DEVELOPMENT OF METHODS FOR THE ANALYSIS OF DEEP SEQUENCING DATA; APPLICATIONS TO THE DISCOVERY OF TARGETS OF RNA-BINDING PROTEINS

Inauguraldissertation

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

von

Biter Bilen
aus der Türkei

Basel, 2014

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Mihaela Zavolan und Asst. Prof. Štěpánka Vaňáčová

Basel, den 11. December 2012

Prof. Jörg Schibler, Dekan

Acknowledgements

I gratefully acknowledge Mihaela Zavolan, Štěpánka Vaňáčová, Walter Keller, and my colleagues for rewarding discussions and collaborations and my family for their continuous support. Additionally, I thank my friends and colleagues Piotr Balwierz and Evgeniy Ozonov for the fruitful discussions.

Development Of Methods For The Analysis Of Deep Sequencing Data; Applications To The Discovery Of Targets Of RNA-binding Proteins

by

Biter Bilen

With the recent advances in nucleotide sequencing technologies, it became easy to generate tens of millions of reads with genome- or transcriptome-wide distribution with reduced cost and high accuracy. One of the applications of deep sequencing is the determination of the repertoire of targets of RNA-binding proteins. The method, called CLIP (for UV crosslinking and immune-precipitation) is now widely used to characterize a variety of proteins with regulatory as well as enzymatic functions. Here we focus on the statistical analysis of data obtained through a variant of CLIP, called PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced CLIP), which was applied to three different RNA binding proteins whose function was previously not well characterized: PAPD5 (PAP associated domain containing 5), DIS3L2 (DIS3 mitotic control homolog (S. cerevisiae)-like 2), and EWSR1 (Ewing sarcoma breakpoint region 1). Our computational analysis was instrumental for the definition of the main *in vivo* substrates of these proteins, which were confirmed by additional experiments. In the analysis, we also used extensively publicly available high-throughput data sets that enabled us make inferences about the function of the proteins. The main results of biological significance were as follows. We determined ribosomal RNAs are the main targets of PAPD5 and that the main substrates of the DIS3L2 nuclease are tRNAs and found that the tRNA-derived fragments processed by DIS3L2 could be loaded in the RNA silencing complex and be involved in gene silencing. Finally, we determined that EWSR1preferentially binds to RNAs that originate from instability-prone regions like sub-telomeres, known to be hotspots of genomic rearrangements, as well as other genes located in internal regions of chromosomes, that have been implicated in genomic translocations. These include EWSR1's own pre-mRNA. All together this dissertation illustrates the point that when coupled with proper statistical analysis, CLIP is able to reveal targets of RNA-binding proteins that were difficult to study with other methods and that and integration of public domain datasets is very powerful in deciphering complex RNA-protein and regulatory RNA networks implicated in post-transcriptional gene regulation.

Table of Contents

Chapter 1: Introduction

1.1 Recent Advances in Sequencing Technologies And Their Applications

12 years after the completion of the draft of the human genome in (Venter et al., 2001), a myriad of applications for sequencing-based high–throughput methods have been found. The ability to detect variations, yet unknown transcripts, high throughput and accuracy and reduced cost play a big role in this revolution. At the time writing this thesis, the Ensembl database contained ~70 fully sequenced vertebrate genomes (http://www.ensembl.org/info/about/species.html). The target is currently moving towards delineating variations in genomes of individuals of the same species and the 1000 Genomes Project (www.1000genomes.org) is a good example; it was launched in 2008 to decipher human genetic variations (Consortium, 2010). In parallel to the whole genome sequencing efforts, measurements of transcriptomes are also taking place. Sequencing has also all but rapidly replaced microarray-based expression profiling. Furthermore, the ENCODE (ENCyclopedia Of DNA Elements) (http://genome.ucsc.edu/ENCODE/) project is set up to define all functional elements in the human genome and to catalogue the transcriptomes of several human cell lines (2004). modENCODE (http://www.modencode.org/) is a variant of the initial ENCODE project and aims to identify functional elements in selected model organisms, specifically *Drosophila melanogaster* and *Caenorhabditis elegans*; model organisms are of great value, permitting experimental validation of findings in a manner that would not be possible in humans (Consortium et al., 2010). Although very exciting, these advances brought with them computational challenges. Optimal data structures, some of which highly distributed, are being developed to store the massive amounts of generated data

and novel algorithms with extremely good performance are needed to operate on these data.

1.2 Transcriptome-wide Sequencing Methods

Sequencing can be used to quantify gene expression transcriptome-wide. Among the most widely used methods is mRNAseq (mature RNA sequencing), which aims to profile all polyadenylated RNA species and is extensively applied to measure the gene expression levels (Mortazavi et al., 2008). Other methods have been developed to quantify not only transcript levels but also alternative splice form, alternative promoter and polyA site usage in different cells (Guttman et al., 2011, Trapnell et al., 2010). Quantification of individual transcripts levels is a non-trivial problem. Surprisingly, quite substantial differences can be found by applying different methods to one data set, reflecting different assumptions and heuristics of the programs (Katz et al., 2010). High-throughput sequencing has also been integrated recently into methods for determination of binding sites of RNA-binding proteins. For example, CLIP (Crosslinking and Immunoprecipitation) is a recently developed method that relies on ultraviolet light-induced crosslinking, followed by immunoprecipitation of the protein of interest with a specific antibody and sequencing of the associated RNAs. CLIP was initially applied in low throughput to identify mRNA targets of a neuron specific splicing factor, Nova (Ule et al., 2003). Integration of high-throughput sequencing lead to development of the HITS-CLIP (High Throughput Sequencing CLIP) variant, which was first applied to a rather challenging problem, namely determination of microRNA targets transcriptome-wide (Chi et al., 2009). Our group contributed to the development of the PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) approach. PAR-CLIP is an alternative to HITS-CLIP, conceived to increase the efficiency of crosslinking through the use of a photo-reactive nucleoside variant that also allows reduction of the UV light-induced damage because the energies that are employed are typically lower. PAR-CLIP was applied initially to the determination of targets of several RNA binding proteins including Argonaute (Chi et al., 2009). iCLIP (individual-nucleotide resolution CLIP) is the most recent variant of CLIP that aims to exploit what is in fact a problem in the previous CLIP approaches. This is the fact that crosslinking positions are usually an obstacle for the reverse transcriptase that dissociates at these positions, leading to the loss of the majority of templates at the stage of reverse transcription; iCLIP was applied to decipher among others, the role of heterogeneous nuclear ribonucleoprotein C in splicing regulation (Konig et al., 2011). Comparisons of the efficiency and accuracy of these methods have also started to emerge (Kishore et al., 2011, Sugimoto et al., 2012). A study from our group determined that the type of nuclease and the intensity of digestion could considerably bias the set of identified RNA targets (Kishore et al., 2011). Similar to CLIP, ribosome profiling is a method that takes advantage of ribosome binding to RNAs to determine actively translated regions protected by ribosome along the mRNA and to measure translation rates (Ingolia et al., 2009).

1.3 Initial Processing of the Sequencing Data

The product of a high-throughput transcriptome-wide sequencing is typically 10s of millions of 25-100 nucleotide-long sequence-reads. When a reference genome assembly is available, the first step of the analysis is to align these reads to the reference genome, which may be in the order of several giga-bases long. To find all occurrences of a read in the reference genome, one needs to exhaustively scan each read against the reference. Initially, hash based methods (either for the reference genome or the reads) were used to reduce the search space (surveyed in (Fonseca et al., 2012)). The huge amount of data that is obtained today already makes this process too time-consuming. The fact that the reads don't have perfect complementarity to the reference exponentially increases the search space. The use of algorithms from the computer science field for file compression has essentially solved this problem (Burrows and Wheeler, 1994). Currently, it is possible to map several millions of varying length reads length reads within hours.

1.4 Functional Annotation Of The Sequence Reads

Functional annotation of the short reads, similar to expressed sequence tags (ESTs) and mRNAs before, is usually done on the basis of previously known elements that collocate with the read in the same genomic locus. It has become apparent however, that gene structures in mammalian genomes are quite complex, with multiple genomic elements frequently originating from the same genomic locus. Additionally, RNAs can be extensively modified or nucleotides could be added at the ends of the RNAs post-transcriptionally (Thornton et al., 2012, Trippe et al., 2003, Betat et al., 2010). The first complication causes redundant annotation of reads, while the latter might hinder accurate mapping of the reads. In our group, in the CLIPZ web server that we maintain for annotation of various types of samples (Trippe et al., 2003), we deal with the first issue by setting an annotation hierarchy based on the relative abundance of RNA species within cells. We handle the second issue by allowing increased error rates in the short read-to-genome alignments (typically 1 error in 10 nucleotides) and trimming non-template nucleotides that we know are added to specific transcripts (e.g. CCA for tRNAs) prior to mapping the reads to the reference. More sophisticated methods could be applied depending on the problem setting but this method performs well for common settings.

