenol blue dye migrated approximately 10-18 cm into the gels, which were dried prior exposure. For the rescue experiments shown in Figure 7D, all cell lines were grown for six days under the presence of 1 µg/mL doxycycline, which was added freshly every second day.

Detection of radioactive signals

The detection of radioactive signals from Northern blot, radioactive PCR and EMSA assays occurred with storage phosphor screens which were scanned with the Personal Molecular Imager (Bio-Rad) upon exposure. If indicated, signal intensities were quantified from three biological replicates using Quantity One Software (version 4.6.9, Bio-Rad). Error bars display ± standard deviations of the normalized signals.

In vitro pull down of biotinylated snoRNA-U6 snRNA-LARP7 complexes

For pull down experiments, *in vitro* transcribed SNORD8 RNA was biotinylated at its 3′ end using pCp-biotin (Jena Bioscience) and the T4 RNA Ligase 1 (New England Biolabs) according to the manufacturer’s instructions. The enzyme was heat-inactivated and unincorporated pCp-biotin was removed via G-25 gel filtration columns (GE Healthcare). The biotinylated SNORD8 RNA was thermally refolded and diluted 1:2 with binding buffer (25 mM HEPES pH 7.2, 150 mM NaCl, 5 mM $MgCl_2$, 2 mM DTT, 0.05 mg/mL bovine serum albumin and 0.005% NP-40 alternative) prior immobilization on magnetic streptavidin beads (Dynabeads M-270, Invitrogen). Following 20 min incubation at 20°C under constant shaking (850 rpm), the beads with the bound SNORD8 RNA were collected with a magnetic rack and were subsequently washed three times with binding buffer.

Approximately 0.3 µM of the immobilized SNORD8 RNA were incubated with 0.3 µM *in vitro* transcribed and thermally refolded U6 RNA in the presence or absence of 0.1 µM recombinant LARP7 protein variants. For the experiment shown in the upper panel of Figure 2F the assembly reaction occurred in a total volume of 50 µl, for the experiment shown in the lower panel of Figure 2F the volume was increased to 100 µl maintaining the same RNA and protein concentrations. Of note, the critical concentration of recombinant LARP7 protein required to discriminate between the characteristics of the different LARP7 variants, was carefully determined in preceding experiments.

Following assembly of the complexes for 20 min on ice, the beads were collected with a magnetic rack and were washed three times with wash buffer (25 mM HEPES pH 7.2, 1 M NaCl, 5 mM $MgCl_2$, 2 mM DTT, 0.05 mg/mL bovine serum albumin and 0.005% NP-40 alternative). For Northern blot analyses, the beads were resuspended after the last washing step in 15 µl PBS, incubated for 2 min at 95°C and the RNA was finally extracted using TRIzol reagent (Invitrogen). Samples for western blotting were resuspended directly in 1x Laemmli buffer following the last washing step.

Spliceosome assembly/*in vitro* splicing assay

The nuclear extracts from HEK293 and LARP7 knockout cells for the spliceosome assembly and *in vitro* splicing assays were prepared as described by Dignam et al. (1983). Briefly, the trypsinized cells were collected after washing twice with PBS. The cells were resuspended in low-salt buffer containing 10 mM HEPES pH 7.9, 10 mM KCl, 1.5 mM $MgCl_2$, cocktail of protease inhibitors and RNase inhibitor. The cells were allowed to swell on ice for 10 min. Then, the cell suspension was passed through the type B pestle douncer for 12 times. The nuclei were pelleted by centrifugation at 9,500 rpm for 5 min at 4°C. The nuclei were again washed with low salt buffer to remove any cytoplasmic debris. The nuclei were then resuspended in high-salt buffer containing 20 mM HEPES pH 7.9, 420 mM NaCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 25% glycerol, cocktail of protease inhibitors and RNase inhibitor. The nuclei were homogenized by passing through a douncer for 20 times and stirring on ice for 40 min. The lysate was centrifuged at 12,300 rpm for 30 min at 4°C to pellet the debris. The supernatant was dialyzed twice in 20 volumes of buffer containing 20 mM HEPES pH 7.9, 100 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 0.5 mM DTT, 10% glycerol. Following dialysis, the nuclear lysate was centrifuged at 7,200 rpm for 2 min at 4°C before flash freezing in liquid nitrogen and storing at −80°C.

Splicing assays were performed in a reaction volume of 25 μL containing 60% of nuclear extract, 8 mM creatine phosphate, 1.6 mM ATP, 60 mM KCl, 12 mM HEPES pH 7.9, 0.12 mM EDTA, 0.3 mM DTT and 12% glycerol. The samples were incubated with 20 counts of ^{32}P labeled *in vitro* transcribed MINX pre-mRNA (kind gift from Dr. Elmar Wolf) at 30°C for 0, 5, 15, 30, 60, 90 min and the reaction was stopped by addition of heparin. Half of the reaction volume was used to analyze the spliceosomal assembly in 2% agarose gels. For checking the efficiency of splicing, the other half of the reaction volume was treated with Proteinase K and the RNA was isolated using TRIzol reagent (Invitrogen). The RNA was resolved on 15% polyacrylamide (acrylamid/bisacrylamid 19:1) urea gels.

Fractionation of nuclear extracts by sucrose gradient centrifugation

Six cell culture dishes (150 mm diameter) each of HEK293 and LARP7 knockout cell lines were harvested to prepare nuclear extracts according to Dignam et al. (1983) but with some modifications to the protocol described above. Pelleted cells were resuspended in five packed cell pellet volumes of low-salt buffer (10 mM HEPES pH 7.9, 10 mM KCl, 1.5 mM $MgCl_2$, 1 mM DTT and 0.5 mM AEBSF). Following swelling for 15 min on ice, cells were collected by low-speed centrifugation and were resuspended in low-salt buffer (two packed cell pellet volumes). Cells were lysed in a douncer applying several strokes with the type A pestle until 80%–90% of the cells resulted to be lysed as determined microscopically upon staining with trypan blue. The homogenate was centrifuged at 3,000 rpm for 10 min at 4°C and the cytoplasmic extract was discarded. The pellet containing the intact nuclei was washed 3 to 4 times with two packed cell pellet volumes of low-salt buffer and was then resuspended in 2/3 volume of the original packed cell pellet high-salt buffer (20 mM HEPES pH 7.9, 420 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 5% glycerol, 1 mM DTT and 0.5 mM AEBSF). The suspension was transferred to a douncer and the nuclei were homogenized by several strokes with a type B pestle. Nuclear extracts were finally obtained by 20,000xg centrifugation for 30 min at 4°C.

The fractionation of the nuclear extracts was performed on 15% (w/v) to 55% (w/v) sucrose gradients prepared in 25 mM Tris pH 7.5, 150 mM KCl, 2 mM EDTA, 1 mM DTT and 0.5 mM AEBSF. RNPs were resolved according to their sedimentation rate by centrifuging the gradients in a SW 40 Ti rotor at 30,000 rpm for 18 h at 4°C. Subsequently, 500 μl fractions were collected manually and half of the volume of each fraction was used for RNA extraction with the TRIzol LS reagent (Invitrogen).

SDS-PAGE and western blotting

Western blot samples were mixed with 4x Laemmli buffer and incubated at 95°C for 5 min. Proteins were separated by SDS-PAGE on 15% gels, for FBL and LSM4 western blots, as well as for the Coomassie staining of the recombinant LARP7 C-terminal domain constructs. For western blotting of the RNA Pol II subunit B1 (POLR2A), protein samples were either resolved on a NuPAGE 4%–12% Bis-Tris protein gel (Invitrogen) using a MOPS buffer system (Figure 5B, left panels and Figure S7D) or by conventional SDS-PAGE on a 6% gel (Figure 5B, right panels). For the detection of all other proteins 10% SDS-gels were used. The primary antibodies used for immunodetection were diluted as follows: rabbit polyclonal anti-LARP7 (serum SY7863) 1:500, rabbit polyclonal anti-FBL (Bethyl Laboratories) 1:1,000, rabbit polyclonal anti-LSM4 (Antibody Genie) 1:1,000, rabbit polyclonal anti-CDK9 (Bethyl Laboratories) 1:1,000, rat monoclonal anti-RNA polymerase II subunit B1 (phospho CTD Ser-2) (clone 3E10, Merck Millipore) 1:1000, rabbit polyclonal anti-SART3 (serum 1631) 1:500, mouse monoclonal anti-beta actin (clone AC-15, GeneTex) 1:10,000, mouse monoclonal anti-HA (clone 16B12, Covance Research Products) 1:2,000. Secondary antibodies (goat polyclonal anti-Rabbit IgG IRDye 800CW conjugated antibody, goat polyclonal anti-Mouse IgG IRDye 800CW conjugated antibody, goat polyclonal anti-Rat IgG IRDye 800CW conjugated antibody and goat anti-Mouse IgG IRDye 680RD conjugated antibody) were obtained from LI-COR Biosciences and were diluted 1:15,000. Signals were detected with the Odyssey Infrared Imaging System (LI-COR Biosciences).

The specificity of the anti-RNA polymerase II subunit B1 (phospho CTD Ser-2) (clone 3E10, Merck Millipore) antibody was tested by treating the lysate of HEK293 cells with the lambda protein phosphatase (New England Biolabs) according to the manufacturer's instructions. For this experiment, cells were lysed in PBS supplemented with 0.5% (v/v) NP-40 alternative, 1 mM DTT and 0.5 mM AEBSF.

Coomassie staining of SDS-gels was performed according to standard procedures and gels were scanned with the Odyssey Infrared Imaging System (LI-COR Biosciences).

Silver stainings were performed with samples separated on a NuPAGE 4%–12% Bis-Tris protein gels (Invitrogen) using MES buffer at 200V. The gel was briefly rinsed with water and was fixed for 1 h at room temperature in 50% methanol and 12% acetic acid containing formaldehyde. After incubation, the gel was washed with 50% ethanol for three times for 20 min each. To sensitize the gel, a 0.02% sodium thiosulfate solution was used. Immediately, the gel was washed with water and was incubated in 0.2% silver nitrate solution containing formaldehyde. After rinsing the gel with water, it was developed with a 6% anhydrous sodium carbonate solution containing formaldehyde. The reaction was stopped using 2.5% acetic acid.

Mass spectrometric analysis

Proteins were separated on a 4%–12% NUPAGE Bis-Tris gel (Invitrogen) using a MOPS buffer system. The gel was stained with Simply Blue colloidal Coomassie blue G250 (Invitrogen). For mass spectrometric (MS) analysis of the proteins a gel lane was cut into consecutive slices. The gel slices were then transferred into 2 mL tubes and washed with 50 mM NH_4HCO_3, 50 mM NH_4HCO_3/acetonitrile (3:1) and 50 mM NH_4HCO_3/ acetonitrile (1:1) while shaking gently in an orbital shaker. Gel pieces were lyophilized after shrinking by 100% acetonitrile. To block cysteines, reduction with DTT was carried out for 30 min at 57°C followed by an alkylation step with iodoacetamide for 30 min at room temperature in the dark. Subsequently, gel slices were washed and lyophilized again as described above. Proteins were subjected to in gel tryptic digest overnight at 37°C with approximately 2 μg trypsin per 100 μL gel volume (Trypsin Gold, mass spectrometry grade, Promega). Peptides were eluted twice with 100 mM NH_4HCO_3 followed by an additional extraction with 50 mM NH_4HCO_3 in 50% acetonitrile. Finally, the combined eluates were lyophilized. Prior to LC-MS/MS analysis, lyophilized peptides were reconstituted in 20 μL of 1% formic acid. Separation of peptides by reversed-phase chromatography was carried out on an UltiMate 3000 RSLCnano System (Thermo Scientific), which was equipped with a C18 Acclaim Pepmap100 preconcentration column (100 μm i.D.x 20 mm, Thermo Fisher) in front of an Acclaim Pepmap100 C18 nano column (75 μm i.d. × 250 mm, Thermo Fisher). A linear gradient of 4% to 40% acetonitrile in 0.1% formic acid over 90 min was used to separate peptides at a flow rate of 300 nl/ min. The LC-system was coupled on-line to a maXis plus UHR-QTOF System (Bruker Daltonics) via a CaptiveSpray nanoflow electrospray source (Bruker Daltonics). Data-dependent acquisition of MS/MS spectra by CID fragmentation was performed at a resolution of minimum 60000 for MS and MS/MS scans. The MS spectra rate of the precursor scan was 2 Hz processing a mass range between m/z 175 and m/z 2000. Via the Compass 1.7 acquisition and processing software (Bruker Daltonics) a dynamic method with a fixed cycle time of 3 s and an m/z dependent collision energy adjustment between 34 and 55 eV was applied. Raw data processing was performed in Data Analysis 4.2 (Bruker Daltonics), and Protein Scape 3.1.3 (Bruker Daltonics) in connection with Mascot 2.5.1 (Matrix Science) facilitated database searching of the UniProtKB/Swiss-Prot *Homo sapiens* database (Breuza et al., 2016) (release-2017_09, 20238 entries). Search parameters were as follows: enzyme specificity trypsin with 2 missed cleavage allowed, precursor tolerance 0.02 Da, MS/MS tolerance 0.04 Da, carbamidomethylation or propionamide modification of cysteine, oxidation of methionine, deamidation of asparagine and glutamine were set as variable modifications. Protein list compilation was done using the Protein Extractor function of Protein Scape.

GO term enrichment analysis

For the GO term enrichment analysis MS data were further processed as follows: keratin and immunoglobulin entries were manually removed, as well as overlapping entries with a ratio of the Mascot peptide ion-score between the FH-LARP7 and control IP datasets lower than two. Hits with a Mascot peptide ion-score below 100 as well as proteins detected with only one peptide were removed. The resulting list was cleaned for known contaminants of anti-FLAG immunoprecipitations performed from HEK293 cell lysates using the CRAPome (Mellacheruvu et al., 2013) web tool. Entries with a FC_A scores $\geq$ 1.0 were maintained. The resulting list (Table S2) was used as input for the GO term enrichment analysis which was performed with the AmiGO2 web tool using the PANTHER overrepresentation test (released 2018-11-13), GO ontology database (released 2019-01-01) and Bonferroni correction for multiple testing (Mi et al., 2017). The GO terms for the category molecular function containing the proteins depicted in Figure S1D were selected from the top hits represented in the hierarchical output view. These terms were enriched more than ten-fold with a P value < 0.05. The hierarchy served as well to group related GO terms in the gray clouds minimizing redundancies in the proteins shown. FBL was detected in our MS analysis and is shown in gray within the GO term "snoRNA binding." This assignment occurred manually, since FBL is the catalytic subunit of C/D box snoRNPs but is not present in the GO term "snoRNA binding."

QUANTIFICATION AND STATISTICAL ANALYSIS

Please refer to the Figure Legends or the Experimental Model and Subject Details for description of sample size and statistical analyses.

DATA AND CODE AVAILABILITY

The raw and analyzed sequencing data have been deposited in NCBI's Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/gds) under the ID codes GEO: GSE126911.

The Alazami-related LARP7 variant has been deposited in the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/) under ID code: RCV000678485.1.

Appendix D: CFIm-mediated alternative polyadenylation remodels cellular signaling and miRNA biogenesis (Published)

3096–3114 *Nucleic Acids Research, 2022, Vol. 50, No. 6* *Published online 2 March 2022*
https://doi.org/10.1093/nar/gkac114

CFIm-mediated alternative polyadenylation remodels cellular signaling and miRNA biogenesis

Souvik Ghosh [1], Meric Ataman [1,2], Maciej Bak [1,2], Anastasiya Börsch [1,2], Alexander Schmidt [3], Katarzyna Buczak [3], Georges Martin [1], Beatrice Dimitriades [1], Christina J. Herrmann [1,2], Alexander Kanitz [1,2] and Mihaela Zavolan [1,2,*]

[1] Computational and Systems Biology, Biozentrum, University of Basel, Spitalstrasse 41, 4056 Basel, Switzerland, [2] Swiss Institute of Bioinformatics, Biozentrum, University of Basel, Spitalstrasse 41, 4056 Basel, Switzerland and [3] Proteomics Core Facility, Biozentrum, University of Basel, Spitalstrasse 41, 4056 Basel, Switzerland

Received August 05, 2021; Revised January 31, 2022; Editorial Decision February 02, 2022; Accepted February 04, 2022

ABSTRACT

The mammalian cleavage factor I (CFIm) has been implicated in alternative polyadenylation (APA) in a broad range of contexts, from cancers to learning deficits and parasite infections. To determine how the CFIm expression levels are translated into these diverse phenotypes, we carried out a multi-omics analysis of cell lines in which the CFIm25 (NUDT21) or CFIm68 (CPSF6) subunits were either repressed by siRNA-mediated knockdown or over-expressed from stably integrated constructs. We established that >800 genes undergo coherent APA in response to changes in CFIm levels, and they cluster in distinct functional classes related to protein metabolism. The activity of the ERK pathway traces the CFIm concentration, and explains some of the fluctuations in cell growth and metabolism that are observed upon CFIm perturbations. Furthermore, multiple transcripts encoding proteins from the miRNA pathway are targets of CFIm-dependent APA. This leads to an increased biogenesis and repressive activity of miRNAs at the same time as some 3′ UTRs become shorter and presumably less sensitive to miRNA-mediated repression. Our study provides a first systematic assessment of a core set of APA targets that respond coherently to changes in CFIm protein subunit levels (CFIm25/CFIm68). We describe the elicited signaling pathways downstream of CFIm, which improve our understanding of the key role of CFIm in integrating RNA processing with other cellular activities.

INTRODUCTION

Most human genes have multiple sites where pre-mRNA 3′ end processing can occur to generate alternative transcript isoforms in different cell types and conditions (1). Alternative polyadenylation is a main contributor to the observed transcriptome diversity (2–5). Consistently, data from The Cancer Genome Atlas (TCGA) indicate that APA holds the highest prognostic value among all types of isoform variation in hepatocellular carcinoma (6).

A large class of APA isoforms are those that differ in the length of their 3′ untranslated regions (3′ UTRs). Mammalian cleavage factor I, a 3′ end processing complex conserved in multicellular organisms but absent from yeast (7), is one of the main regulators of 3′ UTR length (8–10). A CFIm tetramer composed of two 25 kDa subunits (CFIm25/CPSF5/NUDT21) and two larger subunits of 59 and/or 68 kDa (CFIm59/CPSF7 and CFIm68/CPSF6) associates with the RNA polymerase II (RNAPII) in the initial stages of transcription (11). Crosslinking and immunoprecipitation revealed that within individual genes, the most prominent peaks of CFIm binding are located in the vicinity of those poly(A) sites (PAS) that are ultimately used for the maturation of the messenger RNA (mRNA), indicating that the interaction of CFIm with high-affinity target sites promotes the 3′ end cleavage (9). These interaction sites are typically located distally in 3′ UTRs, in regions enriched in UGUA motifs (9,12). The Fip1 3′ end processing factor stabilizes the interactions of CFIm with the RNA (12), while the ubiquitination of the PCF11 component of the 3′ end processing complex by an ectopically activated MAGE-A11 ubiquitin ligase in cancer leads to the dissociation of CFIm25 from the complex (13).

The depletion of CFIm25 or CFIm68 subunits of CFIm leads to systematically shortened 3′ UTRs (9,10), whereas CFIm59 does not seem to impact the 3′ UTR length (9,14). The number of reported targets varies between tens to over a thousand among studies (9,10,15,16). The phenotypes observed upon perturbation of CFIm expression have been attributed to growth regulators in glioblastoma (15), chromatin-regulatory factors in somatic cell reprogramming (16), and metabolic enzymes in the activation

[*] To whom correspondence should be addressed. Tel: +41 61 207 1577; Email: mihaela.zavolan@unibas.ch

of hematopoietic stem cells (17). While CFIm has emerged as an important regulator of cell fate decisions in normal and pathological contexts, its involvement in cancers is not fully understood (13,18). Reduced CFIm expression was reported to increase cell proliferation and promote glioblastoma and hepatocellular carcinoma formation (15,19,20), while opposite effects, reduced proliferation and increased apoptosis, have been reported in K562 leukemia cells (21).

To better understand how cells respond to the pervasive remodeling of the RNA pool that follows fluctuations in CFIm levels, we carried out a multi-omics characterization of HEK293 cells in which the CFIm25 and CFIm68 components of the CFIm complex were either stably overexpressed, or transiently repressed via siRNA-mediated knockdown. We demonstrate that hundreds of transcripts undergo reciprocal changes in 3′ UTR length in the knockdown (KD) and overexpression (OE) conditions, changes that are also very consistent between perturbations of CFIm25 and CFIm68. These targets cluster in specific cellular pathways, including stress signaling, cell cycle, RNA processing and miRNA-mediated repression. Many kinase-encoding transcripts undergo CFIm-dependent APA, which led us to globally estimate the changes in kinase activities upon CFIm perturbations from phosphoproteome measurements. Among the kinases whose activity changes we validate here, the stress-related ERK directly traces the CFIm expression, leading to expected changes in cell metabolism and growth. Specifically, real time growth estimates revealed that the ectopic expression of CFIm subunits promotes cellular proliferation, while the siRNA-mediated knockdown reduces growth in multiple cell lines. By regulating the processing of transcripts encoding miRNA pathway proteins, CFIm also modulates the miRNA activity. Interestingly, while the CFIm knockdown results in transcripts with shortened 3′ UTRs that can escape miRNA-dependent regulation, it also activates the biogenesis of miRNAs and their repressive activity on reporter constructs. Our study thus identifies key signaling pathways downstream of CFIm, explaining the impact of this 3′ end processing factor on fundamental cellular processes such as metabolism and growth.

MATERIALS AND METHODS

Cell culture, transfections, treatments and common reagents

For most experiments, wild type HEK293 cells were cultured as described before (22). For the overexpression of CFIm25 and CFIm68, the cDNAs were cloned into pcDNA5/FRT from Invitrogen. These were then stably integrated in the Flp-in recombination site of HEK293 cells (Flp-In™-293 Cell Line #R75007, Invitrogen). For RNAi, HEK293 cells were seeded at a density of 20% in six-well plates and all subsequent steps were done according to the “forward method” from the RNAiMAX protocol (Invitrogen). Following a 48 hr incubation, double-stranded siRNAs (starting from 30 pmol, from Dharmacon and Microsynth) were incubated with Lipofectamine RNAiMAX (Invitrogen) and added to the wells. Cells were harvested after 72 hours for further analysis. Western blotting was performed as described earlier (22). The HRP-labelled secondary antibodies were developed with SuperSignal™ West Pico PLUS Chemiluminescent Substrate (ThermoFisher Scientific #34580) or with SuperSignal™ West Femto Maximum Sensitivity Substrate (ThermoFisher Scientific #34095). LICOR IR680/800nm dye-labelled secondary antibodies were used for multiplexing several antibodies on the same membrane. All western blot images were documented with Azure c600 Gel documentation system equipped with a 8.3 MP CCD camera. Western blot quantifications were performed using the ImageJ software by quantifying the pixels of each band and normalizing against a housekeeping control. For comparison between conditions, all samples were normalised to the average levels of the corresponding control samples. Note that the loading control proteins (GAPDH/ACTIN) are shown multiple times in several figures, whenever a single membrane was re-utilised for staining of multiple candidates. Detailed information regarding antibodies and primers/oligos used for the study are listed in Supplementary Data.

Transcriptome profiling with poly(A)-enriched mRNA-seq

Total RNA was quality-checked on a Bioanalyser instrument (Agilent Technologies, Santa Clara, CA, USA) using the RNA 6000 Nano Chip (Agilent, Cat# 5067–1511) and quantified by spectrophotometry using a NanoDrop ND-1000 Instrument (NanoDrop Technologies, Wilmington, DE, USA). 1μg total RNA was used for library preparation with the TruSeq Stranded mRNA Library Prep Kit High Throughput (Cat# RS-122-2103, Illumina, San Diego, CA, USA). Libraries were quality-checked on the Fragment Analyser (Advanced Analytical, Ames, IA, USA) using the Standard Sensitivity NGS Fragment Analysis Kit (Cat# DNF-473, Advanced Analytical). The average concentration was 128 ± 12 nmol/l. Samples were pooled in equal molarity. Each pool was quantified by PicoGreen fluorometric measurement to be adjusted to 1.8 pM and used for clustering on a NextSeq 500 instrument (Illumina). Samples were sequenced using a NextSeq 500 High Output Kit 75-cycles (Illumina, Cat# FC-404-1005). Primary data analysis was performed with the Illumina RTA version 2.4.11 and base calling software version bcl2fastq-2.20.0.422.

Quantification of gene expression by poly(A)-enriched mRNA-seq

Human protein-coding and lincRNA genes from the Ensembl (23) release 90 annotation were stringently filtered for transcripts whose splice junctions ‘are supported by at least one non-suspect mRNA’ (Ensembl transcript support level 1). To minimize the chance of erroneous estimates of gene expression due to large changes in transcript length by 3′ UTR shortening the 3′-terminal exons of each transcript were discarded. Then, for every gene, we identified those regions that are annotated as belonging to an exon in all of the retained transcripts of that gene. Raw sequencing data in FASTQ format were processed with standard tools: cutadapt (version 1.16) (24) to remove adapters and poly(A)-tails from the reads, and STAR aligner (version 2.7.1a) (25) to map resulting fragments to the genome (assembly version

GRCh38 with splice junction annotations derived from Ensembl release 90). The alignments were sorted and indexed with SAMtools (version 1.10) (26) and later used for plotting coverage profiles of distinct gene loci in RNA-seq samples with the Gviz R package (version 1.28.0; `R` software version 3.6.0) (27). For every gene, all reads with a single best reported alignment ('unique mappers') whose alignment start positions overlapped with any of the exclusively 'exonic' regions of that gene, prepared as described above, were summed up. The resulting gene count tables for each sequencing library were used as input for differential gene expression analyses with the `edgeR` (28) package (version 3.34.0; `R` version 4.1.0). First, genes with low counts were discarded by applying `edgeR`'s `filterByExpr()` function with default parameters across samples of all conditions to ensure that the same genes would be called for each comparison. Then, differentially expressed genes were identified in a pairwise manner between treated and control ('wild type') libraries, by applying the `calcNormFactors()`, `estimateDisp()`, `exactTest()` and `topTags()` functions with default parameters, yielding fold changes ($\log_2$) and corresponding *P* values and Benjamini-Hochberg-corrected (29) false discovery rates for every gene and for each comparison. The scripts used for this analysis are available from the github repository https://github.com/zavolanlab/CFI2021.

Quantification of relative PAS expression and average relative terminal exon lengths from poly(A)-enriched RNA-seq data

To quantify the relative usage of distinct poly(A) sites we applied the PAQR tool (30). The values were aggregated at the level of individual terminal exons to obtain the proportion of transcripts ending at individual positions in individual terminal exons. From these values we calculated a weighted average relative terminal exon length as the sum over all 3′ ends in the terminal exon, relative usage of the 3′ end multiplied by the length of the terminal exon ending at the respective site. We obtained quantification for 1′750 terminal exons with multiple poly(A) sites. The PAQR code is available from https://github.com/zavolanlab/PAQR2 and the source code for target identification and analysis from https://github.com/zavolanlab/CFI2021.

Quantification of relative PAS expression and average relative terminal exon lengths from 3′ end sequencing data

To identify targets of CFIm-mediated 3′ end processing based on 3′ end sequencing data, we used poly(A) site quantifications in relevant cellular systems from the PolyASite database (31). Briefly, this database contains 3′ end sequencing data from control and CFIm25/CFIm68-depleted HEK293 cells as well as control and CFIm25/CFIm68-depleted HeLa cells, both obtained with the A-seq method for 3′ end sequencing (32). PolyASite also contains data for HeLa control and CFIm68-depleted samples, generated with the PAPERCLIP method for 3′ end sequencing (33). Based on the ENSEMBL90 gene annotation, we extracted all annotated terminal exons, intersected the quantified poly(A) sites from the PolyASite database in the samples mentioned above, and then carried out the terminal exon length calculation as described in the previous section.

Selection of CFIm targets

We applied Principal Component Analysis (PCA) to per-sample average terminal exon lengths and calculated the projection on, as well as correlation of each terminal exon (treated as a vector in the space of samples) with principal component 1. We then selected those transcripts and genes whose exons exhibited higher than 0.9 correlation and higher than 10 projection scores (both in absolute values) as CFI targets. Almost all (855 of 858) of the transcripts underwent 3′ UTR shortening upon CFIm KD. These were the focus of our study.

Selection of CFIm targets from 3′ end sequencing datasets

We applied a similar analysis to the terminal exon data from the 3′ end sequencing experiments mentioned above. The threshold on the correlation value was set such as to obtain a number of targets similar to that obtained from RNA-seq data. Specifically, the thresholds were 0.9 for HEK293 A-seq data, 0.8 for HeLa A-seq data and 0.95 for HeLa PAPERCLIP data. This yielded 867, 879, and 1071 target transcripts, respectively, for the three datasets.

UGUA frequency analysis

Terminal exons with exactly two poly(A) sites quantified by PAQR were used for the motif frequency analysis. First, we extracted sequences of 401 nucleotides (200 on each side of the PAS) from both proximal and distal PAS in each TE. We traversed each sequence recording the presence/absence of the UGUA motif at each position and then tabulated the counts at each position across all sites. These were plotted using a running average of 30 nucleotides, sliding by 1 nucleotide at a time.

Frequency analysis of UGUA motifs in genes from specific functional categories

From the genes whose terminal exon lengths we quantified with PAQR, we extracted those that were annotated with the Gene Ontology terms 'Cellular Response to Stress' and 'Protein Transport' (according to the STRING server (34)). We then separated these sets into CFIm targets and non-targets, and then carried out the UGUA motif analysis as described in the previous section.

Global proteome and phosphoproteome analysis by shotgun LC-MS

For each sample, 5 × 10^6 cells were washed twice with ice-cold 1x phosphate-buffered saline (PBS) and lysed in 100 μl urea lysis buffer (8 M urea (AppliChem), 0.1 M Ammonium Bicarbonate (Sigma), 1x PhosSTOP (Roche)). Samples were vortexed, sonicated at 4°C (Hielscher), shaken for 5 min on a thermomixer (Eppendorf) at room temperature and centrifuged for 20 min at 4°C full speed. Supernatants were collected and protein concentration was measured with BCA Protein Assay kit (Invitrogen). Per sample, a total of 300 μg of protein mass was reduced with tris(2-carboxyethyl)phosphine (TCEP) at a final concentration of 10 mM at 37°C for 1 hour, alkylated with 20 mM chloroacetamide (CAM, Sigma) at 37°C for 30 min and incubated for 4 h with Lys-C endopeptidase (1:200 w/w). After diluting samples with 0.1 M ammonium bicarbonate to a final urea concentration of 1.6 M, proteins were further digested overnight at 37°C with sequencing-grade modified trypsin (Promega) at a protein-to-enzyme ratio of 50:1. Subsequently, peptides were desalted on a C18 Sep-Pak cartridge (VAC 3cc, 500 mg, Waters) according to the manufacturer's instructions, split in peptide aliquots of 200 and 25 μg, dried under vacuum and stored at −80°C until further use.

For proteome profiling, sample aliquots containing 25 μg of dried peptides were subsequently labeled with an isobaric tag (TMT 10-plex, Thermo Fisher Scientific) following a recently established protocol (35). To control for ratio distortion during quantification, a peptide calibration mixture consisting of six digested standard proteins mixed in different amounts were added to each sample before TMT labeling. After pooling the TMT labeled peptide samples, peptides were again desalted on C18 reversed-phase spin columns according to the manufacturer's instructions (Macrospin, Harvard Apparatus) and dried under vacuum. TMT-labeled peptides were fractionated by high-pH reversed phase separation using a XBridge Peptide BEH C18 column (3,5 μm, 130 Å, 1 mm × 150 mm, Waters) on an Agilent 1260 Infinity HPLC system. Peptides were loaded on column in buffer A (ammonium formate (20 mM, pH 10) in water) and eluted using a two-step linear gradient starting from 2% to 10% in 5 min and then to 50% (v/v) buffer B (90% acetonitrile / 10% ammonium formate (20 mM, pH 10) over 55 min at a flow rate of 42 μl/min. Elution of peptides was monitored with a UV detector (215 nm, 254 nm). A total of 36 fractions were collected, pooled into 12 fractions using a post-concatenation strategy as previously described (36), dried under vacuum and subjected to LC–MS/MS analysis.

For phosphoproteome profiling, sample aliquots containing 200 μg of dried peptides were subjected to phosphopeptide enrichment using IMAC cartridges and a BRAVO AssayMAP liquid handling platform (Agilent) as recently described (37).

The setup of the μRPLC-MS system was described previously (35). Chromatographic separation of peptides was carried out using an EASY nano-LC 1000 system (Thermo Fisher Scientific), equipped with a heated RP-HPLC column (75 μm × 30 cm) packed in-house with 1.9 μm C18 resin (Reprosil-AQ Pur, Dr. Maisch). Aliquots of 1 μg total peptides were analysed per LC×MS/MS run using a linear gradient ranging from 95% solvent A (0.15% formic acid, 2% acetonitrile) and 5% solvent B (98% acetonitrile, 2% water, 0.15% formic acid) to 30% solvent B over 90 minutes at a flow rate of 200 nl/min. Mass spectrometry analysis was performed on a Q-Exactive HF mass spectrometer equipped with a nanoelectrospray ion source (both Thermo Fisher Scientific) and a custom made column heater set to 60°C. 3E6 ions were collected for MS1 scans for no >100 ms and analysed at a resolution of 120 000 FWHM (at 200 m/z). MS2 scans were acquired of the 10 most intense precursor ions at a target setting of 100 000 ions, accumulation time of 50 ms, isolation window of 1.1 Th and at resolution of 30 000 FWHM (at 200 *m/z*) using a normalized collision energy of 35%. For phosphopeptide enriched samples, the isolation window was set to 1.4 Th and a normalized collision energy of 28% was applied. Total cycle time was ∼1–2 s.

For proteome profiling, the raw data files were processed using the Mascot and Scaffold software and TMT reporter ion intensities were extracted. Phosphopeptide enriched samples were analysed by label-free quantification. Therefore, the acquired raw-files were imported into the Progenesis QI software (v2.0, Nonlinear Dynamics Limited), which was used to extract peptide precursor ion intensities across all samples applying the default parameters.

Quantitative analysis results from label-free and TMT quantification were further processed using the SafeQuant R package v.2.3.2. (https://github.com/eahrne/SafeQuant/) to obtain protein relative abundances. This analysis included global data normalization by equalizing the total peak/reporter areas across all LC–MS runs, summation of peak areas per protein and LC–MS/MS run, followed by calculation of protein abundance ratios. Only isoform specific peptide ion signals were considered for quantification. The summarized protein expression values were used for statistical testing of differences in expression of abundant proteins between conditions. Here, empirical Bayes moderated *t*-tests were applied, as implemented in the limma package (http://bioconductor.org/packages/release/bioc/html/limma.html) of R/Bioconductor. The resulting per protein and condition comparison *P*-values were adjusted for multiple testing using the Benjamini–Hochberg method.