1.5 Statistical Analysis Of mRNAseq Data For (Differential) Gene Expression Estimation

Only a fraction of mRNA genes are expressed in a cellular state or conditions. This is also reflected in the estimation of mRNA expression based on mRNA-seq experiments; the normalized expression values appear to be well described by a mixture of two components, one corresponding to clearly expressed mRNA species, and the other to very rare mRNAs or to spurious mappings of reads to genomic loci that are not those from which the reads originated (Hebenstreit et al., 2011). The noise in the estimated expression values is higher for low-expressed genes, hence it is important to first determine the expressed genes before carrying out differential expression analysis. Following this, one can simply calculate the expression fold changes in between two different treatments and identify those genes with highest absolute fold change as the differentially expressed genes. Depending on the inherent noise in the data (i.e. whether or not the biological replicates for the same condition behave consistently), one can apply more sophisticated, model based approaches, and call the genes as differentially expressed only when they have a lower variance in the same condition than across different conditions (Hardcastle and Kelly, 2010).

1.6 Statistical Analysis Of CLIP Data

Initial CLIP methods developed in the Darnell lab used the enrichment of reads obtained in CLIP with respect to the reads obtained by *in silico* dicing of transcripts assumed to be expressed at a level given by measurements obtained with microarrays,, for the identification of the targets of the RNA binding proteins (Chi et al., 2009, Hebenstreit et al., 2011). It is unclear though how to best model the biases introduced in the various steps of CLIP. PAR-CLIP and to some extent CLIP introduce however mutations that are indicative of crosslinking. Using more sophisticated, model-based scoring methods that take advantage of crosslink-diagnostic mutations, one can accurately identify the targets of RNA binding proteins (Jaskiewicz et al., 2012), without the need of an mRNA-sequencing sample from which to estimate the relative expression of putative RNA targets.

1.7 APPLICATIONS

In this thesis I describe the methods (Chapter 2) that we developed and the analyses that we carried out to identify targets of various RNA binding proteins. We applied PAR-CLIP to identify RNA targets of PAPD5 (PAP associated domain containing 5), DIS3L2 (DIS3 mitotic control homolog (S. cerevisiae)-like 2), and EWSR1 (Ewing sarcoma breakpoint region 1) as described in Chapter 3-5, respectively.

PAP domain is found in poly(A) polymerases and has been shown to have polynucleotide adenylyltransferase activity in yeast (Motamedi et al., 2004). PAPD5 is the human homolog of yeast Trf4; however, *in vivo* function of it was not known at the time when we initiated the studies described in Chapter 3. We identified rRNAs as the main *in vivo* targets of PAPD5.

Our second application of PAR-CLIP was to DIS3L2, one of the three human homologs of exosome complex exonuclease Dis3 (Astuti et al., 2012, Hebenstreit et al., 2011). Among the Dis3 homologs, DIS3L2 was one of the least characterized. Through computational analysis of DIS3L2-CLIP data, we identified tRNAs as the main DIS3L2 substrates. Additional experiments then suggested that DIS3L2 might be the main exonuclease involved in the biogenesis of tRNA-derived small RNAs, which are subsequently loaded in the RNA silencing complex and function in gene regulation.

Our most recent study aimed to characterize the function of EWSR1, an RNA-binding protein best known for the frequent translocations that its locus shows in specific cancers (reviewed in (Kovar, 2011)). SR1 RNA binding domain is usually lost in the fusion proteins, presumably leading to a loss of function phenotypes in sarcomas. We therefore applied PAR-CLIP to assess the importance of the EWSR1 RNA binding domain. We have additionally integrated analysis of several different dataset to finally establish that EWSR1 binds RNAs originating from instability prone regions including its own pre-mRNA, which is involved in translocations.

1.8 References

2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*.

ASTUTI, D., MORRIS, M. R., COOPER, W. N., STAALS, R. H. J., WAKE, N. C., FEWS, G. A., GILL, H., GENTLE, D., SHUIB, S., RICKETTS, C. J., COLE, T., VAN ESSEN, A. J., VAN LINGEN, R. A., NERI, G., OPITZ, J. M., RUMP, P., STOLTE-DIJKSTRA, I., MÜLLER, F., PRUIJN, G. J. M., LATIF, F. & MAHER, E. R. 2012. Germline mutations in DIS3L2 cause the Perlman syndrome of overgrowth and Wilms tumor susceptibility. *Nat Genet*.

BETAT, H., RAMMELT, C. & MÖRL, M. 2010. tRNA nucleotidyltransferases: ancient catalysts with an unusual mechanism of polymerization. *Cell Mol Life Sci*.

BURROWS, M. & WHEELER, D. 1994. A block-sorting lossless data compression algorithm. *Citeseer*

CHI, S. W., ZANG, J. B., MELE, A. & DARNELL, R. B. 2009. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*.

CONSORTIUM, G. P. 2010. A map of human genome variation from population-scale sequencing. *Nature*.

CONSORTIUM, M., ROY, S., ERNST, J., KHARCHENKO, P. V., KHERADPOUR, P., NEGRE, N., EATON, M. L., LANDOLIN, J. M., BRISTOW, C. A., MA, L., LIN, M. F., WASHIETL, S., ARSHINOFF, B. I., AY, F., MEYER, P. E., ROBINE, N., WASHINGTON, N. L., DI STEFANO, L., BEREZIKOV, E., BROWN, C. D., CANDEIAS, R., CARLSON, J. W., CARR, A., JUNGREIS, I., MARBACH, D., SEALFON, R., TOLSTORUKOV, M. Y., WILL, S., ALEKSEYENKO, A. A., ARTIERI, C., BOOTH, B. W., BROOKS, A. N., DAI, Q., DAVIS, C. A., DUFF, M. O., FENG, X., GORCHAKOV, A. A., GU, T., HENIKOFF, J. G., KAPRANOV, P., LI, R., MACALPINE, H. K., MALONE, J., MINODA, A., NORDMAN, J., OKAMURA, K., PERRY, M., POWELL, S. K., RIDDLE, N. C., SAKAI, A., SAMSONOVA, A., SANDLER, J. E., SCHWARTZ, Y. B., SHER, N., SPOKONY, R., STURGILL, D., VAN BAREN, M., WAN, K. H., YANG, L., YU, C., FEINGOLD, E., GOOD, P., GUYER, M., LOWDON, R., AHMAD, K., ANDREWS, J., BERGER, B., BRENNER, S. E., BRENT, M. R., CHERBAS, L., ELGIN, S. C. R., GINGERAS, T. R., GROSSMAN, R., HOSKINS, R. A., KAUFMAN, T. C., KENT, W., KURODA, M. I., ORR-WEAVER, T., PERRIMON, N., PIRROTTA, V., POSAKONY, J. W., REN, B., RUSSELL, S., CHERBAS, P., GRAVELEY, B. R., LEWIS, S., MICKLEM, G., OLIVER, B., PARK, P. J., CELNIKER, S. E., HENIKOFF, S., KARPEN, G. H., LAI, E. C., MACALPINE, D. M., STEIN, L. D., WHITE, K. P. & KELLIS, M. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*.

FONSECA, N. A., RUNG, J., BRAZMA, A. & MARIONI, J. C. 2012. Tools for mapping high-throughput sequencing data. *Bioinformatics*.

GUTTMAN, M., DONAGHEY, J., CAREY, B. W., GARBER, M., GRENIER, J. K., MUNSON, G., YOUNG, G., LUCAS, A. B., ACH, R., BRUHN, L., YANG, X., AMIT, I., MEISSNER, A., REGEV, A., RINN, J. L., ROOT, D. E. & LANDER, E. S. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*.

HARDCASTLE, T. J. & KELLY, K. A. 2010. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*.

HEBENSTREIT, D., FANG, M., GU, M., CHAROENSAWAN, V., VAN OUDENAARDEN, A. & TEICHMANN, S. A. 2011. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol Syst Biol*.

INGOLIA, N. T., GHAEMMAGHAMI, S., NEWMAN, J. R. S. & WEISSMAN, J. S. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*.

JASKIEWICZ, L., BILEN, B., HAUSSER, J. & ZAVOLAN, M. 2012. Argonaute CLIP - A method to identify in vivo targets of miRNAs. *Methods*.

KATZ, Y., WANG, E. T., AIROLDI, E. M. & BURGE, C. B. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Meth*.

KISHORE, S., JASKIEWICZ, L., BURGER, L., HAUSSER, J., KHORSHID, M. & ZAVOLAN, M. 2011. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Meth*.

KONIG, J., ZARNACK, K., ROT, G., CURK, T., KAYIKCI, M., ZUPAN, B., TURNER, D. J., LUSCOMBE, N. M. & ULE, J. 2011. iCLIP--transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. *J Vis Exp*.

KOVAR, H. 2011. Dr. Jekyll and Mr. Hyde: The Two Faces of the FUS/EWS/TAF15 Protein Family. *Sarcoma*.

MORTAZAVI, A., WILLIAMS, B. A., MCCUE, K., SCHAEFFER, L. & WOLD, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*.

MOTAMEDI, M. R., VERDEL, A., COLMENARES, S. U., GERBER, S. A., GYGI, S. P. & MOAZED, D. 2004. Two RNAi complexes, RITS and RDRC, physically interact and localize to noncoding centromeric RNAs. *CELL*.

SUGIMOTO, Y., KÖNIG, J., HUSSAIN, S., ZUPAN, B., CURK, T., FRYE, M. & ULE, J. 2012. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol*.

THORNTON, J. E., CHANG, H.-M., PISKOUNOVA, E. & GREGORY, R. I. 2012. Lin28-mediated control of let-7 microRNA expression by alternative TUTases Zcchc11 (TUT4) and Zcchc6 (TUT7). *RNA*.

TRAPNELL, C., WILLIAMS, B. A., PERTEA, G., MORTAZAVI, A., KWAN, G., VAN BAREN, M. J., SALZBERG, S. L., WOLD, B. J. & PACHTER, L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*.

TRIPPE, R., RICHLY, H. & BENECKE, B.-J. 2003. Biochemical characterization of a U6 small nuclear RNA-specific terminal uridylyltransferase. *Eur J Biochem*.

ULE, J., JENSEN, K. B., RUGGIU, M., MELE, A., ULE, A. & DARNELL, R. B. 2003. CLIP identifies Nova-regulated RNA networks in the brain. *Science*.

VENTER, J. C., ADAMS, M. D., MYERS, E. W., LI, P. W., MURAL, R. J., SUTTON, G. G., SMITH, H. O., YANDELL, M., EVANS, C. A., HOLT, R. A., GOCAYNE, J. D., AMANATIDES, P., BALLEW, R. M., HUSON, D. H., WORTMAN, J. R., ZHANG, Q., KODIRA, C. D., ZHENG, X. H., CHEN, L., SKUPSKI, M., SUBRAMANIAN, G., THOMAS, P. D., ZHANG, J., GABOR MIKLOS, G. L., NELSON, C., BRODER, S., CLARK, A. G., NADEAU, J., MCKUSICK, V. A., ZINDER, N., LEVINE, A. J., ROBERTS, R. J., SIMON, M., SLAYMAN, C., HUNKAPILLER, M., BOLANOS, R., DELCHER, A., DEW, I., FASULO, D., FLANIGAN, M., FLOREA, L., HALPERN, A., HANNENHALLI, S., KRAVITZ, S., LEVY, S., MOBARRY, C., REINERT, K., REMINGTON, K., ABU-THREIDEH, J., BEASLEY, E., BIDDICK, K., BONAZZI, V., BRANDON, R., CARGILL, M., CHANDRAMOULISWARAN, I., CHARLAB, R., CHATURVEDI, K., DENG, Z., DI FRANCESCO, V., DUNN, P., EILBECK, K., EVANGELISTA, C., GABRIELIAN, A. E., GAN, W., GE, W., GONG, F., GU, Z., GUAN, P., HEIMAN, T. J., HIGGINS, M. E., JI, R. R., KE, Z., KETCHUM, K. A., LAI, Z., LEI, Y., LI, Z., LI, J., LIANG, Y., LIN, X., LU, F., MERKULOV, G. V., MILSHINA, N., MOORE, H. M., NAIK, A. K., NARAYAN, V. A., NEELAM, B., NUSSKERN, D., RUSCH, D. B., SALZBERG, S., SHAO, W., SHUE, B., SUN, J., WANG, Z., WANG, A., WANG, X., WANG, J., WEI, M., WIDES, R., XIAO, C., YAN, C., et al. 2001. The sequence of the human genome. *Science*.

Chapter 2: Argonaute CLIP – A method to identify in vivo targets of miRNAs

Lukasz Jaskiewicz, Biter Bilen, Jean Hausser, Mihaela Zavolan*

Biozentrum, University of Basel, Klingelbergstrasse 50-70, 4056 Basel, Switzerland

*Corresponding author (Email: mihaela.zavolan@unibas.ch, Phone: +41 612671577, Fax: +41 612671585)

Running title: Argonaute CLIP

Keywords: Argonaute; Ago; miRNA; microRNA; miRNA targets; CLIP; PAR-CLIP; Computational biology

2.1 Summary

microRNAs are important regulators of gene expression that guide translational repression and degradation of target mRNAs. Only relatively few miRNA targets have been characterized, and computational prediction is hampered by the relatively small number of nucleotides that seem to be involved in target recognition. Argonaute (Ago) crosslinking and immunoprecipitation (CLIP) in combination with next-generation sequencing proved to be a successful method for identifying targets of endogenous cellular miRNAs on a transcriptome-wide scale. Here we review various approaches to Ago CLIP, describe in detail the PAR-CLIP method and provide an outline of the necessary computational analysis for identification of in vivo miRNA binding sites.

2.2. Overview of CLIP Methods

The rapid development of high throughput technologies during the last decade made it possible to investigate cellular processes on a global scale. Crosslinking and immunoprecipitation (CLIP) became a method of choice for identifying target sites of individual RNA-binding proteins. Initially employing classical sequencing, CLIP has been recently combined with next-generation sequencing to generate comprehensive catalogues of binding sites. Several variants are currently in use, most notable being high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) (Licatalosi et al., 2008), Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) (Hafner et al., 2010) and individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP) (Konig et al., 2010). Although initially applied to the problem of identifying direct RNA-binding protein (RBP)-RNA interactions, recent studies (Hafner et al., 2010, Chi et al., 2009) demonstrated that CLIP can be used in more complex situations, as for example identification of mRNA targets of microRNAs (miRNAs). These target sites are part of ternary complexes (RISC for RNA-induced silencing complex) that contain, beside the mRNA, an Argonaute (Ago) protein and a guiding miRNA.

Elucidation of miRNA targets had been a challenge since the discovery of these regulatory molecules. Although it quickly became apparent, particularly from computational analyses, that a single miRNA has hundreds of targets (Grimson et al., 2007), experimental studies typically focus on a single or at best several targets of a given miRNA. Because the interaction between a miRNA and its target can involve as few as 7-8 nucleotides (Bartel, 2009), it is difficult to accurately predict miRNA targets transcriptome-wide with computational methods. To circumvent this problem and obtain extensive sets of miRNA targets from which additional specificity determinants can be inferred, a number of groups applied Ago-CLIP and identified miRNA/Argonaute targets in mouse tissues and cells (Chi et al., 2009, Leung et al., 2011), human cells (Hafner et al., 2010, Gottwein et al., 2011, Kishore et al., 2011, Stark et al., 2012), and in the worm Caenorhabditis elegans (Zisoulis et al., 2010) in close-to-physiological conditions.

A typical CLIP protocol is quite complex. It involves the UV-induced crosslink formation, cell lysis, immunoprecipitation of the protein of interest along with the crosslinked RNA, ribonuclease (RNase) treatment in order to reduce the RNA targets to approximately the protein binding sites, RNA isolation, adaptor ligation, reverse transcription and library generation by PCR and finally, deep sequencing. It has been argued (Hafner et al., 2010, Kishore et al., 2011) that the resulting libraries yield quantitative information about the miRNAs present in the cell as well as the sequences of their target sites. Recently, we have performed a quantitative analysis of variants of CLIP and PAR-CLIP protocols and determined that the RNase fragmentation step used during sample preparation can introduce a substantial bias in the set of isolated RNAs, while the use of photoactivatable nucleotide analogue moderately improves the CLIP performance (Hafner et al., 2010).

Here we outline a typical Ago CLIP or PAR-CLIP experiment aiming at the isolation and analysis of miRNAs and their targets in vivo and discuss the computational analysis of the resulting sequence data (Figure 1).

2.3 Detailed CLIP protocol

2.3.1 Incorporation of photoactivatable nucleotide analog and UV crosslinking

Incorporation of modified, photoreactive nucleotides like 4-thiouridine (4SU) or 6-thioguanosine (6SG) is mandatory for PAR-CLIP and can result in higher crosslinking efficiency (Hafner et al., 2010, Kishore et al., 2011). Furthermore, the crosslinked modified nucleotides lead to the introduction of characteristic mutations (T-to-C for 4SU and G-to-A for 6SG, (Hafner et al., 2010)) during reverse transcription, providing a means to pinpoint the exact nucleotide that was contacted by the protein (Hafner et al., 2010, Kishore et al., 2011). Subsequent studies demonstrated that classical crosslinking at

254 nm in the absence of a modified nucleotide as in HITS-CLIP also leads to specific mutations that can be exploited in the same way as the T-to-C or G-to-A mutations in PAR-CLIP to pinpoint the binding sites (Kishore et al., 2011, Zhang and Darnell, 2011), although the frequency of these mutations is substantially lower than in PAR-CLIP. While both PAR-CLIP and HITS-CLIP achieve nucleotide resolution through diagnostic mutations that are introduced as the reverse transcriptase (RT) transcribes over a crosslinked nucleotide, the propensity of RT to actually stop transcribing at these nucleotides has been exploited in another CLIP method called iCLIP (Konig et al., 2010).

Adherent or suspension cells can be grown in presence of the modified nucleoside. Concentrations of 100 μM 4SU or 10-100 μM 6SG are typically used, and labeling times are usually 12 – 14 hours. Afterward, the cells are crosslinked with 365 nm UV light (typically with 150 mJ/cm2), pelleted by centrifugation, flash frozen in liquid nitrogen and stored at -80 °C until use.

2.3.2 Cell lysis and immunoprecipitation

Depending on the experimental design, cytoplasmic, nuclear or whole cell extract can be prepared. The antibody selection is critical for an efficient and accurate CLIP experiment. For human Ago2 CLIP and PAR-CLIP we successfully used the 11A9 monoclonal antibody (Rudel et al., 2008), while for mouse Ago2 2A8 monoclonal (Nelson et al., 2007) and 2D4 (Wako) were reported to work in HITS-CLIP (Chi et al., 2009, Leung et al., 2011). If specific antibodies are not available, an alternative strategy is to express a FLAG-tagged version of the protein and perform the immunoprecipitation with anti-FLAG antibodies. The latter approach has been employed to determine the targets of all four Ago proteins in human cells (Hafner et al., 2010).