Inference of kinase activity from phosphoproteome data

We used the Kinase Set Enrichment Analysis (KSEA) as described by Hernandez-Armenta *et al.* (38) and implemented in the R-package KSEA (https://github.com/evocellnet/ksea) to predict the kinase activity changes across conditions. The software takes as input the $\log_2$ fold-change in intensity of each phosphopeptide between two conditions, as well as kinase-substrate associations. It then determines whether the substrates of any of the kinases are enriched among the phosphopeptides with the largest change between conditions, and reports the $-\log_{10}$ of the *P*-value as a proxy of kinase regulatory activity (38). As only ∼6% of the quantified phosphopeptides in our data set have associated kinases in the PhosphoSitePlus database (39), we used weight matrix models of kinase substrate specificity to predict further associations as follows. Considering all of the peptide sequences S_i obtained in an experiment, the likelihood of a sequence S_i to have a binding site for a kinase k can be written as:

$$P(S_i|k) = \sum_{j=0}^{l_i-l_k} P(S_i[0..j-1]|B)\,P(S_i[j..j+l_k-1]|W_k) \times P(S_i[j+l_k..l_i-1]|B),$$

where l_i is the length of the peptide, l_k is the length of the weight matrix W_k corresponding to kinase k, and B

is the background model for the relative occurrence of amino acids (AA) in peptides (here we used the overall frequency of each AA in all peptides in the dataset). We constructed the weight matrices W_k from all known kinase-substrate associations, taking a window of length $l_k = 15$ AA (±7 AA around the phospho site) for each kinase k from the PhosphoSitePlus database (39). For completeness, we also included the possibility that the peptide does not correspond to any of the known WMs, i.e. explaining the peptide sequence entirely by the background model, $P(S_i|B) = P(S_i[0..l_i - 1|B])$. From Bayes' theorem, we have that the probability of a phosphorylated peptide S_i being explained by kinase k is given by $P(k|S_i) = \frac{P(S_i|k)P(k)}{P(S_i)} = \frac{P(S_i|k)P(k)}{\sum_{k'=1}^{N} P(S_i|k')P(k')}$, where $P(k)$ is the prior probability that a randomly selected phosphopeptide from the data is explained by kinase k, and N is the number of kinases for which we have sequence specificity information (including the "background"). As we do not have prior information on $P(k)$, we assumed a uniform distribution, i.e. $P(k) = 1/N$. Finally, we have assigned to each phosphopeptide the kinase that had the highest posterior probability of explaining the peptide.

Real time proliferation assay

Cell growth was assayed using the xCELLigence system (RTCA, ACEA Biosciences, San Diego, CA). The background impedance of the xCELLigence system was measured for 12 s using 50 μl of cell culture media at room temperature in each well of an E-plate 16. After reaching 75% confluence, the cells were washed with PBS and detached from the flasks using a short treatment with trypsin/EDTA. Ten thousand cells were dispensed into each well of an E-plate 16. Growth and proliferation were monitored every 15 min up to 48 hrs via the incorporated sensor electrode arrays of the xCELLigence system, using the RTCA-integrated software according to the manufacturer's parameters. For the siRNA treatments, a lower number (3000) of cells were seeded and allowed to grow without interruption for a minimum of 42–48 h before the assay was briefly interrupted for the addition of the siRNA mixes or lipofectamine (for the mock treatment) to the corresponding wells. For the ERK inhibition assays, we used Ravoxertinib hydrochloride (GDC-0994 hydrochloride). This compound was validated as an orally bioavailable, selective inhibitor of ERK kinase activity, with a half-maximal inhibitory concentration (IC50) of 6.1 nM. We used a 10 mM solution in 1 ml of DMSO obtained from Medchem Express (Cat. No.: HY-15947A). The final concentration of the inhibitor used for seeding of cells was 6.1 nM in complete growth media. As control, an equivalent amount of DMSO was added to the cell culture medium. Ten thousand cells were counted from their culture flasks and mixed with Ravoxertinib hydrochloride or DMSO and seeded into the xCelligence plates as per standard protocol. All measurements were done with a minimum of five biological replicates.

qRT-PCR to estimate the abundance of RNAs and miRNAs

For mRNA quantifications, 50 ng of total RNA was used for reverse transcription following the manufacturer's protocol and cycling conditions (High-Capacity cDNA Reverse Transcription Kit, Thermo Fisher Scientific). Subsequently, the RT reaction was diluted 4-fold with water and subjected to q-PCR in a 96-well format, using primers specific to individual genes and GoTaq® qPCR Master Mix (Promega). The incubation and cycling conditions were set as described in the kit and the plates were analysed in a StepOnePlus Real-Time PCR System (Thermo Scientific). GAPDH was used as housekeeping control for relative estimation. Real-time analyses by two-step RT–PCR were carried out to quantify miRNA expression using the Thermo Scientific TaqMan chemistry-based miRNA assay system as performed earlier (40). Briefly, 25 ng of cellular RNA were used along with specific primers for human let7-a (assay ID 000377), miR-92a (assay ID 000431), miR-16 (assay ID 000391) and miR-19b (assay ID 000396). U6 snRNA (assay ID 001973) was used as an endogenous control. One third of the reverse transcription mix was subjected to PCR amplification with TaqMan® Universal PCR Master Mix No AmpErase (Thermo Scientific) and the respective TaqMan® reagents for target miRNA. Samples were analysed in PCR triplicates from at least two biological replicates of each condition, processed independently. The comparative Ct method which included normalization by the U6 snRNA, was used for each cell type for plotting of mean values with S.D.

Microscopy analysis

Stellaris® FISH Probes, Custom Assay with CAL Fluor® Red 590 Dye targeting the Dicer Long Isoform and Stellaris® FISH Probes, Custom Assay with Quasar®670 Dye, targeting the common region of the transcript were obtained by utilizing the Stellaris RNA FISH Probe Designer (Biosearch Technologies, IncPetaluma, CA) available online at www.biosearchtech.com/stellarisdesigner). Cells were grown on coverslips coated with gelatin and subsequently fixed as done previously (22). FISH was performed as described on the website of the manufacturer derived from protocols established previously (https://biosearchassets.blob.core.windows.net/assets/bti_stellaris_protocol_adherent_cell.pdf) (41,42). Samples were imaged on a fast and stable inverted wide field microscope equipped with a MORE frame and enclosure, motorized XY-stage. Images were captured using a Hamamatsu ORCA flash 4.0 cooled sCMOS with the following parameters: Effective number of pixels: 2048 × 2048, Dynamic range: 16-bit, Quantum efficiency (peak): >70%, Read out noise: 1.9 electrons rms. The Objective used was a 60× TIRF APON with numerical aperture (NA) equal to 1.49. Illumination of the dyes was performed with 395/25, 550/15, 631/28 (nm) solid state light sources. The software used for the purpose of documentation was Live Acquisition 2.5. Images were exported to OMERO for documentation. Detection and analysis of spots were performed using automated pipelines developed in image analysis software IMARIS (BITPLANE). Prior to counting, the signal was deconvoluted using Huygens deconvolution software as per protocol recommended by the in-house imaging facility. Subsequently, images were transformed in IMARIS using SPOT and SURFACE detection modules according to the software-recommended steps. Following creation of spots and surfaces on a control image, the parameters were extrapolated for all other analysed images. Nuclear surface-overlapping spots were counted as nuclear signals, while all the others were counted towards cytosolic numbers. Each field of view was counted in aggregate and then normalised to the number of DAPI stained surfaces (after segmentation). For simplicity, we rounded the number of spots per nucleus to the closest integer before using these numbers for further calculations. A total of nine different fields of view from two independent biological replicates were utilised for statistics. Imaging of paraspeckles were performed using standard IF protocols as performed earlier (22) on an inverted Axio Observer Zeiss microscope (Zeiss) using a Plan Apochromat N 63×/1.40 oil DIC M27 objective with a Photometrics Prime 95B camera. Z-stack images were deconvoluted using ZEN software and further processed using the OMERO client server web tool for generating figures.

Luciferase assays

psiCHECK-2 Vector (Promega; C8021) was digested at XhoI–NotI restriction sites for insertion of the binding regions for the miRNA targets used in our analysis. Specifically, oligonucleotide constructs harboring a perfect match to the candidate miRNAs (hsa-miR-16–5p and hsa-miR-92a-3p) were inserted into the psiCHECK-2 vector MCS between XhoI and NotI for use as reporters. The sequence of the oligos used for the reporter constructs were TTGTAGTATTTTGCGCCAATATTTAC GTGCTGCTAGTCGACCATTGTTAATC for the miR-16 Reporter and TGTAGTATTTTGACAGGCCGGG ACAAGTGCAATAGTCGACCATTGTTAATC for the miR-92a Reporter. For the luciferase assay, 50 ng of the miRNA reporter construct or the undigested parent vector were transfected into HEK293 cells. siRNA treatment with oligos against CFIm25 or CFIm68 was performed 24 hours prior to transfection of the reporter plasmids. Cells were lysed at the 48 h mark post transfection using Passive Lysis Buffer (Promega) and 5 μl of each lysate was used for quantification of Renilla and Firefly luciferase expression. Firefly-normalised Renilla luciferase expression levels were used to compute fold-repression as described earlier (22).

Seahorse XF Real-Time ATP rate assay

For the seeding of cells, cell counting was performed and around 2650 cells were seeded in each well of a Agilent Seahorse XF96 Cell Culture Microplates. The plate was incubated for 72 h before the siRNA treatment was done. Measurement of ATP production rate in cells was performed using the Seahorse XF Real-Time ATP Rate Assay Kit according to the manufacturer's instructions. Briefly, Seahorse XF96 fluxpak cartridges were hydrated using Seahorse XF Calibrant Solution, 24 h pre-measurement. On the day of measurement, the culture medium was replaced with Seahorse XF DMEM medium (2 mM glutamine, 1 mM pyruvate, 10 mM glucose) and cells were incubated for 1 h at 37°C without additional CO_2. Measurement was performed using the standard program for the ATP rate assay kit (Oligomycin injection after 18min, Rotenone/Antimycin A injection after 36 min). Acquired data were normalized to cell numbers via Hoechst33342 staining. Measurement of fluorescence intensity was performed using a Tecan Infinite® M1000 PRO.

Statistics

Samples were compared using the GRAPHPAD PRISM software *t*-test unless otherwise mentioned in the text. A *P*-value of less than or equal to 0.05 was considered significant and indicated on plots wherever applicable.

RESULTS

Wide-spread reciprocal changes in 3′ UTR length in CFIm KD and OE

As the overlap of CFIm targets reported in different studies is limited, we took advantage of prior observations that CFIm25 and CFIm68 have largely similar effects on 3′ UTR length (8–10) to establish a reference set of CFIm targets, specifically by identifying mRNAs whose 3′ UTRs undergo (1) similar changes in length upon perturbation of either CFIm25 or CFIm68, as well as (2) reciprocal length changes when the expression of these factors is reduced or increased. We therefore analysed HEK293 cell lines in which CFIm25 or CFIm68 were depleted by siRNA-mediated knockdown (Figure 1A and Supplementary Figure S1A) as well as HEK293 cell lines expressing FLAG-fusion constructs of either of the two CFIm subunits stably integrated into their genomes (9) (Figure 1B). After sequencing polyadenylated RNAs from these cell lines in 2–3 biological replicates per condition, we quantified the usage of tandem poly(A) sites in terminal exons (TE) with the PAQR tool (30) (Supplementary Figure S1B). The cumulative density functions (CDF) of average terminal exon length revealed the expected trend toward proximal PAS usage and short 3′ UTRs in CFIm KD cell lines (9,10,12,43), and a milder trend in the opposite direction in the OE conditions (Figure 1C). Principal component analysis of terminal exon length showed the expected condition-dependent clustering of the samples, and also that CFIm25 and CFIm68 affect the terminal exon (TE) length in similar ways. The first principal component (PC1), which explains over 90% of the variance in TE length data, reflects the level of CFIm expression (Figure 1D), as samples from CFIm25/68 KD and OE conditions are located at negative and positive coordinates on PC1, respectively. Therefore, we extracted our reference set of CFIm-dependent APA targets as those whose TE length aligned very well with PC1 (correlation > 0.9 in absolute value and projection > 10 in absolute value). We obtained 858 transcripts that satisfied these criteria, 855 of which had shorter 3′ UTRs upon CFIm25/68 KD (Supplementary Table S1, Figure 1E). The consistency of these results with previously reported effects of CFIm25/68 (9,10) supports the validity of our approach to CFIm target selection. Analysis of 3′ UTRs with exactly two PAS used across conditions showed that the CFIm-binding UGUA motif is more prevalent upstream of the distal PAS of transcripts that respond to CFIm perturbations (APA targets) compared to both the proximal PAS of these targets, as well as the proximal and

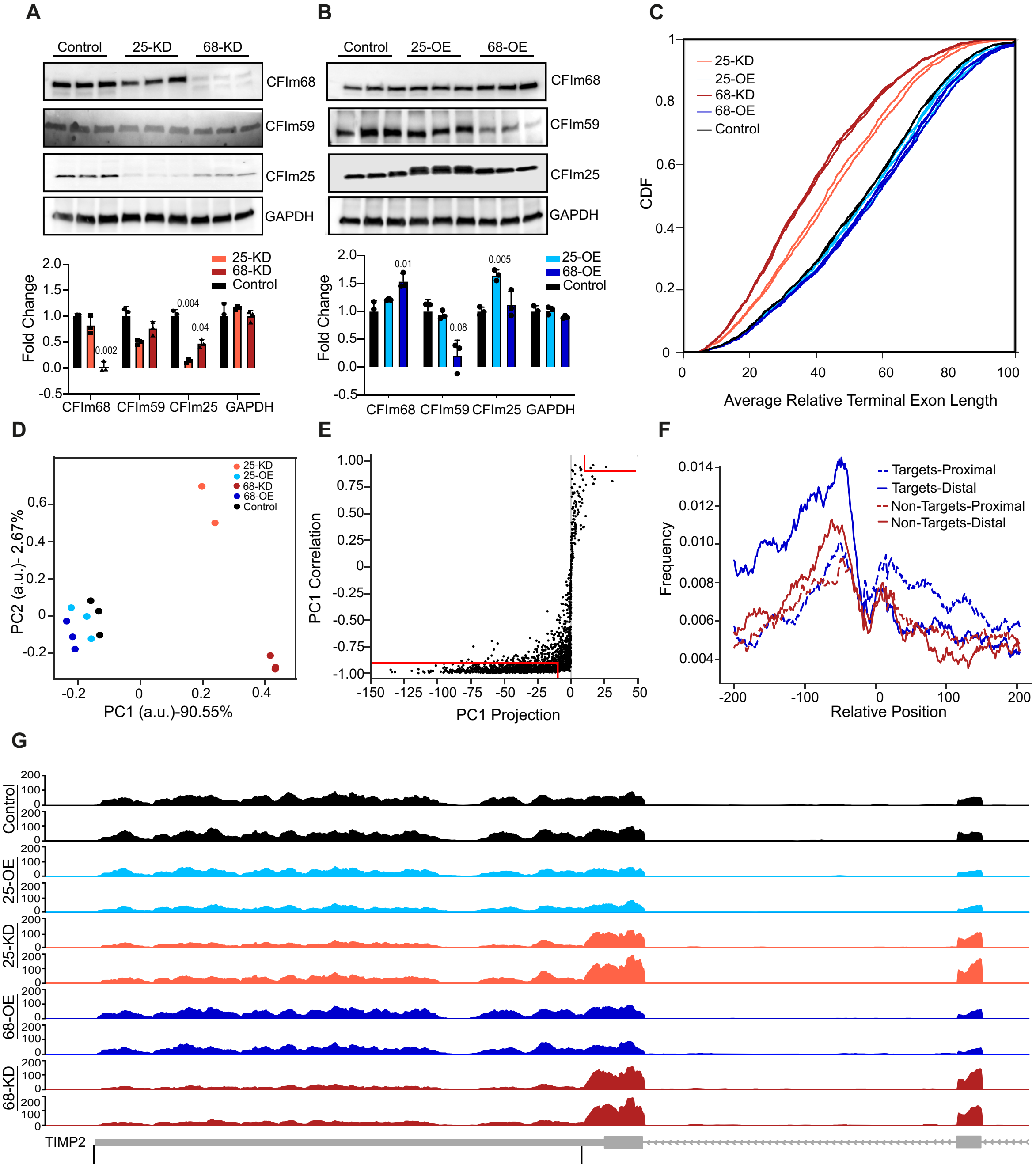


Figure 1. Inference of a reference set of CFIm APA targets. (A, B) Western blots demonstrating the reduced expression of CFIm25 and CFIm68 in the KD (**A**) and increased expression in OE (**B**) HEK293 cell lines. Three biological replicates were generated for each condition. GAPDH was used as loading control. The expression of the 59 kDa component of CFIm, which does not influence the length of 3′ UTRs, was also measured. Shown are also quantifications of protein levels normalized to the mean expression in Control samples (±S.D.). Significant (<0.05) *P*-values computed from a two tailed *t*-test comparing each condition to Control are marked above individual columns. (**C**) Cumulative distributions of average terminal exon length, relative to the maximum given by the annotation (see Materials and Methods), in the different cell lines. *P*-values from two-sample KS-tests for the difference between the CDFs of the average TE length in CFIm25 OE and WT: 0.16, in CFIm68 OE and WT: 0.008, in CFIm25 KD and WT: 1.65e−28, and in CFIm68 KD and WT: 5.42e−64). (**D**) Principal component analysis of TE length. Each dot corresponds to a sample in the space defined by the first two principal components. (**E**) Selection of APA targets of CFIm: the vectors representing average length of individual TEs in all samples were projected onto the first principal component (from panel D) and the length of the projection (x-axis), as well as the correlation of these vectors (y-axis) were calculated. TEs for which both of these values were large in absolute value (marked by the red lines) were considered APA targets of CFIm. (**F**) Position-dependent frequency of occurrence of the CFIm-binding UGUA motif in the vicinity of proximal (dashed lines) and distal (full lines) sites of CFIm targets (blue) and non-targets (red). The curves represent running averages computed over 30 consecutive positions. (**G**) Genome browser tracks showing the coverage of the TE of *TIMP2* (shown in the bottom track) by RNA-seq reads from two replicate experiments for each condition, with the two PAS that were quantified for this gene marked by black lines. The conditions are color-coded (color scheme as in panels C and D) and indicated on the y-axis.The y-axis shows the smoothened number of reads mapping along the TE, calculated by the GViz R package.

distal PAS of transcripts that do not respond (non-targets, Figure 1F). Similar observations were made based on CFIm binding sites determined by crosslinking and immunoprecipitation (9). TIMP2, a previously documented target of CFIm (9,14,15), showed the preferential usage of a proximal poly(A) site in CFIm25/68 KD samples (Figure 1G). These results demonstrate that a large number of transcripts undergo coherent changes in PAS usage upon perturbation in CFIm25 and CFIm68 expression, forming a reference set of CFIm targets.

The CFIm knockdown increases the biogenesis and activity of miRNAs

As 3′ UTR shortening enables mRNAs to escape the repressive effect of miRNAs (44,45), we were intrigued by the conspicuous presence of key components of the miRNA pathway (DICER1, AGO2) among CFIm targets (Supplementary Table S1, Figure 2A, Supplementary Figure S2A and B). To verify changes in *DICER1* isoform expression and further determine whether they occur in the nucleus as a result of APA, as opposed to the cytoplasm as a result of other mechanisms, we visualized the abundance and distribution of *DICER1* 3′ UTR isoforms within cells by single molecule RNA FISH (Figure 2B) in control and CFIm KD (25-KD and 68-KD) conditions. We used probes that selectively bind either to the distal end of the long 3′ UTR isoform (green), or to a region that is shared by the long and short 3′ UTR isoforms (red). The probes are expected to co-localize on the long isoform, which will appear yellow, whereas the short isoform, which lacks the sequence that can hybridize to the green probe, will only fluoresce in the red channel. The quantitative analysis of the relative number of RNA molecules hybridizing to the different probes in CFIm 25-KD/68-KD cells revealed a marked depletion of the long isoform already in the nuclear region, indicating the increased usage of the proximal poly(A) site of the *DICER1* transcript upon CFIm knockdown. The longer isoform was also depleted in the cytosol in these conditions, where the overall number of *DICER1* molecules was markedly higher than in the nucleus (Figure 2C and Supplementary Figure S2D). Western blotting showed that DICER1 protein expression also increases upon CFIm KD (Figure 2D), matching closely both the increased counts of *DICER1* transcripts estimated from RNA-seq analysis and the imaging data (Figure 2C).

As DICER1 upregulation is predicted to increase the production of miRNAs, we measured the levels of three randomly-selected, ubiquitously-expressed miRNA by real time PCR, finding that they were indeed higher in CFIm KD cells relative to Control (Figure 2E). In contrast, despite the shortening of *AGO2* 3′ UTR as a result of CFIm KD, the *AGO2* mRNA level only increased by 42/72% in CFIm25/68 KD relative to control (Supplementary Table S3), and the protein level changes measured by TMT proteomics were even smaller (27/24% in the same conditions). These differences were not detectable when AGO2 protein levels were compared by western blotting (Supplementary Figure S2A). To determine whether the increased miRNA biogenesis translates into increased miRNA-mediated repression, we measured the activity of dual luciferase reporters for two ubiquitously-expressed miRNAs, miR-16 and miR-92a. The reporter expression showed an increased miRNA activity in CFIm25/68 KD cells compared to mock-transfected Control samples, indicating that AGO2 levels were not limiting upon CFIm knockdown (Figure 2F). These results demonstrated the coherent effects of CFIm on the biogenesis and activity of miRNAs whereby the reduction in CFIm expression leads to increased miRNA-mediated repression of target reporters.

CFIm modulates signaling via CMGC kinases

To identify the molecular pathways whose components are APA targets of CFIm, we performed Gene Ontology enrichment analysis (Supplementary Table S1) with the clusterProfiler R package (46). Most enriched in CFIm targets were processes such as cellular response to stress and protein transport and modification (Figure 3A). Genes from these functional categories that we identified as targets exhibited the expected enrichment of the UGUA motif relative to those that are not CFIm targets according to our analysis (Supplementary Figure S3), indicating a sequence-specific effect of CFIm (Supplementary Figure S3). To further map the signaling events in which these targets participate, we measured the abundance of phosphopeptides by phosphoproteomics with IMAC enrichment (see Materials and Methods) in all of the HEK293 cell lines used in this study (Supplementary Table S2). Principal component analysis of the normalized phosphopeptide intensities showed that the OE samples separate well from Control as well as between CFIm components, while the KD samples separate well from Control, but less well between CFIm components (Figure 3B). Of the 22’707 phosphopeptides that were measured, 4′536 showed condition-dependent changes. We then sought to apply a recently developed method, kinase activity enrichment analysis (KSEA) (38) to identify kinases whose activity changes in a CFIm-dependent manner. KSEA is similar to the broadly-used Gene Set Enrichment Analysis (GSEA) (47), quantifying whether phophopeptides associated with a specific kinase are enriched among the phosphopeptides that undergo the largest change in abundance between two conditions. As described by Hernandez-Armenta *et al.* (38), we used the $-\log_{10}$ of the *P*-value, calculated from KSEA, as a proxy of the change in kinase activity with the sign indicating the direction of change of its associated phophopeptides between conditions. Along with changes in phosphopeptide levels, KSEA uses kinase-substrate interactions as input. Finding that only 6.7% of the phosphopeptides that we identified in our experiments are represented among known kinase-substrate interactions in the reference PhosphoSitePlus database (39), we first predicted additional kinase-substrate relationships using position-dependent weight matrix models of kinase substrate specificity (see Materials and Methods). KSEA then revealed pronounced changes in kinase activity in CFIm25/68 KD conditions and milder changes upon OE (Figure 3C). The more pronounced effects of the KD relative to OE on kinase activities mirror the response of 3′ UTR length to these perturbations (Figure 1C). Interestingly, 14 of 35 kinases with a significant activity change (KS-test $P < 0.01$)

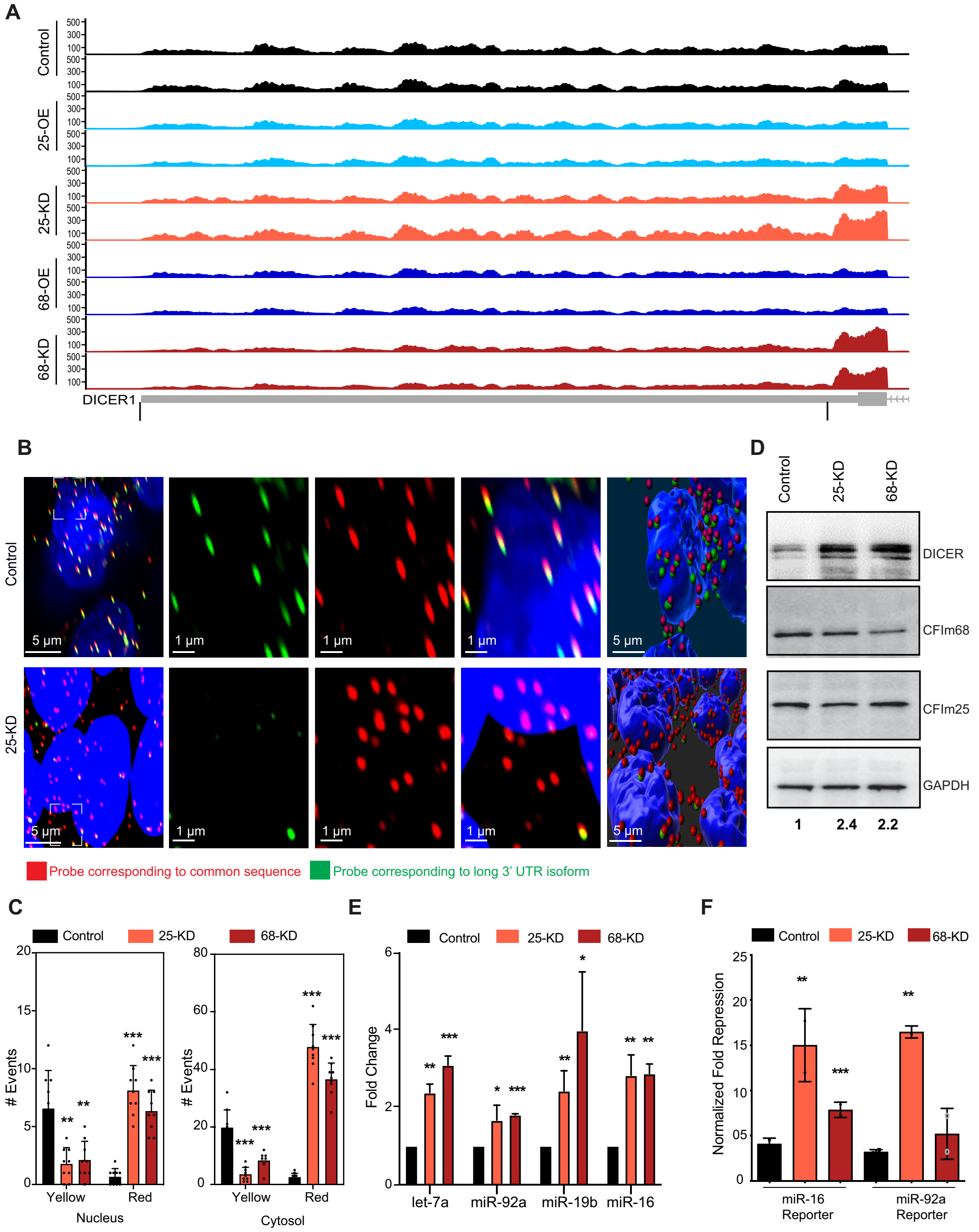


Figure 2. CFIm KD increases the activity of miRNAs. (**A**) Genome browser tracks showing the coverage of *DICER1* TE (bottom track) by RNA-seq reads from two replicate experiments for each condition, with the two PAS that were quantified for this gene marked by black lines. The conditions are color-coded (as in Figure 1) and also indicated on the y-axis. Y-axis shows the smoothened number of reads mapping along the TE, calculated by the GViz R package. (**B**) RNA fluorescence *in situ* imaging of *DICER1* isoforms in Control and CFIm25 KD HEK293 cells with probes corresponding to the common region of the long and short 3′ UTRs (red) or to the region between the proximal and distal cleavage sites, thus present exclusively in the long 3′ UTR (green). Nuclei are marked with DAPI. Zoom-ins of the regions marked with dashed boxes are further shown both with the individual and merged channels. A snapshot of a digital representation of the actual image as processed in IMARIS is also depicted for reference. (**C**) Quantification of the copy number of the long and short 3′ UTR isoforms of *DICER1* in the nucleus (left plot) and cytoplasm (right plot) of Control, CFIm25 and CFIm68 KD cells. Colocalization of the red and green signals reveals the presence of the long 3′ UTR isoform (yellow) whereas the signal from the red probe only reveals the presence of the shorter 3′ UTR isoform. mRNA copy numbers were estimated separately from the nucleus (overlapping with DAPI) and cytosol. Segregation of the signal was performed with IMARIS (see Methods). (**D**) Representative western blot showing the DICER1 expression in the Control, CFIm25 and CFIm68 KD cells. The quantification is relative to GAPDH. (**E**) qPCR measurements of let-7, miR-92a, miR-16 and miR-19b expression in CFIm25/68 KD cells relative to Control. ΔΔct values were calculated relative to U6 snRNA and then relative to the Control cells (where the ratio was set to 1). (**F**) Normalized Renilla luciferase expression of reporter mRNAs carrying binding sites for miR-16 and miR-92a in their 3′ UTRs, in Control, CFIm25 and CFIm68 KD cells, respectively. The Firefly luciferase expressed from the same construct was used as normalization control.

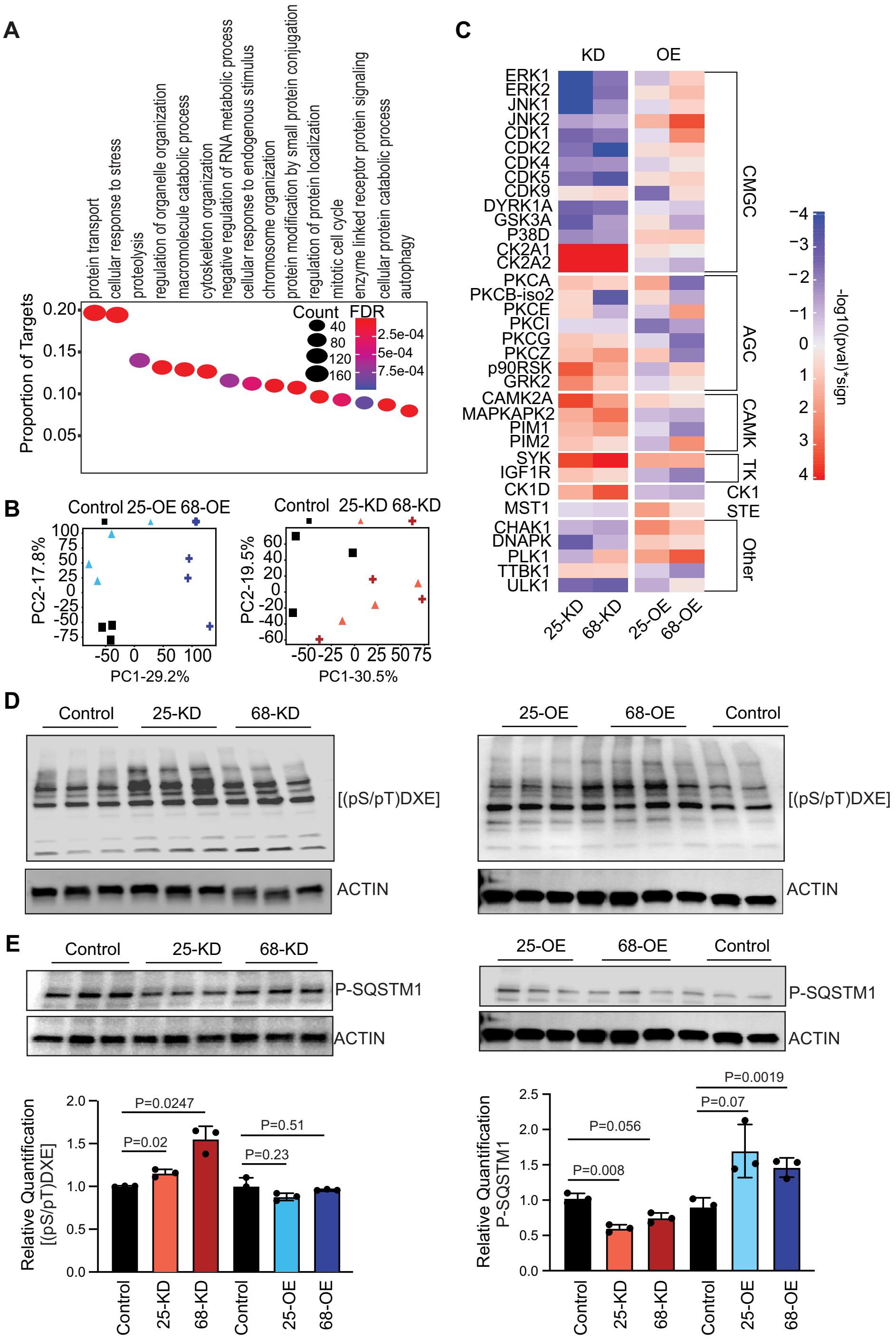


Figure 3. CFIm has a large impact on the intra-cellular signaling landscape. (**A**) Gene Ontology analysis with the clusterProfiler R package (46) identifies biological processes that are significantly enriched in CFIm targets (FDR < 0.01). The y-axis shows the proportion of CFIm targets with a specific biological process annotation and the size of the oval is proportional to the absolute number. The color indicates the significance of the enrichment (FDR value). The 15 most significantly enriched GO categories are depicted. (**B**) Principal component analysis of phosphopeptide intensity data, showing the projection of Control, CFIm25/68 KD and OE samples on the first two principal components. (**C**) Kinase activity changes in KD and OE conditions, computed with the KSEA algorithm (see Methods). Shown are kinases estimated to have a significant activity change (KS-test *P*-value < 0.01) in at least one condition. The scale indicates both the statistical significance of the difference relative to Control samples (log10(*P*-value) of the KS-test) and the direction of the change (indicated by the sign). The sign is that of the mean of log2 fold-change in phosphopeptide intensity between conditions, taken over all phosphopeptides associated with a given kinase. (**D**, **E**) Western blots showing the response of the CK2A1 target motif (D) and of the phosphorylated SQSTM1 (E) to CFIm25/68 KD and OE, with associated quantifications (±S.D., *P*-values in the two-sided *t*-test). Values are calculated relative to the actin control, and then the ratios to the mean of Control samples are used to construct the bar graphs. The same blot was reprobed in the bottom panels of D and E.

in at least one condition belonged to the CMGC family, which includes cyclin-dependent kinases (**C**DKs), mitogen-activated protein kinases (**M**AP kinases), glycogen synthase kinases (**G**SK) and **C**DK-like kinases (48). Only 21 of 173 kinases without a significant change were part of this family.

KSEA predicted an increase in the activity of CK2A1/CK2A2, SYK, MAPKAPK2, CAMK2 and CK1D kinases upon CFIm25/68 KD (Figure 3C), but only CK2A2, MAPKAPK2, and CK1D exhibited a reciprocal change in activity upon CFIm25/68 OE. Focusing on the kinases whose activity changes reflected the change in CFIm expression, we evaluated the KSEA predictions by checking the patterns of phosphorylation of specific substrates. While we did not have any site that could be unambiguously attributed to CK1D in our data, we did find one of the best characterized substrates of MAPKAPK2, the Ser82 residue of the heat shock protein 27 (HSP27) (49). The phosphorylation level of this site was higher in CFIm25/68 KD (1.4-fold, Supplementary Table S2) and lower in CFIm25/68 OE (0.85/0.75, Supplementary Table S2) conditions compared to Control cells, consistent with the overall MAPKAPK2 activities predicted by KSEA. Taking advantage of an antibody that selectively labels all instances of the CK2 substrate consensus sequence, pS/pTD/EXD/E (the most crucial residues being those at positions +3 and +1 with respect to the phosphorylation site (50)), we also sought to independently validate the changes in this kinase's activity in our experimental conditions. Quantitative western blot analysis of the lysates obtained from KD and OE cells revealed an upregulation of total phosphorylation levels in the KD samples relative to the Control cell lysate, and no significant change upon OE, in agreement with the results obtained from the KSEA analysis (Figure 3D).

KSEA also predicted reciprocal changes in the activity of CMGC family kinases such as the mitogen-activated protein kinases JNK1/2 (MAPK8/9), P38D (MAPK13), ERK1/2 (MAPK3/1), and cyclin dependent-kinases CDK1/2/5 upon CFIm KD/OE, the activity decreasing in the KD and increasing in OE conditions (Figure 2C). Of these, transcripts corresponding to *MAPK1*, *MAPK9* and *MAPK13* are also in our reference set of CFIm targets (Supplementary Table S1). To independently validate the changes in MAPK13 activity we focused on its known target, Sequestosome-1, also known as ubiquitin-binding protein p62, which undergoes MAPK13-dependent phosphorylation at Thr269 and/or Ser272 in response to proteasomal stress (51). Both of these sites responded as expected in our phosphoproteome data, with decreased phosphorylation in the CFIm25/68 KD samples (fold-changes relative to Control 0.42/0.55 at Thr269, and 0.47/0.64 at Ser272, Supplementary Table S2) and increased phosphorylation in the CFIm25/68 OE conditions (fold-changes relative to Control 1.56/2.64 at Thr269, and 1.37/1.96 at Ser272, Supplementary Table S2). We observed similar changes in western blots, using an antibody that recognizes SQSTM1 only when phosphorylated at Thr269 and/or Ser272 (Figure 3E). Altogether, these results demonstrate that the level of CFIm is linked to the activity of CMGC kinases, some of which are encoded by transcripts that undergo CFIm-dependent APA.

CFIm-induced changes in cell proliferation reflect the activity of ERK1/2 kinases

The ERK/MAPK signaling pathway plays a key role in cell proliferation, differentiation and apoptosis (52). Activated by endoplasmic reticulum stress and unfolded protein response (53), this pathway can have both tumorigenic (54) and anti-tumorigenic (55) effects. These contrasting roles are reminiscent of the divergent changes in CFIm expression reported in various cancers (18). To validate the predicted change in ERK1/2 activity in our system, we estimated the levels of phosphorylated (Tyr202/Tyr204) ERK1/2 by western blotting. We found them to indeed be positively correlated with the CFIm25/68 expression level (Figure 4A). We then used a real time assay to determine the effect of CFIm on cell proliferation, which was also reported to differ between cell types (15,21). We found that the KD of CFIm25/68 reduced and the OE increased the growth of HEK293 cells (Figure 4B). To ascertain a compelling role of ERK signaling in the increased proliferative state of the CFIm25/68 OE cells, we used an inhibitor of ERK signaling, Ravoxertinib hydrochloride, at reported IC-50 concentrations (56). Cells seeded in the presence of the inhibitor had a conspicuous growth arrest and the growth patterns of the OE cells traced that of Control cells. In contrast, the DMSO treatment had no effect on the growth patterns (Figure 4C). We also verified the reduced growth phenotype of CFIm KD in other cell types, HeLa and LN-18 glioblastoma (Figure 4B), although for these cell lines the effects were milder than those observed in HEK293 cells.

Both the upregulation (57) and downregulation (58) of ERK1/2 activity have been linked to the Warburg-like effect, the switch from oxidative phosphorylation to glycolysis in cellular energy production that is a hallmark of cancer (59). To determine whether the metabolic activity in our cell systems is consistent with changes in ERK1/2 activity, we compared the ATP production by glycolysis and oxidative phosphorylation in all conditions (WT, CFIm KD and OE) in a Seahorse ATP real-time rate assay (60). Indeed, we found the switch from oxidative phosphorylation to glycolysis in ATP production in both CFIm KD and OE cells compared to Control (Figure 4D), as reported for changes in ERK1/2 activity. The main enzyme that drives the carbon flux into mitochondria for the TCA cycle and oxidative phosphorylation is pyruvate dehydrogenase (PDH), whose inhibition leads to the Warburg effect (61). By western blotting, we found that the level of the inhibitory phosphorylation on Ser293 of PDH, known to be catalyzed by pyruvate dehydrogenase kinase 1 and 2 (PDHK1-2), was increased (Figure 4E) in both CFIm KD and OE conditions. Thus, perturbation in CFIm expression leads to metabolic shifts that are consistent with ERK1/2 activity changes converging on PDHK1-2 and PDH.

To better understand how CFIm KD and OE induce divergent changes in cell numbers but convergent changes in the metabolism of cells, we performed endpoint western blot analysis of cleaved PARP (Figure 4F). The cleavage of PARP-1 by caspases is a hallmark of apoptosis (62). The anti-correlation of cleaved PARP with CFIm expression levels indicates that the reduced cell culture growth in CFIm25/68 KD conditions is due to increased levels

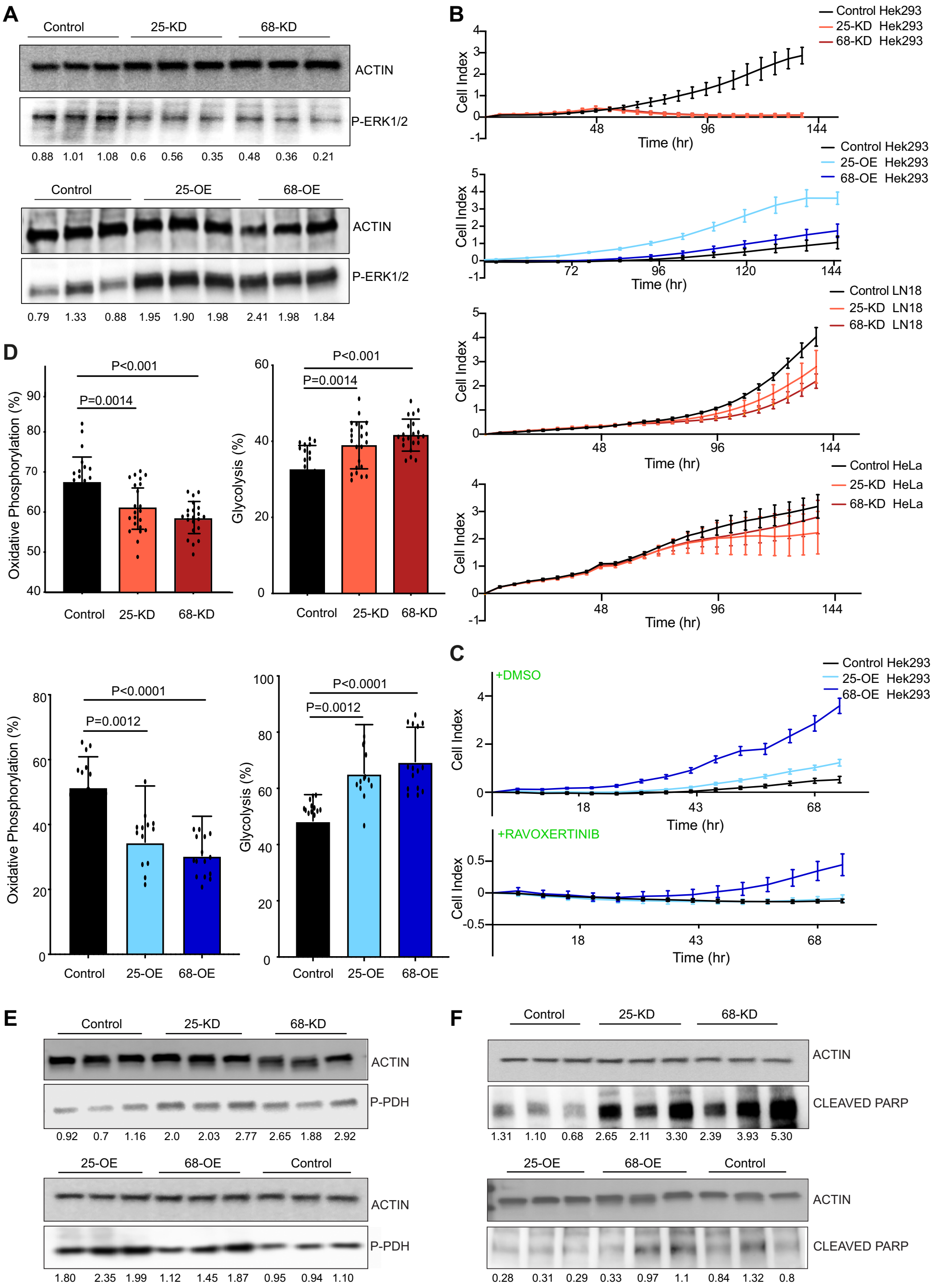


Figure 4. CFIm expression level is linked to the activity of the ERK signaling pathway. (**A**) Western blot analysis of phospho-ERK1/2 (P-ERK1/2) levels in cell lysates from the KD and OE cell lines. The numbers below the blots indicate the levels of phospho(P)-ERK1/2 relative to the actin loading control. For the comparison across conditions, the ratio of P-ERK1/2 to actin in all samples was normalized to the average of the values from Control samples. (**B**) Proliferation index of cells with perturbed CFIm expression and Control cells. Graphs show the observed cell index values in a standard xCelligence real time growth analysis of Control and CFIm25/68 KD and OE HEK293 lines. LN18 and HeLa cells with siRNA-induced depletion of CFIm components were also measured. (**C**) The proliferation index of CFIm OE lines were compared to that of the Control line in the presence of ERK1/2 inhibitor ravoxertinib hydrochloride added at 6.1 nM final concentration in the growth medium. DMSO-containing medium was used as control, to ascertain the growth defects following the ERK1/2 inhibition. (**D**) Real time quantification of the rate of adenosine triphosphate (ATP) production via glycolysis and oxidative phosphorylation in live cells. Bar plots show the percentage of ATP produced via the glycolytic and oxidative phosphorylation routes in HEK293 cells in CFIm KD (top) and OE (bottom) relative to Control cells, estimated with the Seahorse ATP assay (see Methods). (**E**) Western blot estimation of the phosphorylated pyruvate dehydrogenase levels in CFIm25/68 KD/OE and Control HEK293 cells. (**F**) Estimation of cleaved PARP in lysates obtained from CFIm25/68 KD/OE and Control cells. The quantifications below the blots represent the fold changes of the respective targets in the various conditions in CFIm25/68 KD/OE relative to Control. The relative levels were calculated first by dividing the signal intensity to that of the actin loading control and then further normalized to the mean ratio in Control samples. The fold changes of the respective targets in CFIm25/68 KD/OE relative to Control were statistically significant (*P*-value in the two tailed *t*-test ≤ 0.05).

of apoptosis. In summary, these results demonstrate that CFIm promotes the growth of multiple cell types, primarily by suppressing apoptosis. Furthermore, both the cell proliferation and metabolic phenotype are consistent with changes in the ERK1/2 kinase activity.

3′ end sequencing reveals similar CFIm targets

We finally assessed the reproducibility of the CFIm target set with respect to the cellular system and the method for quantifying the 3′ end usage. Perturbations of CFIm25 and/or CFIm68 expression have been carried out not only in HEK293, but also in HeLa cells, where distinct methods for quantifying 3′ end usage were applied. We extracted 3′ end usage data upon CFIm KD in these systems from the PolyASite database (31), and, by applying the same target selection method (Supplementary Figure S4, also see Materials and Methods), we obtained ∼850–1000 genes whose TEs became shorter upon CFIm25/68 KD (Figure 5A–C) from each of these datasets. We then visualized their relationship in a Venn diagram (Figure 5D). The overlap of targets obtained by two distinct methods for PAS usage quantification or in two different cell systems was ∼20–40%. The majority (476 of 853) of genes in our core set are also identified as CFIm targets in another data set, while 51 are common to all data sets (significance of overlap from the SuperExactTest (63) *P*-value: 3.66e−101). Notably, this latter set includes kinases and kinase regulators such as MAPK1 (ERK2), Serine/threonine-protein kinase Chk1 (CHEK1), AMP-activated protein kinase 1 and 2 (AMPKA1-2), C-Jun-amino-terminal kinase-interacting protein 4 (SPAG9) and Receptor-interacting serine/threonine-protein kinase 2 (RIPK2) (Supplementary Table S4). These results indicate that the inferences we made based on the RNA-seq data in HEK293 data were robust, and in line with the growth phenotypes that we assessed above (Figure 4B).

DISCUSSION

Our study makes two main contributions to the expanding field of alternative polyadenylation. First, we provide a reference set of APA targets of the CFIm 3′ end processing factor, the main regulator of 3′ UTR length known to date. These targets were stringently selected based on their consistent and coherent 3′ UTR length changes upon KD/OE of both CFIm25 and CFIm68 components of CFIm. They provide a basis for future analyses in other cell systems, and especially in cancers, where CFIm has been already implicated, with somewhat divergent roles (18). Second, our study uncovered a so-far uncharted layer of regulation downstream of the CFIm 3′ end processing complex, revealing that signaling pathways are extensively remodeled upon perturbations in CFIm expression. The activity of the ERK pathway essentially traces the CFIm expression level and can explain the proliferation, apoptosis and metabolic responses of cells to CFIm perturbations. Beyond these main findings, our results expand the knowledge of the interplay between RNA 3′ end processing and other cellular processes such as miRNA-mediated repression, as detailed below.

In spite of CFIm68 having similar 3′ UTR length regulatory functions, most prior studies focused on CFIm25, reporting a range of CFIm-dependent APA targets that varied ∼100-fold (15,64). To identify conserved functions of CFIm-dependent RNA processing in cell biology, here we constructed a reference set of CFIm targets by carrying out both the KD and the OE of not only CFIm25 but also the CFIm68 subunits of CFIm. We identified 858 transcripts with a highly consistent response across all of these conditions (Figure 1), 855 of which (from 853 genes) exhibited 3′ UTR shortening upon the KD of CFIm factors. The majority of these transcripts are also identified in other cellular systems or with other methods for poly(A) site usage quantification (Figure 5). The set includes well known CFIm targets like *TIMP2* (9,14,65), *DICER1* (15) and *MECP2* (15,16) and, interestingly, paralogs of some reported targets, e.g. *CCND2* and *CHD6* in place of *CCND1* and *CHD9* (15,16). Along with the intersection of targets obtained by 3′ end sequencing data from HEK293 and HeLa cells being only partial (Figure 5), this latter finding may indicate that a subset of CFIm targets is cell type-specific. However, the targets identified in the same cellular system by distinct 3′ end sequencing methods are also not identical (Figure 5D), suggesting that differences in target sets could also be due to differences in the experimental design (e.g. PAPERCLIP-based 3′ end processing data was only available for the CFIm68 KD, and not for the CFIm25 KD). The changes in TE length estimated from the RNA-seq or the 3′ end sequencing data were well correlated, with Pearson correlation coefficients in the 0.5–0.6 range (Figure 5E-F), as observed before (30), underscoring the robustness of the target set.

The frequency of CFIm-binding UGUA motifs is ∼1.5–2-fold higher at the distal PAS of CFIm targets compared to the proximal PAS, in contrast to non-targets, where the two sites are not clearly distinguishable by the UGUA motif frequency (Figure 1). These results are in line with prior observations (9,12,64). The UGUA motif is strongly depleted downstream of the distal PAS, while at the proximal sites UGUA motifs occur with comparable frequency both upstream and downstream of the PAS. This supports a recently proposed model, whereby binding of CFIm to UGUA motifs flanking the proximal PAS leads to the looping of the RNA around the proximal PAS (66) and to an unproductive interaction with FIP1, which masks the site from cleavage. In contrast, the UGUA motif is only present upstream of the distal site of CFIm targets, leading to productive interaction with the other components of the 3′ end processing complex and 3′ end cleavage (64).

It was noted in a previous study that cell cycle-related genes such as cyclin D1 are targets of CFIm (15), linking APA to cellular signaling. Here, we found a strong enrichment of signaling-related proteins among the reference set of CFIm targets (Figure 3). We further predicted changes in the activity of many kinases upon perturbations in CFIm expression, particularly those from the CMGC family. We focused primarily on ERK1/2, because it can link the perturbation in CFIm level to multiple phenotypes reported in the context of cancer, including proliferation, apoptosis and glucose metabolism. ERK activity was positively correlated with the expression level of CFIm, consistent with the effects described in CFIm25-depleted K562 cells (21). Downregulation of ERK/MAPK has a negative effect on cellular

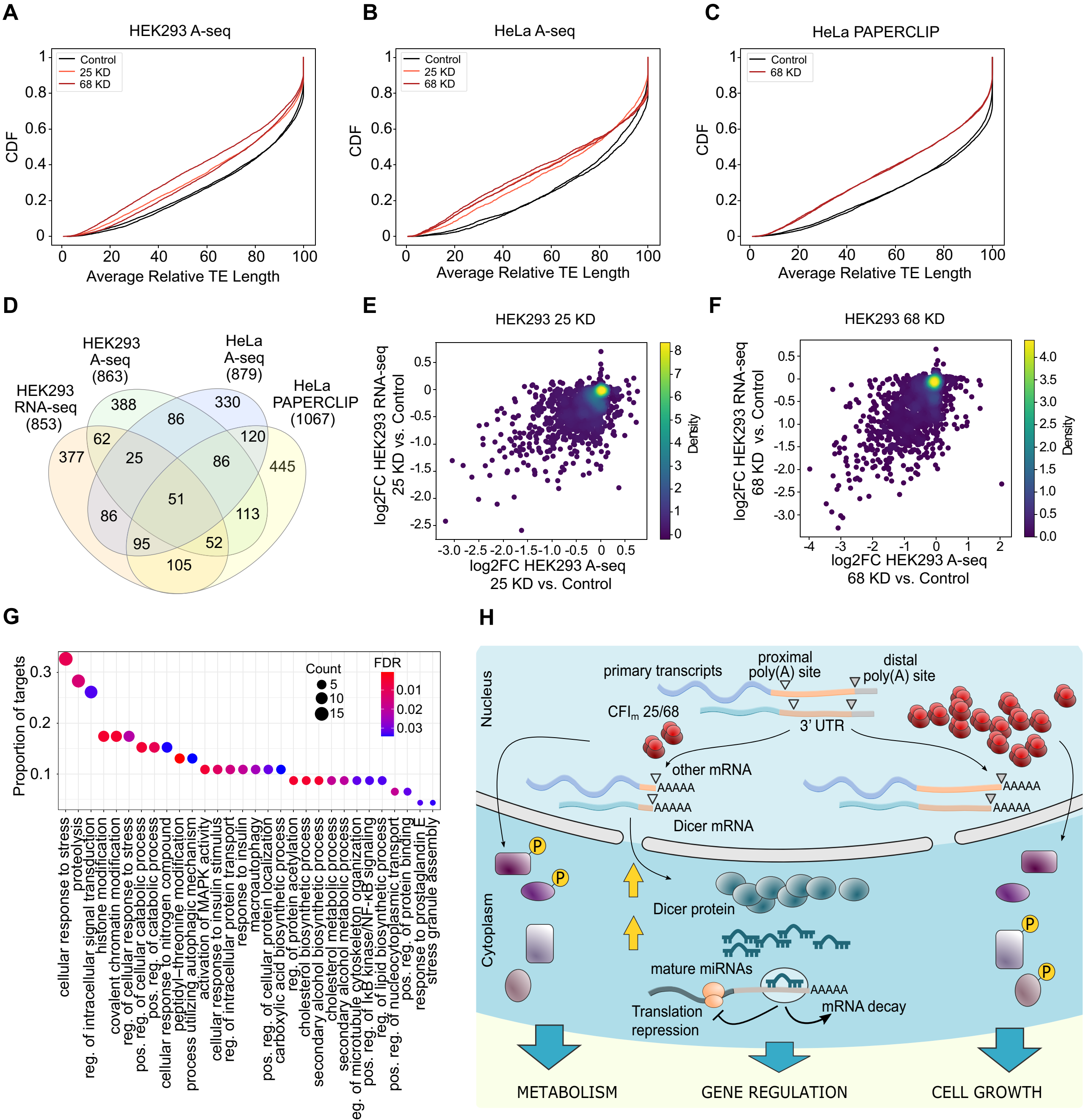


Figure 5. Overlap of CFIm targets identified with several methodologies from different cell types. (A, B) CDF of average relative TE length inferred from A-seq data in HEK293 cells (**A**) and HeLa cells (**B**). The *P*-value in the two-sample KS-test comparing the HEK293 CFIm25 KD and Control conditions is 3.04e−23, while for CFIm68 KD and Control conditions is 1.70e−43. The *P*-value for the two-sample KS-test comparing the HeLa CFIm25 KD and Control conditions is 8.75e−22, while for CFIm68 KD and Control conditions is 1.52e−28. (**C**) CDF of average relative TE length inferred from PAPERCLIP data in HeLa cells; the *P*-value of the two-sample KS-test comparing HeLa CFIm68 KD and Control conditions is 1.60e−58. (**D**) Venn diagram (94) showing the overlap of targets obtained from HEK293 RNA-seq, HEK293 A-seq, HeLa A-seq and HeLa PAPERCLIP data sets. The *P*-value of the overlap of 51 genes among all data sets is 3.66e-101 (SuperExactTest (63)). (**E**) 2D density plot of log2FC in average relative TE length measured with A-seq or RNA-seq in CFIm25 KD relative to Control cells. The Pearson and Spearman correlation coefficients are, respectively, 0.53 and 0.46 (*P*-values 6.27e−109 and 3.94e−81). (**F**) 2D density plot of log2FC in average relative TE length measured with A-seq or RNA-seq in the CFIm68 KD compared to Control cells. Pearson and Spearman correlation coefficients are, respectively, 0.59 and 0.54 (*P*-values 1.41e−142 and 3.09e−112), (**G**) Gene Ontology enrichment analysis with the clusterProfiler R package (46) identifies biological processes that are significantly enriched in CFIm targets that are common among 4 cell lines and 2 techniques for quantifying PAS usage (FDR < 0.05). The y-axis shows the proportion of CFIm targets with a specific biological process annotation and the size of the circle is proportional to the absolute number. The color indicates the significance of the enrichment (FDR value). (**H**) A graphical model of CFIm's impact on several cellular processes.

growth and leads to a metabolic switch that favors glycolysis over oxidative phosphorylation (58), both of which we were able to demonstrate in CFIm KD conditions (Figure 4). Conversely, with real-time assays and cell-cycle analysis we showed that OE of CFIm promotes growth, as expected given the phosphorylation-dependent ERK/MAPK activation (67). Surprisingly, the ERK activity was anti-correlated with the activity of CK2A1, a kinase reported to contribute positively to ERK signaling, as part of the Kinase Suppressor of Ras 1 scaffolding complex (68) and as mediator of ERK nuclear translocation (69). Increased CK2A1 activity was observed in cancer cells (70), associated with drug resistance (71) and resistance to apoptosis (72). Although the mechanism behind seemingly discordant CK2A1 and ERK activities in our system remains to be determined, the upregulation of CK2A1 activity could contribute to the glycolytic switch that we paradoxically observed in CFIm KD conditions (73).

The activity of another MAP kinase, MAPK13 (P38D), was also correlated with the CFIm25/68 levels (Figure 3). MAPK13 participates in key processes such as cell proliferation, differentiation, transcription regulation and development, and is overexpressed in a large set of human breast cancers (74,75). A main target of MAPK13 is SQSTM1 (p62), which MAPK13 phosphorylates on Ser269 and Thr272 (51), as we also found here (Figure 3E). Interestingly, our data includes 5 additional sites on SQSTM1, all 5 with increased phosphorylation in CFIm25/68 OE (Supplementary Table S2). As hyperphosphorylation of p62 is a marker for chemotherapy resistance in ovarian cancer cells (76), the increased phosphorylation of these sites is consistent with the increased growth of CFIm OE cells.

How do changes in 3′ UTR length lead to a remodeling of cellular signaling? Consistent with prior expectations (44), our data shows a small but significant increase in gene expression levels of CFIm targets compared to non-targets in the CFIm KD cells, indicating a small tendency toward increased stability of short 3′ UTR isoforms (77,78) (Supplementary Figure S5 and S6). We further calculated the Pearson correlation coefficient of gene expression changes with both changes in terminal exon length and changes in the proximal/distal PAS usage ratio. While small (<0.23), these correlation coefficients were significant (P-values < 0.01) and had the expected trend (negative correlation of terminal exon length and positive of proximal/distal ratio with gene expression, Supplementary Figure S6). The changes at the protein level were smaller (Supplementary Figure S5), for reasons that remain unclear. The OE of CFIm components did not lead to an opposite effect, namely reduction in target levels, which likely reflects the milder effect of OE on 3′ UTR length compared to the KD of CFIm. This is not surprising because the distal PAS are already preferentially used in HEK293 cells under control culture conditions (9), leading to limited lengthening of 3′ UTRs upon CFIm overexpression. Focusing on signaling-related targets, some of the key regulators that we analysed here, such as MAPK1 (ERK2), and MAPK13 did not show significant gene expression changes, while many others were upregulated upon CFIm25/68 KD. These include, for e.g. CK2A1, MAPK9 (JNK2), the TOR signaling pathway regulator (TIPRL) that negatively impacts JNK signaling by binding to MKK7 (79), the TAK1-interacting protein 27 (JAZF1), which inhibits cell proliferation and enhances apoptosis through its negative control of the TAK1/NF-KB signaling pathway (80), the MAP2K4/MKK4, an upstream activator of JNK signaling (81), and the NDFIP1 ubiquitin ligase activator involved in the ubiquitination of upstream activators of the JNK signaling pathway (82). In contrast to the RNAs, the abundance of the corresponding proteins was less affected by CFIm perturbations. This may be due to a lower sensitivity of the measurement technology, as the ∼2-fold change in abundance of DICER1 that we measured by WB was not apparent in the proteomics data. However, protein level changes were also not uniformly detectable for targets that we measured by WB (Figure 2D, Supplementary Figure S2C), suggesting additional post-transcriptional control of CFIm targets. The overall small protein-level changes could indicate that the phenotypic changes observed upon CFIm perturbations are due to a cumulative effect of small changes in many targets rather than to a small number of targets whose expression is strongly altered.

The similar 3′ UTR shortening in cancer and in CFIm KD conditions make CFIm a very appealing candidate for explaining the cancer-related remodeling of 3′ UTRs. Indeed, in glioblastoma and hepatocellular carcinoma (15,19,20), reduced CFIm expression has been implicated in 3′ UTR shortening and tumorigenesis. However, this relationship does not appear to be universal (18) and, in fact, the expression of 3′ end processing factors is typically higher in tumors compared to matched control tissues (Supplementary Figure S7A, (83)). Kaplan–Meier analysis (84) shows that the levels of CFIm25/68 are also not good predictors of cancer-free survival and that in the majority of cancer cohorts where a significant (P-value < 0.05) association between CFIm25/68 expression and survival can be detected, it is the high, not the low expression of CFIm that represents a risk factor (Supplementary Figure S7B, C). What could account for seemingly contrasting results regarding the CFIm expression and function in cancers? A hypothesis that can reconcile these observations is that the level of CFIm *per se* is not sufficient to predict the pattern of RNA processing in cancers, and that the RNA processing load of cells plays an equally important role; an increased processing load in proliferating cells may lead to transient CFIm deficiency in spite of its increased overall expression and this relationship may further be cell type-specific. This scenario has been reported for the U1 snRNA during neuronal activation, also leading to APA at proximal PAS (85). Of course, technical artifacts such as the variable degree of RNA degradation among samples may also lead to divergent results, due to erroneous estimates of CFIm expression levels (30). Nevertheless, the relationship between RNA processing demand and availability of 3′ end processing factors, especially in the context of cancer, warrants further studies.

Finally, our data show a complex effect of CFIm-dependent APA on the miRNA-mediated gene regulation. Much of the work on APA in the past decade has been motivated by the observation that 3′ UTRs become shorter in proliferating cells (44), presumably leading to the escape of the corresponding transcripts from miRNA regulation (44,45). That DICER1, the key enzyme in miRNA biogenesis, would be regulated in this manner to increase the production of miRNAs in these conditions is counterintuitive, even though it has been observed before (15,45). Here, we found that CFIm KD leads to reduced expression of the long *DICER1* 3′ UTR isoform already in the nucleus, presumably via APA. The DICER1 protein expression increases in parallel to the transcript level. A further contribution to the increased miRNA biogenesis in CFIm KD condition may come from the reorganization of paraspeckles (PS), nuclear condensates that form around the long non-coding RNA (lncRNA) *NEAT1* (86). Reduced CFIm levels lead to the production of a long isoform of *NEAT1*, called *NEAT1_2*, which nucleates the PS (87). *NEAT1_2* also recruits the Drosha/DGCR8 Microprocessor complex to PS, where primary miRNAs interact closely with the NONO-PSF protein dimer (88), leading to increased miRNA biogenesis. Indeed, immunofluorescence analysis with an antibody targeting the PS-essential NONO protein (89) confirms the relationship between the CFIm level and the *NEAT1_2*-dependent size and organization of the PS (90) (Supplementary Figure S7D, E). The effector component of the miRNA pathway, *AGO2*, also undergoes CFIm-dependent APA, but without detectable protein-level changes (Supplementary Figure S2C, Supplementary Table S3). AGO2 does not seem to be limiting for miRNA repression in our systems, as the increased miRNA biogenesis upon CFIm KD is accompanied by a corresponding increase in their repressive activity, as demonstrated with reporter genes (Figure 2). These results demonstrate that CFIm organizes the miRNA-dependent regulatory layer by modulating both miRNA biogenesis and the subset of transcripts that are susceptible to miRNA-dependent regulation. It has been reported, for instance, that uncapped RNAs that are downstream products of cleavage at proximal PAS stably persist in the cell (91). These may alter the cellular milieu to trigger some of the signaling events that we observed, while the increased miRNA activity may serve to clear out some of these RNA species and counteract the cellular stress that they induce. How cells deal with globally increasing or decreasing RNA processing load is an important question to address in future studies.

In conclusion, our study has revealed a novel layer of CFIm-dependent gene regulation, mediated by numerous kinases, especially from the CMGC family. The ERK/JNK/MAPK pathways can explain many of the observed phenotypic changes caused by CFIm expression perturbations, including in cell proliferation, apoptosis and metabolism. We provide a reference set of transcripts that respond in a consistent manner to both KD and OE of CFIm25 and CFIm68 proteins and likely underlie the roles of CFIm in various cellular systems. Finally, we found that CFIm largely promotes cell growth, consistent with some, but not all of the previous studies of cancer systems. Given that the expression of 3′ end processing factors, and especially of CFIm is positively correlated with proliferation, our study integrates a variety of prior observations into a consistent framework. The exact mechanism that bridges the CFIm-mediated APA of several transcripts to the alteration in kinase activities is an interesting topic for future studies (Figure 5F), especially when considering the simultaneous changes in the miRNA biogenesis and, as suggested by our reporter assays, in the activity of the miRNA pathway.

DATA AVAILABILITY

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (92) partner repository with the dataset identifier PXD027034 and 10.6019/PXD027034. The source code to replicate all the analyses presented in this study is available at https://github.com/zavolanlab/CFI2021. Raw and processed RNA-seq data are available at Gene Expression Omnibus (GEO) (93) (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE179630.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Philipp Demougin from the Quantitative Genomics Facility for help with RNA sequencing and the sciCORE facility of the University of Basel for maintaining the high-performance computing network where all the computations have been carried out. M.B. is a recipient of the Fellowship for Excellence at the Biozentrum, University of Basel. We also like to acknowledge the support from Kai Schleicher and Laurent Guerard IMCF for their help in setting up a pipeline for IMARIS Quantifications and Konstantin Schneider-Heieck from the Christoph Handschin group for Seahorse assays.

Author contributions: Project design: S.G. and M.Z.; Experimental analyses: S.G. with help from B.D. (western blots) and G.M. (generation of OE lines); Computational analyses: M.A., M.B., A.B., A.K; with help from C.J.H. Proteomics measurements: A.S, K.B.; Manuscript writing: S.G., M.A., M.Z. with contributions from all authors; Funding acquisition: M.Z.

FUNDING

SNF Grant [310030_189063]. Funding for open access charge: SNF.

Conflict of interest statement. None declared.

REFERENCES

1. Tian,B., Hu,J., Zhang,H. and Lutz,C.S. (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.*, **33**, 201–212.
2. Zhang,H., Lee,J.Y. and Tian,B. (2005) Biased alternative polyadenylation in human tissues. *Genome Biol.*, **6**, R100.
3. Shepard,P.J., Choi,E.-A., Lu,J., Flanagan,L.A., Hertel,K.J. and Shi,Y. (2011) Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*, **17**, 761–772.
4. 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 Res.*, **22**, 1173–1183.
5. Reyes,A. and Huber,W. (2018) Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.*, **46**, 582–592.