2.3.3 Nuclease treatment

RNase treatment is an essential step in CLIP library preparation. The type, amount and time of RNase digestion must be carefully optimized to yield RNA fragments whose length distribution is in a range that enables their unambiguous assignment to appropriate genomic loci and that are minimally affected by the sequence bias of the nuclease. In CLIP/HITS-CLIP/iCLIP experiments either a mixture of RNase A and RNase T1 (Chi et al., 2009, Ule et al., 2005) or bacterial RNase I are used (Wang et al., 2009), while in PAR-CLIP extensive RNase T1 digestion has been employed (Hafner et al., 2010, Kishore et al., 2011). The high specificity of RNase T1 of cleaving after guanosine residues, coupled with a high efficiency of this nuclease can lead to isolation of RNA fragments that are essentially devoid of G nucleotides. This bias can be diminished substantially by titrating the amount of nuclease in the experiment (Kishore et al., 2011). Alternatively, a nuclease of the complementary specificity can be used, like micrococcal nuclease (MNase) that exhibits preference for A and U nucleotides (Kishore et al., 2011, Zisoulis et al., 2010).

In a typical Ago2 CLIP experiment we use 5 U/ml of RNase T1 in the extract prior to immunoprecipitation and 20 U/μl on the beads for 15 min at 22 °C in or 0.2 gel U/μl of MNase for 5 min at 37 °C (Kishore et al., 2011). In these conditions RNA is partially digested into 20-100 nucleotide fragments (Jaskiewicz, Kishore and Zavolan, unpublished).

2.3.4 SDS-PAGE and recovery of the immunoprecipitated RNA

The fragmented RNA that is still covalently linked to the protein is then dephoshorylated by alkaline phosphatase and radioactively labeled using 32P-γ-ATP and T4 polynucleotide kinase. This step allows the identification of the immunoprecipitated protein-RNA complexes after SDS-PAGE. It is worth noting that the protein is not in a denatured state during the phosphorylation reaction, so the nature of the protein-RNA interaction can influence the yield of radiolabeled RNA. miRNAs are bound by Ago in a specific fashion: their 5’ phosphate is positioned within the protein binding pocket (Wang et al., 2008) thereby being partially inaccessible for enzymatic reactions, in particular ligation to the 5’ adaptor (Jaskiewicz, Kishore and Zavolan, unpublished data), which can be important when the adaptor ligation is performed on the beads prior to SDS-PAGE.

After labeling, protein-RNA complexes are eluted from the antibodies with SDS-loading buffer and subjected to SDS-PAGE. The gel is subsequently visualized by autoradiography and the sections that correspond to the RNA-bound fragments are excised and further processed. In the PAR-CLIP protocol the recovered sections of the gel are subjected to electroelution, proteinase K treatment to remove proteins, phenol-chlorophorm extraction and RNA precipitation. Studies with the CLIP protocol (Ule et al., 2005) found that the transfer from the SDS-PAGE to a nitrocellulose membrane significantly improves the signal:noise ratio by eliminating free RNA fragments that happen to migrate at the same apparent size as immunoprecipitated protein-RNA complexes. The isolated RNA-protein complexes can also be visualized by autoradiography, and the membrane can be similarly subjected to excision of the relevant sections, proteinase K treatment to liberate immunoprecipitated RNA fragments, phenol-chlorophorm extraction and precipitation of the RNA.

2.3.5 Ligation of adaptors

As with most other steps, adaptor ligation has also been performed in variety of scenarios in the various CLIP protocols. The original CLIP (Ule et al., 2005) called for ligation of one of the adaptors prior to SDS-PAGE, allowing an easier separation of ligated and unligated adaptor-RNA products. Care must be taken however that only fragments that are above the position at which the protein is expected to migrate on the gel are recovered. In addition, as detailed in the previous paragraph, the conformation of the protein-RNA complex might negatively influence the efficiency of the ligation. For these reasons the PAR-CLIP protocol calls for ligation of the protein-free RNA fragments after gel electrophoresis (or after gel transfer). To minimize the fraction adaptor-adaptor ligation products, sequential ligation of adaptors is recommended. The 3' preadenylated adaptor is ligated first with truncated T4 RNA ligase 2 K227Q (New England Biolabs) and ATP-free buffer, maximizing the efficiency of RNA-adaptor ligation and preventing concatenation of RNA fragments (Hafner et al., 2008). After 3' adaptor ligation the RNA is resolved by denaturing 15% urea-PAGE, and the ligated fragments are excised, eluted, precipitated and ligated to the 5' adaptor using T4 RNA ligase 1 in ATP-containing buffer. The ligation reaction is then resolved by 12% urea-PAGE, eluted and precipitated.

2.3.6 Reverse transcription, PCR amplification and introduction of indexes for multiplexing

The nucleotide-level resolution in binding site identification has been achieved by exploiting in various ways reverse transcription through crosslinked sites. Reverse transcriptase has a tendency to terminate at the position of crosslink (Urlaub et al., 2002), although the frequency with which this happens in the conditions of the CLIP experiment is unknown. The iCLIP protocol (Konig et al., 2010) has been developed based on this premise yielding sequence reads whose 5' ends should map immediately downstream of crosslinks. HITS-CLIP and PAR-CLIP protocols on the other hand rely precisely on the reads that are generated when the RT does transcribe through a crosslinked site. It appears that nucleotide misincorporation occurs frequently at such sites, especially in PAR-CLIP, leading to diagnostic mutations and nucleotide-level resolution in the identification of binding sites.

The cDNA resulting from reverse transcription is PCR-amplified, generating a library which is then submitted to next-generation sequencing. Because extensive PCR amplification is known to bias the composition of the library (Pinard et al., 2006, Dohm et al., 2008), various CLIP protocols take precautions to minimize such effects. We opt for minimizing the number of PCR cycles used in the library amplification. Libraries generated with Ago2 PAR-CLIP require typically 14-18 amplification cycles to yield sufficient material for sequencing. This does not appear to introduce strong biases because the sites derived by taking into account the actual copy number of the reads that are derived from unique fragments are similar in the frequency of expected protein-specific sequence motifs (Kishore et al., 2011). In HITS-CLIP and iCLIP, the PCR amplification is more extensive (25-35 cycles), but the introduction of random barcodes before amplification allows one to identify sequences that originated from the same RNA fragment (Konig et al., 2010, Chi et al., 2009). These will be collapsed to a single sequence during annotation leading to appropriate quantification of reads.

Current sequencing depth allows for sample multiplexing without sacrificing the library complexity. We typically multiplex 4-6 samples in a sequencing lane of a HiSeq 2000 instrument. Different reverse primer sequences need to be used, as per Illumina indexing protocol.

The step-by-step Ago2 CLIP protocol is provided in Extended Methods.

2.4 SITE DEFINITION AND EXTRACTION BASED ON ENRICHMENT RELATIVE TO MRNA-SEQ

Reads obtained from deep sequencing need to be annotated and used for the extraction of individual Ago binding sites. For this purpose we use an automated procedure which we implemented in the CLIPZ server (www.clipz.unibas.ch (Khorshid et al., 2011)). For each of the steps of this procedure various solutions have been proposed and implemented, but they have not been so far bundled in stand-alone analysis packages that can be readily used on the raw sequence data. After discarding low-quality sequences (based on the number of nucleotides that could not be unambiguously called) we use an in-house implementation of an ends-free alignment algorithm to identify 3’ adaptor fragments that are present in the reads and remove them. Reads that are at least 15 nucleotides in length after adaptor removal are then mapped to the genome assembly using a short read alignment program. Recently, the Burrows-Wheeler transform has been used to develop very efficient short read alignment programs (Langmead et al., 2009, Li and Durbin, 2009). In our analyses we use the Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009), which allows us to find not only ungapped but also gapped alignments, which in the case of reads derived from CLIP experiments are quite common due to nucleotide deletions that are introduced in the cDNAs as the RT transcribes over a nucleotide that was crosslinked (Kishore et al., 2011). We then assign the reads with equal probability to the possible genomic loci, but for most analyses focus on the subset of reads that map uniquely to the genome.

Typical CLIP experiments yield a large number of reads that appear to originate in many different types of RNAs. Initial studies, including ours (Hafner et al., 2010, Kishore et al., 2011) have focused on mRNA-derived reads. Furthermore, to be able to quantify and compare the relative abundance of various sites in CLIP data, we typically select one representative mRNA for every gene, defined as the longest mRNA in the Refseq database (Pruitt et al., 2009) that has annotated 5' UTR, coding domain and 3' UTR. The mappings of CLIP reads to these representative mRNAs serve as the basis for the binding site identification described below.

Binding sites are identified as peaks in the read density profile along transcripts. A number of variants of this procedure have been proposed (Chi et al., 2009, Kishore et al., 2011, Corcoran et al., 2011). Our approach was to identify non-overlapping regions of identical length (we typically used 40 nucleotides-long regions) with high coverage by reads. In a given transcript we start with the region with highest coverage by reads, then identify the window with the next-highest coverage that is not overlapping with the first, etc.

Different binding sites presumably have different affinities to Ago. The high affinity binding sites, in the vicinity of which Ago is present a large fraction of the time, should have the highest propensity to crosslink with Ago during the CLIP experiment and should subsequently yield a large number of CLIP reads. One could thus argue that ranking binding sites by the number of reads would allow one to discover the highest affinity and perhaps most relevant binding sites. However, the number of CLIP reads mapping to a given binding site depends on other factors, among which the abundance of the mRNA in which the site resides. If two sites s_A and s_B of equal affinity to Ago are located on different mRNAs m_A and m_B, the abundance of m_A being much higher than that of m_B, more sequences of type m_A will be available for crosslinking than sequences of type m_B resulting into more CLIP reads of type s_A than of type s_B. A few approaches have again been used to remove the confounding effect of mRNA abundance in the ranking of biding sites. Chi et al. (Chi et al., 2009) use an in silico fragmentation procedure to estimate the relative abundance of mRNA fragments based on mRNA expression data obtained through microarray experiments. We used mRNA-seq data obtained from the same cell line in which the CLIP experiments were performed to the same end (Kishore et al., 2011).

Namely, by computing the ratio between the number of CLIP reads mapping to a putative binding site and the number of reads that would be expected to map to that site if there was no specific selection of RNA fragments by the Ago CLIP, we can obtain an estimate of the affinity of the site for the RISC complex. Ranking sites by this ratio is also not entirely problem-free. For instance, the enrichment ratio for CLIP sites located on mRNAs that were not captured in mRNA-seq is infinity, irrespective of the number of reads obtained from this site in CLIP. To deal with this issue and obtain meaningful estimates of enrichment, we currently use a Bayesian probabilistic model in which the fraction p_i of CLIP reads expected to map to site i given mRNA abundance alone can be written as

$$p_i = \frac{n_{\mu(i)} + 1}{l_{\mu(i)} \sum_{j=1}^{c} \frac{n_{\mu(j)}+1}{l_{\mu(j)}}} \tag{1}$$

where μ(i) denotes the mRNA that carries site i, $n_{\mu(i)}$ is the number of mRNA-seq reads that map to the mRNA on which site i is located, $l_{\mu(i)}$ is the length of that mRNA and c is the total number of CLIP sites. Consequently, if r_i out of a total of r CLIP reads map to site i, the enrichment ratio for site i is $\frac{r_i}{p_i r}$. Based on the enrichment ratio p_i, one can rank CLIP sites located on mRNAs of widely different abundances in a consistent fashion, using a single metric. To further reflect our expectation that at equal enrichment ratios we would be more confident in the estimate of enrichment for the site supported by more reads, we compute the posterior probability $P(H_1|r_i, r)$ that a given site is enriched compared to what would be expected from the abundance of the corresponding mRNA alone. This probability is given in terms of the reverse cumulative of a beta distribution as

$$P(H_1|r_i, r) = \int_{p_i}^{1} \frac{\Gamma(r+2)}{\Gamma(r_i+1)\Gamma(r-r_i+1)} \rho^{r_i}(1-\rho)^{r-r_i} d\rho \tag{2}$$

where p_i is the fraction of CLIP reads expected to map to site i (see Equation 1) and r is the total number of CLIP reads, r_i of which map to the CLIP site i (Hausser and Zavolan, unpublished).

The tests that we carried out with various measures for ranking CLIP sites of Ago and HuR proteins indicate that the probability $P(H_1|r_i, r)$ of a site being enriched over what is expected from mRNA abundance is most accurate. It is also an intuitive, general method that can be applied equally to PAR-CLIP data (Hafner et al., 2010, Kishore et al., 2011), which typically feature a large number of diagnostic mutations, and to HITS-CLIP data (Chi et al., 2009) in which diagnostic mutations are scarcer (Kishore et al., 2011). Nonetheless, when mRNA-seq data is not available, bona fide binding sites can still be obtained by ranking sites according to the number of CLIP reads mapping to them or by the frequency of crosslink-diagnostic mutations. These mutations can also be used a posteriori to identify the position on the mRNA where the crosslink occurred with nucleotide resolution, as we will examine in the next section.

2.5 Ranking sites based on crosslink-induced mutations

T-to-C mutations in the cDNAs have been used to identify crosslinked sites since the introduction of the PAR-CLIP method (Hafner et al., 2010). Recently, kernel density estimates of these events were utilized to identify binding sites (Corcoran et al., 2011). Our analyses ((Kishore et al., 2011) and Bilen, Hausser and Zavolan, unpublished) indicate that although such mutations do indeed pinpoint binding sites at nucleotide resolution, measures of the density of crosslink-diagnostic mutations do not necessarily outperform that based on the enrichment in reads. Nonetheless, particularly for situations when the binding sites are located in non-protein coding transcripts whose abundance is difficult to estimate, we developed a model to estimate the probability that the frequency of crosslink-diagnostic mutations in reads originating from a given site is typical for a crosslinked site. This is because apparent T-to-C mutations can be also due to sequencing errors, mis-mapped reads or to single nucleotide polymorphisms, while crosslink-induced mutations in a given sample appear to occur at a characteristic frequency (Figure 2).

Formally, let us assume that we observe k T-to-C mutations at a position covered by n reads. Under the assumption that the reads were independently sampled and that T-to-C transitions occur at rate μ, the number of mutations k can be modeled to follow a binomial distribution and thus the probability of the data can be written as

$$P(k|n,\mu) = \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)}\mu^{k}(1-\mu)^{n-k} \tag{3}$$

where $0 \leq \mu \leq 1$ and $\Gamma(x) = (x-1)!$ is the gamma function. Further assuming that the unknown probability of T-to-C mutations, μ, has a prior probability of the form

$$P(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{4}$$

where $a > 0$ and $b > 0$, we can use Bayes' theorem to calculate the posterior probability distribution of μ as

$$P(\mu|k,n) = \frac{P(k|\mu,n)P(\mu|n)}{\int_0^1 P(k|\mu,n)P(\mu|n)d\mu}. \tag{5}$$

Substituting $P(\mu|n) = P(\mu|a,b)$ from Eq. 4 leads to

$$P(\mu|k,l,a,b) = \frac{\Gamma(k+a+l+b)}{\Gamma(k+a)\Gamma(l+b)}\mu^{k+a-1}(1-\mu)^{l+b-1} \tag{6}$$

where $l = n - k$. Finally, as a measure of our confidence that the given position represents a crosslink site we compute the probability that the rate of T-to-C transitions μ, is within a certain range defined by the lower and upper bounds μ_l and μ_u:

$$P(\mu > \mu_l \; and \; \mu < \mu_u|k,l,a,b) = \frac{\int_{\mu_l 0}^{\mu_u} x^{k+a-1}\,(1-x)^{l+b-1}dx}{\int_0^1 x^{k+a-1}\,(1-x)^{l+b-1}dx} \tag{7}$$

To be able to estimate the parameters a and b of this distribution as well as the cutoffs μ_l and μ_u, we analyzed the distribution of the frequency of T-to-C mutations as a function of the coverage of a given nucleotide in the data. We found that nucleotides with a reasonably high coverage have a very narrowly distributed frequency of T-to-C mutation (around 0.1 in the Ago PAR-CLIP datasets that we obtained, Figure 2). Thus, we use a "trusted set" of nucleotides that we believe were crosslinked in the experiment to estimate the a and b parameters by maximum likelihood. Defining the set of trusted nucleotides is done on the basis of the data. As lower and upper bounds μ_l and μ_u, we use the values that correspond to the 0.001 and 0.999 quantiles of the prior distribution of μ defined by formula (Chi et al., 2009). Finally, we calculate the posterior probability of the observed T-to-C mutations being within the boundaries that we expect for a crosslinked position with formula (Leung et al., 2011). More specific information can be used to further reduce the rate of false positives. For example, single-nucleotide polymorphisms can be identified based on RNA-seq data and explicitly excluded from the set of putative crossliked positions. Similarly, reads that come from repetitive genomic regions and are likely to be mis-mapped can be excluded from the analysis.

2.6 Identification of miRNA binding sites in Ago-CLIP sites

The methods described above allow us to identify genomic regions to which the Argonaute protein is crosslinked with high probability. They do not immediately reveal which miRNA guided the interaction of Argonaute with the mRNA. A trivial solution to this problem is to intersect the Ago-binding sites with predicted miRNA-binding sites. This approach has of course the limitation that no novel binding sites would be discovered. Furthermore, the most accurate target prediction programs enforce the requirement that ~7 nucleotides from the 5' end of miRNA (so-called 'seed') are perfectly complementary to the target site but many Ago-CLIP sites do not seem to be explained by this type of sequence complementarity. In cases when a miRNA is very highly and specifically expressed and at the same time has a particular type of non-canonical (i.e. not a seed-matching) site that occurs with high frequency in the data (Chi et al., 2012), de novo motif finding with programs such as MEME (Bailey et al., 2009) can reveal the miRNA binding site. Nonetheless, a principled method for identifying the guide miRNA associated with each Ago-CLIP site is still to be developed.