6. Wu,F., Chen,Q., Liu,C., Duan,X., Hu,J., Liu,J., Cao,H., Li,W. and Li,H. (2020) Profiles of prognostic alternative splicing signature in hepatocellular carcinoma. *Cancer Med.*, **9**, 2171–2180.
7. Darmon,S.K. and Lutz,C.S. (2012) mRNA 3′ end processing factors: a phylogenetic comparison. *Comp. Funct. Genomics*, **2012**, 876893.
8. Kubo,T., Wada,T., Yamaguchi,Y., Shimizu,A. and Handa,H. (2006) Knock-down of 25 kDa subunit of cleavage factor im in hela cells alters alternative polyadenylation within 3′-UTRs. *Nucleic Acids Res.*, **34**, 6264–6271.
9. Martin,G., Gruber,A.R., Keller,W. and Zavolan,M. (2012) Genome-wide analysis of pre-mRNA 3′ end processing reveals a decisive role of human cleavage factor i in the regulation of 3′ UTR length. *Cell Rep.*, **1**, 753–763.
10. Li,W., You,B., Hoque,M., Zheng,D., Luo,W., Ji,Z., Park,J.Y., Gunderson,S.I., Kalsotra,A., Manley,J.L. *et al.* (2015) Systematic profiling of poly(a)+ transcripts modulated by core 3′ end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. *PLoS Genet.*, **11**, e1005166.
11. Venkataraman,K., Brown,K.M. and Gilmartin,G.M. (2005) Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. *Genes Dev.*, **19**, 1315–1327.
12. Zhu,Y., Wang,X., Forouzmand,E., Jeong,J., Qiao,F., Sowd,G.A., Engelman,A.N., Xie,X., Hertel,K.J. and Shi,Y. (2018) Molecular mechanisms for CFIm-mediated regulation of mRNA alternative polyadenylation. *Mol. Cell*, **69**, 62–74.
13. Yang,S.W., Li,L., Connelly,J.P., Porter,S.N., Kodali,K., Gan,H., Park,J.M., Tacer,K.F., Tillman,H., Peng,J. *et al.* (2020) A cancer-specific ubiquitin ligase drives mRNA alternative polyadenylation by ubiquitinating the mRNA 3′ end processing complex. *Mol. Cell*, **77**, 1206–1221.
14. Kim,S., Yamamoto,J., Chen,Y., Aida,M., Wada,T., Handa,H. and Yamaguchi,Y. (2010) Evidence that cleavage factor im is a heterotetrameric protein complex controlling alternative polyadenylation. *Genes Cells*, **15**, 1003–1013.
15. 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**, 412–416.
16. Brumbaugh,J., Di Stefano,B., Wang,X., Borkent,M., Forouzmand,E., Clowers,K.J., Ji,F., Schwarz,B.A., Kalocsay,M., Elledge,S.J. *et al.* (2018) Nudt21 controls cell fate by connecting alternative polyadenylation to chromatin signaling. *Cell*, **172**, 629–631.
17. Sommerkamp,P., Altamura,S., Renders,S., Narr,A., Ladel,L., Zeisberger,P., Eiben,P.L., Fawaz,M., Rieger,M.A., Cabezas-Wallscheid,N. *et al.* (2020) Differential alternative polyadenylation landscapes mediate hematopoietic stem cell activation and regulate glutamine metabolism. *Cell Stem Cell*, **26**, 722–738.
18. Jafari Najaf Abadi,M.H., Shafabakhsh,R., Asemi,Z., Mirzaei,H.R., Sahebnasagh,R., Mirzaei,H. and Hamblin,M.R. (2019) CFIm25 and alternative polyadenylation: conflicting roles in cancer. *Cancer Lett.*, **459**, 112–121.
19. Han,T. and Kim,J.K. (2014) Driving glioblastoma growth by alternative polyadenylation. *Cell Res.*, **24**, 1023–1024.
20. Tan,S., Li,H., Zhang,W., Shao,Y., Liu,Y., Guan,H., Wu,J., Kang,Y., Zhao,J., Yu,Q. *et al.* (2018) NUDT21 negatively regulates PSMB2 and CXXC5 by alternative polyadenylation and contributes to hepatocellular carcinoma suppression. *Oncogene*, **37**, 4887–4900.
21. Zhang,L. and Zhang,W. (2018) Knockdown of NUDT21 inhibits proliferation and promotes apoptosis of human K562 leukemia cells through ERK pathway. *Cancer Manag. Res.*, **10**, 4311–4323.
22. Ghosh,S., Bose,M., Ray,A. and Bhattacharyya,S.N. (2015) Polysome arrest restricts miRNA turnover by preventing exosomal export of miRNA in growth-retarded mammalian cells. *Mol. Biol. Cell*, **26**, 1072–1083.
23. Cunningham,F., Achuthan,P., Akanni,W., Allen,J., Amode,M.R., Armean,I.M., Bennett,R., Bhai,J., Billis,K., Boddu,S. *et al.* (2019) Ensembl 2019. *Nucleic Acids Res.*, **47**, D745–D751.
24. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.
25. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
26. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
27. Hahne,F. and Ivanek,R. (2016) Visualizing genomic data using gviz and bioconductor. *Methods Mol. Biol.*, **1418**, 335–351.
28. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
29. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **57**, 289–300.
30. Gruber,A.J., Schmidt,R., Ghosh,S., Martin,G., Gruber,A.R., van Nimwegen,E. and Zavolan,M. (2018) Discovery of physiological and cancer-related regulators of 3′ UTR processing with KAPAC. *Genome Biol.*, **19**, 44.
31. Herrmann,C.J., Schmidt,R., Kanitz,A., Artimo,P., Gruber,A.J. and Zavolan,M. (2020) PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Res.*, **48**, D174–D179.
32. Martin,G., Schmidt,R., Gruber,A.J., Ghosh,S., Keller,W. and Zavolan,M. (2017) 3′ end sequencing library preparation with A-seq2. *J. Vis. Exp.*, **128**, 56129.
33. Hwang,H.-W., Park,C.Y., Goodarzi,H., Fak,J.J., Mele,A., Moore,M.J., Saito,Y. and Darnell,R.B. (2016) PAPERCLIP identifies MicroRNA targets and a role of cstf64/64tau in promoting Non-canonical poly(A) site usage. *Cell Rep.*, **15**, 423–435.
34. Szklarczyk,D., Morris,J.H., Cook,H., Kuhn,M., Wyder,S., Simonovic,M., Santos,A., Doncheva,N.T., Roth,A., Bork,P. *et al.* (2016) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.*, **45**, D362–D368.
35. Ahrné,E., Glatter,T., Viganò,C., Schubert,C., Nigg,E.A. and Schmidt,A. (2016) Evaluation and improvement of quantification accuracy in isobaric mass tag-based protein quantification experiments. *J. Proteome Res.*, **15**, 2537–2547.
36. Wang,Y., Yang,F., Gritsenko,M.A., Wang,Y., Clauss,T., Liu,T., Shen,Y., Monroe,M.E., Lopez-Ferrer,D., Reno,T. *et al.* (2011) Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics*, **11**, 2019–2026.
37. Post,H., Penning,R., Fitzpatrick,M.A., Garrigues,L.B., Wu,W., MacGillavry,H.D., Hoogenraad,C.C., Heck,A.J.R. and Altelaar,A.F.M. (2017) Robust, sensitive, and automated phosphopeptide enrichment optimized for low sample amounts applied to primary hippocampal neurons. *J. Proteome Res.*, **16**, 728–737.
38. Hernandez-Armenta,C., Ochoa,D., Gonçalves,E., Saez-Rodriguez,J. and Beltrao,P. (2017) Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics*, **33**, 1845–1851.
39. Hornbeck,P.V., Zhang,B., Murray,B., Kornhauser,J.M., Latham,V. and Skrzypek,E. (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.*, **43**, D512–D20.
40. Ghosh,S., Guimaraes,J.C., Lanzafame,M., Schmidt,A., Syed,A.P., Dimitriades,B., Börsch,A., Ghosh,S., Mittal,N., Montavon,T. *et al.* (2020) Prevention of dsRNA-induced interferon signaling by AGO1x is linked to breast cancer cell proliferation. *EMBO J.*, **39**, e103922.
41. Raj,A., van den Bogaard,P., Rifkin,S.A., van Oudenaarden,A. and Tyagi,S. (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods*, **5**, 877–879.
42. Femino,A.M., Fay,F.S., Fogarty,K. and Singer,R.H. (1998) Visualization of single RNA transcripts in situ. *Science*, **280**, 585–590.
43. Gruber,A.R., Martin,G., Keller,W. and Zavolan,M. (2012) Cleavage factor imis a key regulator of 3′ UTR length. *RNA Biol.*, **9**, 1405–1412.
44. Sandberg,R., Neilson,J.R., Sarma,A., Sharp,P.A. and Burge,C.B. (2008) Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science*, **320**, 1643–1647.
45. Mayr,C. and Bartel,D.P. (2009) Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, **138**, 673–684.
46. Yu,G., Wang,L.-G., Han,Y. and He,Q.-Y. (2012) clusterProfiler: an r package for comparing biological themes among gene clusters. *OMICS*, **16**, 284–287.

47. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
48. Hanks,S.K. and Hunter,T. (1995) The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification 1. *FASEB J.*, **9**, 576–596.
49. Stokoe,D., Campbell,D.G., Nakielny,S., Hidaka,H., Leevers,S.J., Marshall,C. and Cohen,P. (1992) MAPKAP kinase-2; a novel protein kinase activated by mitogen-activated protein kinase. *EMBO J.*, **11**, 3985–3994.
50. Meggio,F. and Pinna,L.A. (2003) One-thousand-and-one substrates of protein kinase CK2?*FASEB J.*, **17**, 349–368.
51. Zhang,C., Gao,J., Li,M., Deng,Y. and Jiang,C. (2018) p38δ MAPK regulates aggresome biogenesis by phosphorylating SQSTM1 in response to proteasomal stress. *J. Cell Sci.*, **131**, jcs216671.
52. Guo,Y.-J., Pan,W.-W., Liu,S.-B., Shen,Z.-F., Xu,Y. and Hu,L.-L. (2020) ERK/MAPK signalling pathway and tumorigenesis. *Exp. Ther. Med.*, **19**, 1997–2007.
53. Darling,N.J. and Cook,S.J. (2014) The role of MAPK signalling pathways in the response to endoplasmic reticulum stress. *Biochim. Biophys. Acta*, **1843**, 2150–2163.
54. Smalley,I. and Smalley,K.S.M. (2018) ERK inhibition: a new front in the war against MAPK pathway–driven cancers?*Cancer Discov.*, **8**, 140–142.
55. Gulmann,C., Sheehan,K.M., Conroy,R.M., Wulfkuhle,J.D., Espina,V., Mullarkey,M.J., Kay,E.W., Liotta,L.A. and Petricoin,E.F. 3rd (2009) Quantitative cell signalling analysis reveals down-regulation of MAPK pathway activation in colorectal cancer. *J. Pathol.*, **218**, 514–519.
56. Blake,J.F., Burkard,M., Chan,J., Chen,H., Chou,K.-J., Diaz,D., Dudley,D.A., Gaudino,J.J., Gould,S.E., Grina,J. *et al.* (2016) Discovery of (S)-1-(1-(4-Chloro-3-fluorophenyl)-2-hydroxyethyl)-4-(2-((1-methyl-1H-pyrazol-5-yl)amino)pyrimidin-4-yl)pyridin-2(1H)-one (GDC-0994), an extracellular signal-regulated kinase 1/2 (ERK1/2) inhibitor in early clinical development. *J. Med. Chem.*, **59**, 5650–5660.
57. Papa,S., Choy,P.M. and Bubici,C. (2019) The ERK and JNK pathways in the regulation of metabolic reprogramming. *Oncogene*, **38**, 2223–2240.
58. Grassian,A.R., Metallo,C.M., Coloff,J.L., Stephanopoulos,G. and Brugge,J.S. (2011) Erk regulation of pyruvate dehydrogenase flux through PDK4 modulates cell proliferation. *Genes Dev.*, **25**, 1716–1733.
59. Vander Heiden,M.G., Cantley,L.C. and Thompson,C.B. (2009) Understanding the warburg effect: the metabolic requirements of cell proliferation. *Science*, **324**, 1029–1033.
60. Mookerjee,S.A., Gerencser,A.A., Nicholls,D.G. and Brand,M.D. (2018) Quantifying intracellular rates of glycolytic and oxidative ATP production and consumption using extracellular flux measurements. *J. Biol. Chem.*, **293**, 12649–12652.
61. McFate,T., Mohyeldin,A., Lu,H., Thakar,J., Henriques,J., Halim,N.D., Wu,H., Schell,M.J., Tsang,T.M., Teahan,O. *et al.* (2008) Pyruvate dehydrogenase complex activity controls metabolic and malignant phenotype in cancer cells. *J. Biol. Chem.*, **283**, 22700–22708.
62. Kaufmann,S.H., Desnoyers,S., Ottaviano,Y., Davidson,N.E. and Poirier,G.G. (1993) Specific proteolytic cleavage of poly(ADP-ribose) polymerase: an early marker of chemotherapy-induced apoptosis. *Cancer Res.*, **53**, 3976–3985.
63. Wang,M., Zhao,Y. and Zhang,B. (2015) Efficient test and visualization of multi-set intersections. *Sci. Rep.*, **5**, 16923.
64. Schwich,O.D., Blümel,N., Keller,M., Wegener,M., Setty,S.T., Brunstein,M.E., Poser,I., De Los Mozos,I.R., Suess,B., Münch,C. *et al.* (2021) SRSF3 and SRSF7 modulate 3′UTR length through suppression or activation of proximal polyadenylation sites and regulation of CFIm levels. *Genome Biol.*, **22**, 82.
65. Kubo,T., Wada,T., Yamaguchi,Y., Shimizu,A. and Handa,H. (2006) Knock-down of 25 kDa subunit of cleavage factor im in hela cells alters alternative polyadenylation within 3′-UTRs. *Nucleic Acids Res.*, **34**, 6264–6271.
66. Yang,Q., Coseno,M., Gilmartin,G.M. and Doublié,S. (2011) Crystal structure of a human cleavage factor CFIm25/CFIm68/RNA complex provides an insight into poly(a) site recognition and RNA looping. *Structure*, **19**, 368–377.
67. Zhang,W. and Liu,H.T. (2002) MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Res.*, **12**, 9–18.
68. Ritt,D.A., Zhou,M., Conrads,T.P., Veenstra,T.D., Copeland,T.D. and Morrison,D.K. (2007) CK2 is a component of the KSR1 scaffold complex that contributes to raf kinase activation. *Curr. Biol.*, **17**, 179–184.
69. Plotnikov,A., Chuderland,D., Karamansha,Y., Livnah,O. and Seger,R. (2019) Nuclear ERK translocation is mediated by protein kinase CK2 and accelerated by autophosphorylation. *Cell. Physiol. Biochem.*, **53**, 366–387.
70. Ruzzene,M. and Pinna,L.A. (2010) Addiction to protein kinase CK2: a common denominator of diverse cancer cells? *Biochim. Biophys. Acta*, **1804**, 499–504.
71. Borgo,C. and Ruzzene,M. (2019) Role of protein kinase CK2 in antitumor drug resistance. *J. Exp. Clin. Cancer Res.*, **38**, 287.
72. Ahmad,K.A., Wang,G., Unger,G., Slaton,J. and Ahmed,K. (2008) Protein kinase CK2–a key suppressor of apoptosis. *Adv. Enzyme Regul.*, **48**, 179–187.
73. Silva-Pavez,E. and Tapia,J.C. (2020) Protein kinase CK2 in cancer energetics. *Front. Oncol.*, **10**, 893.
74. Wada,M., Canals,D., Adada,M., Coant,N., Salama,M.F., Helke,K.L., Arthur,J.S., Shroyer,K.R., Kitatani,K., Obeid,L.M. *et al.* (2017) P38 delta MAPK promotes breast cancer progression and lung metastasis by enhancing cell proliferation and cell detachment. *Oncogene*, **36**, 6649–6657.
75. Tan,F.L.-S., Ooi,A., Huang,D., Wong,J.C., Qian,C.-N., Chao,C., Ooi,L., Tan,Y.-M., Chung,A., Cheow,P.-C. *et al.* (2010) p38delta/MAPK13 as a diagnostic marker for cholangiocarcinoma and its involvement in cell motility and invasion. *Int. J. Cancer*, **126**, 2353–2361.
76. Nguyen,E.V., Huhtinen,K., Goo,Y.A., Kaipio,K., Andersson,N., Rantanen,V., Hynninen,J., Lahesmaa,R., Carpen,O. and Goodlett,D.R. (2017) Hyper-phosphorylation of sequestosome-1 distinguishes resistance to cisplatin in patient derived high grade serous ovarian cancer cells. *Mol. Cell. Proteomics*, **16**, 1377–1392.
77. Spies,N., Burge,C.B. and Bartel,D.P. (2013) 3′ UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.*, **23**, 2078–2090.
78. Gruber,A.R., Martin,G., Müller,P., Schmidt,A., Gruber,A.J., Gumienny,R., Mittal,N., Jayachandran,R., Pieters,J., Keller,W. *et al.* (2014) Global 3′ UTR shortening has a limited effect on protein abundance in proliferating t cells. *Nat. Commun.*, **5**, 5465.
79. Song,I.S., Jun,S.Y., Na,H.-J., Kim,H.-T., Jung,S.Y., Ha,G.H., Park,Y.-H., Long,L.Z., Yu,D.-Y., Kim,J.-M. *et al.* (2012) Inhibition of MKK7-JNK by the TOR signaling pathway regulator-like protein contributes to resistance of HCC cells to TRAIL-induced apoptosis. *Gastroenterology*, **143**, 1341–1351.
80. Huang,L., Cai,Y., Luo,Y., Xiong,D., Hou,Z., Lv,J., Zeng,F., Yang,Y. and Cheng,X. (2019) JAZF1 suppresses papillary thyroid carcinoma cell proliferation and facilitates apoptosis via regulating TAK1/NF-κB pathways. *Onco. Targets. Ther.*, **12**, 10501–10514.
81. Lee,S., Rauch,J. and Kolch,W. (2020) Targeting MAPK signaling in cancer: mechanisms of drug resistance and sensitivity. *Int. J. Mol. Sci.*, **21**, 1102.
82. Mund,T. and Pelham,H.R.B. (2010) Regulation of PTEN/Akt and MAP kinase signaling pathways by the ubiquitin ligase activators ndfip1 and ndfip2. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 11429–11434.
83. Schmidt,R., Ghosh,S. and Zavolan,M. (2018) The 3′ UTR landscape in cancer. *eLS*, https://doi.org/10.1002/9780470015902.a0027958.
84. Nagy,Á., Munkácsy,G. and Győrffy,B. (2021) Pancancer survival analysis of cancer hallmark genes. *Sci. Rep.*, **11**, 6047.
85. Berg,M.G., Singh,L.N., Younis,I., Liu,Q., Pinto,A.M., Kaida,D., Zhang,Z., Cho,S., Sherrill-Mix,S., Wan,L. *et al.* (2012) U1 snRNP determines mRNA length and regulates isoform expression. *Cell*, **150**, 53–64.
86. Visa,N., Puvion-Dutilleul,F., Bachellerie,J.P. and Puvion,E. (1993) Intranuclear distribution of U1 and U2 snRNAs visualized by high resolution in situ hybridization: revelation of a novel compartment containing U1 but not U2 snRNA in hela cells. *Eur. J. Cell Biol.*, **60**, 308–321.
87. Naganuma,T., Nakagawa,S., Tanigawa,A., Sasaki,Y.F., Goshima,N. and Hirose,T. (2012) Alternative 3′-end processing of long noncoding

RNA initiates construction of nuclear paraspeckles: LncRNA processing for nuclear body architecture. *EMBO J.*, **31**, 4020–4034.

88. Jiang,L., Shao,C., Wu,Q.-J., Chen,G., Zhou,J., Yang,B., Li,H., Gou,L.-T., Zhang,Y., Wang,Y. *et al.* (2017) NEAT1 scaffolds RNA-binding proteins and the microprocessor to globally enhance pri-miRNA processing. *Nat. Struct. Mol. Biol.*, **24**, 816–824.
89. Yamazaki,T., Souquere,S., Chujo,T., Kobelke,S., Chong,Y.S., Fox,A.H., Bond,C.S., Nakagawa,S., Pierron,G. and Hirose,T. (2018) Functional domains of NEAT1 architectural lncRNA induce paraspeckle assembly through phase separation. *Mol. Cell*, **70**, 1038–1053.
90. Yamazaki,T., Yamamoto,T., Yoshino,H., Souquere,S., Nakagawa,S., Pierron,G. and Hirose,T. (2021) Paraspeckles are constructed as block copolymer micelles. *EMBO J.*, **40**, e107270.
91. Malka,Y., Steiman-Shimony,A., Rosenthal,E., Argaman,L., Cohen-Daniel,L., Arbib,E., Margalit,H., Kaplan,T. and Berger,M. (2017) Post-transcriptional 3′-UTR cleavage of mRNA transcripts generates thousands of stable uncapped autonomous RNA fragments. *Nat. Commun.*, **8**, 2029.
92. Perez-Riverol,Y., Csordas,A., Bai,J., Bernal-Llinares,M., Hewapathirana,S., Kundu,D.J., Inuganti,A., Griss,J., Mayer,G., Eisenacher,M. *et al.* (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.*, **47**, D442–D450.
93. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
94. Heberle,H., Meirelles,G.V., da Silva,F.R., Telles,G.P. and Minghim,R. (2015) InteractiVenn: a web-based tool for the analysis of sets through venn diagrams. *BMC Bioinf.*, **16**, 169.

Appendix E: Inferring binding sites of RNA-binding proteins with bindz (Manuscript)

Inferring binding sites of RNA-binding proteins with bindz

Bak Maciej[1,2], Agarwal Krish[3], Katsantoni Maria[1,2], Zavolan Mihaela[1,2]

1 Biozentrum, Universitat Basel, Basel, Switzerland
2 Swiss Institute of Bioinformatics
3 Vishwakarma Institute of Technology, Pune, India

Abstract

Maturation of mRNA is a key step in gene expression in eukaryotes with a significant role of alternative splicing and alternative polyadenylation. Similarly to transcription factors binding to DNA sequences and activating/silencing transcription, RNA-binding proteins (RBPs) provide additional layers of regulation by acting at various steps of mRNA maturation. It is therefore crucial to precisely pinpoint the location where these proteins bind on the pre-mRNA molecules and whether these interactions are affected by genetic mutations. Here we have implemented a bioinformatics tool, bindz, to predict binding sites of RBPs for which the sequence specificity is known on RNA sequences. Bindz is implemented as an efficient Snakemake pipeline which can be easily executed on a personal laptop with a Unix-like operating system to help the design of mutagenesis experiments.

Introduction

Maturation of primary transcripts is an essential step in the gene expression in eukaryotes. The processing of precursor mRNAs involves capping, splicing, cleavage and polyadenylation at the 3'end of the sequence. Mechanisms of alternative splicing as well as alternative polyadenylation are well described in the literature (Wang et al. 2015; Lutz 2008) and they lead to distinct isoforms of genes being expressed. Proteins which are translated from these isoforms may differ in their chemical properties like solubility or hydrophobicity but also in their biological function in situations where the structure of the proteins is altered, for example by the inclusion or exclusion of a certain protein domain (Liu and Altman 2003). The impact of RNA processing on the transcriptome is significant; the vast majority of multi-exonic genes in humans undergo both alternative splicing and alternative polyadenylation (Jiang and Chen 2021; Gruber and Zavolan 2019). Currently, how these post-transcriptional regulatory mechanisms are controlled on individual transcripts is still not well understood. However, a large number of RNA-binding proteins (RBPs) is now known to be involved. RBPs bind to sequence motifs in both precursor and mature RNAs to modulate not only their splicing and polyadenylation, but also localization, transport and translation (García-Mauriño et al. 2017). Therefore the key challenge in studying regulation of mRNA processing and in deciphering the impact of genetic mutations is the identification of potential binding sites for these RBPs on the pre-mRNAs. To address this task we have developed bindZ - an easy-to-use bioinformatics tool for predicting binding sites for RBPs whose binding specificity is known in transcript sequences.

bindz

At the core of our work we incorporated MotEvo (Arnold et al. 2012) - a Bayesian probabilistic method for the prediction of binding probabilities between a selected short motif (specified in a Position Weight Matrix format) and a given nucleotide sequence. While MotEvo was developed to study the control of transcription by transcription factors, here we apply the tool to RBP-RNA interactions. The source of PWMs was the largest database of RBP binding motifs, ATtRACT (Giudice et al. 2016). The main output of bindz is a TSV-formatted table containing coordinates of predicted binding site, RBP IDs, predicted target sequences as well as posterior probability for the binding event. Moreover, the information is graphically presented as heatmaps. From the software engineering perspective bindz is implemented as a Snakemake (Mölder et al. 2021) workflow and the user needs to provide all the necessary information in a configuration text file in a YAML format. We ensure the reproducibility and reliability of our computations with the conda technology for dedicated virtual environments for data processing (“Anaconda Software Distribution” 2020). bindz is designed for both computational biologists as well as experimentalists with basic bioinformatics skills of working in a text shell. Its primary use case is in the design of experiments with sequence variants that might differ in their ability to bind RNA-binding proteins due to specific point mutations that would create or destroy binding sites. The immediate future goal is to enable global screening for binding events over a wide range of pre-mRNA sequences, rendering bindz useful for more general studies on the regulation of gene expression at the post-transcriptional level.

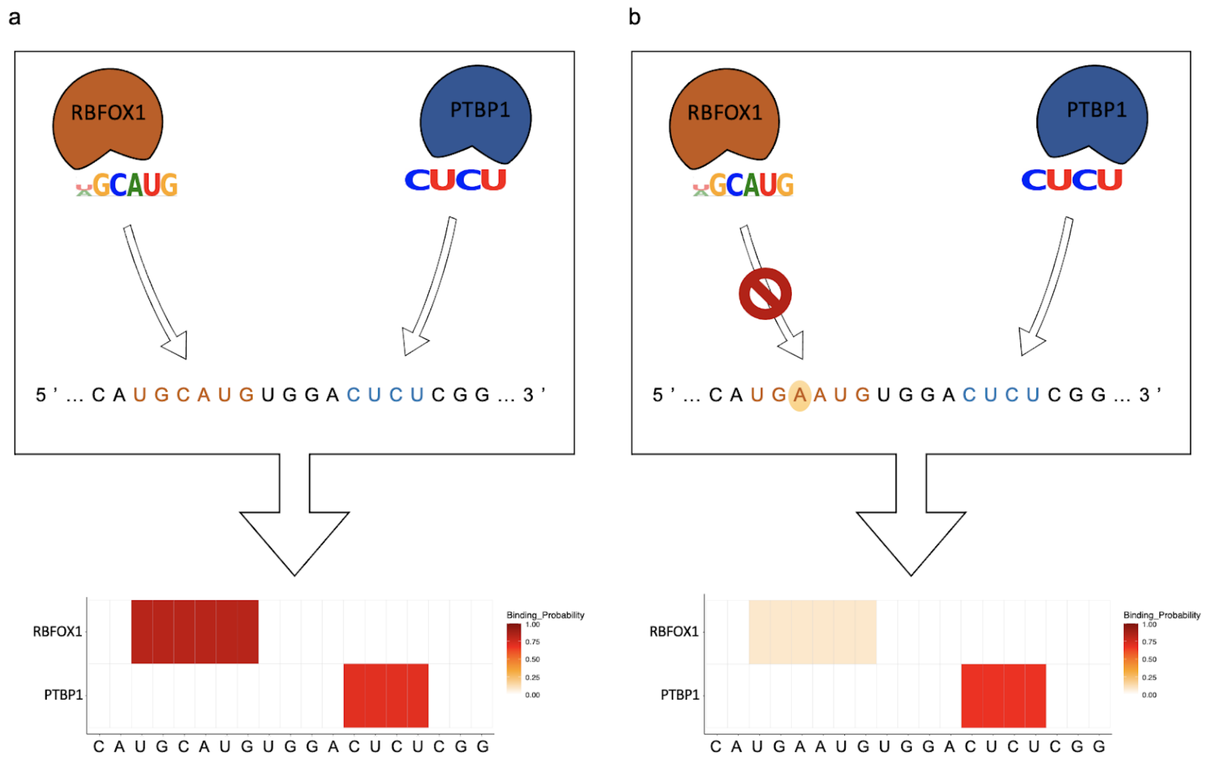

a
RBFOX1
GCAUG
PTBP1
CUCU
5 ' ... C A U G C A U G U G G A C U C U C G G ... 3 '
RBFOX1
PTBP1
Binding_Probability
1.00
0.75
0.50
0.25
0.00
C A U G C A U G U G G A C U C U C G G
b
RBFOX1
GCAUG
PTBP1
CUCU
5 ' ... C A U G A A U G U G G A C U C U C G G ... 3 '
RBFOX1
PTBP1
Binding_Probability
1.00
0.75
0.50
0.25
0.00
C A U G A A U G U G G A C U C U C G G

Figure 1: bindz predicts sites of interaction between RNA-binding proteins and RNA sequences, to facilitate the design of point mutants with distinct RBP interactomes: (a) wildtype (WT) pre-mRNA subsequence with two binding sites for RBFOX1 and PTBP1 regulators. (b) Mutated subsequence where the site for RBFOX1 has been destroyed.

In terms of computational resources bindz may be easily executed on a personal computer with a UNIX-based operating system. Each step of the analysis consumes less than 1GB RAM. We have tested our workflow on a MacBook Pro 2017 machine with macOS Big Sur 11.0.1 operating system. We processed all 1195 PWMs annotated to human RBPs from the ATtRACT database at the time of writing and with 2 cores provided, the pipeline finished in under 1h on a sequence of length 19nt. Moreover, execution time scales linearly with the number of PWMs and since the workflow is parallelized at the level of motifs, the whole runtime might be greatly reduced by providing more processors.

Software availability

A permanent snapshot of bindz version 1.0.1 has been uploaded to zenodo, doi: 10.5281/zenodo.5607105. The tool is hosted and developed on GitHub: https://github.com/zavolanlab/bindz. All community contributions in forms of comments, issues and pull requests are warmly welcome.

Acknowledgments

We would like to express our wholehearted gratitude towards the sciCORE Team at the University of Basel. We are very thankful for the computational infrastructure they have provided us with as well as their time and effort while aiding us with their IT expertise.

Competing interests

None.

References

"Anaconda Software Distribution." 2020. *Anaconda Documentation*. Anaconda Inc. https://docs.anaconda.com/.

Arnold, Phil, Ionas Erb, Mikhail Pachkov, Nacho Molina, and Erik van Nimwegen. 2012. "MotEvo: Integrated Bayesian Probabilistic Methods for Inferring Regulatory Sites and Motifs on Multiple Alignments of DNA Sequences." *Bioinformatics* 28 (4): 487–94.

García-Mauriño, Sofía M., Francisco Rivero-Rodríguez, Alejandro Velázquez-Cruz, Marian Hernández-Vellisca, Antonio Díaz-Quintana, Miguel A. De la Rosa, and Irene Díaz-Moreno. 2017. "RNA Binding Protein Regulation and Cross-Talk in the Control of AU-Rich mRNA Fate." *Frontiers in Molecular Biosciences* 4 (October): 71.

Giudice, Girolamo, Fátima Sánchez-Cabo, Carlos Torroja, and Enrique Lara-Pezzi. 2016. "ATtRACT-a Database of RNA-Binding Proteins and Associated Motifs." *Database: The Journal of Biological Databases and Curation* 2016 (April). https://doi.org/10.1093/database/baw035.

Gruber, Andreas J., and Mihaela Zavolan. 2019. "Alternative Cleavage and Polyadenylation in Health and Disease." *Nature Reviews. Genetics* 20 (10): 599–614.

Jiang, Wei, and Liang Chen. 2021. "Alternative Splicing: Human Disease and Quantitative Analysis from High-Throughput Sequencing." *Computational and Structural Biotechnology Journal* 19: 183–95.

Liu, Shuo, and Russ B. Altman. 2003. "Large Scale Study of Protein Domain Distribution in the Context of Alternative Splicing." *Nucleic Acids Research* 31 (16): 4828–35.

Lutz, Carol S. 2008. "Alternative Polyadenylation: A Twist on mRNA 3' End Formation." *ACS Chemical Biology* 3 (10): 609–17.

Mölder, Felix, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, et al. 2021. "Sustainable Data Analysis with Snakemake." *F1000Research* 10 (January): 33.

Wang, Yan, Jing Liu, B. O. Huang, Yan-Mei Xu, Jing Li, Lin-Feng Huang, Jin Lin, et al. 2015. "Mechanism of Alternative Splicing and Its Regulation." *Biomedical Reports* 3 (2): 152–58.

Appendix F: MAPP unravels frequent co-regulation of splicing and polyadenylation by RBPs and their dysregulation in cancer (Manuscript)

MAPP unravels frequent co-regulation of splicing and polyadenylation by RBPs and their dysregulation in cancer

Maciej Bak[1,2*], Erik van Nimwegen[1,2], Ian U. Kouzel[3], Ralf Schmidt[1,2], Mihaela Zavolan[1,2] and Andreas J. Gruber[3*§]

[1] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

[2] Biozentrum, University of Basel, 4056 Basel, Switzerland

[3] Department of Biology, University of Konstanz, D-78464 Konstanz, Germany

* These authors made equal first author contributions

§ To whom correspondence should be addressed: gruber@uni-konstanz.de.

Abstract

Maturation of eukaryotic pre-mRNAs via splicing and polyadenylation is modulated across cell types and conditions by a variety of RNA-binding proteins (RBPs). Although there exist over 1,500 RBPs in human cells, their binding motifs and functions still remain to be elucidated, especially in the complex environment of tissues and in the context of diseases. To overcome the lack of methods for the systematic and automated detection of sequence motif-guided pre-mRNA processing regulation based on RNA-seq data we have developed MAPP. Applying MAPP to RBP knock-down experiments unravels that many RBPs regulate both splicing and polyadenylation of nascent transcripts by acting on similar sequence motifs. MAPP not only infers these sequence motifs, but also reveals the position-dependent impact of the RBPs on pre-mRNA processing. Interestingly, all investigated RBPs that act on both splicing and 3’ end processing exhibit a consistently repressive or activating effect on both processes, providing a first glimpse on the underlying mechanism. Applying MAPP to normal and malignant brain tissue samples unveils that the motifs bound by the PTBP1 and RBFOX RBPs coordinately drive the oncogenic splicing program active in glioblastomas demonstrating that MAPP paves the way for characterizing pre-mRNA processing regulators under physiological and pathological conditions.

Introduction

Splicing and 3' end processing of nascent RNAs are crucial steps in eukaryotic precursor messenger RNA (pre-mRNA) maturation, also responsible for transcriptome diversification through the generation of transcript isoforms. Both processes are modulated by various RNA-binding proteins (RBPs), whose expression varies across tissues. To date, a few dozen regulators have been described to modulate splicing [1,2], whereas only a handful were reported to impact both splicing and 3' end processing. The Poly(rC) Binding Protein 1 (PCBP1) RBP is a known splicing regulator [3], which has also been reported to regulate the cleavage and polyadenylation (poly(A)) of transcript 3' ends by binding to C-rich sequences that are located in close proximity to poly(A) sites [4]. Further, in previous studies we have shown that the well-known splicing factors HNRNPC (Heterogeneous nuclear ribonucleoproteins C) [5] and PTBP1 (Polypyrimidine Tract Binding Protein 1) [6] regulate alternative cleavage and polyadenylation (APA) by binding to sequence motifs that are located within -200 to +100 and -25 to +75 nucleotides (nt), respectively, relative to the regulated poly(A) sites. ELAVL1 (ELAV Like RNA Binding Protein 1) is another RBP that was reported to impact splicing [7] and polyadenylation [8]. Also the TAR DNA Binding Protein (TARDBP) is known to act as a regulator of alternative splicing (AS) [9] and APA [10]. While these examples indicate that RBPs coordinately regulate splicing and 3' end processing, the sparse characterization of RBP binding specificities of the more than 1,500 RBPs encoded in the human genome [11] has limited these studies. To circumvent this problem we have developed MAPP (Motif Activity on pre-mRNA processing). MAPP enables the identification of RBP-specific sequence motifs that can explain global patterns of both alternative splicing and alternative polyadenylation events quantified from standard RNA sequencing (RNA-seq) experiments. MAPP further unravels the type of regulation (repressive or activating) as well as the binding site position dependency and by charting RBP impact maps MAPP provides a panoramic view on the regulation of alternative splicing and polyadenylation by specific RBPs. We have benchmarked MAPP using data sets in which RBPs with well-characterized impact on splicing and/or 3' end processing have been overexpressed or depleted by siRNA-mediated knock-down, showing that MAPP identifies not only the correct sequence motif, but also the binding site position-dependent impact of the RBP on mRNA processing. Applying MAPP to >400 RBP knock-down experiments from the ENCODE project we have identified multiple pyrimidine motif-binding RBPs that seemingly explain changes in both exon inclusion and poly(A) site choice. The corresponding RBP impact maps provide first insights into patterns that are common to pre-mRNA processing regulation by RBPs. Finally, to demonstrate the ability of MAPP to capture meaningful signals from tissues, we have applied MAPP to glioblastoma (GBM), a cancer type in which large numbers of pre-mRNA processing changes were reported previously [12,13]. MAPP reveals that the PTBP1 and RBFOX RBPs co-regulate the splicing of hundreds of cassette exons, some of which have already been reported to drive GBM development and progression. In summary, in this study we have developed MAPP and demonstrated that "*MAPPing*" RNA-seq experiments enables to identify key pre-mRNA regulators, their binding motifs and functions as well as their role in healthy and diseased cellular states.