2.7 Conclusions

CLIP approaches are instrumental in the discovery of miRNA target sites and for gaining a mechanistic understanding of their interaction with RNA-induced silencing complexes. However, the obtained data is complex, and accurate methods, that need not make assumptions currently made by miRNA target prediction programs, are needed to identify bona fide miRNA binding sites. The large collection of experimentally identified sites can then be used to learn more about miRNA-target interactions, because only 40-60% of sites can typically be explained through matching to the seed region of abundantly expressed miRNAs (Hafner et al., 2010, Kishore et al., 2011). In the very near future, these approaches are expected to yield comprehensive catalogs of miRNAs and their targets in specific cells, tissues and organisms.

Table 1. Comparison of different Ago2 CLIP methods

	Chi et al., 2008	Hafner et al., 2010	Kishore et al., 2011
Organism, cell type	mouse, brain	human, cell line	human, cell line
Immunoprecipitated protein	Endogenous	FLAG-tagged, overexpressed	endogenous
Protocol	CLIP (HITS-CLIP)	PAR-CLIP	PAR-CLIP / CLIP
Modified ribonucleoside	None	4-thio-uridine	4-thio-uridine/none
Crosslinking UV wavelength	254 nm	365 nm	365 nm / 254 nm
Nuclease	RNase A	RNase T1	RNase T1, MNase
Nuclease digestion	Limited	extensive	extensive / limited
Radioactive labeling	3' adaptor	immunoprecipitated RNA	immunoprecipitated RNA
3' adaptor ligation	on beads, prior to SDS-PAGE	after SDS-PAGE	after SDS-PAGE
Preadenylated 3' adaptor and truncated RNA ligase	No	yes	yes
Nitrocellulose transfer after SDS-PAGE	Yes	no	no
Crosslink-induced mutation analysis for site identification	No	yes	yes, for both PAR-CLIP and CLIP

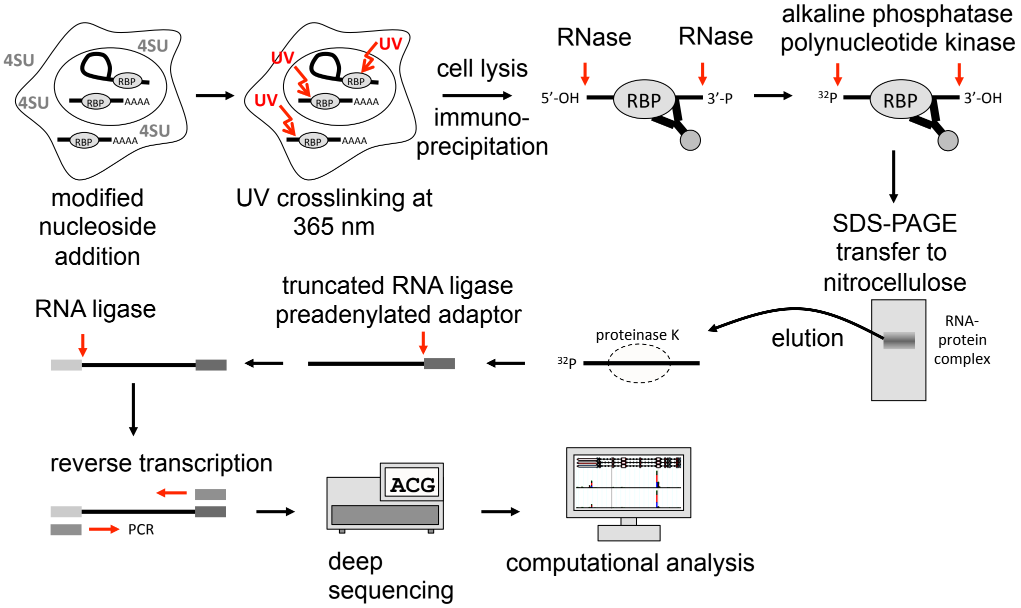


Figure 1. Sketch of the CLIP protocol. Incubation with a photoactivatable nucleoside analog like 4SU is performed in PAR-CLIP and cells 365 nm UV light. In CLIP the incubation with modified nucleoside is omitted and the crosslinking is performed at 254 nm. Optimization of ribonuclease treatment conditions is recommended in order to avoid too extensive RNA digestion which can deplete bona fide binding sites. After immunoprecipitation, radioactively labeled RNA–protein complexes are resolved by SDS–PAGE and blotted to nitrocellulose membrane. This step reduces the background of non-crosslinked RNA fragments. After adaptor ligation and reverse transcription, a pilot PCR should be performed to determine the number of amplification cycles in the preparative-scale PCR. The number of cycles should be chosen to allow library amplification with minimal PCR amplification bias.

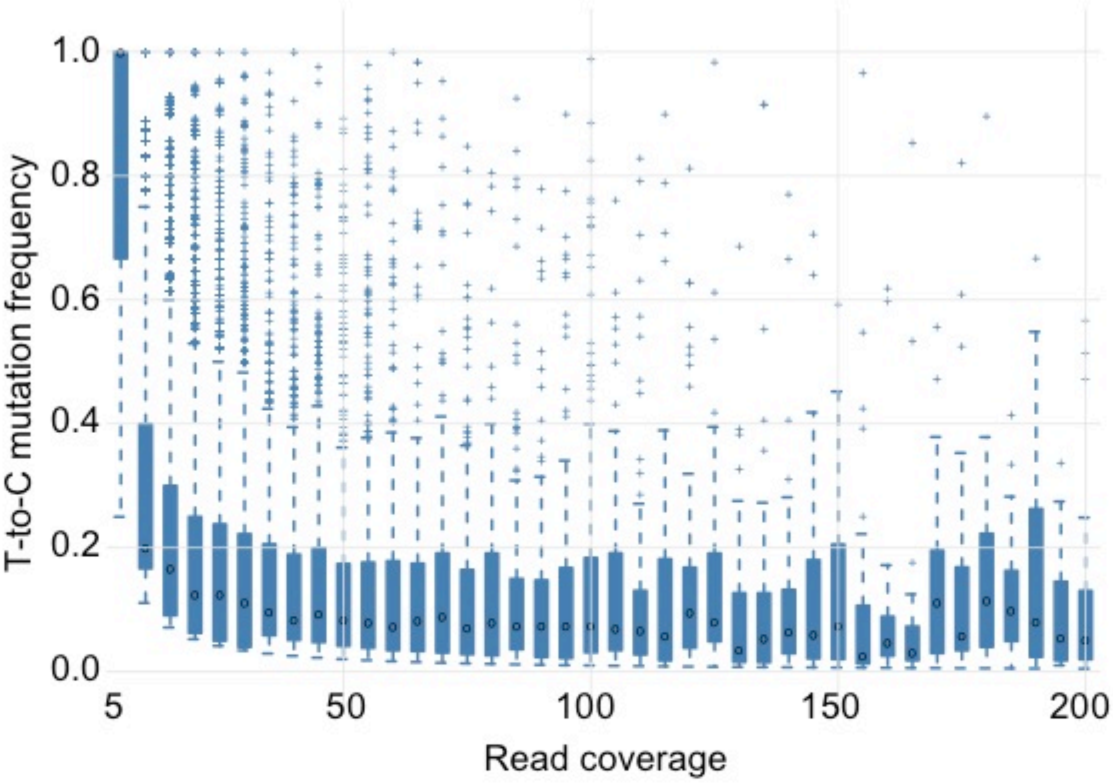


Fig. 2. Frequency of T-to-C mutations at individual genomic positions that are covered by mRNA-annotated reads obtained in an Argonaute2 PAR-CLIP experiment (Kishore et al., 2011). The reads obtained in the experiment were annotated with the CLIPZ server (www.clipz.unibas.ch). Reads that were annotated as mRNAs were used to identify genomic positions at which T-to-C mutations occurred, and to compute the frequency of T-to-C mutations (number or reads carrying the T-to-C mutation divided by the total number of reads covering the given genomic position). Genomic positions were then divided into bins, based on the number of mRNA-annotated reads that covered them. The first bin corresponded to [0,5) reads covering an individual genomic position, the second bin corresponded to [5,10) reads per position etc. Finally, we constructed box-plots of the mutation frequency across all genomic positions in a bin. The box indicates the inter-quartile range, the circle the median mutation frequency and the crosses show the outliers.

2.8 EXTENDED METHODS

2.8.1 Step-by-step Ago2 PAR-CLIP protocol

- Grow 5x108 – 109 cells. At ~60% confluency, add 4SU to final concentration of 100 μM and grow for additional 12-14 hours. Wash cells with ice cold PBS and crosslink on ice with 365 nm UV at 0.15 J/cm2 in a crosslinker (e.g. Stratalinker 2400 from Stratagene). Collect cells by centrifugation at 2000 rpm for 5 min at 4 °C, flash freeze and store at -80 °C until use.
- Add 3 ml of NP40 lysis buffer per g of cell pellet, incubate 10 min on ice, centrifuge 14.000 g at 4 °C and filter supernatant through 0.45 μm filter.
- Perform initial RNase T1 digestion with 5 U/ml RNase T1 (Fermentas) at 22 °C for 15 min.
- Preincubate protein G magnetic beads (Dynabeads, Life technologies) with Ago2 antibody (generally 10 μl of beads per ml of cell lysate) with Ago2-specific antibody (final concentration 0.25 mg/ml) at 4 °C for 1 hour, wash twice with 1 ml of citrate phosphate buffer, add to lysate and incubate for 2 hours at 4 °C.
- Collect beads on ice using magnetic rack, wash 3 times with 1 ml of IP washing buffer and resuspend in original amount of IP washing buffer containing 20 U/μl of RNase T1 (Fermentas), incubate at 22 °C for 15 min. Alternatively, use MNase (NEB) at 0.2 gel U/μl final concentration instead of RNase T1. As MNase activity is very sensitive to the salt concentration, MNase digestion should be carried out in the buffer supplied with the enzyme and incubated at 37 °C for 5 min.
- Wash beads 3 times with 1 ml of high salt buffer and resuspend beads in the original bead volume of 1x NEB buffer 3 with calf intestinal phosphatase (NEB) at a final concentration of 0.5 U/μl and incubate at 37 °C for 10 min with intermittent shaking.