Results

MAPP infers impact maps for pre-mRNA processing regulators

Whereas RBPs have long been known to orchestrate pre-mRNA splicing (e.g. [14]), their impact on 3' end processing has only recently started to become apparent [4,5,15,16], and the question of whether RBP regulators act on splicing and 3' end processing in a coordinated manner arose [6,16]. A bottleneck in addressing this question is that compared to other types of regulators, such as transcription or epigenetic factors, the fraction of RBPs with well-characterized binding specificities is relatively minor. In addition, even for those RBPs for which binding specificities have recently been characterized with high-throughput experiments, the impact and mode of action on pre-mRNA processing remain speculative. To address such questions, we have developed MAPP (**Fig. 1**).

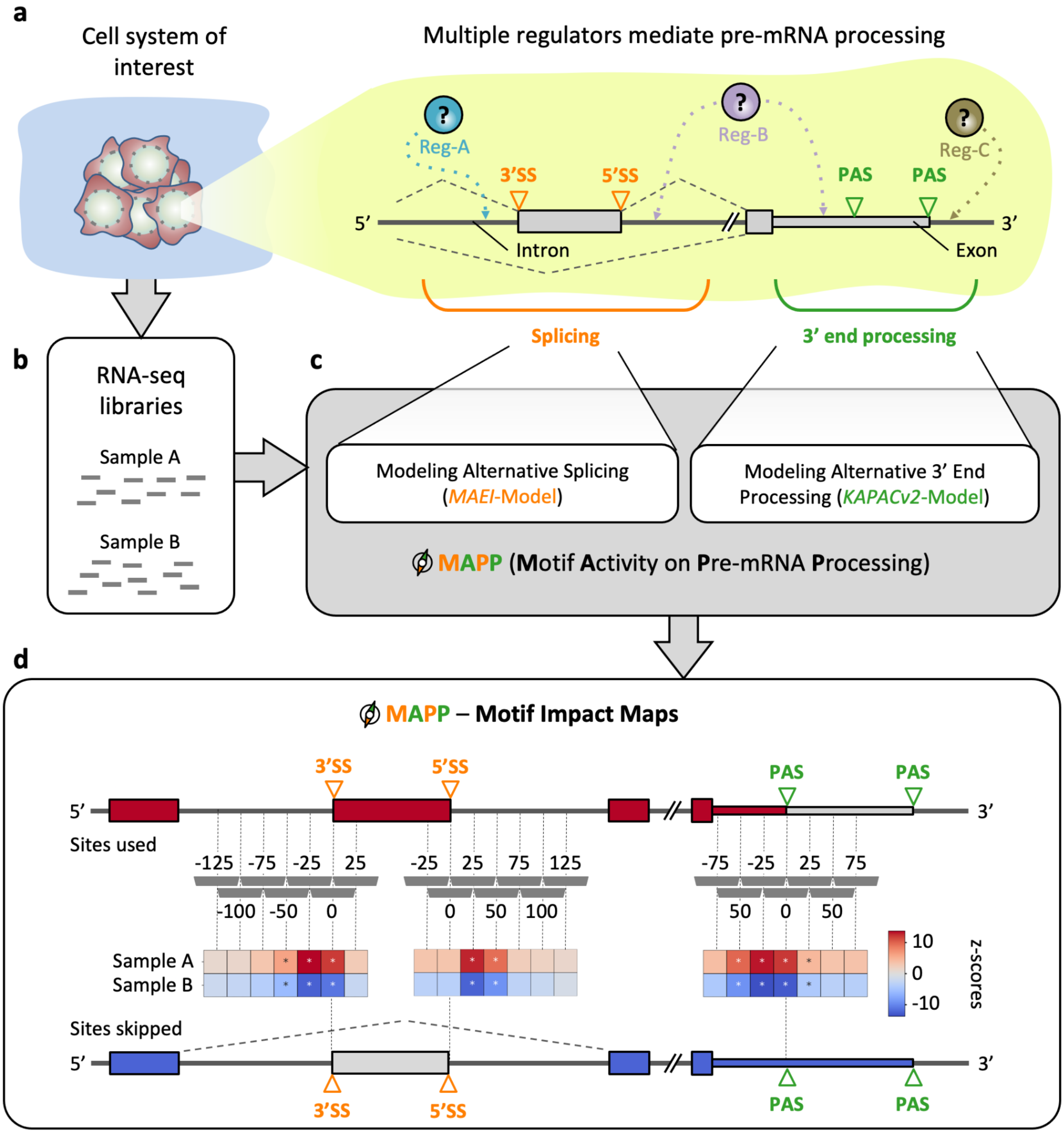


Fig. 1 | **Inferring maps of RBP impact on splicing and 3' end processing with MAPP.** **a |** Sketch illustrating how regulators (Reg) bind pre-mRNAs to influence the usage of splice sites (SS) and / or poly(A) sites (PAS). **b |** RNA-sequencing (RNA-seq) libraries are available or can be created for most cellular systems of interest. **c |** MAPP analyzes the splicing and 3' end processing patterns apparent in the RNA-seq data with the MAEI (Motif Activity on Exon Inclusion) and KAPACv2 (K-mer Activity on PolyAdenylation site Choice version 2.0) models, respectively. **d |** MAPP infers regulatory motifs for RBPs and reports detailed maps of their position-dependent impact on cassette exon inclusion and poly(A) site usage, respectively, by applying the models to genomic windows located at specific distances relative to the RNA processing sites (dashed gray vertical bars).

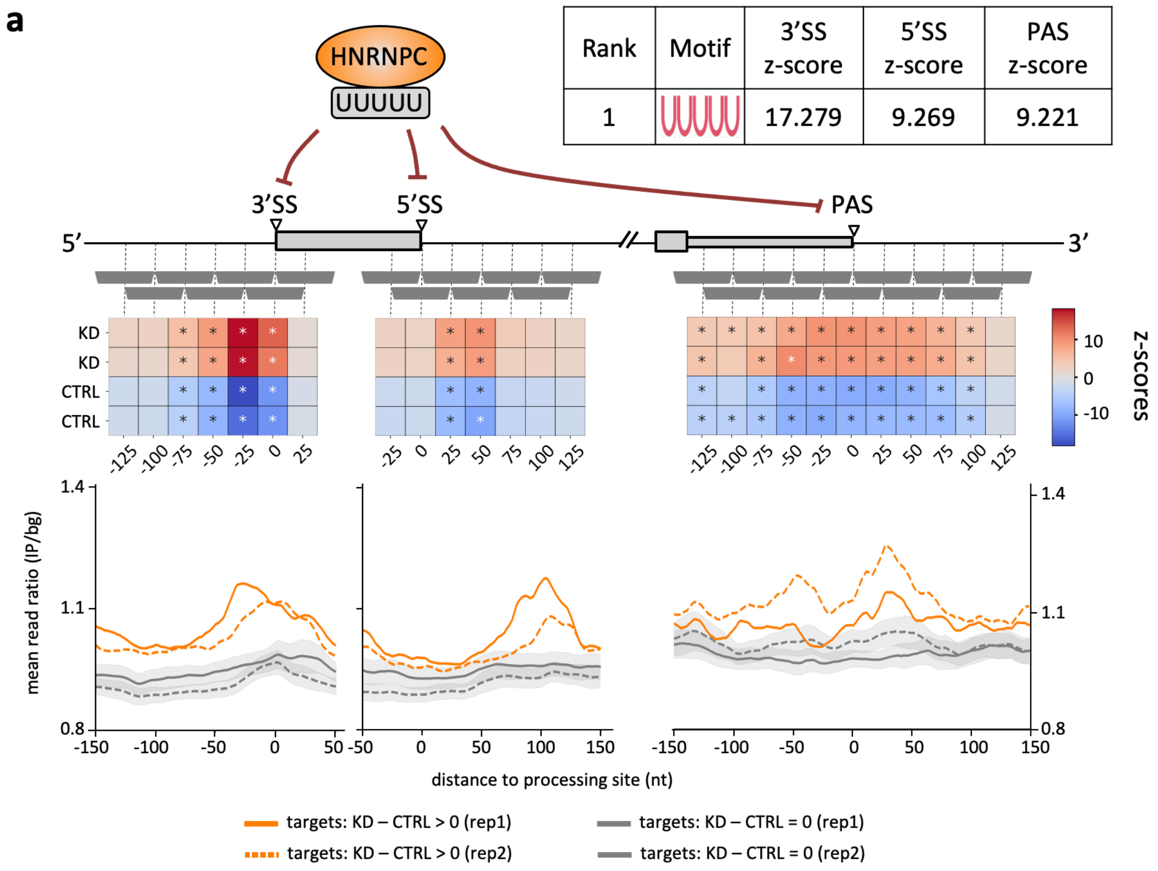
MAPP makes use of a powerful functional concept that we have previously exploited in our KAPAC tool [6], namely explaining relative expression levels of transcript isoforms across samples with sequence motifs located throughout nascent transcripts. In contrast to KAPAC, MAPP provides an end-to-end solution to the inference of motifs, known or not to bind specific RBPs, that impact splicing, 3' end cleavage or both processes.

MAPP includes a novel model, MAEI (Motif Activity on Exon Inclusion), designed to infer the position-dependent activity of sequence elements on cassette exon inclusion, along the KAPACv2.0 model that infers the activity of motifs on poly(A) site processing which builds upon our previously-described KAPAC approach [6]. While similarly to KAPAC, MAPP considers the entire space of sequence motifs, modeled as k-mers, that could impact pre-mRNA processing, its functionality is more general, as it can also work with position-dependent weight matrices (PWMs, see Methods) representing known binding specificities of RBPs. The two components model changes in exon inclusion and poly(A) site usage across genes as functions of the motif counts within regions located at various distances relative to these events. More specifically, given RNA-seq data from a cellular system of interest (**Fig. 1a,b**), MAPP first infers the level of inclusion of alternatively spliced exons and the usage of distinct poly(A) sites. For the latter it makes use of our previously developed PAQR tool [6] (**Supplementary Figures S1** and **S2**). Then, the MAEI and KAPACv2.0 models are fitted to the corresponding pre-mRNA processing event data to identify sequence motifs that can explain global splicing and poly(A) site usage patterns, respectively (**Fig. 1c**). By applying the models to sequence windows located at specific distances relative to pre-mRNA processing sites (for all our analyses unless specified otherwise: 50 nt in length, slided by 25 nt), position-dependent activity z-scores are inferred for each motif. MAPP ranks the sequence elements based on their significance and reports the position-dependent z-scores in the form of impact maps [6] (**Fig. 1d**), which provide detailed insights into the activity (activating or repressive) as well as the position dependency of specific RBPs. Importantly, as MAPP can infer impact maps for motifs known to correspond to specific regulators, as well as for motifs that have not yet been linked to a specific RBP, it is able to unravel the impact of any regulator that regulates pre-mRNA processing in a sequence-specific manner, even if the existence of the regulator and it's binding specificity is unknown to date.

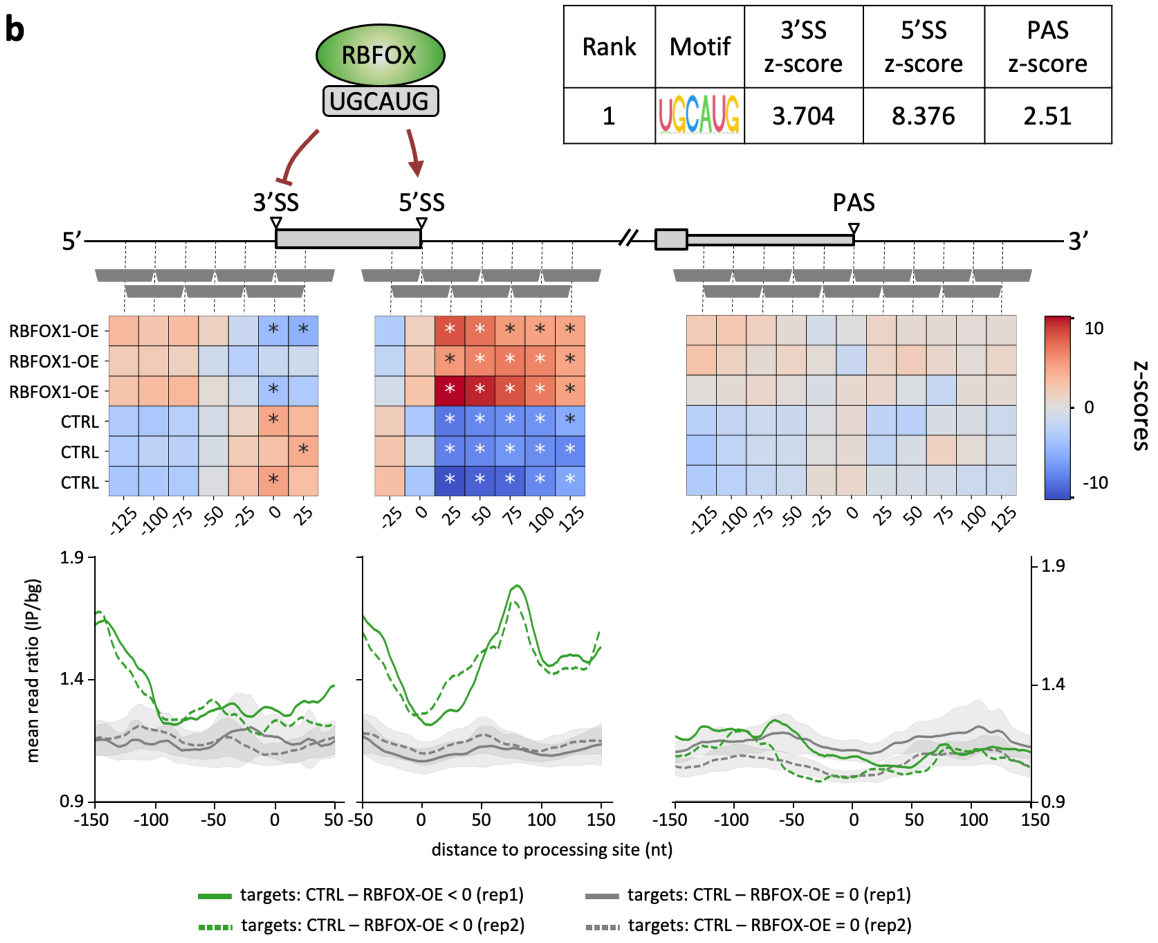
Both the binding specificity and the position-specific impact of known regulators are uncovered *de novo* by MAPP

To validate MAPP, we applied it to data sets from experiments where proteins with a known effect on splicing/polyadenylation were perturbed. We started with the well characterized HNRNPC RBP and found that the sequence motif most significantly associated with both the measured changes in exon inclusion as well as poly(A) site usage is penta-U, the motif that was previously confirmed by multiple studies to be the primary binding motif of HNRNPC [17–19] (**Fig. 2a**, top panel). The PWM representing this motif had the largest combined z-score out of the 344 PWMs that we curated from the ATtRACT database (see Methods). Also, the impact map inferred by MAPP recovers the reported regulation of splicing and 3’ end processing by the HNRNPC RBP. That is, in control (CTRL) samples, where the expression of HNRNPC is high, MAPP infers a repressive effect (marked as blue squares) on 3’ splice site (3’SS), 5’ splice site (5’SS) and polyadenylation site (PAS) processing. Conversely, these sites are processed and thus the activity of the penta-U motif is positive in the knock-down cells, where the expression of HNRNPC is low. These results are supported by a multitude of studies (e.g. refs [5,19,20]). To determine whether the differentially processed sites are indeed bound by the suggested RBPs, we further analyzed data from enhanced crosslinking and immunoprecipitation (eCLIP) experiments from the ENCODE project [21,22]. Towards this, we have selected the top 200 3’SS, 5’SS and PAS whose usage changes most in the expected direction, upon HNRNPC knock-down, as well as the 1,000 sites that change least. For these sites we have constructed position-dependent coverage profiles for HNRNPC eCLIP data. The resulting profiles indicate that HNRNPC is indeed regulating splicing and polyadenylation via direct interaction with the RNAs at the regions inferred by MAPP (**Fig. 2a**, bottom panel).

We next turned to a well characterized splicing regulator, the RBFOX1 RBP. Analyzing data from an experiment where the RBFOX1-dependency of exons was determined in RBFOX2-deficient HEK293 cells in which RBFOX1 was inducibly expressed from a Flp-in locus [23], MAPP ranks the previously described RBFOX1-binding sequence, UGCAUG, as the most significant in explaining exon inclusion, further inferring that it has an activating activity when located downstream of 5’SS [24] (**Fig. 2b**, top panel). MAPP also highlights the opposite activity near the 3’SS, where RBFOX1-binding motifs are associated with reduced exon inclusion. While this repressive effect appears to be much weaker compared to the activating effect of motifs located downstream of 5’SS, it is interesting to infer simply from the RNA-seq data that RBFOX1, like other RBPs [25], can have opposing impacts depending on the location of binding sites. These results demonstrate that by making use of standard RNA-seq experiments only, MAPP enables fine grained insights into the binding-specificity and position-dependent impact of RBPs on splicing and 3' end processing.



Rank	Motif	3'SS z-score	5'SS z-score	PAS z-score
1	UUUUU	17.279	9.269	9.221

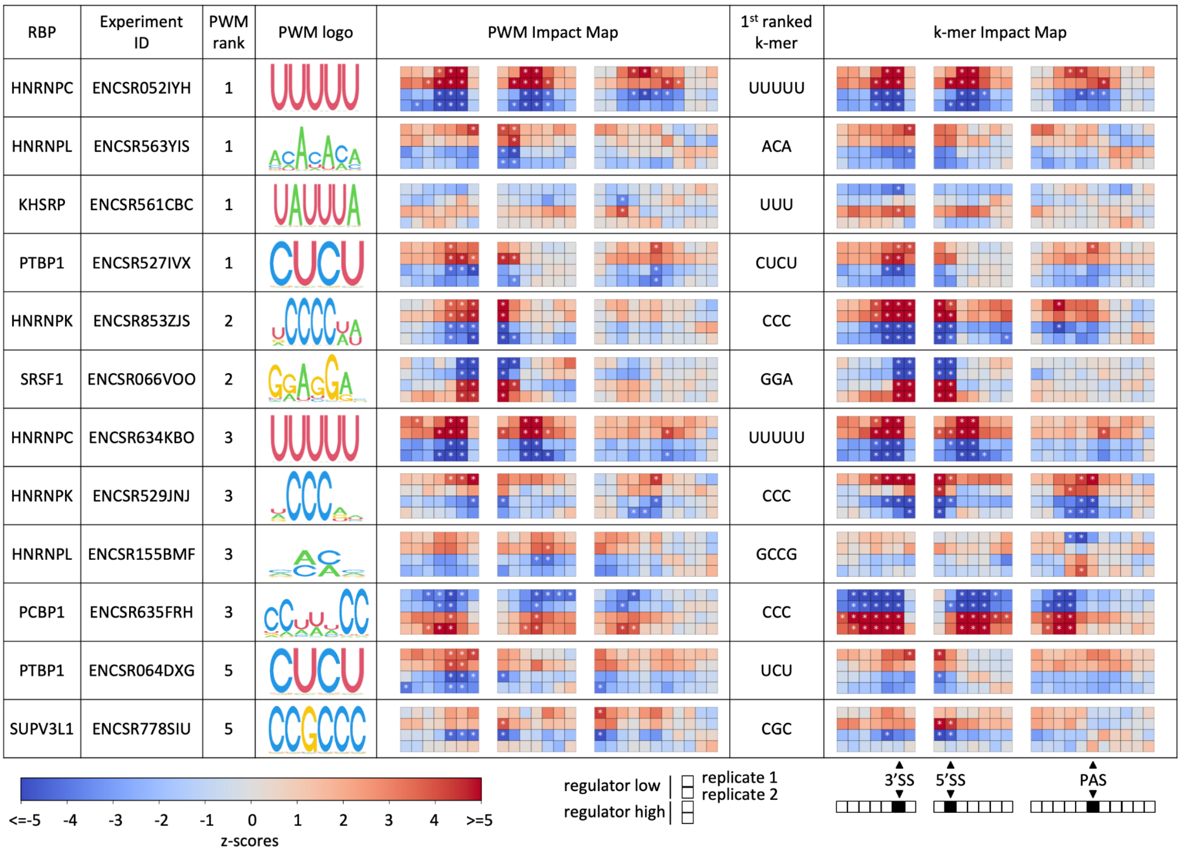


Rank	Motif	3'SS z-score	5'SS z-score	PAS z-score
1	UGCAUG	3.704	8.376	2.51

Fig. 2 | **MAPP infers the known regulatory impact of the HNRNPC and the RBFOX RBPs on splicing and 3' end processing from RBP expression perturbation data sets.** **a |** Top panel: z-scores of activity changes in the vicinity of 3' splice sites (3'SS), 5' splice sites (5'SS) and poly(A) sites (PAS) inferred from a HNRNPC knock-down data set [18]. The PWM with the highest inferred combined z-score of all PWMs has the penta-U motif as consensus. By fitting the splicing and 3' end processing models of MAPP to overlapping windows (horizontal gray bars) located at specific distances relative to splice and poly(A) sites, position-dependent activity z-scores are inferred. Statistically significant z-scores are marked with an asterisk. Bottom panel: Smoothened (+/- 5 nt) HNRNPC eCLIP-based coverage profiles in the vicinity of the top 200 3'SS, 5'SS and PAS, whose usage is most upregulated (red) or does change least (gray) upon HNRNPC knock-down. **b |** Top panel: MAPP results as described in **a**, but here applied to a RBFOX2-deficient HEK293 cell line with induced expression of RBFOX1 which is known to regulate splicing at 5'SS by binding to UGCAUG sequences [24]. Bottom panel: eCLIP profiles as in **a**, but for the RBFOX2 RBP in the vicinity of the top 200 3'SS, 5'SS and PAS, whose usage is most upregulated (blue) or does change least (gray) upon RBFOX1 over-expression.

MAPP impact maps unveil the regulation code of multiple RBPs

Next, we used the large array of RBP knock-down datasets available from the ENCODE project to comprehensively infer the sequence specificity, binding site position-dependent impact and activating or repressive mode of action of human RBPs on pre-mRNA processing. Applying MAPP to 456 RBP knock-down experiments available in ENCODE we found that the tool is also here able to identify the motif known from the ATtRACT database to correspond to the protein whose expression was altered in the experiment. **Fig. 3** shows summary results for samples for which the ATtRACT-provided PWM for the targeted RBP was ranked among the top 5 most significant motifs. As the ATtRACT-provided PWMs corresponding to the perturbed proteins were not always the most significant motifs in explaining the RNA processing alterations, we also ran MAPP in the k-mer mode, to determine which sequence elements explain best the observed changes. For some RBPs, such as PCBP1 and HNRNPK, the k-mer based results are more significant and consistent compared to the PWM-based results, indicating that the inferred k-mer better represents the RBPs binding specificity than the PWM available in public databases. Interestingly, MAPP uncovers a promoting activity of the general splicing factor SRSF1 and the PCBP1 RBP on splice site processing, while other RBPs (e.g. HNRNPC, PTBP1, HNRNPK) appear to have a repressive role. Importantly, half of the RBPs (HNRNPC, PTBP1, PCBP1, HNRNPK) regulate both splicing and polyadenylation by acting on similar sequence motifs. Another interesting observation is that the RBPs generally have the same type of activity on exon inclusion - activating or repressive - when located at both 5’ and 3’ splice sites, but also in those cases where they act on cleavage and polyadenylation. Thus, RBPs inferred by MAPP to have a dual role, on splicing and polyadenylation, appear to predominantly act as either activators or repressors on both pre-mRNA processing steps. This may hint to a concerted regulation of alternative terminal exons by individual regulators, but it must go beyond the regulation of terminal exons, because in many cases the motifs have similar activity around the 5’SS, which does not occur in terminal exons. Finally, MAPP also infers that RBPs with similar sequence specificity can exert their regulatory roles by binding to the pre-mRNA in different positions relative to the pre-mRNA processing sites. For instance, both PCBP1 and HNRNPK bind a ‘CCC’ sequence element to regulate splicing and polyadenylation. However, while the impact of HNRNPK seems to be focused on the immediate vicinity of processing sites, PCBP1 appears to activate splicing from a broader region.



RBP	Experiment ID	PWM rank	PWM logo	PWM Impact Map	1st ranked k-mer	k-mer Impact Map
HNRNPC	ENCSR052IYH	1			UUUUU	
HNRNPL	ENCSR563YIS	1			ACA	
KHSRP	ENCSR561CBC	1			UUU	
PTBP1	ENCSR527IVX	1			CUCU	
HNRNPK	ENCSR853ZJS	2			CCC	
SRSF1	ENCSR066VOO	2			GGA	
HNRNPC	ENCSR634KBO	3			UUUUU	
HNRNPK	ENCSR529JNJ	3			CCC	
HNRNPL	ENCSR155BMF	3			GCCG	
PCBP1	ENCSR635FRH	3			CCC	
PTBP1	ENCSR064DXG	5			UCU	
SUPV3L1	ENCSR778SIU	5			CGC	

Fig. 3 | **MAPP reveals the concerted impact of pre-mRNA processing regulators on splicing and polyadenylation.** For each RBP we determined the first motif (in the order of MAPP-provided significance) that is assigned in the ATtRACTdb to the RBP that was depleted in each experiment. Column 3 shows the rank of that motif as inferred by MAPP. The table contains RBPs where the known binding motif was among the top 5 reported by MAPP. The activity profiles are shown similarly to those in **Fig. 2**, the top two rows indicating the knock-down and the bottom two the control conditions. Windows within regions around 3'SS, 5'SS and PAS are set to the same ranges as done in **Fig. 2**. The central window sliding through a given RNA processing site (-25nt,+25nt) was marked as black square in the legend (bottom right). Furthermore, in addition to the PWM-based MAPP runs, we have carried out a similar analysis exploring all possible k-mers of length 3 to 5. The top-ranked k-mer is reported for each experiment alongside the corresponding PWM result.

While the proteins with known PWMs shown in **Fig. 3** have been implicated in splicing before, we also investigated cases where MAPP identified a significant k-mer, but not a PWM of a known regulator as being able to explain the pre-mRNA processing changes. One interesting example is the Poly(U) Binding Splicing Factor 60 (PUF60) RBP. ENCODE provides knock-down experiments for this protein in two cell lines: K5643 (*ENCSR558XNA*) and HepG2 (*ENCSR648BSC*). The two experiments that exhibit the most significant MAPP results consistently infer a highly similar U-rich sequence element (**Supplementary Fig. S3**), which is also the motif inferred to be bound by PUF60 *in vitro*, in RNA Bind'n-seq experiments [21,22]. As PUF60 exhibited a narrow window of activity upstream of 3' splice sites, we have rerun MAPP at a higher resolution, using windows of 20 nts slided by only 10 nts. The resulting MAPP impact maps charts the regulation of 3' splice site usage by PUF60 at a fine level of detail. The PUF60 RBP is only active when binding to U-rich regions located within a very narrow window (30 to 10 nt) upstream of the 3' splice site (**Supplementary Fig. S3**). While this is consistent with a previous report of PUF60 activating exon inclusion by binding to U-rich regions upstream of 3' splice sites [26], MAPP reveals the very narrow window of PUF60 activity, the intronic region of ~30-10 nts upstream of the 3' splice site, to be much more position-specific than other regulators mentioned above (**Fig. 3**). These results illustrate the utility of MAPP in elucidating the position- and sequence-dependent regulation of pre-mRNA processing by RBPs.

MAPP unravels RBPs that drive the oncogenic splicing program active in glioblastomas

As key factors in the post-transcriptional regulation of gene expression, RBPs have been reported to play an important role in numerous diseases, including cancer [27]. In a previous study we have uncovered that the PTBP1 RBP best explains the global remodeling of 3' UTR length in glioblastoma [6]. Importantly, PTBP1 was previously mainly studied in the context of splicing and the results from our ENCODE screening suggest that the PTBP1 RBP indeed does significantly impact splicing (**Fig. 3**). To follow this further up we applied MAPP to a high quality PTBP1 knock-down dataset without PTBP2 background [28] confirming that PTBP1 does act as global splicing and 3' end processing regulator (**Fig. 4a**, bottom panel). Specifically, in addition to its repressive activity on poly(A) site usage, PTBP1 represses the processing of 3'SS and to some extent 5'SS. Moreover, from the MAPP impact maps we can conclude that PTBP1-binding motifs located within the cassette exon itself or the first ~75 intronic nt upstream of the 3'SS are associated with reduced exon inclusion when the expression levels of PTBP1 are high, i.e. in control conditions. These inferences are also supported by PTBP1 eCLIP data (**Fig. 4a**, top panel).

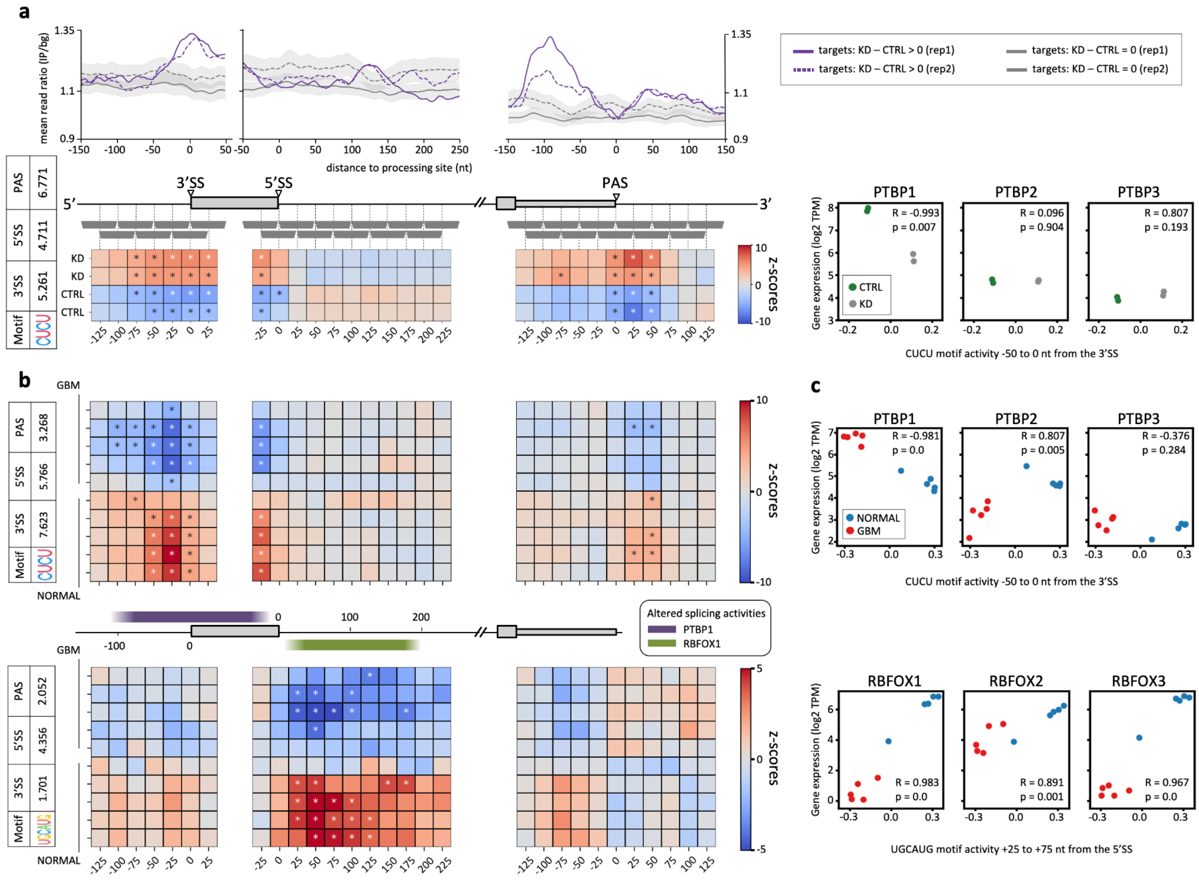


Fig 4. | **MAPP unravels the joint effect of the PTBP1 and RBFOX RBPs on splicing and 3' end processing in glioblastoma. a |** PTBP1 eCLIP densities around pre-mRNA processing sites (top panel) and impact maps for the PTBP-bound CUCU motif as inferred by MAPP (bottom left panel) from control cells and cells depleted of both PTBP1 and PTBP2 by siRNA-mediated knock-down. Bottom right panel: PTBP1/2/3 expression versus the activity of the CUCU motif within 50 nt upstream of 3' splice sites (3'SS). **b |** MAPP results for glioblastoma (GBM) and normal brain (NORMAL) samples for the PTBP-bound CUCU motif (top panel) as well as for the RBFOX-bound UGCAUG motif (bottom panel). Regions with statistically significant CUCU motif activity (purple) or UGCAUG motif activity (green), respectively, are highlighted in the cartoon (mid panel). MAPP was run without a minimum exon length constraint in order to also account for micro-exons prevalent in neurons.
c | Scatter plots of the RBP mRNA expression levels versus the MAPP inferred activities for the region windows indicated.

To uncover which regulators can best explain splicing in glioblastomas we next applied MAPP to cancer samples [29], where it inferred that the PTBP-binding motif has the most significant activity on pre-mRNA processing (**Supplementary Tables S4 and S5**) with a motif ranking and position-dependent activity that matches the profile obtained from the PTBP1/2 knock-down data (**Fig. 4a,b; Supplementary Table S6**). The combined activity of PTBP1 on splicing and polyadenylation in glioblastoma strengthens the case for PTBP1 as a main regulator of pre-mRNA processing in glioblastomas, where PTBP1 is highly overexpressed. Also, as expected for a regulator that acts as a repressor of pre-mRNA processing, the expression of PTBP1 anti-correlates with its motif activity (**Fig. 4c**). Interestingly, besides the PTBP1 motif, the RBFOX-associated motif was also identified by MAPP as being differentially active in glioblastomas compared to normal brain samples (**Fig. 4b**). Moreover, a k-mer-based MAPP run confirmed that in addition to PTBP1-associated CU-rich k-mers, the GCAUG sequence bound by the RBFOX RBPs is also among the significant 5'SS-proximal k-mers that regulate exon inclusion (**Supplementary Table S4**). Consistently with the known role of RBFOX RBPs as activators of splice site usage, the MAPP-inferred activity correlates remarkably well with RBFOX expression (**Fig. 4c**).

Multiple oncogenic splicing events take place downstream of the PTBP1 and the RBFOX RBPs in glioblastoma

Investigating the percent-spliced-in (PSI) of exons having binding sites for the PTBP1 RBP, the RBFOX RBPs, or for both within the MAPP-inferred regions we found that cassette exons being co-regulated by both RBPs exhibit the most prominent differences in PSI when comparing glioblastoma to normal brain tissue (**Fig. 5a**). Importantly, the average change in exon inclusion increased with the minimum binding site probability required in our analysis to be considered a target of PTBP1 and RBFOX, respectively (**Supplementary Fig. S6**). Gene ontology (GO) analysis of genes with cassette exons that are differentially expressed in glioblastoma versus normal brain tissue and that have binding sites for the PTBP1 and the RBFOX RBPs within the corresponding regions inferred by MAPP, reveal a highly significant enrichment of genes involved in synaptic signaling (**Fig. 5b**). This suggests that cassette exons that are spliced-in within normal brain tissue due to the low and high splicing activity of the PTBP1 and the RBFOX1 RBPs, respectively, are largely involved in neuron-specific functions. Importantly, both RBPs were previously reported to regulate brain-specific micro-exon inclusion in healthy brain tissue [30], suggesting that the dysregulation of these two factors in glioblastoma leads to a pattern of exon inclusion that is less brain-specific, and probably more akin to that of an undifferentiated state, which is a hallmark of many cancers [31,32].