- Wash beads twice with 1 ml of crosslink washing buffer and twice with 1 ml of PNK buffer.
- Resuspend beads in the original volume of PNK buffer containing 5 mM DTT, add γ32P-ATP to a final concentration of 0.5 μCi/μl and T4 PNK (NEB) to a final concentration of 1 U/μl and incubate at 37 °C for 30 min with intermittent shaking. Next, add ATP to a final concentration of 100 μM and incubate at 37 °C for additional 5 min.
- Wash beads 5 times with PNK buffer without DTT.
- Resuspend beads in 1x SDS-loading buffer, incubate at 90 °C for 3 min and load the supernatant on NuPAGE 4-12% Bis-Tris gel in 1x MOPS running buffer (Life technologies), resolve for 1 hour 15 min at 150 mA, 200 V.
- Incubate gel in 2x NuPAGE transfer buffer for 5 min and transfer to nitrocellulose membrane using semi-dry blotter at 20 V, 400 mA for 2 to 3 hours.
- After transfer rinse the nitrocellulose in 1x PBS, wrap membrane in plastic wrap and expose to phosphoimager screen. Depending on the signal expose for 10 min to several hours.
- Excise membrane fragments corresponding to the migration of Ago2-RNA complexes (~100-130 kDa), add in proteinase K at final concentration of 1.2 mg/ml in 1x proteinase K buffer to the membrane fragments and incubate at 55 °C for 30 min, spin down and recover the supernatant.
- Add 1 vol of phenol-chlorophorm-isoamyl alcohol (25:24:1) solution, vortex and centrifuge at 14000 g for 15 min at room temperature. Transfer aqueous phase to fresh tubes, add 1 vol of chlorophorm, vortex and centrifuge at 14000 g for 15 min at room temperature.

- Recover the aqueous phase into fresh tubes, add 1/10 volume of 3M NaCl, 1 μl of GlycoBlue (Ambion) and 3 volumes of absolute ethanol. Precipitate at -80 °C for 30 min and centrifuge at 14000 g for 30 min at 4 °C. Discard supernatant and wash pellets with 70% ethanol. Centrifuge at 10000 g for 10 min at 4 °C, discard supernatant and briefly (3-5 min) air-dry pellets at room temperature.
- Resuspend pellets in 19 μl of 3'-adaptor ligation mixture containing 2.5 μM preadenylated adaptor RA3 (Illumina) and 1x RNA ligase buffer without ATP (NEB), denature at 90 °C for 30 s, place on ice, add 1 μl of 1 μg/μl T4 RNA ligase 2 truncated K227Q (NEB) and incubate for 2-3 hours at 22 °C. As an efficiency control, a reaction containing ligation mixture and 19 nt RNA oligonucleotide should be separately performed.
- Add 1 vol of 2x RNA loading buffer (Fermentas), denature by incubating at 90 °C for 30 s and resolve on 15% urea-PAGE in 1x TBE. Wrap gel in plastic wrap and expose to phosphorimager screen for 20 – 60 min.
- Excise gel fragments containing bands corresponding to the ligated product size, generally between 40 and 70 nt, elute RNA by incubating in 0.4 M NaCl overnight at 4 °C on a rotating wheel and precipitate with 3 vol of absolute ethanol at -80 °C for 30 min and centrifuge at 14000 g for 30 min at 4 °C. Discard supernatant and wash pellets with 70% ethanol. Centrifuge at 10000 g for 10 min at 4 °C, discard supernatant and briefly (3-5 min) air-dry pellets at room temperature.
- Resuspend pellets in 18 μl of 5' ligation mix containing 5 μM 5' adaptor (RA5, Illumina), and 1x RNA ligase buffer with 0.2 mM ATP, denature for at 90 °C for 30 s , place on ice, add 2 μl of T4 RNA ligase 1 (Fermentas) and incubate at 37 °C for 1 hour.

- Add 1 vol of 2x RNA loading buffer (Fermentas), denature by incubating at 90 °C for 30 s and resolve on 12% urea-PAGE in 1x TBE. Wrap gel in plastic wrap and expose to phosphorimager screen for 30 – 180 min.
- Excise gel fragments containing fragments corresponding to the ligated product size, generally between 60 and 90 nt, elute RNA by incubating in 0.4 M NaCl with 1 μl of 100 μM reverse transcription primer (RTP, Illumina) overnight at 4 °C on a rotating wheel and precipitate with 3 vol of absolute ethanol at -80 °C for 30 min and centrifuge at 14000 g for 30 min at 4 °C. Discard supernatant and wash pellets with 70% ethanol. Centrifuge at 10000 g for 10 min at 4 °C, discard supernatant and briefly (3-5 min) air-dry pellets at room temperature.
- Dissolve pellets in 5.6 μl of water, denature at 90 °C for 30 s, add 5x first strand buffer (Life technologies), dNTPs (final concentration 2.5 mM), DTT to final concentration of 10 mM, 0.75 μl Superscript III (Life technologies) reverse transcriptase and incubate at 42 °C for 1 hour.
- Add 2U of RNase H (Fermentas) and incubate at 37 °C for 20 min. Bring the volume to 20 μl.
- Perform a 100 μl pilot PCR with 2 μl of cDNA by collecting 12 μl aliquots of the PCR reaction every other cycle, starting at cycle 14. Evaluate the pilot PCR by 2.5% agarose electrophoresis in 1x TBE and identify the cycle number where the product above the band corresponding to adaptor-adaptor starts appearing.
- Perform a 400 μl large scale PCR reaction with the remaining cDNA, after ethanol precipitation resuspend in 1x DNA loading buffer, resolve by 2.5% agarose electrophoresis in 1x TBE, extract the PCR product corresponding to the ligated product, precipitate and subject to deep sequencing.

2.8.2 Buffers used in CLIP protocol

NP40 lysis buffer: 50 mM HEPES, pH 7.5; 150 mM KCl; 2 mM EDTA; 1 mM NaF; 0.5% (v/v) NP40; 0.5 mM DTT; protease inhibitor cocktail (Roche)

citrate phosphate buffer: 4.7 g /l citric acid; 9.32 g/l Na2HPO4; pH 5.0

IP washing buffer: 50 mM HEPES, pH 7.5; 300 mM KCl; 0.05% (v/v) NP40; 0.5 mM DTT; protease inhibitor cocktail (Roche)

high salt wash buffer: 50 mM HEPES, pH 7.5; 500 mM KCl; 0.05% (v/v) NP40; 0.5mM DTT; protease inhibitor cocktail (Roche)

crosslink washing buffer: 50 mM TrisHCl, pH 7.5; 20 mM EGTA; 0.5% (v/v) NP40

PNK buffer: 50 mM TrisHCl, pH 7.5; 50 mM NaCl; 10 mM MgCl2; 5 mM DTT

2x proteinase K buffer: 100 mM TrisHCl, pH 7.5; 150 mM NaCl; 12,5mM EDTA; 20% (w/v) SDS

2.9 References

BAILEY, T. L., BODEN, M., BUSKE, F. A., FRITH, M., GRANT, C. E., CLEMENTI, L., REN, J., LI, W. W. & NOBLE, W. S. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res, 37, W202-8.

BARTEL, D. P. 2009. MicroRNAs: target recognition and regulatory functions. *Cell,* 136**,** 215-33.

CHI, S. W., HANNON, G. J. & DARNELL, R. B. 2012. An alternative mode of microRNA target recognition. *Nat Struct Mol Biol,* 19**,** 321-7.

CHI, S. W., ZANG, J. B., MELE, A. & DARNELL, R. B. 2009. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature,* 460**,** 479-86.

CORCORAN, D. L., GEORGIEV, S., MUKHERJEE, N., GOTTWEIN, E., SKALSKY, R. L., KEENE, J. D. & OHLER, U. 2011. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol,* 12**,** R79.

DOHM, J. C., LOTTAZ, C., BORODINA, T. & HIMMELBAUER, H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res,* 36**,** e105.

GOTTWEIN, E., CORCORAN, D. L., MUKHERJEE, N., SKALSKY, R. L., HAFNER, M., NUSBAUM, J. D., SHAMULAILATPAM, P., LOVE, C. L., DAVE, S. S., TUSCHL, T., OHLER, U. & CULLEN, B. R. 2011. Viral microRNA targetome of KSHV-infected primary effusion lymphoma cell lines. *Cell Host Microbe,* 10**,** 515-26.

GRIMSON, A., FARH, K. K., JOHNSTON, W. K., GARRETT-ENGELE, P., LIM, L. P. & BARTEL, D. P. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell,* 27**,** 91-105.