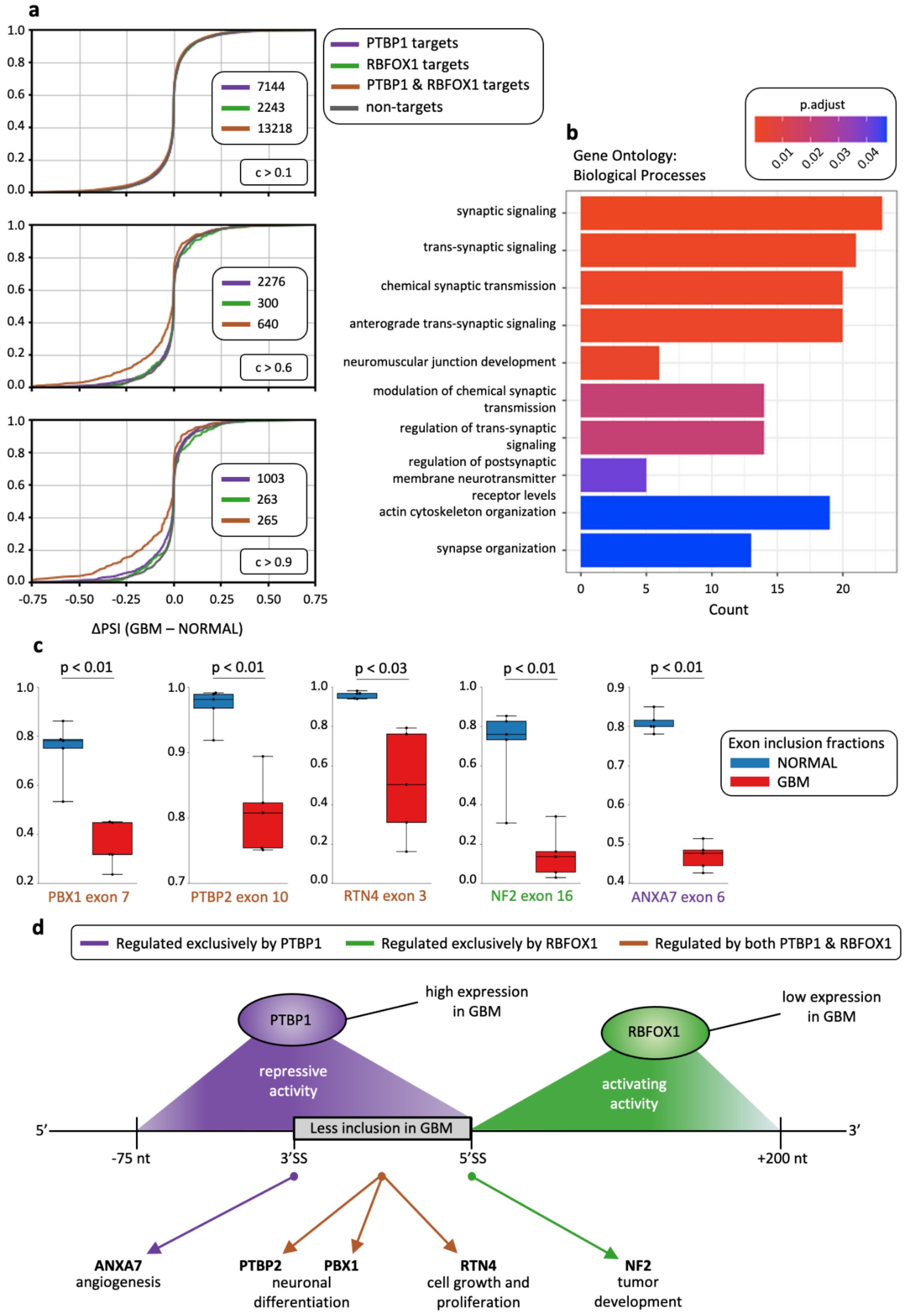


Fig. 5 | **Concerted effects of the PTBP1 and RBFOX RBPs regulate cassette exon skipping in glioblastomas. a |** Distribution of inclusion score differences observed in normal and malignant brain samples for groups of differentially-regulated exons at increasing levels of binding probability cutoffs. **b |** Top 10 enriched gene ontology terms in the category Biological Processes as inferred from cassette exons having binding sites for both RBPs (with binding probability > 0.6) and are differentially expressed in glioblastomas compared to normal brain tissue. **c |** Differences in inclusion scores for five exons which have been previously associated with cancer-related isoforms. **d |** Graphical summary of cassette exon co-regulation by the PTBP1 and RBFOX RBPs in glioblastoma and their reported downstream effects on tumour development and progression.

Interestingly, there are multiple splicing events among the cassette exons that are differentially spliced-in in normal versus malignant brain tissue and contain binding sites for the PTBP1 RBP, the RBFOX RBPs or both (**Fig. 5c; Supplementary Table S7**), thereby providing a link between the dysregulation of the PTBP1 and RBFOX RBPs in glioblastomas and the downstream effects on malignant cellular behavior.

PTBP1 is known to regulate a neuronal isoform expression program, which, amongst others, ensures the inclusion of the PBX Homeobox 1 (PBX1) exon 7 in neurons. Importantly, the opposite effect takes place in glioblastoma samples (**Fig. 5c**), in which the PTBP1 and RBFOX RBPs are up- and down-regulated, respectively (**Fig. 4c**), and MAPP infers a consistent switch in their splicing activity (**Fig. 4b**). Exon 7 contains binding sites for both RBPs within the regions inferred by MAPP (**Supplementary Table S7**). Consistently, PTBP1 was previously shown to repress PBX1 exon 7 in mouse embryonic stem cells (ESCs). Induced expression of exon 7 of the PBX1 transcription factor in ESCs activates the transcription of neuronal genes [33]. Thus, the high expression of PTBP1 in glioblastoma relative to normal brain tissue might drive neurons into a more stem cell-like, i.e. undifferentiated state, which was suggested to be the origin of glioblastoma [34], and is also one of the general hallmarks of cancer [31].

In line with this, we observe PTBP1 being upregulated, and PTBP2 being downregulated in GBM compared to normal brain samples (**Fig. 4c**). It is well known that PTBP1 is abundantly expressed in undifferentiated neural stem cells and is downregulated during neuronal differentiation, while its paralog PTBP2 is upregulated, leading to the increased inclusion of neuronal exons [35,36]. Importantly, it was shown that high expression of PTBP1 in neural stem cells and undifferentiated precursors promotes skipping of the PTBP2 exon 10, which has binding sites for both, the PTBP1 and the RBFOX RBPs (**Supplementary Table S7**). PTBP2 exon 10 is significantly less included in the GBM samples relative to normal brain tissues (**Fig. 5c**) and skipping of exon 10 was demonstrated to result in transcript isoforms that contain a premature stop codon and are thus subject to degradation by nonsense-mediated mRNA decay [37]. Consistently, we observe reduced expression of PTBP2 in GBM relative to the normal brain tissue samples (**Fig. 4c**).

Another interesting exon having binding sites in the MAPP-inferred regions for PTBP1 and RBFOX is exon 3 of the RTN4 gene (**Supplementary Table S7**). Consistent with our observations in GBM (**Fig. 5c**), high levels of PTBP1 were reported to cause exon 3 skipping and overexpression of the RTN4 splice isoform that contains exon 3 was shown to decrease cell proliferation of glioma cells, whereat skipping of RTN4 exon 3 contributes significantly to their rapid growth characteristics [38]. Even though many exons skipped in GBM have binding sites for both PTBP1 and RBFOX RBPs within the MAPP-inferred regions (**Fig. 5a; Supplementary Fig. S6**), there exist also candidates that appear to be under regulation of only one of the RBPs. For instance, exon 16 of the NF2 gene is much less included in GBM (**Fig. 5c**) and has only binding sites for RBFOX RBPs, but not PTBP1 (**Supplementary Table S7**). Previous studies have shown that there exist two major NF2 isoforms, isoform 1, which does not contain exon 16, and isoform 2, which does. Even though the exact role of these isoforms is still a matter of debate, both of them have been reported to play roles in cancer development [39,40]. Finally, there exist also exons that are regulated and have only binding sites for PTBP1, but not RBFOX RBPs (**Fig. 5a; Supplementary Fig. S6**). An example is exon 6 of the ANXA7 gene (**Supplementary Table S7**), which is skipped in GBM (**Fig. 5c**). Importantly, ANXA7 exon 6 skipping was shown to promote the progression of GBM by fostering angiogenesis [41] and thus provides another link of the MAPP-inferred regulators and the molecular properties of GBM (**Fig. 5d**).

Discussion

By binding to sequence elements in transcripts, RBPs regulate gene expression at co-transcriptional and post-transcriptional levels. In particular, they can affect both splicing and 3' end processing, key steps in the maturation of mRNAs. Additionally, the interaction of RBPs with mature mRNAs can regulate the transport, localization and translation of these mRNAs [42]. Understanding of the global and concerted effect of various RBPs on the cellular transcriptome is undoubtedly key to understanding how gene expression is dysregulated in various pathological conditions, including cancer [43,44]. In this study we presented a novel computational approach for inferring the regulatory impact of various RBPs on splicing and 3' end processing.

We have validated our method on data pertaining to proteins with well-established roles in splicing and/or polyadenylation. Specifically, the MAPP-inferred activities of the HNRNPC RBP are in line with its previously reported role in preventing exonization of cryptic *Alu* elements [20,45]. Many *Alu* elements have evolved to become cassette exons, and the potentially deleterious inclusion of these exons needs to be tightly regulated. The impact maps constructed by MAPP are fully consistent with this role of HNRNPC (**Fig. 2a**). MAPP also recovers the previously-noted position-specific regulation of exon inclusion by RBFOX RBPs (**Fig. 2b**), whereat binding of RBFOX upstream of cassette exons results in their exclusion, while binding downstream of such exons promotes their inclusion [46]. While our results are consistent with this model, they also provide higher granularity in the binding site position-dependent effects of RBFOX. Specifically, they indicate a higher impact of the downstream, inclusion-promoting sites. Furthermore, MAPP indicates that binding sites that are located further upstream in the introns also have an overall inclusion-promoting effect, consistent with an earlier report [24]. Thus MAPP provides direct and broad insight into the activity of RBP binding sites from individual RNA-seq data sets, without a need for stratifying the data or determining the binding sites with methods such as crosslinking and immunoprecipitation. MAPP's position-dependent impact maps thus enable an efficient and improved understanding of how RBPs exert their roles both globally and on individual targets.

After benchmarking MAPP on RBPs with known impact on pre-mRNA processing, we turned to the >400 RBP knock-down data sets available from the ENCODE project and revealed that multiple regulators affect exon inclusion and 3' end processing (**Fig. 3**). Once again, thanks to the sliding window approach of MAPP we found that distinct regulators differ not only in their role (which for all investigated RBPs seem to be the same in the two processes, i.e. to either enhance or repress) but also in their position specificity. For example, MAPP not only highlighted the opposite effect of two proteins, HNRNPK and PCBP1, which bind the same 'CCC' sequence, on cassette exon inclusion, but also that the distance range of their impact differs despite binding motifs obviously exist elsewhere: PCBP1 acts more broadly in the introns flanking the cassette exon, while HNRNPK acts in a more focused manner, at the exon-intron boundaries. Interestingly, our results indicate that the majority of investigated RBPs act as repressors of pre-mRNA processing. Given that the sequence elements that are involved in processing are typically short, our results could indicate that many repressors are needed to mask the many decoy processing sites across the genome [47].

Of course, a knock-down is an artificial condition, while within tissues multiple RBPs likely vary in concentration in a concerted manner, leading to more complex patterns of regulation of mRNA maturation [48]. Importantly, applying MAPP to normal brain and glioblastoma samples we uncovered that many exons are co-regulated by the PTBP1 and RBFOX RBPs (**Fig. 4**), two regulators that were reported previously to act in concert [22,30]. MAPP analysis of glioblastoma samples yields impact maps that strikingly resemble those obtained from RBP knock-down data (**Fig. 4a,b**). In glioblastomas, the splicing activating RBFOX RBPs are downregulated, whereas the PTBP1 splicing repressor is highly expressed compared to normal brain tissue (**Fig. 4c**). The usage of many cassette exons alternatively spliced in glioblastomas are repressed at their 5'SS by the highly abundant PTBP1 RBP, whereas the usage of their 3'SS lacks splicing due to the lack of RBFOX RBPs (**Fig. 4b**). Thus, the oncogenic splicing program of glioblastomas is a result of both overexpression of PTBP1 and the downregulation of RBFOX RBPs compared to healthy brain tissues (**Fig 5**). Notably, multiple of the cassette exons that are differentially spliced in glioblastoma compared to normal brain tissue and that have binding sites for the PTBP1 RBP, the RBFOX RBPs or both, were previously validated experimentally to drive cells into a more malignant state (**Fig. 5c,d**). Examples are skipping of PBX1 exon 7 and PTBP2 exon 10. Both of these splicing events were reported to contribute to less differentiated cellular states [33,35,36], which was suggested to drive glioblastoma development [34] and is a general hallmark of cancers [31]. Further, skipping of the RTN4 exon 3 was demonstrated to increase cell proliferation of glioma cells [38] and reduced inclusion of exon 6 of the ANXA7 gene was reported to promote glioblastoma progression [41]. Besides the already experimentally validated oncogenic splicing events, among the large number of cassette exon skipping events taking place in glioblastomas (**Fig. 5a**), there are most probably further candidates that remain to be characterized towards their involvement in brain tumour development and progression. Importantly, the identification of RBPs that broadly impact mRNA processing in specific conditions and in particular in individual cancers is highly relevant, as it can provide novel entry points for the development of therapies. Targeting of mRNAs and mRNA-RBP interactions with antisense oligonucleotides [49,50] or small molecules [51] holds much promise for medical applications. As MAPP is a fully automated workflow, the task of identifying regulators of pre-mRNA processing from novel RNA-seq datasets is considerably facilitated.

Other groups have investigated binding site location-dependent effects of RBPs, specifically proposing the concept of "RNA maps" [52], which summarize the density of RBP binding sites in the vicinity of various types of landmarks (exon and transcript boundaries), where RBPs exert regulatory roles. For instance, binding of the Nova RBP upstream of a cassette exon is associated with the skipping of that exon, while the binding downstream of the cassette exon is associated with the exon inclusion. The impact maps that MAPP constructs provide complementary information. They do not rely on direct information about the location of the binding site of the RBP (usually obtained with CLIP), nor on specific thresholds for defining regulated events such as exon inclusions. Rather, MAPP makes use of the quantitative information in the inclusion level of each exon or PAS as well as in the number of predicted binding sites in the vicinity of these exons. As a result, MAPP provides quantitative information about the impact of motifs on RNA processing, circumventing issues regarding the coverage of the binding sites by CLIP in targets with different levels of expression. Also interesting to note is the increasing use of massively parallel assays for exploring the dependence of RNA processing on specific motifs [53]. These provide information more analogous to MAPP's impact maps, but are limited to a small number of conditions, a small number of targets and regions within these targets, and have been so far used to characterize general principles of RNA processing. In contrast, MAPP's utility comes primarily in exploring a broad range of conditions and identifying condition/tissue-specific regulators. Thus, MAPP extends the RNA biologist's toolbox to enable the functional characterization of RBP-RNA interactions and their consequences at an increasing level of detail.

In conclusion we developed a powerful computational approach to identify regulators of splicing and 3' end processing, which are frequently coordinated. MAPP has been developed using modern principles of high-quality scientific software engineering, facilitating further development by a broad community of developers.

Methods

Datasets

We validated our method on publicly available RNA-seq data with perturbed levels of RBPs with known impact on splicing, and, to some extent, polyadenylation. The full list of samples' and records' IDs is included as **Supplementary Table S1**. Similarly, **Supplementary Table S2** lists all RNA-Seq data sets related to RBP knock-downs we have obtained from the ENCODE project. To apply MAPP we require that samples meet minimal criteria of quality. For example, we require a sufficiently high Transcript Integrity Number (>50, typically > 70) [54], high proportion of uniquely mapped reads (>0.95), high proportion of high-quality mapped reads (>0.85; following RNA-SeQC's documentation: proportion of properly paired reads with less than 6 mismatched bases and a perfect mapping quality out of all mapped reads), low level of rRNA contamination (<0.05) and low proportion of reads mapped to intergenic regions (<0.1), as reported by RNA-SeQC [55]. BAM files with mapped RNA-seq reads of normal and tumor sample pairs from TCGA were obtained from the Genomic Data Commons (GDC) data portal [56]. The selection of normal-tumor pairs from glioblastoma data was done as described previously (**Supplementary Table S3**) [6]. Additional transcriptomic alignments were generated by first unmapping and then re-aligning RNA-Seq reads, utilizing Samtools and STAR with proper command line options.

MAPP

MAPP, standing for Motif Activity on Pre-mRNA Processing, is implemented as a modular snakemake workflow [57] with distinct standalone sub-workflows dedicated to separate functionalities. These are: RNA-Seq data preprocessing, selection of cassette exons, selection of tandem poly(A) sites, quantification of exon inclusion, quantification of poly(A) site usage, generation of motif count matrices (PWMs/k-mers) in each window around each site, the MAEI model (splicing), the KAPACv2.0 model (polyadenylation), and the summary of results. Each of these modules is described in detail in the Supplementary Methods. MAPP supports two distinct software technologies: Conda environments [58] and Singularity containers [59].

MAEI

MAEI, which stands for Motif Activity on Exon Inclusion, is a novel model designed to infer the impact of short sequence motifs on the differential inclusion of cassette exons. In order to prevent the confounding effect of sites located within intronic regions, by default MAPP fits activities for windows of 50 nt length and considering only exons that are at least 50 nt in length. As input the MAEI model uses, for each exon *e*, the expression levels of transcripts including and excluding the exon across a set of samples *s*, together with a matrix *N* whose entries $N_{e,m}$ correspond to the motif counts of each motif *m* in a window around the mRNA processing sites of interest, i.e. 5'SS or 3'SS, for each exon *e*. The motifs can either be specified as PWMs or k-mers. We model the inclusion fractions $f_{e,s}$ (i.e. the fraction of transcripts including the cassette exon *e* among transcripts for which *e* was included in the pre-mRNA, see **Supplementary Methods**) of every exon *e* in every sample *s* using a logistic function: $\Theta_{e,s} = \frac{e^{X_{e,s}}}{1+e^{X_{e,s}}}$, where $X_{e,s} = b_s + c_e + N_{e,m} * A_{m,s}$ is a linear function of the model parameters: b_s - the baseline inclusion rate of all exons in sample *s*, c_e - the baseline inclusion rate of exon *e* in all samples and $A_{m,s}$ - the 'activity' of a motif *m* in a sample s. $N_{e,m}$ denotes the number of binding sites of motif *m* in the proximity of exon *e*, which is either given by the sum of site probabilities predicted with the PWM or the raw k-mer counts). We fit this logistic regression model using a Bayesian approach resulting in inferred motif activities $A_{m,s}$ with corresponding error bars $\sigma_{m,s}$ and finally obtain for each motif *m* in sample *s* a z-score $Z_{m,s} = \frac{A_{m,s}}{\sigma_{m,s}}$.

The activity z-scores are then presented visually on the impact maps. See the **Supplementary Methods** for more details on all calculations.

In order to distinguish motifs with statistically significant z-scores from those with z-scores from a Gaussian background distribution we use a Gaussian mixture model to renormalize the z-scores and transform them into p-values from a standard normal distribution. Statistical significance is then finally assessed upon Bonferroni-correction of these p-values. Again, we refer the reader to the **Supplementary Methods** for more details on the procedure.

KAPACv2.0

KAPACv2.0, standing for K-mer Activity on PolyAdenylation site Choice version 2.0, implements a more general version of our previously published KAPAC tool [6]. KAPACv2.0

models genome-scale changes in 3' end usage to infer sequence motifs that can explain 3' end site usage across samples. In contrast to KAPAC, KAPACv2.0 can use both binding sites predicted with position-dependent weight matrices (PWMs) as well as k-mer counts. Also, while the first version of KAPAC was designed to run on sample contrasts, such as knockdown versus control samples, KAPACv2.0 does not require contrasts but can be applied to any set of samples, such as different tissues or a time series. First, we define the relative usage of poly(A) site *p* in sample *s* as $u_{p,s}$. KAPACv2.0 then models the relative usage $u_{p,s}$ with respect to the mean of all samples as a linear function of the occurrence of PWM binding sites or k-mer counts and the unknown 'activity' of these PWMs / k-mers: $log_2\left(u_{p,s}\right) = N_{p,k} * A_{k,s} + c_p + c_{s,e} + \varepsilon$, where $N_{p,k}$ is the number of binding sites (predicted with the PWM or by k-mer counting) around poly(A) site *p,* c_p is the mean log_2 relative usage of poly(A) site *p* across all samples, $c_{s,e}$ is the mean log_2 relative usage of the poly(A) site from exon *e* in sample s and ε is the residual error. Finally, $A_{k,s}$ is the activity of the PWM / k-mer *k* in sample *s,* which determines how much the PWM / k-mer contributes to the relative usage of the poly(A) site. KAPACv2.0 calculates for every PWMs or k-mer, respectively, a z-score z = $A_{k,s}$ / $\sigma_{k,s}$, whereas $\sigma_{k,s}$ are the fitting errors of the activities $A_{k,s}$. Background correction and ranking of PWMs / k-mers is done as described for the MAEI approach above (see **Supplementary Methods** and ref. [6] for further information).

Curation of PWMs of RBPs binding motifs

ATtRACT is a publicly available database of RNA-binding proteins and associated motifs [60]. On 20 August 2021 we downloaded the zip file containing all available RBP motifs in the format of position-dependent weight matrices (giving the probability of observing any of the four bases at each position of the binding site) as well as their corresponding metadata (ATtRACT_db.txt). From the ATtRACT_db.txt we first selected motifs annotated with the species *Homo sapiens* (3256 records) and from these only those that corresponded to *wild-type* proteins ("Mutated" field having the value "no", 3178 records). We next selected only one of the records that had the same gene ID, PWM ID and experiment description (where for experiment description, records that contained the word 'SELEX' were considered as having the same description). This procedure resulted in 1120 records. Next we clustered records for which the entries in the PWMs (position-dependent frequencies of nucleotide occurrence) were identical. If the cluster with identical PWM entries contained multiple RBPs, we discarded them all, as we could not unambiguously assign the PWM to one RBP.

If the cluster contained multiple records for the same RBP, we kept only one of them. This step left 523 records. We further determined the length of the core motif for each PWM, that is, the longest motif such that the first and last position had a non-zero information content and discarded those records where core motifs were not in the range of 4 to 7 nucleotides. This step left 346 records. Finally, for each PWM we calculated the total motif entropy and discarded those that were too degenerate (with an entropy higher than 10). This procedure yielded 344 PWMs for the MAPP analyses.

Coverage profiles of RNA-binding proteins

To gain additional confidence in MAPP's inferences, we constructed coverage profiles for distinct RNA-binding proteins based on CLIP data in HepG2 and K562 cells, publicly available as a part of the ENCODE project (experiment IDs: ENCSR550DVK, ENCSR987FTF, ENCSR384KAN, ENCSR249ROI, ENCSR756CKJ, ENCSR981WKN). For **Fig. 2a**,**b** and **4a** we have used the experiments conducted in HepG2 cells (similar plots for K562 cells can be found in **Supplementary Figs. S4 and S5**) we selected the group of the top 200 targets with the highest change in alternative splicing as well as alternative polyadenylation into the expected direction based on the average quantified exon inclusion fraction and poly(A) site usage, respectively. We have extended the margins around these sites so that the eCLIP analysis matches the regions covered by our MAPP sliding windows. For every RNA processing site separately we have calculated foreground/background ratios of library-size-normalized position-wise CLIP read coverages (foreground being CLIP reads from the RBP pulldown experiment and background being the corresponding control pulldown experiment). We have plotted the position-wise mean ratio over all sites (smoothened by the -5/+5nt of each position). Additionally to the target set we selected a group of 1,000 sites with the least change in RNA processing and treated them as background to estimate coverage profiles for non-targets. From this set there were randomly sampled 200 non-targets for 100 times, each time following the same procedure as described above for the non-random sites in order to obtain 100 background coverage profiles. These random profiles were used to plot the (smoothened) per-position mean of means together with a confidence boundary which reflects the per-position standard deviation of the means. The data processing notebook is available in the supplementary data (see **Data availability** section).

Selection of ENCODE experiments, reported motifs and k-mers

We have downloaded and analyzed RNA-seq samples linked to 472 knock-down experiments of RBPs, publicly available as a part of the ENCODE project; 16 of these did not pass the quality-control step as defined in the "**Datasets**" section. We used the remaining 456 data sets for the analysis shown in **Fig. 3**. Briefly, each ENCODE experiment has been analyzed with MAPP in both PWM-based and k-mer-based approaches. For each of the knock-down experiments we selected the lowest rank of any ATtRACT PWM associated with the perturbed RBP for which MAPP found statistically significant impact on any of the signals (statistical significance annotated with the "abs" strategy, please see **Supplementary Methods**). Such obtained "PWM rank" is the key by which the results table is sorted in descending order. For only 12 ENCODE experiments we found that the PWM associated with the perturbed RBP had a rank of maximum 5 (out of 344 curated PWMs). For these, we report the PWM ranks and their impact maps, as inferred by MAPP. We then checked whether the appropriate motif is also recovered in the k-mer mode. For this, we selected the k-mer with the highest overall statistically significant activity z-score, averaged over all of the processing sites (labeled as "1st ranked k-mer") from each experiment. Alongside with the previously described PWM-based results we report the 1st ranked k-mer and its impact map. Data processing scripts are available in the supplementary data (see **Data availability** section).

Data availability

The results generated in this study are available in the supplementary data, which are accessible from Zenodo under doi: https://doi.org/10.5281/zenodo.5789986. The accession numbers for used datasets are available from **Supplementary Tables S1 and S2**.

Code availability

The MAPP code is available at https://github.com/gruber-sciencelab/MAPP under the Apache 2.0 open-source license.

Author contributions

M.B. and A.G. designed the MAPP pipeline, M.B. implemented the pipeline with contributions from A.G. I.K. helped testing the pipeline and M.B. prepared the GitHub repository. E.N. and M.B. designed the MAEI model and background correction and M.B. implemented it. A.G. designed and implemented the KAPACv2.0 model. R.S. developed the TPA module of the MAPP pipeline. M.B. and A.G. analyzed the data and created the figures. M.B. created the supplementary data record. I.K. carried out the GO analysis. A.G. conceived the study. A.G. and M.Z. designed and supervised the project. A.G., M.B. and M.Z wrote the manuscript.

Acknowledgements

We express our wholehearted gratitude to the sciCORE facility of the University of Basel and to Stefan Gerlach from the Scientific Compute Cluster of the University of Konstanz (SCCKN) for maintaining the high-performance computing clusters where we have carried out all the computations and analyses. This work was supported by the Swiss National Science Foundation (grant 310030_189063 to M.Z.). M.B. was a recipient of a PhD scholarship “Fellowship for Excellence” of the Biozentrum, University of Basel.

References

1. Fredericks, A. M., Cygan, K. J., Brown, B. A. & Fairbrother, W. G. RNA-Binding Proteins: Splicing Factors and Disease. *Biomolecules* **5**, 893–909 (2015).
2. Zheng, D. & Tian, B. RNA-binding proteins in regulation of alternative cleavage and polyadenylation. *Adv. Exp. Med. Biol.* **825**, 97–127 (2014).
3. Meng, Q. *et al.* Signaling-dependent and coordinated regulation of transcription, splicing, and translation resides in a single coregulator, PCBP1. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 5866–5871 (2007).
4. Ji, X., Wan, J., Vishnu, M., Xing, Y. & Liebhaber, S. A. αCP Poly(C) binding proteins act as global regulators of alternative polyadenylation. *Mol. Cell. Biol.* **33**, 2560–2573

(2013).

5. Gruber, A. J. *et al.* 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 Res.* **26**, 1145–1159 (2016).
6. Gruber, A. J. *et al.* Discovery of physiological and cancer-related regulators of 3’ UTR processing with KAPAC. *Genome Biol.* **19**, 44 (2018).
7. Chang, S.-H. *et al.* ELAVL1 regulates alternative splicing of eIF4E transporter to promote postnatal angiogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 18309–18314 (2014).
8. Dai, W., Zhang, G. & Makeyev, E. V. RNA-binding protein HuR autoregulates its expression by promoting alternative polyadenylation site usage. *Nucleic Acids Res.* **40**, 787–800 (2012).
9. Tollervey, J. R. *et al.* Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat. Neurosci.* **14**, 452–458 (2011).
10. Rot, G. *et al.* High-Resolution RNA Maps Suggest Common Principles of Splicing and Polyadenylation Regulation by TDP-43. *Cell Rep.* **19**, 1056–1067 (2017).
11. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15**, 829–845 (2014).
12. Siddaway, R. *et al.* Splicing is an alternate oncogenic pathway activation mechanism in glioma. *Nat. Commun.* **13**, 588 (2022).
13. Larionova, T. D., Kovalenko, T. F., Shakhparonov, M. I. & Pavlyukov, M. S. The Prognostic Significance of Spliceosomal Proteins for Patients with Glioblastoma. *Dokl. Biochem. Biophys.* **503**, 71–75 (2022).
14. Van Nostrand, E. L. *et al.* Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. *Genome Biol.* **21**, 90 (2020).

15. Masuda, A. *et al.* Position-specific binding of FUS to nascent RNA regulates mRNA length. *Genes Dev.* **29**, 1045–1057 (2015).

16. Lee, S. *et al.* ELAV/Hu RNA binding proteins determine multiple programs of neural alternative splicing. *PLoS Genet.* **17**, e1009439 (2021).

17. Cieniková, Z., Damberger, F. F., Hall, J., Allain, F. H.-T. & Maris, C. Structural and mechanistic insights into poly(uridine) tract recognition by the hnRNP C RNA recognition motif. *J. Am. Chem. Soc.* **136**, 14536–14544 (2014).

18. Liu, N. *et al.* N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature* **518**, 560–564 (2015).

19. König, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* **17**, 909–915 (2010).

20. Zarnack, K. *et al.* Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* **152**, 453–466 (2013).

21. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).

22. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).

23. Damianov, A. *et al.* Rbfox Proteins Regulate Splicing as Part of a Large Multiprotein Complex LASR. *Cell* **165**, 606–619 (2016).

24. Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat. Struct. Mol. Biol.* **20**, 1434–1442 (2013).

25. Ule, J. *et al.* An RNA map predicting Nova-dependent splicing regulation. *Nature* **444**, 580–586 (2006).

26. Královičová, J. *et al.* PUF60-activated exons uncover altered 3′ splice-site selection by germline missense mutations in a single RRM. *Nucleic Acids Res.* **46**, 6166–6187 (2018).

27. Müller-McNicoll, M., Rossbach, O., Hui, J. & Medenbach, J. Auto-regulatory feedback by RNA-binding proteins. *J. Mol. Cell Biol.* **11**, 930–939 (2019).

28. Gueroussov, S. *et al.* An alternative splicing event amplifies evolutionary differences between vertebrates. *Science* **349**, 868–873 (2015).

29. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

30. Li, Y. I., Sanchez-Pulido, L., Haerty, W. & Ponting, C. P. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. *Genome Res.* **25**, 1–13 (2015).

31. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **12**, 31–46 (2022).

32. Cao, Y. Tumorigenesis as a process of gradual loss of original cell identity and gain of properties of neural precursor/progenitor cells. *Cell Biosci.* **7**, 61 (2017).

33. Linares, A. J. *et al.* The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation. *Elife* **4**, e09268 (2015).

34. Friedmann-Morvinski, D. *et al.* Dedifferentiation of neurons and astrocytes by oncogenes can induce gliomas in mice. *Science* **338**, 1080–1084 (2012).

35. Li, Q. *et al.* The splicing regulator PTBP2 controls a program of embryonic splicing required for neuronal maturation. *Elife* **3**, e01201 (2014).

36. Keppetipola, N., Sharma, S., Li, Q. & Black, D. L. Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2. *Crit. Rev. Biochem. Mol. Biol.* **47**, 360–378 (2012).

37. Kim, J.-H. *et al.* SON drives oncogenic RNA splicing in glioblastoma by regulating PTBP1/PTBP2 switching and RBFOX2 activity. *Nat. Commun.* **12**, 5551 (2021).

38. Cheung, H. C. *et al.* Splicing factors PTBP1 and PTBP2 promote proliferation and migration of glioma cell lines. *Brain* **132**, 2277–2288 (2009).

39. Sherman, L. *et al.* Interdomain binding mediates tumor growth suppression by the NF2

gene product. *Oncogene* **15**, 2505–2509 (1997).

40. Zoch, A. *et al.* Merlin Isoforms 1 and 2 Both Act as Tumour Suppressors and Are Required for Optimal Sperm Maturation. *PLoS One* **10**, e0129151 (2015).

41. Ferrarese, R. *et al.* Lineage-specific splicing of a brain-enriched alternative exon promotes glioblastoma progression. *J. Clin. Invest.* **124**, 2861–2876 (2014).

42. García-Mauriño, S. M. *et al.* RNA Binding Protein Regulation and Cross-Talk in the Control of AU-rich mRNA Fate. *Front Mol Biosci* **4**, 71 (2017).

43. Qi, F. *et al.* Significance of alternative splicing in cancer cells. *Chin. Med. J.* **133**, 221–228 (2020).

44. Jain, B. P. The role of alternative polyadenylation in cancer progression. *Gene Reports* **12**, 1–8 (2018).

45. Attig, J. *et al.* Splicing repression allows the gradual emergence of new Alu-exons in primate evolution. *Elife* **5**, (2016).

46. Sun, S., Zhang, Z., Fregoso, O. & Krainer, A. R. Mechanisms of activation and repression by the alternative splicing factors RBFOX1/2. *RNA* **18**, 274–283 (2012).

47. Gruber, A. J. & Zavolan, M. Reply to ‘A different perspective on alternative cleavage and polyadenylation’. *Nature reviews. Genetics* vol. 21 63–64 (2020).

48. Dassi, E. Handshakes and Fights: The Regulatory Interplay of RNA-Binding Proteins. *Front Mol Biosci* **4**, 67 (2017).

49. Roberts, T. C., Langer, R. & Wood, M. J. A. Advances in oligonucleotide drug delivery. *Nat. Rev. Drug Discov.* **19**, 673–694 (2020).

50. Bennett, C. F. Therapeutic Antisense Oligonucleotides Are Coming of Age. *Annu. Rev. Med.* **70**, 307–321 (2019).

51. Desterro, J., Bak-Gordon, P. & Carmo-Fonseca, M. Targeting mRNA processing as an anticancer strategy. *Nat. Rev. Drug Discov.* **19**, 112–129 (2020).

52. Ule, J. *et al.* CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**,

1212–1215 (2003).

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

54. Wang, L. *et al.* Measure transcript integrity using RNA-seq data. *BMC Bioinformatics* **17**, 58 (2016).

55. DeLuca, D. S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).

56. Evans, B. J. Genomic Data Commons. *Governing Medical Knowledge Commons* 74–101 Preprint at https://doi.org/10.1017/9781316544587.005 (2017).

57. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Res.* **10**, 33 (2021).

58. Anaconda Software Distribution. *Anaconda Documentation* Preprint at https://docs.anaconda.com/ (2020).

59. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PLoS One* **12**, e0177459 (2017).

60. Giudice, G., Sánchez-Cabo, F., Torroja, C. & Lara-Pezzi, E. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database* **2016**, (2016).

Supplementary Materials

Maciej Bak[1,2*], Erik van Nimwegen[1,2], Ian U. Kouzel[3], Ralf Schmidt[1,2], Mihaela Zavolan[1,2] and Andreas J. Gruber[3*§]

1 Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
2 Biozentrum, University of Basel, 4056 Basel, Switzerland
3 Department of Biology, University of Konstanz, D-78464 Konstanz, Germany
∗ These authors made equal first author contributions
§ To whom correspondence should be addressed: gruber@uni-konstanz.de

1 Supplementary Figures

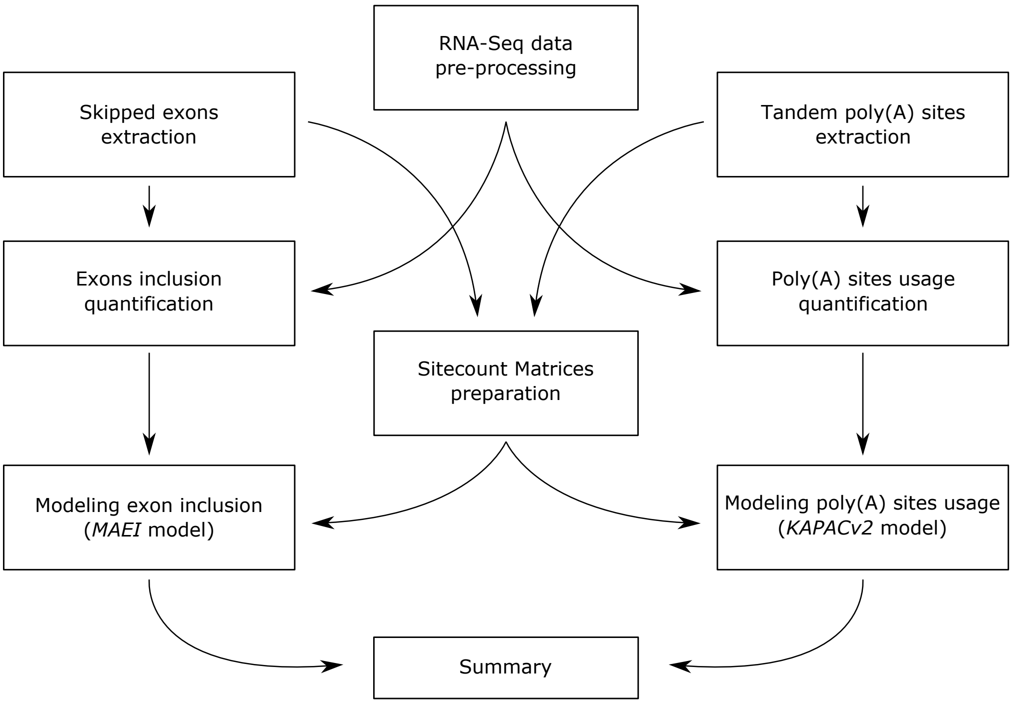

RNA-Seq data pre-processing
Skipped exons extraction
Tandem poly(A) sites extraction
Exons inclusion quantification
Poly(A) sites usage quantification
Sitecount Matrices preparation
Modeling exon inclusion (*MAEI* model)
Modeling poly(A) sites usage (*KAPACv2* model)
Summary

Figure S1. High-level overview of MAPP. The pipeline may be decomposed into nine separate functional sub-modules, each of which can be executed individually on its own. Modules on the left-hand side are related to the analysis of alternative splicing, whereas the right-hand side is dedicated to alternative polyadenylation. The modules in the middle preprocess and prepare the RNA-seq data for the splicing and the polyadenylation modules (top), create the sitecount matrixes (middle) and summarize the results in the form of a report (bottom). Accordingly, MAPP starts from three independent entry points.

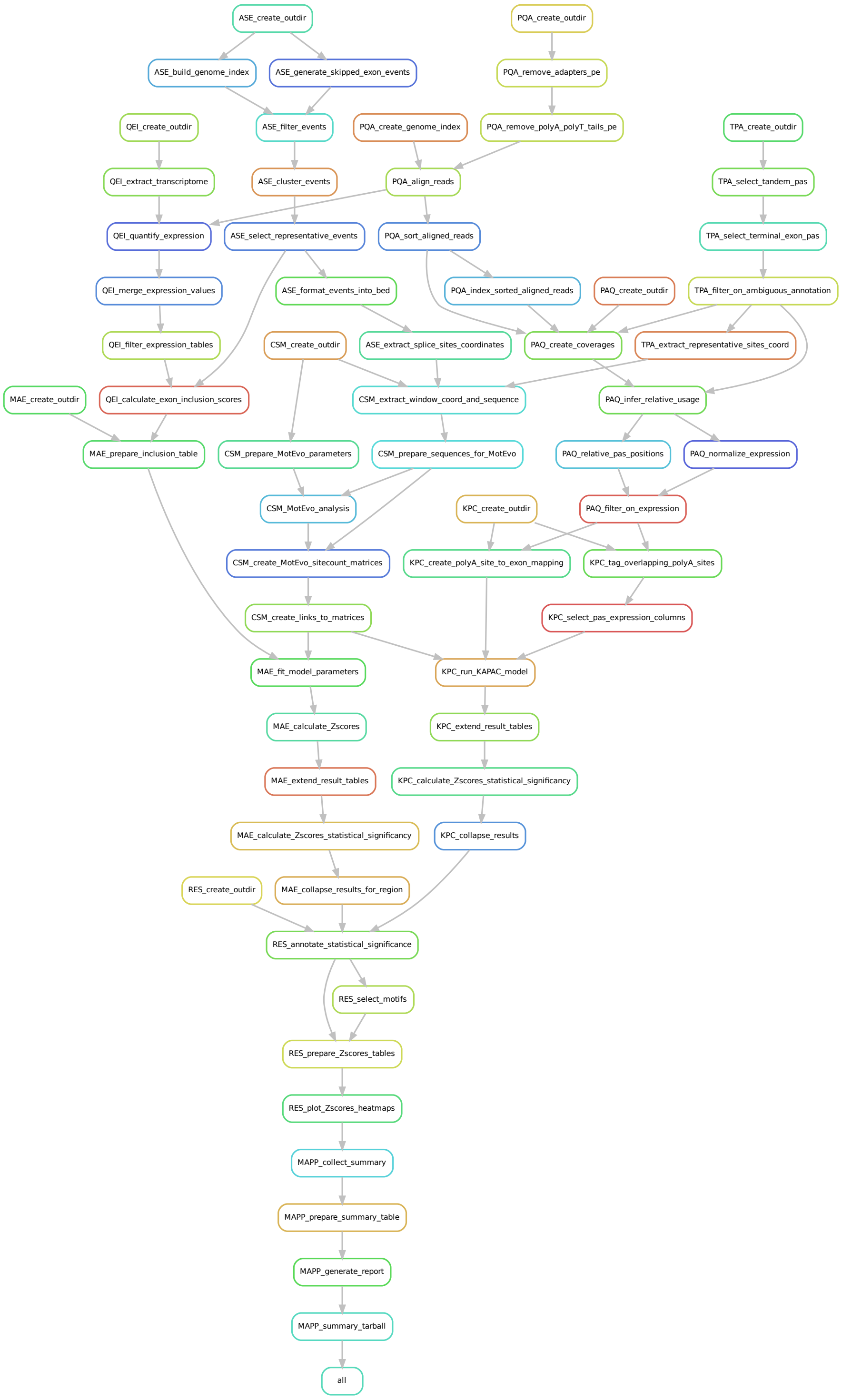
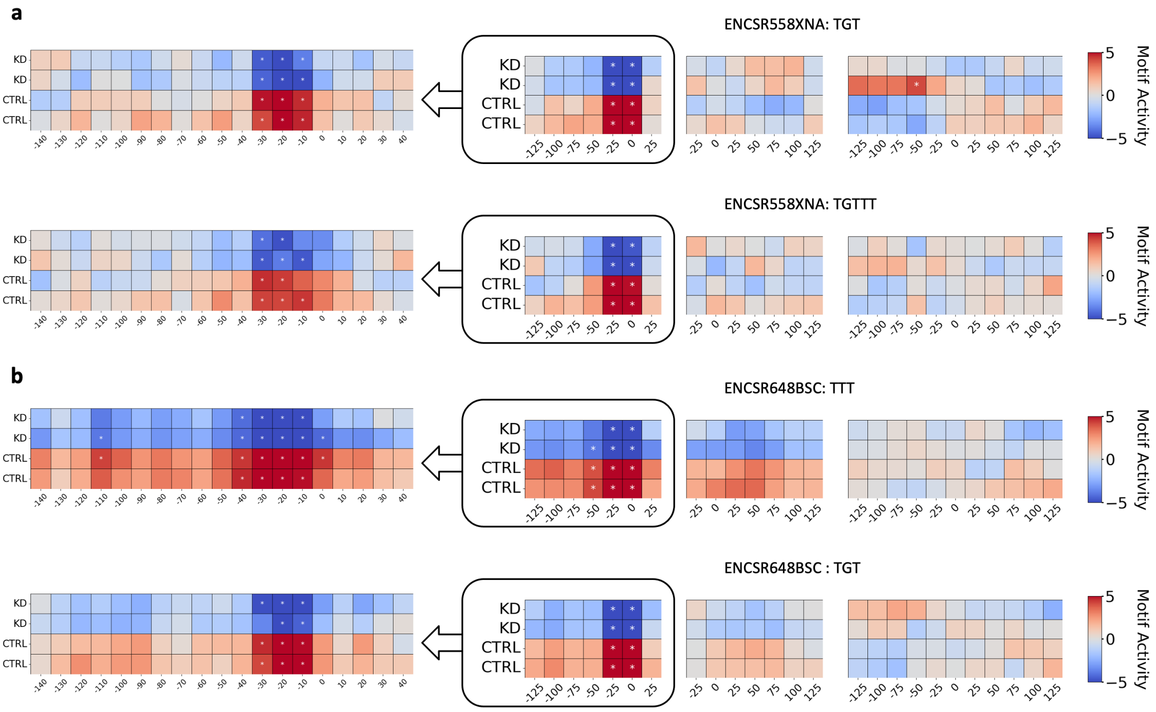

ASE_create_outdir
PQA_create_outdir
ASE_build_genome_index
ASE_generate_skipped_exon_events
PQA_remove_adapters_pe
QEI_create_outdir
ASE_filter_events
PQA_create_genome_index
PQA_remove_polyA_polyT_tails_pe
TPA_create_outdir
QEI_extract_transcriptome
ASE_cluster_events
PQA_align_reads
TPA_select_tandem_pas
QEI_quantify_expression
ASE_select_representative_events
PQA_sort_aligned_reads
TPA_select_terminal_exon_pas
QEI_merge_expression_values
ASE_format_events_into_bed
PQA_index_sorted_aligned_reads
PAQ_create_outdir
TPA_filter_on_ambiguous_annotation
QEI_filter_expression_tables
CSM_create_outdir
ASE_extract_splice_sites_coordinates
PAQ_create_coverages
TPA_extract_representative_sites_coord
MAE_create_outdir
QEI_calculate_exon_inclusion_scores
CSM_extract_window_coord_and_sequence
PAQ_infer_relative_usage
MAE_prepare_inclusion_table
CSM_prepare_MotEvo_parameters
CSM_prepare_sequences_for_MotEvo
PAQ_relative_pas_positions
PAQ_normalize_expression
CSM_MotEvo_analysis
KPC_create_outdir
PAQ_filter_on_expression
CSM_create_MotEvo_sitecount_matrices
KPC_create_polyA_site_to_exon_mapping
KPC_tag_overlapping_polyA_sites
CSM_create_links_to_matrices
KPC_select_pas_expression_columns
MAE_fit_model_parameters
KPC_run_KAPAC_model
MAE_calculate_Zscores
KPC_extend_result_tables
MAE_extend_result_tables
KPC_calculate_Zscores_statistical_significancy
MAE_calculate_Zscores_statistical_significancy
KPC_collapse_results
RES_create_outdir
MAE_collapse_results_for_region
RES_annotate_statistical_significance
RES_select_motifs
RES_prepare_Zscores_tables
RES_plot_Zscores_heatmaps
MAPP_collect_summary
MAPP_prepare_summary_table
MAPP_generate_report
MAPP_summary_tarball
all

Figure S2. Automatically generated snakemake rule graph of the MAPP pipeline. Rules of distinct sub-modules of the workflow are prefixed with a three-letter code. Four additional rules are added at the end of the workflow (prefix: *MAPP*) which generate a compressed HTML-formatted report of the MAPP result.


ENCSR558XNA: TGT

ENCSR558XNA: TGTTT

ENCSR648BSC: TTT

ENCSR648BSC : TGT

Figure S3. Impact maps of the top two most significant k-mers reported by MAPP after analyzing two PUF60 knock-down experiments from the ENCODE project. (a) PUF60 knock-down in the K5643 cell line and (b) in the HepG2 cell line. Right side: Results based on the windows used by MAPP per default. Left side: In order to obtain more fine-grained insight into the position-specific activity of the RBP we rerun MAPP using more narrow sliding windows around the 3'SS (20nt in length, slided by 10nt).

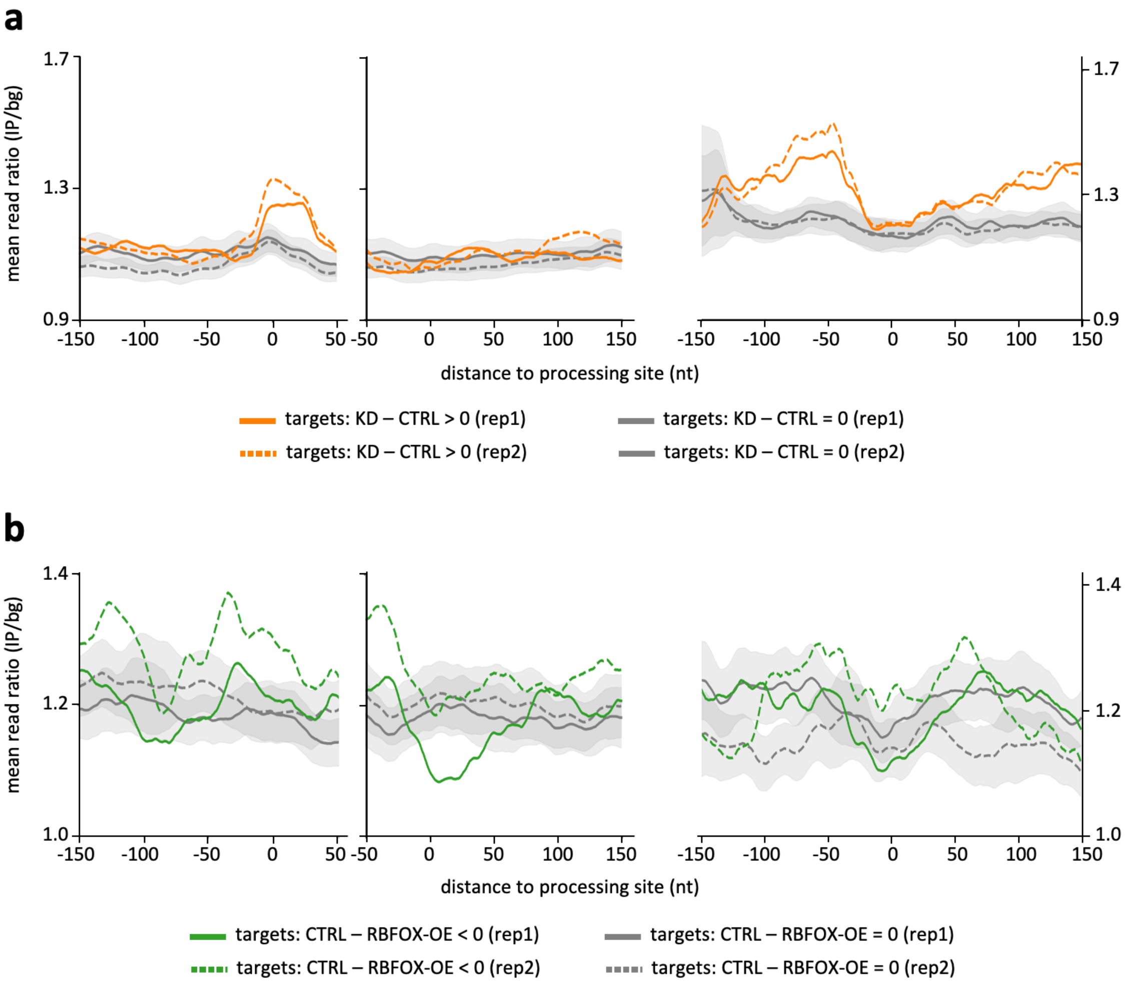


Figure S4. eCLIP-based coverage profiles for distinct RNA-binding proteins as in Fig 2 of the main text but in K562 cell line: (a) HNRNPC eCLIP (b) RBFOX2 eCLIP.

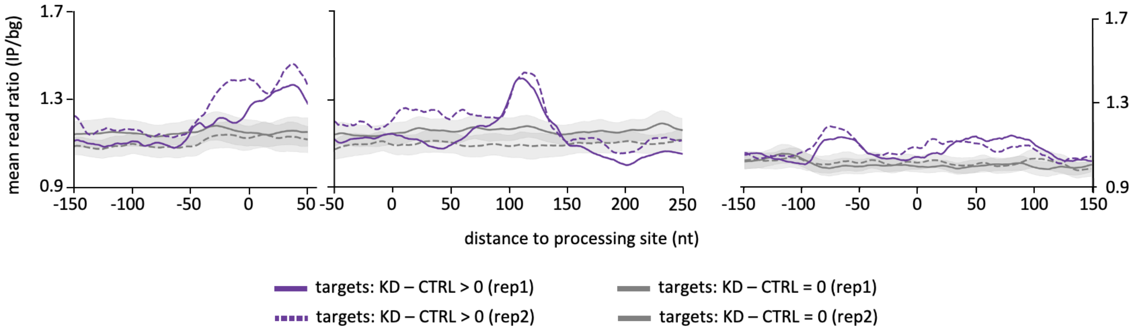


Figure S5. eCLIP-based coverage profiles for PTBP1 RBP as in Fig 4 of the main text but in K562 cell line.

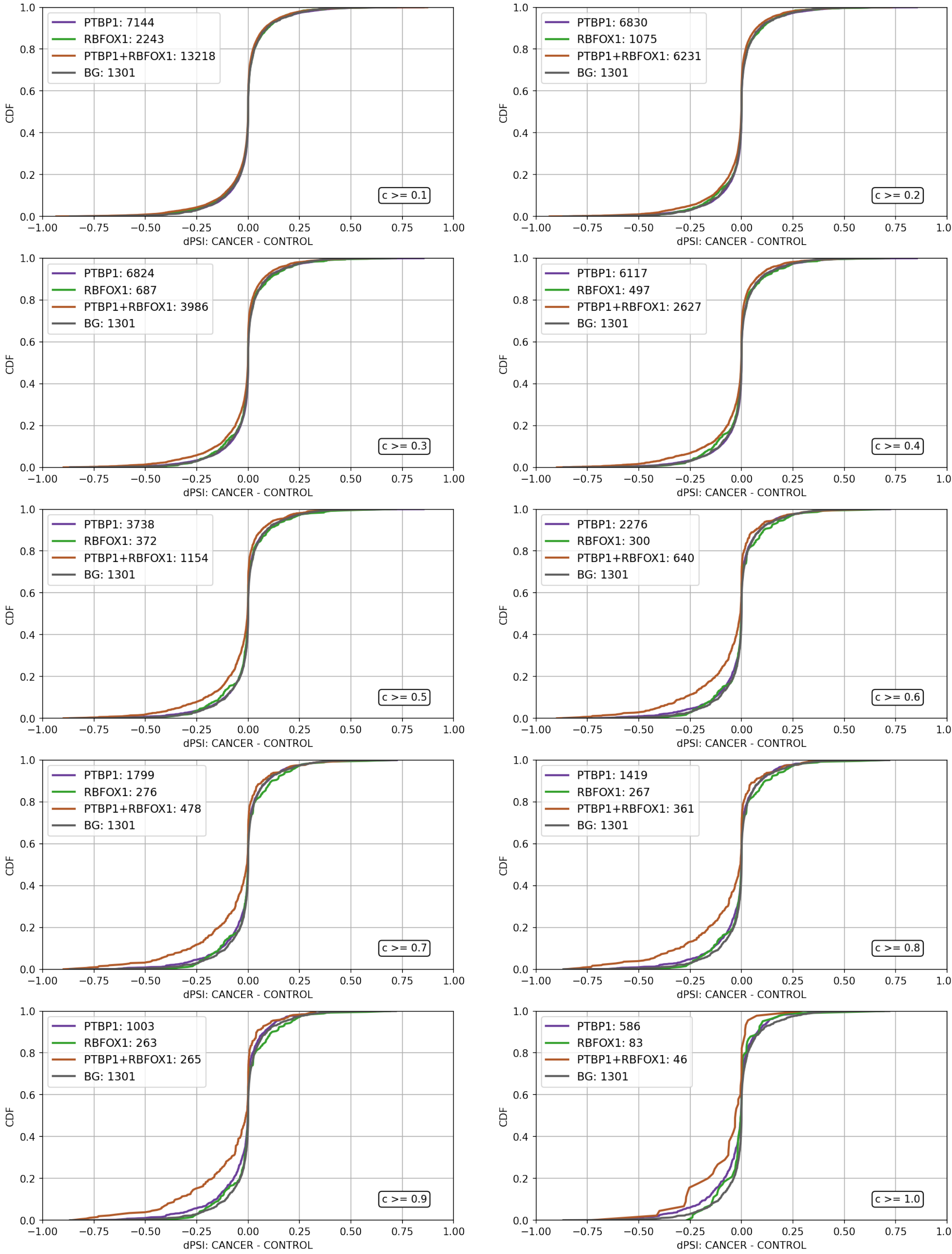


Figure S6. CDF plots of the differences in average percent-spliced-in (dPSI) scores between conditions in the glioblastoma dataset. Distinct colors denote cassette exons annotated as regulated by PTBP1 and/or RBFOX. Targets are assigned based on the binding probability inferred with MotEvo [1] in the core region of regulation: -75nt upstream to +25nt around 3'SS for PTBP1 and +25nt to +125nt downstream of 5'SS for RBFOX. CDFs are presented for increasing binding probability cutoffs *c*. The number of cases is indicated in the legend.

2 Supplementary Methods

2.1 Data preprocessing

Processing of raw sequencing reads in FASTQ/FASTA format starts with two consecutive runs of cutadapt tool [2]. First, the sequences of adapters are removed and subsequently we trim poly(A) tails. Following that the reads are mapped to both genome and transcriptome using STAR aligner [3], which also - based on the provided resources - builds a genomic index prior aligning the reads. Obtained alignments in BAM format were then sorted and indexed with samtools [4]. If appropriate parameters are set MAPP will also carry out quality control of the data and, prior to the downstream analysis, filter-out such RNA-Seq samples which do not meet specified criteria. For the quality analysis we use metrics provided by RNA-SeQC [5] and TIN score calculated as in RSeQC package [6]. Additionally all samples are handled over to FastQC [7] which generates additional per-sample summary report.
Snakemake rules which belong to this module are marked with a three-letter namespace: *PQA*.

2.2 Selection of cassette exons

We select a set of cassette (also known as 'skipped') exons based solely on the standard ENSEMBL genomic annotation (version: hg38). We first run SUPPA2 [8] to generate all skipped exon events and then filter them for minimal length sufficient for the downstream analyses. We focus only on records annotated as *protein_coding*. Obtained events are further clustered according to a mutual coverage dissimilarity measure:

$$d(e_1, e_2) = 1 - min(\frac{len(e_1 \cap e_2)}{len(e_1)}, \frac{len(e_1 \cap e_2)}{len(e_2)}) \tag{1}$$

Where e_1 and e_2 denote two exons and with $\cap$ we take only the overlap between then. We applied a hierarhical clustering with a maximum linkage strategy and set 0.05 as a cluster linkage cutoff parameter. From every cluster we selected one event with the highest number of distinct transcripts that support it. Such an event we call a representative exon. For all representative exons we keep the information which transcripts of a given gene include it as well as the list of all transcripts for a given gene (both provided by SUPPA2). We also save the coordinates of 3' and 5' splice-sites of these representatives.
Snakemake rules which belong to this module are marked with a three-letter namespace: *ASE*.

2.3 Extraction of tandem poly(A) sites

We select a set of poly(A) sites which may be classified as proximal/distal within a given terminal exon based on a provided BED-formatted poly(A) site atlas as well as GTF-formatted genomic annotation. Throughout the following study we use PolyASite 2.0 [9] and ENSEMBL annotation, version hg38. We focus on all sites supported by at least one protocol (atlas-specific information) but filter for such which are located only on *protein-coding* transcripts. We also discard all sites which could be ambiguously annotated to distinct genes. We export coordinates of the resulting tandem polyA sites into a BED-formatted list.
Snakemake rules which belong to this module are marked with a three-letter namespace: *TPA*.

2.4 Quantification of exon inclusion

In order to quantify transcripts' expression based on transcriptomic alignments we run Salmon [10]. Having obtained per-sample, per-transcript TPM-normalized expression values we use the previous information to collapse the scores into a more exon-centric level: for each representative exon (section 2.2) in every sample we calculate two values: summed up total TPM expression of all transcripts which include this exon ($i_{e,s}$) as well as summed up total TPM expression of all transcripts of a given gene ($t_{e,s}$).
Snakemake rules which belong to this module are marked with a three-letter namespace: *QEI*.

2.5 Quantification of poly(A) sites' expression

In order to quantify the expression of distinct tandem poly(A) sites we employed our previously developed tool: PAQR [11]. The method takes as input genomic alignments of RNA-Seq reads in BAM format and a BED-formatted list of poly(A) sites of interest. It infers expression of distinct sites directly from the coverage profiles in their proximity (please see original publication for more details). As an output it provides TPM-normalized expression of poly(A) sites in all samples and, additionally, a list of their relative position within respective terminal exons. Furthermore, we filter the output table to keep tandem sites only located on such exons for which all their sites are considered as expressed in all analyzed samples. Snakemake rules which belong to this module are marked with a three-letter namespace: *PAQ*.

2.6 Sitecount Matrices

Both statistical models for discovering regulators of casette exon inclusion and poly(A) sites' usage require tables containing quantified information about binding sites of distinct RNA-binding proteins within a certain distance relative to the site of interest. We call such tables 'sitecount matrices'. In case of alternative splicing analysis we define a range around 3'SS and 5'SS. For the alternative polyadenylation analysis we focus around tandem poly(A) sites. We employ a 'sliding-window' strategy in order to gain a better resolution into the positional-dependent effect of RBPs binding sites on the mRNA maturation process. Thus, the pipeline generates multiple sitecount matrices for 3'SS, 5'SS and poly(A) sites - each one corresponding to a unique window defined relatively to the site of interest. MAPP pipeline can be run in two modes: *kmer*- and *pwm-based*, depending on how the user chooses to generate sitecount matrices. In both cases the module reads in BED-formatted files with the positions of 3'SS and 5'SS of cassette exons as well as locations of tandem poly(A) sites (all previously generated) and produces files with coordinates of distinct windows around them. We extract genomic sequences of the regions encoded in these windows with Pybedtools [12]. In the former mode we sum up the occurrences of distinct kmers over the sequence of the whole window (making sure not to overcount short overlapping homomeric subsequences) and the resulting sitecount matrix contains raw counts. In the *pwm-based* approach we utilise MotEvo [1], a probabilistic method which quantifies binding probabilities between nucleotide sequences and distinct motifs (in PWM format). MotEvo parameters were set to: prior for background binding probability as 0.99 (which corresponds to an expectation of 1 site every 100 bp), a minimum binding posterior probability to consider a binding event as 0.01, and the Markov order of the background model to 1. In this case the output sitecount matrices contain summed posterior probabilities for each RBP binding to each region. Regardless of the approach selected the resulting information is stored in a per-exon, per-motif matrix ($N_{e,m}$), for each window separately. Throughout the following study whenever we run MAPP in *kmer-based* mode we count all 3-mers, 4-mers and 5-mers. In case of *pwm-based* analyses we infer binding probabilities of pre-filtered subset of motifs deposited in the ATtRACT databse [13].
Snakemake rules which belong to this module are marked with a three-letter namespace: *CSM*.

2.7 Modeling exons' inclusion (MAEI)

Our aim is to model previously quantified inclusion of cassette exons with the information stored in sitecount matrices (either raw counts of k-mers or binding probabilities for distinct PWMs). Please note that the following procedure is applied to each of the sliding windows separately.

For every exon e we model its inclusion fraction $f_{e,s}$ in sample s with a logistic function $\Theta_{e,s}$:

$$f_{e,s} = \frac{i_{e,s}}{t_{e,s}} \sim \Theta_{e,s} = \frac{e^{b_s + c_e + N_{e,m} \times A_{m,s}}}{1 + e^{b_s + c_e + N_{e,m} \times A_{m,s}}}, \tag{2}$$

where $t_{e,s}$ is the total expression in sample s of the gene containing exon e, and $i_{e,s}$ the expression of transcripts containing exon e. The model parameters are the baseline inclusion level b_s of all exons in sample s, the baseline inclusion level c_e of exon e, the number of binding sites $N_{e,m}$ for motif m at exon e, and the activity $A_{m,s}$ of motif m in sample s. The motif activities $A_{m,s}$ are the key quantities of interest that account for the effect of short sequence motifs on the differential exon inclusion process and are inferred from the expression data, i.e. $i_{e,s}$ and $t_{e,s}$, and the computationally predicted site counts $N_{e,m}$. Note that, as specified, the model is redundant in that the motif activities $A_{m,s}$, exon inclusions c_e and sample inclusions b_s can be shifted so as to leave all probabilities $\Theta_{e,s}$ invariant. To remove this redundancy, we demand that the mean sample inclusion b_s is zero, i.e. $\sum_s b_s = 0$, and that the mean activity of each motif m is zero across the samples, i.e. $\sum_s A_{m,s} = 0$ for each m.
If we had observed $i_{e,s}$ exon inclusion transcripts out of a total $t_{e,s}$ for each exon e in each sample s, then given the model M (as parametrized by the logistic functions $\Theta_{e,s}$) the probability of observing all the quantified data obtained from the RNA-Seq experiment would simply equal a product of consecutive Bernoulli trials where exon inclusion is treated as ”success” and exclusion as ”failure”:

$$P(D|M) = \prod_{s,e} \Theta_{e,s}^{i_{e,s}} \times (1 - \Theta_{e,s})^{t_{e,s}-i_{e,s}} \tag{3}$$

Using this and noting that $i_{e,s} = t_{e,s} f_{e,s}$ we can write the log-likelihood as

$$LL = \sum_{s,e} t_{e,s} \times \left[f_{e,s} \times (b_s + c_e + N_{e,m} \times A_{m,s}) - \log(1 + e^{b_s+c_e+N_{e,m}\times A_{m,s}})\right], \tag{4}$$

which clearly brings out that the ‘weight’ of exon e is sample s in the fitting is simply given by the total number of transcripts $t_{e,s}$ of exon e in sample s. However, in general we cannot meaningfully estimate such absolute transcript numbers, i.e. the estimated total expression levels $t_{e,s}$ are only proportional to these observed transcript numbers, with an unknown proportionality constant. In addition, given that absolute expression levels of transcripts typically vary over $4 - 5$ orders of magnitude in RNA-seq, the likelihood (4) will be dominated by the fitting of the most highly expressed genes, which is clearly not desirable. To address both these issues, we will replace $t_{e,s}$ with renormalized expression level $R_{e,s}$ that is directly proportional to $t_{e,s}$ at low expression levels but saturates to a constant at expression levels significantly above a critical level t_s^{crit}, so that the weight of the highest expressed genes in the fitting remains limited. In particular, we define

$$R_{e,s} = C_s \frac{t_{e,s}}{t_{e,s} + t_s^{crit}} \tag{5}$$

where the pre-factor C_s is set so that the sum of all expression levels $\sum_{e,s} t_{e,s}$ remains invariant, i.e.

$$C_s = \frac{\sum_e t_{e,s}}{\sum_e \frac{t_{e,s}}{t_{e,s}+t_s^{\text{crit}}}}. \tag{6}$$

The parameter t_s^{crit} determines at what expression level $R_{e,s}$ starts to saturate and we chose to set t_s^{crit} to the median of the $t_{e,s}$ across all exons by default.

In a number of applications, such as knock-down experiments of RNA processing factors or other perturbations that are mostly targeted toward RNA processes such as splicing, there often is relatively little change in the absolute expression levels $t_{e,s}$ across the samples. In those cases it can be desirable to replace the sample-dependent statistical weights $R_{e,s}$ with sample-averaged weights R_e. This ensures that observations on the inclusion frequency of a given exon e are weighted equally across all samples s. Although sample-dependent weights $R_{e,s}$ (as defined above) are utilised with default settings of MAEI, the user can also specify to use sample-independent weights R_e that are useful when running on experiments where absolute levels $t_{e,s}$ vary relatively little. We define

$$R_e = C\frac{\langle t_e\rangle}{\langle t_e\rangle + t^{crit}}, \tag{7}$$

where $\langle t_e\rangle$ is the average expression of exon e across all samples and the factor C is given by:

$$C = \frac{\sum_e \langle t_e\rangle}{\sum_e \frac{\langle t_e\rangle}{\langle t_e\rangle + t^{crit}}}, \tag{8}$$

Finally, we can express the log-likelihood of our model as:

$$LL = \sum_{s,e} R \times [f_{e,s} \times (b_s + c_e + N_{e,m} \times A_{m,s}) - \log(1 + e^{b_s + c_e + N_{e,m} \times A_{m,s}})], \tag{9}$$

with R being set to either $R_{e,s}$ (default) or R_e as defined above.

We find maximum likelihood estimates of the model parameters using an EM algorithm where we iteratively calculate partial derivatives with respect to model parameters ($\frac{\partial LL}{\partial A_s}, \frac{\partial LL}{\partial b_s}, \frac{\partial LL}{\partial c_e}$), update their values and re-calculate the likelihood until it converges. In order to ensure a successful procedure we demand that at each iteration of the algorithm $\sum_s A_{m,s} = 0$ and $\sum_s b_{m,s} = 0$. Having enforced such constraints we let all parameters c_e adjust accordingly to preserve the likelihood at it's current value.

We obtain standard deviations of motif activities ($A_{m,s}$) from the Hessian matrix of the log-likelihood function at its optimum. Its negative inverse is an estimator of the covariance matrix of the model parameters. We use these estimates to standardize the activities: for every sample s we calculate a per-sample motif activity z-score: $Z_{m,s} = \frac{A_{m,s}}{\sigma_{m,s}}$.

In order to distinguish motifs with statistically significant z-scores from those with z-scores expected under a Gaussian background model we fit the distribution of observed z-scores to a mixture of a uniform (foreground) and Gaussian (background) distribution:

$$P(D|M) = \rho \times \frac{1}{\max Z_{m,s} - \min Z_{m,s}} + (1-\rho) \times \frac{1}{\sqrt{2\pi}\sigma} \times e^{\frac{-(Z_{m,s}-\mu)^2}{2\sigma^2}}, \tag{10}$$

where the max and min functions are over all motifs m for a given sample s.
We find the maximum likelihood estimates for the parameters (ρ, μ, σ) of the model using an EM algorithm and then use the fitted parameters of the Gaussian background distribution, i.e. μ and σ, to 'renormalize' the z-scores as $Z^{\#}_{m,s} = \frac{Z_{m,s}-\mu}{\sigma}$. After this, we then finally transform these into p-values using these z-scores derive from a standard normal distribution. In order to assess statistical significance at $\alpha = 0.05$ level we also apply a Bonferroni correction where we adjust by the total number of motifs. In this way we obtain p-values $p_{m,s}$ for every motif m in every sample s (separately for every sliding window).

Snakemake rules which belong to this module are marked with a three-letter namespace: *LSM.*

2.8 Modeling poly(A) site usage (KAPACv2.0)

KAPACv2.0, standing for K-mer Activity on PolyAdenylation site Choice version 2.0, builds upon our previously published KAPAC approach [11], whereas for the needs of MAPP we have implemented a new version of KAPAC, KAPACv2.0, which does not depend on the definition of sample contrasts, such as tumor versus normal, but can be applied to any set of samples. Another new features of KAPACv2.0 is that it is capable of running on both raw k-mer counts or binding sites predicted from position weight matrices, similar to the MAEI model. That is, for a given sequence window relative to a poly(A) site p, KAPACv2.0 considers either the sum of site probabilities predicted with a PWM representing a motif k or the raw counts of a k-mer k. It uses then these counts ($N_{p,k}$) to model the relative usage $U_{p,s}$ of each poly(A) site p in sample s as follows (further details are provided in ref. [11]):

$$\log_2(U_{p,s}) = N_{p,k} * A_{k,s} + p_s + p_{s,e} + \epsilon, \tag{11}$$

whereas $c_{s,e}$ is the mean $\log_2$ relative usage of the poly(A) site p from exon e in sample s, c_s is the mean $\log_2$ relative usage of poly(A) site p across all samples, ϵ is the residual error, and the relative usage $U_{p,s}$ of a poly(A) site p from a terminal exon with I poly(A) sites in sample s is calculated from its usage $R_{p,s}$ as follows:

$$U_{p,s} = \frac{R_{p,s}}{\sum_{i=1}^{I} R_{i,s}} \tag{12}$$

KAPACv2.0 solves for the unknown activity $A_{k,s}$ of PWM / k-mer k in sample s and the corresponding error $\sigma_{k,s}$ using an ordinary least squares approach. Similar to the MAEI model (see above), KAPACv2.0 calculates then for every activity $A_{k,s}$ of PWM / k-mer k in sample s and its corresponding error $\sigma_{k,s}$ the z-score $z = \frac{A_{k,s}}{\sigma_{k,s}}$ and performs background correction as done for the MAEI z-scores (see above).

Snakemake rules which belong to this module are marked with a three-letter namespace: *KPC*.

2.9 Analysis summary

Following both statistical models and having inferred motif activites ($A_{m,s}$), their z-scores ($Z_{m,s}$) and p-values ($p_{m,s}$) in all samples and in every window we proceed to summarize the analysis, select those results which we consider statistically significant and visualise them with heatmaps of activity z-scores, which we refer to as 'Impact Maps'. In the last module of the workflow we implemented two distinct strategies to filter motifs based on statistical significance, applicable to different analyses types. In the *avg* mode we call a given motif m statistically significant if and only if there exist a window w within which for at least half of the samples $p^w_{m,s}$ are below a previously defined cutoff (0.05). This strategy is designed for common comparative analyses of two biological conditions, each being sequenced in multiple replicates. The other approach - *max* - requires m to be called as statistically significant in only one sample in order to annotate significance to the whole motif. The rationale behind this was to provide an insightful way of investigating datasets which consist of multiple distinct conditions. For every motif called as statistically significant we plot an Impact Map as a visual summary of that motifs activity on exon inclusion and poly(A) site usage.

Snakemake rules which belong to this module are marked with a three-letter namespace: *RES*.

2.10 Final report

At the end of the workflow we designed a few additional steps which prepare an output directory with the most important text files and tables as well as a summary report in HTML format. Summary directory is also compressed into .tar.gz format to facilitate reproducibility and usability: users may exchange their results as well as run configurations easily. The final HTML report contains a sorted table of motifs whose activity z-scores were reported to be statistically significant in at least one sliding window around any of the processing sites. The columns of that table include: motif ID, sequence logo (in case of a

PWM-based MAPP execution), three maximum activity z-scores (with the maximum taken over all sliding windows around each of the analyzed sites: 3'SS, 5'SS, PAS), ranking score (set to an average of the three aforementioned z-scores) and the Impact Map.

Top-level snakemake rules related to the final summary are marked with a four-letter namespace: *MAPP*.

2.11 Gene ontology analysis

Prior to gene ontology (GO) analysis, differentially expressed cassette exons were obtained by performing t-tests of normal brain vs. glioblastoma samples using R [14] (v.4.2.0) and "matrixTests" library (v.0.1.9.1). P-values were adjusted for multiple comparisons using Benjamini-Hochberg (BH) method. Genes containing differentially expressed cassette exons with binding sites (binding probability > 0.6) for the PTBP1 and the RBFOX RBPs within the corresponding regions inferred by MAPP. For the PTBP1 RBP 3'SS the considered windows reached from -125 to +50 nt and for the 5'SS there was only one window considered from -50 to 0 nt. For the RBFOX RBPs 5'SS the considered regions reached from 0 to +200 nt. The exons having counts in both of the windows were used for gene ontology (GO) analysis which was performed with the "enrichGO" function from the Bioconductor package "clusterProfiler" [15] (v.4.4.1) for each of the GO categories: biological process (BP), molecular function (MF) and cellular component (CC). As minimal number of genes annotated per ontology term (minGSSize) we used 10 and an adjusted p-value cutoff (pvalueCutoff) for enrichment tests was 0.1. P-values for enriched GO terms were adjusted for multiple comparisons using BH method. All genes having MAPP quantified cassette exons in the dataset served as a background ("universe") for GO analysis. Full GO analysis in the form of a compiled HTML-report as well as the Rmarkdown scripts and R session info are provided in the GitHub repository: `https://github.com/gruber-sciencelab/MAPP_GO_KEGG_htmlbook`. The HTML-report was generated with the "Bookdown" R package [16] (v.0.29) and can be viewed by cloning the repository and opening the "index.html" file in a browser from the "_book" folder.

References

1. Arnold, P., Erb, I., Pachkov, M., Molina, N. & van Nimwegen, E. Motevo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of dna sequences. *Bioinformatics* **28**, 487–494 (2012).

2. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10–12 (2011).

3. Dobin, A. *et al.* Star: ultrafast universal rna-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

4. Li, H. *et al.* The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).

5. DeLuca, D. S. *et al.* Rna-seqc: Rna-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).

6. Wang, L., Wang, S. & Li, W. Rseqc: quality control of rna-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).

7. Andrews, S. *et al.* FastQC. Babraham Institute (2012).

8. Trincado, J. L. *et al.* Suppa2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome biology* **19**, 1–11 (2018).

9. Herrmann, C. J. *et al.* Polyasite 2.0: a consolidated atlas of polyadenylation sites from 3 end sequencing. *Nucleic acids research* **48**, D174–D179 (2020).

10. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods* **14**, 417–419 (2017).

11. Gruber, A. J. *et al.* Discovery of physiological and cancer-related regulators of 3 utr processing with kapac. *Genome biology* **19**, 1–17 (2018).

12. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).

13. Giudice, G., Sánchez-Cabo, F., Torroja, C. & Lara-Pezzi, E. Attract—a database of rna-binding proteins and associated motifs. *Database* **2016** (2016).

14. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2022). URL `https://www.R-project.org/`.

15. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology* **16**, 284–287 (2012).

16. Xie, Y. *bookdown: Authoring Books and Technical Documents with R Markdown* (Chapman and Hall/CRC, Boca Raton, Florida, 2016). URL `https://bookdown.org/yihui/bookdown`. ISBN 978-1138700109.

Supplementary Table S1: RNA-Seq samples downloaded from NCBI:GEO and analyzed with MAPP.
Identifiers and metadata of publicly available datasets with perturbed expression of selected RBPs.

Experiment Name	Series ID	Sample ID	Description
HNRNPC KD	GSE56010	GSM1502498	RNA-seq-control-rep1
HNRNPC KD	GSE56010	GSM1502499	RNA-seq-control-rep2
HNRNPC KD	GSE56010	GSM1502500	RNA-seq-HNRNPCKD-rep1
HNRNPC KD	GSE56010	GSM1502501	RNA-seq-HNRNPCKD-rep2
PTBP1 KD	GSE69656	GSM1705375	293 control knockdown_rep1
PTBP1 KD	GSE69656	GSM1705376	293 control knockdown_rep2
PTBP1 KD	GSE69656	GSM1705377	293 PTBP1 and PTBP2 double knockdown_rep1
PTBP1 KD	GSE69656	GSM1705378	293 PTBP1 and PTBP2 double knockdown_rep2
RBFOX1 OE	GSE71468	GSM2026221	control_noRbfox_1
RBFOX1 OE	GSE71468	GSM2026222	control_noRbfox_2
RBFOX1 OE	GSE71468	GSM2026223	control_noRbfox_3
RBFOX1 OE	GSE71468	GSM2026224	control_+Rbfox1_1
RBFOX1 OE	GSE71468	GSM2026225	control_+Rbfox1_2
RBFOX1 OE	GSE71468	GSM2026226	control_+Rbfox1_3

Supplementary Table S2: RNA-Seq data of RBP knock-down experiments available through ENCODE project were analyzed with MAPP.

Experiment identifiers together with targeted RBP name; all data are publicly available.

#	Experiment ID	Target
1	ENCSR599UDS	AARS
2	ENCSR547NWD	AARS
3	ENCSR424YSV	AATF
4	ENCSR973QSV	AATF
5	ENCSR610VTA	ABCF1
6	ENCSR721MXZ	ABCF1
7	ENCSR511SYK	ACO1
8	ENCSR164TLB	ADAR
9	ENCSR104OLN	ADAR
10	ENCSR812TLY	AGGF1
11	ENCSR268JDD	AGO1
12	ENCSR533HXS	AGO1
13	ENCSR495YSS	AGO2
14	ENCSR207QGW	AGO3
15	ENCSR016IDR	AKAP1
16	ENCSR338CON	AKAP1
17	ENCSR000YYN	AKAP8
18	ENCSR958KSY	AKAP8
19	ENCSR807ODB	AKAP8L
20	ENCSR809ISU	AKAP8L
21	ENCSR012DAF	APOBEC3C
22	ENCSR963RLK	APOBEC3C
23	ENCSR624OUI	AQR
24	ENCSR253DCB	ASCC1
25	ENCSR193FFA	ASCC1
26	ENCSR713OLV	ATP5C1
27	ENCSR231DXJ	ATP5C1
28	ENCSR395FYF	AUH
29	ENCSR409CSO	AUH
30	ENCSR570CWH	BCCIP
31	ENCSR606QIX	BCCIP
32	ENCSR410ZPU	BCLAF1
33	ENCSR481AYC	BCLAF1
34	ENCSR775TMW	BOP1
35	ENCSR925SYZ	BOP1
36	ENCSR382QKD	BUD13
37	ENCSR267RHP	BUD13
38	ENCSR040WAK	CALR

39	ENCSR081IAO	CCAR1
40	ENCSR386YEV	CCAR1
41	ENCSR984CLJ	CCAR2
42	ENCSR237IWZ	CCAR2
43	ENCSR874ZLI	CCDC124
44	ENCSR929PXS	CEBPZ
45	ENCSR351CNN	CELF1
46	ENCSR605MFS	CELF1
47	ENCSR695XOD	CELF1
48	ENCSR056QEW	CIRBP
49	ENCSR230ORC	CIRBP
50	ENCSR269SJB	CKAP4
51	ENCSR113PYX	CNOT7
52	ENCSR274KWA	CNOT7
53	ENCSR312HJY	CNOT8
54	ENCSR795VAK	CPEB4
55	ENCSR676EKU	CPSF6
56	ENCSR384BDV	CPSF6
57	ENCSR895BTE	CPSF6
58	ENCSR594DNW	CPSF7
59	ENCSR222LRL	CPSF7
60	ENCSR815JDY	CSTF2
61	ENCSR885YOI	CSTF2
62	ENCSR286OKW	CSTF2T
63	ENCSR914WQV	CSTF2T
64	ENCSR840QFR	DAZAP1
65	ENCSR220TBR	DAZAP1
66	ENCSR907UTB	DAZAP1
67	ENCSR208GPE	DDX1
68	ENCSR070LJO	DDX1
69	ENCSR281IUF	DDX19B
70	ENCSR312SFA	DDX19B
71	ENCSR961WVL	DDX21
72	ENCSR485ZTC	DDX21
73	ENCSR300IEW	DDX24
74	ENCSR067LLB	DDX24
75	ENCSR210RWL	DDX27
76	ENCSR584LDM	DDX27
77	ENCSR205VSQ	DDX28
78	ENCSR222CSF	DDX28
79	ENCSR637JLM	DDX3X
80	ENCSR000KYM	DDX3X
81	ENCSR334HNJ	DDX47

82	ENCSR155EZL	DDX47
83	ENCSR388CNS	DDX47
84	ENCSR808FBR	DDX5
85	ENCSR029LGJ	DDX51
86	ENCSR560RSZ	DDX52
87	ENCSR913ZWR	DDX52
88	ENCSR331DUD	DDX55
89	ENCSR964YTW	DDX55
90	ENCSR856CJK	DDX55
91	ENCSR598GKQ	DDX59
92	ENCSR067AUG	DDX59
93	ENCSR119QWQ	DDX6
94	ENCSR147ZBD	DDX6
95	ENCSR853PBF	DHX30
96	ENCSR345VVZ	DHX30
97	ENCSR494UDF	DKC1
98	ENCSR118KUN	DKC1
99	ENCSR577OVP	DNAJC2
100	ENCSR004OSI	DNAJC2
101	ENCSR079LMZ	DNAJC21
102	ENCSR385KOY	DNAJC21
103	ENCSR624XHG	DROSHA
104	ENCSR477TRX	EEF2
105	ENCSR181RLB	EEF2
106	ENCSR620OKS	EFTUD2
107	ENCSR117WLY	EFTUD2
108	ENCSR546MBH	EIF2S1
109	ENCSR861ENA	EIF2S1
110	ENCSR076PMZ	EIF2S2
111	ENCSR110HAA	EIF2S2
112	ENCSR258VGD	EIF3A
113	ENCSR788HVK	EIF3D
114	ENCSR660ETT	EIF3D
115	ENCSR778AJO	EIF3G
116	ENCSR143UET	EIF3G
117	ENCSR957EEG	EIF4A3
118	ENCSR961YAG	EIF4A3
119	ENCSR774BXV	EIF4B
120	ENCSR313CHR	EIF4B
121	ENCSR712CSN	EIF4G1
122	ENCSR509LIV	EIF4G1
123	ENCSR152MON	EIF4G2
124	ENCSR040FSN	EIF4G2

125	ENCSR077BPR	ESF1
126	ENCSR060IWW	ESF1
127	ENCSR840QOH	ETF1
128	ENCSR831YGP	EWSR1
129	ENCSR532ZPP	EWSR1
130	ENCSR597IYB	EXOSC9
131	ENCSR812EIA	EXOSC9
132	ENCSR492BKM	FAM120A
133	ENCSR047VPW	FAM120A
134	ENCSR728BOL	FASTKD1
135	ENCSR716WZH	FASTKD2
136	ENCSR608IAI	FASTKD2
137	ENCSR511BNY	FIP1L1
138	ENCSR116QBU	FIP1L1
139	ENCSR379VXW	FKBP4
140	ENCSR639LKS	FKBP4
141	ENCSR555LCE	FMR1
142	ENCSR905HID	FMR1
143	ENCSR688GVV	FTO
144	ENCSR389HFU	FTO
145	ENCSR755KOM	FUBP3
146	ENCSR373KOF	FUBP3
147	ENCSR325OOM	FUS
148	ENCSR927JXU	FUS
149	ENCSR009PPI	FXR1
150	ENCSR780YFF	FXR1
151	ENCSR577XBW	FXR2
152	ENCSR139BIJ	FXR2
153	ENCSR792CBM	G3BP1
154	ENCSR074UZM	G3BP1
155	ENCSR945UYL	G3BP2
156	ENCSR246SOU	G3BP2
157	ENCSR771QMJ	GEMIN5
158	ENCSR398GHW	GEMIN5
159	ENCSR874DVZ	GLRX3
160	ENCSR116YMU	GNB2L1
161	ENCSR968YWY	GPKOW
162	ENCSR967QNT	GPKOW
163	ENCSR674KDQ	GRSF1
164	ENCSR835RMN	GRSF1
165	ENCSR850FEH	GRWD1
166	ENCSR528ASX	GRWD1
167	ENCSR188IPO	GTF2F1

168	ENCSR295XKC	GTF2F1
169	ENCSR634KHL	HDGF
170	ENCSR958NDU	HLTF
171	ENCSR010ZMZ	HLTF
172	ENCSR720BPO	HNRNPA0
173	ENCSR552NBS	HNRNPA0
174	ENCSR182DAW	HNRNPA1
175	ENCSR048BWH	HNRNPA1
176	ENCSR794NUE	HNRNPA2B1
177	ENCSR769GES	HNRNPA2B1
178	ENCSR354XQY	HNRNPAB
179	ENCSR778WPL	HNRNPAB
180	ENCSR052IYH	HNRNPC
181	ENCSR470PRV	HNRNPC
182	ENCSR634KBO	HNRNPC
183	ENCSR660MZN	HNRNPD
184	ENCSR392HSJ	HNRNPF
185	ENCSR693MZJ	HNRNPF
186	ENCSR853ZJS	HNRNPK
187	ENCSR529JNJ	HNRNPK
188	ENCSR155BMF	HNRNPL
189	ENCSR563YIS	HNRNPL
190	ENCSR490DYI	HNRNPLL
191	ENCSR746NIM	HNRNPM
192	ENCSR995JMS	HNRNPM
193	ENCSR047IUS	HNRNPU
194	ENCSR732ICL	HNRNPU
195	ENCSR308IKH	HNRNPU
196	ENCSR689ZJC	HNRNPUL1
197	ENCSR034VBA	HNRNPUL1
198	ENCSR222ABK	HSPD1
199	ENCSR243IGA	HSPD1
200	ENCSR629EWX	IGF2BP1
201	ENCSR708GKW	IGF2BP1
202	ENCSR952RRH	IGF2BP2
203	ENCSR478FJK	IGF2BP2
204	ENCSR481YXD	IGF2BP3
205	ENCSR302JQA	IGF2BP3
206	ENCSR710NWE	IGF2BP3
207	ENCSR126ARZ	ILF2
208	ENCSR366FFV	ILF2
209	ENCSR942MBU	ILF3
210	ENCSR269HQA	ILF3

211	ENCSR784FTX	KHDRBS1
212	ENCSR023HWI	KHDRBS1
213	ENCSR561CBC	KHSRP
214	ENCSR850CKU	KHSRP
215	ENCSR182GKG	KIF1C
216	ENCSR823WTA	KIF1C
217	ENCSR542ESY	KRR1
218	ENCSR244SIO	KRR1
219	ENCSR866XLI	LARP4
220	ENCSR744PAQ	LARP4
221	ENCSR770OWW	LARP7
222	ENCSR624FBY	LARP7
223	ENCSR598YQX	LIN28B
224	ENCSR927SLP	LIN28B
225	ENCSR883BXR	LSM11
226	ENCSR762FEO	LSM11
227	ENCSR849STR	MAGOH
228	ENCSR746EKS	MAGOH
229	ENCSR517JHY	MAK16
230	ENCSR105OXX	MARK2
231	ENCSR792XFP	MATR3
232	ENCSR492UFS	MATR3
233	ENCSR222COT	MBNL1
234	ENCSR992JGE	METAP2
235	ENCSR952QDQ	METAP2
236	ENCSR169QQW	MSI2
237	ENCSR896MMU	MSI2
238	ENCSR631RFX	MTPAP
239	ENCSR701GSV	MTPAP
240	ENCSR945GUR	NAA15
241	ENCSR355OQC	NAA15
242	ENCSR030ARO	NCBP2
243	ENCSR361LBE	NCBP2
244	ENCSR939ZRA	NELFE
245	ENCSR201WFU	NELFE
246	ENCSR007XKL	NFX1
247	ENCSR696LLZ	NIP7
248	ENCSR517JDK	NKRF
249	ENCSR227AVS	NOL12
250	ENCSR643UFV	NOL12
251	ENCSR398HXV	NONO
252	ENCSR647NYX	NONO
253	ENCSR346DZQ	NPM1

254	ENCSR016XPB	NPM1
255	ENCSR829EFL	NSUN2
256	ENCSR629RUG	NSUN2
257	ENCSR754RJA	NUFIP2
258	ENCSR584JRB	NUFIP2
259	ENCSR457WBK	NUP35
260	ENCSR927XBT	NUSAP1
261	ENCSR180XTP	NUSAP1
262	ENCSR028YAQ	PA2G4
263	ENCSR309PPC	PA2G4
264	ENCSR910YNJ	PABPC1
265	ENCSR192GBD	PABPC1
266	ENCSR455VZH	PABPC4
267	ENCSR047EEG	PABPC4
268	ENCSR416ZJH	PABPN1
269	ENCSR368ZRP	PAPOLA
270	ENCSR825QXH	PARN
271	ENCSR306IOF	PARN
272	ENCSR635FRH	PCBP1
273	ENCSR545AIK	PCBP1
274	ENCSR028ITN	PCBP2
275	ENCSR648QFY	PCBP2
276	ENCSR496ETJ	PES1
277	ENCSR891DYO	PES1
278	ENCSR912EHP	PES1
279	ENCSR322XVS	PHF6
280	ENCSR681SMT	PHF6
281	ENCSR656DQV	PKM
282	ENCSR978CSQ	PKM
283	ENCSR191VWK	PNPT1
284	ENCSR880DEH	PNPT1
285	ENCSR936TED	POLR2G
286	ENCSR529MBZ	PPIG
287	ENCSR620HAA	PPIG
288	ENCSR556FNN	PPIL4
289	ENCSR851KEX	PPIL4
290	ENCSR844QNT	PPP1R8
291	ENCSR529QEZ	PRPF6
292	ENCSR783LUA	PRPF6
293	ENCSR137HKS	PRPF8
294	ENCSR998MZP	PRPF8
295	ENCSR744YVR	PSIP1
296	ENCSR611LQB	PSIP1

297	ENCSR527IVX	PTBP1
298	ENCSR064DXG	PTBP1
299	ENCSR239BCO	PTBP1
300	ENCSR648BSC	PUF60
301	ENCSR558XNA	PUF60
302	ENCSR945XKW	PUM1
303	ENCSR745WVZ	PUM1
304	ENCSR210DML	PUM2
305	ENCSR118XYK	PUM2
306	ENCSR618IQH	PUS1
307	ENCSR296ERI	PUS1
308	ENCSR330YOU	QKI
309	ENCSR256PLH	QKI
310	ENCSR904BCZ	RAVER1
311	ENCSR576GOW	RAVER1
312	ENCSR767LLP	RBFOX2
313	ENCSR336DFS	RBFOX2
314	ENCSR627NVU	RBM15
315	ENCSR385UPQ	RBM15
316	ENCSR599PXD	RBM15
317	ENCSR898OPN	RBM17
318	ENCSR385TMY	RBM17
319	ENCSR330KHN	RBM22
320	ENCSR947OIM	RBM22
321	ENCSR149DMY	RBM25
322	ENCSR610AEI	RBM25
323	ENCSR222SMI	RBM27
324	ENCSR675KPR	RBM3
325	ENCSR318HAT	RBM34
326	ENCSR560AYQ	RBM34
327	ENCSR678WOA	RBM39
328	ENCSR760EGM	RBM39
329	ENCSR711ZJQ	RBM47
330	ENCSR921KDS	RCC2
331	ENCSR685JXU	RCC2
332	ENCSR572AMC	RECQL
333	ENCSR310VND	RECQL
334	ENCSR014VQS	RECQL
335	ENCSR706SXN	RPL23A
336	ENCSR082YGI	RPLP0
337	ENCSR410UHJ	RPS10
338	ENCSR004RGI	RPS10
339	ENCSR098NHI	RPS19

340	ENCSR486AIO	RPS19
341	ENCSR667RIA	RPS2
342	ENCSR410MIQ	RPS3
343	ENCSR788YGG	RPS3A
344	ENCSR118VQR	RPS3A
345	ENCSR838SMC	RPS5
346	ENCSR210KJB	RRP9
347	ENCSR471GIS	RRP9
348	ENCSR783YSQ	RTF1
349	ENCSR906WTM	RTF1
350	ENCSR770LYW	SAFB2
351	ENCSR110ZYD	SAFB2
352	ENCSR954HAY	SART3
353	ENCSR011BBS	SART3
354	ENCSR219DXZ	SBDS
355	ENCSR343DHN	SBDS
356	ENCSR820ROH	SERBP1
357	ENCSR925RNE	SERBP1
358	ENCSR628JYB	SF1
359	ENCSR562CCA	SF1
360	ENCSR644AIM	SF1
361	ENCSR374NMJ	SF3A3
362	ENCSR454KYR	SF3A3
363	ENCSR896CFV	SF3B1
364	ENCSR047QHX	SF3B1
365	ENCSR148MQK	SF3B4
366	ENCSR081XRA	SF3B4
367	ENCSR782MXN	SFPQ
368	ENCSR535YPK	SFPQ
369	ENCSR519KXM	SLBP
370	ENCSR112YTD	SLBP
371	ENCSR234YMW	SLTM
372	ENCSR185JGT	SLTM
373	ENCSR090UMI	SMN1
374	ENCSR129ROE	SMN1
375	ENCSR995ZGJ	SMNDC1
376	ENCSR408SDL	SMNDC1
377	ENCSR232XRZ	SND1
378	ENCSR398LZW	SND1
379	ENCSR003LSA	SNRNP200
380	ENCSR943LIB	SNRNP200
381	ENCSR635BOO	SNRNP70
382	ENCSR153GKS	SRFBP1

383	ENCSR813NZP	SRFBP1
384	ENCSR312SRB	SRP68
385	ENCSR167JPY	SRP68
386	ENCSR524YXQ	SRPK2
387	ENCSR066VOO	SRSF1
388	ENCSR094KBY	SRSF1
389	ENCSR376FGR	SRSF3
390	ENCSR697GLD	SRSF4
391	ENCSR781YNI	SRSF5
392	ENCSR447UCG	SRSF5
393	ENCSR906RHU	SRSF5
394	ENCSR464ADT	SRSF7
395	ENCSR017PRS	SRSF7
396	ENCSR113HRG	SRSF9
397	ENCSR597XHH	SRSF9
398	ENCSR278CHI	SSB
399	ENCSR891AXF	SSB
400	ENCSR422JMS	SSRP1
401	ENCSR902WSK	SSRP1
402	ENCSR777EDL	STAU1
403	ENCSR124KCF	STAU1
404	ENCSR871BXO	STIP1
405	ENCSR082UWF	STIP1
406	ENCSR997FOT	SUB1
407	ENCSR047AJA	SUB1
408	ENCSR810JYX	SUCLG1
409	ENCSR101OPF	SUCLG1
410	ENCSR837QDN	SUGP2
411	ENCSR192BPV	SUGP2
412	ENCSR281KCL	SUPT6H
413	ENCSR530BOP	SUPT6H
414	ENCSR995RPB	SUPV3L1
415	ENCSR778SIU	SUPV3L1
416	ENCSR611ZAL	TAF15
417	ENCSR031RZS	TAF15
418	ENCSR998RZI	TAF15
419	ENCSR527QNC	TARDBP
420	ENCSR455TNF	TARDBP
421	ENCSR134JRE	TARDBP
422	ENCSR741YCA	TBRG4
423	ENCSR079IPT	TBRG4
424	ENCSR573UBF	TFIP11
425	ENCSR911DGK	TFIP11

426	ENCSR057GCF	TIA1
427	ENCSR694LKY	TIA1
428	ENCSR450VQO	TIAL1
429	ENCSR927TSP	TIAL1
430	ENCSR030GZQ	TRA2A
431	ENCSR916WOI	TRA2A
432	ENCSR300QFQ	TRIM56
433	ENCSR309HXK	TRIM56
434	ENCSR152IWT	TRIP6
435	ENCSR946OFN	TROVE2
436	ENCSR060KRD	TROVE2
437	ENCSR459EMR	TUFM
438	ENCSR602AWR	TUFM
439	ENCSR372UWV	U2AF1
440	ENCSR342EDG	U2AF1
441	ENCSR904CJQ	U2AF2
442	ENCSR622MCX	U2AF2
443	ENCSR426UUG	U2AF2
444	ENCSR424JSU	UBE2L3
445	ENCSR362XMY	UBE2L3
446	ENCSR678MVE	UCHL5
447	ENCSR684HTV	UCHL5
448	ENCSR251ABP	UPF1
449	ENCSR689MIY	UPF1
450	ENCSR318OXM	UPF2
451	ENCSR810FHY	UPF2
452	ENCSR165VBD	UTP18
453	ENCSR269ZAO	UTP18
454	ENCSR910ECL	UTP3
455	ENCSR334BTA	WDR3
456	ENCSR341PZW	WDR43
457	ENCSR165BCF	WRN
458	ENCSR778RWJ	XPO5
459	ENCSR453HKS	XPO5
460	ENCSR732IYM	XRCC5
461	ENCSR715XZS	XRCC5
462	ENCSR232CPD	XRCC6
463	ENCSR500WHE	XRCC6
464	ENCSR717SJA	XRN2
465	ENCSR347ZHQ	XRN2
466	ENCSR306EIU	YBX3
467	ENCSR494VSD	YBX3
468	ENCSR843LYF	YTHDC2

469	ENCSR448JAM	ZC3H8
470	ENCSR518JXY	ZNF622
471	ENCSR850PWM	ZRANB2
472	ENCSR081QQH	ZRANB2

Supplementary Table S3: Used RNA-Seq samples available through TCGA project.

Sample identifiers, hash codes and sample conidtion (healthy control tissue / glioblastoma tissue).

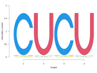
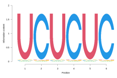
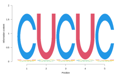
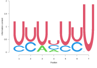
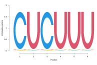
Sample ID	Sample hash	condition
TCGA-06-0681-11A	69c0cd45-79bb-4fec-85d6-f57b3e9ef217	HEALTHY
TCGA-06-0675-11A	efd7c102-95dd-4040-b30a-08ba642ec3b5	HEALTHY
TCGA-06-AABW-11A	1d01565c-c9b3-4182-8823-ad36bfaa095b	HEALTHY
TCGA-06-0680-11A	e41d9335-aad6-4dbb-992e-21a518b20b5e	HEALTHY
TCGA-06-0678-11A	0f38416c-e746-4d3f-a95b-9396ffbb8c59	HEALTHY
TCGA-06-5416-01A	2ae4257e-7fbb-4654-bfb0-7bd815df53cb	GBM
TCGA-06-0644-01A	b44ffe75-508f-49c6-8758-f5feeaebd0c4	GBM
TCGA-32-2621-01A	23768895-9695-4f95-a075-d7a52ac3545f	GBM
TCGA-28-2509-01A	4d32f2c5-4b17-48ec-8a58-a3c582fbd790	GBM
TCGA-06-2567-01A	a9f38529-3b55-48d4-81fe-68fa8c3f7e9c	GBM

Supplementary Table S4: Top 20 significant k-mer activities as inferred by MAPP from glioblastoma and normal brain samples.

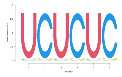
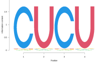
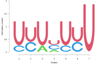
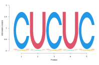
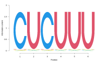
MAPP was run in k-mer mode on the samples listed in Supplementary Table S3.
Only the top 20 significant k-mers are reported for the indicated regions.

3'SS	**5'SS**	**PAS**
CTCT	TCT	CTCTC
TCT	TTTCT	TTT
TCTCT	CTCT	
CTT	TTTC	
TCTC	TTCT	
CTCTT	CTT	
TTCT	CTTT	
AGGTA	TCTCT	
GGTA	CTCTC	
TTC	TTC	
GGTAA	TTT	
CTC	TCTC	
CTCTC	CTTTC	
TTCTC	CCTCT	
CTTTC	GCATG	
TTTC	TCTT	
TTTCT	TTTTC	
TCTT	TGT	
TGCT	TGC	
CTTT	CTCTT	

Supplementary Table S5: Top 5 PWMs ranked by their overall impact on mRNA processing sites as inferred by MAPP from glioblastoma and normal brain samples.

Sequence logo	PWM ID	Ranking score	3'SS Zscore	5'SS Zscore	pas Zscore
	PTBP1_489	5.552	7.623 *	5.766 *	3.268
	PTBP1_992	5.476	6.773 *	5.22 *	4.435 *
	PTBP1_8	5.118	5.96 *	5.072 *	4.323 *
	PTBP1_s100	4.525	5.577 *	5.925 *	2.074
	PTBP1_1000	4.313	6.464 *	4.312 *	2.163

Supplementary Table S6: Top 5 PWMs ranked by their overall impact on mRNA processing sites as inferred by MAPP from PTBP1/2 knock-down experiment.

Sequence logo	PWM ID	Ranking score	3'SS Zscore	5'SS Zscore	pas Zscore
	PTBP1_992	5.638	4.204 *	4.361 *	8.348 *
	PTBP1_489	5.581	5.261 *	4.711 *	6.771 *
	PTBP1_s100	5.225	5.896 *	5.313 *	4.466 *
	PTBP1_8	5.066	3.784 *	3.804 *	7.608 *
	PTBP1_1000	5.006	5.304 *	4.424 *	5.289 *

Supplementary Table S7: Details on selected casette exons regulated by PTBP1 and RBFOX1, as inferred by MAPP from Glioblastoma and normal brain samples.

Comparison of condition-wise differences in quantified inclusion fractions as well as PTBP1/RBFOX1 binding scores reported by MotEvo.

The last two columns contain the maximum binding scores inferred for the windows between -75 to +50 nt from 3'SS and 0 to +200 nt from 5'SS for PTBP1 and RBFOX1 motifs, respectively.

coordinates	gene_id	gene_name	AVG_PSI_HEALTHY	AVG_PSI_GBM	dPSI	dPSI_ttest_pval	max_PTBP1_binding_score	max_RBFOX1_binding_score
1:164820072-164820184:+	ENSG00000185630	PBX1	0.743	0.354	-0.389	0.001	0.521	0.253
2:55025086-55027485:-	ENSG00000115310	RTN4	0.960	0.506	-0.454	0.021	0.999	0.134
1:96806419-96806452:+	ENSG00000117569	PTBP2	0.970	0.806	-0.164	0.001	1.746	1.833
22:29683020-29683064:+	ENSG00000186575	NF2	0.695	0.145	-0.550	0.003	0.056	1.033
10:73396044-73396109:-	ENSG00000138279	ANXA7	0.809	0.470	-0.340	0.000	0.521	0.000