HAFNER, M., LANDGRAF, P., LUDWIG, J., RICE, A., OJO, T., LIN, C., HOLOCH, D., LIM, C. & TUSCHL, T. 2008. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods,* 44**,** 3-12.

HAFNER, M., LANDTHALER, M., BURGER, L., KHORSHID, M., HAUSSER, J., BERNINGER, P., ROTHBALLER, A., ASCANO, M., JR., JUNGKAMP, A. C., MUNSCHAUER, M., ULRICH, A., WARDLE, G. S., DEWELL, S., ZAVOLAN, M. & TUSCHL, T. 2010. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell,* 141**,** 129-41.

KHORSHID, M., RODAK, C. & ZAVOLAN, M. 2011. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res,* 39**,** D245-52.

KISHORE, S., JASKIEWICZ, L., BURGER, L., HAUSSER, J., KHORSHID, M. & ZAVOLAN, M. 2011. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods,* 8**,** 559-64.

KONIG, J., ZARNACK, K., ROT, G., CURK, T., KAYIKCI, M., ZUPAN, B., TURNER, D. J., LUSCOMBE, N. M. & ULE, J. 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol,* 17**,** 909-15.

LANGMEAD, B., TRAPNELL, C., POP, M. & SALZBERG, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol,* 10**,** R25.

LEUNG, A. K., YOUNG, A. G., BHUTKAR, A., ZHENG, G. X., BOSSON, A. D., NIELSEN, C. B. & SHARP, P. A. 2011. Genome-wide identification of Ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs. *Nat Struct Mol Biol,* 18**,** 237-44.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

LICATALOSI, D. D., MELE, A., FAK, J. J., ULE, J., KAYIKCI, M., CHI, S. W., CLARK, T. A., SCHWEITZER, A. C., BLUME, J. E., WANG, X., DARNELL, J. C. & DARNELL, R. B. 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature,* 456**,** 464-9.

NELSON, P. T., DE PLANELL-SAGUER, M., LAMPRINAKI, S., KIRIAKIDOU, M., ZHANG, P., O'DOHERTY, U. & MOURELATOS, Z. 2007. A novel monoclonal antibody against human Argonaute proteins reveals unexpected characteristics of miRNAs in human blood cells. *RNA,* 13**,** 1787-92.

PINARD, R., DE WINTER, A., SARKIS, G. J., GERSTEIN, M. B., TARTARO, K. R., PLANT, R. N., EGHOLM, M., ROTHBERG, J. M. & LEAMON, J. H. 2006. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics,* 7, 216.

PRUITT, K. D., TATUSOVA, T., KLIMKE, W. & MAGLOTT, D. R. 2009. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res,* 37, D32-6.

RUDEL, S., FLATLEY, A., WEINMANN, L., KREMMER, E. & MEISTER, G. 2008. A multifunctional human Argonaute2-specific monoclonal antibody. *RNA,* 14, 1244-53.

STARK, T. J., ARNOLD, J. D., SPECTOR, D. H. & YEO, G. W. 2012. High-resolution profiling and analysis of viral and host small RNAs during human cytomegalovirus infection. *J Virol,* 86, 226-35.

ULE, J., JENSEN, K., MELE, A. & DARNELL, R. B. 2005. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods,* 37, 376-86.

URLAUB, H., HARTMUTH, K. & LUHRMANN, R. 2002. A two-tracked approach to analyze RNA-protein crosslinking sites in native, nonlabeled small nuclear ribonucleoprotein particles. *Methods,* 26, 170-81.

WANG, Y., JURANEK, S., LI, H., SHENG, G., TUSCHL, T. & PATEL, D. J. 2008. Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature,* 456, 921-6.

WANG, Z., TOLLERVEY, J., BRIESE, M., TURNER, D. & ULE, J. 2009. CLIP: construction of cDNA libraries for high-throughput sequencing from RNAs cross-linked to proteins in vivo. *Methods,* 48, 287-93.

ZHANG, C. & DARNELL, R. B. 2011. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotechnol,* 29, 607-14.

ZISOULIS, D. G., LOVCI, M. T., WILBERT, M. L., HUTT, K. R., LIANG, T. Y., PASQUINELLI, A. E. & YEO, G. W. 2010. Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. *Nat Struct Mol Biol,* 17, 173-9

Chapter 3: PAPD5, a noncanonical poly(A) polymerase with an unusual RNA-binding motif

Christiane Rammelt[1,*], Biter Bilen[2], Mihaela Zavolan[2], Walter Keller[2,*]

[1]Institute of Biochemistry and Biotechnology, Martin Luther University Halle-Wittenberg, D-06099 Halle, Germany

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

[*]Corresponding authors (christiane.rammelt@biochemtech.uni-halle.de, walter.keller@unibas.ch)

Running title: The noncanonical poly(A) polymerase PAPD5

Keywords: polyadenylation; noncanonical poly(A) polymerase; TRAMP4; RNA surveillance

3.1 Summary

PAPD5 is one of the seven members of the family of noncanonical poly(A) polymerases in human cells. PAPD5 was shown to polyadenylate aberrant pre-ribosomal RNAs in vivo, similar to degradation-mediating polyadenylation by the noncanonical poly(A) polymerase Trf4p in yeast. PAPD5 has been reported to be also involved in the uridylation-dependent degradation of histone mRNAs. To test whether PAPD5 indeed catalyzes adenylation as well as uridylation of RNA substrates, we analyzed the in vitro properties of recombinant PAPD5 expressed in mammalian cells as well as in bacteria. Our results show that PAPD5 catalyzes the polyadenylation of different types of RNA substrates in vitro. Interestingly, PAPD5 is active without a protein cofactor, whereas its yeast homolog Trf4p is the catalytic subunit of a bipartite poly(A) polymerase in which a

separate RNA- binding subunit is needed for activity. In contrast to the yeast protein, the C terminus of PAPD5 contains a stretch of basic amino acids that is involved in binding the RNA substrate.

3.2 INTRODUCTION

The addition of a poly(A) tail to the 39 end of mRNAs in the nucleus of eukaryotic cells is catalyzed by the classical or canonical poly(A) polymerases (PAPs) and is an important step during mRNA maturation. The poly(A) tail confers stability to the mRNA and allows transport out of the nucleus and efficient translation in the cytoplasm. In contrast, the functions of the so-called noncanonical PAPs or Cid1-like proteins (after the founding member of the family) are only partially understood. Enzymes of this group are involved in cytoplasmic polyadenylation of mRNAs, but also in polyadenylation-mediated degradation of RNAs, and some members of the family even catalyze the addition of uridyl residues to RNA substrates (Schmidt and Norbury, 2010). Like the canonical PAPs, Cid1-like proteins belong to the superfamily of polymerase b-like nucleotidyl transferases. Canonical and noncanonical PAPs have similar catalytic and central domains but differ in their nucleotide- base recognition motif. Interestingly, only a few Cid1-like proteins contain an RNA-binding domain (Martin and Keller, 2007).

One well-studied member of the Cid1-like protein family is Trf4p, a PAP involved in nuclear RNA surveillance in Saccharomyces cerevisiae. Polyadenylation of RNAs by Trf4p facilitates their subsequent degradation by the nuclear exosome, a 39–59 exonuclease complex (Kadaba et al., 2004, Kadaba et al., 2006, Vanacova et al., 2005), or in some cases induces the processing of the RNA by exosome-catalyzed trimming of the 39 end (Egecioglu et al., 2006). A stimulatory effect of polyadenylation on the exosome activity was first shown to allow efficient degradation of a hypomodified, unstable tRNA (LaCava et al., 2005, Vanacova et al., 2005). Later, rRNAs, snRNAs, and snoRNAs, as well as so-called cryptic unstable transcripts (CUTs), spliced-out introns, and different mRNAs were identified as substrates of this polyadenylation- mediated RNA quality-control mechanism (LaCava et al., 2005, Wyers et al., 2005, Davis and Ares, 2006, Dez et al., 2006, Egecioglu et al., 2006, Carneiro et al., 2007, Houseley et al., 2007, San Paolo et al., 2009, Callahan and Butler, 2010).

Trf4p lacks an RNA-binding domain and is therefore inactive on its own. In vivo it is part of the TRAMP4 complex that contains Trf4p as the catalytic subunit, one of the zinc knuckle proteins Air1p or Air2p, thought to confer the RNA-binding function to the complex, and the RNA helicase Mtr4p (LaCava et al., 2005, Vanacova et al., 2005). A similar complex, SpTRAMP, was identified in fission yeast (Win et al., 2006, Buhler et al., 2008). In Drosophila melanogaster, overexpression of a homolog of Trf4p was shown to lead to the polyadenylation of snRNAs (Nakamura et al., 2008), whereas the human homolog PAPD5 is involved in the polyadenylation-mediated degradation of aberrant pre-rRNAs (Shcherbik et al., 2010).

The human PAPD5 protein was also reported to function in the degradation of replication-dependent histone mRNAs. These mRNAs do not have a poly(A) tail but end in a stem–loop structure. Here, however, the degradation pathway does not start with polyadenylation of the RNA, but with the addition of a short oligo(U) tail, which is recognized by the Lsm1-7 complex and subsequently leads to decapping and degradation of the mRNAs. A knock- down of either PAPD5 or of the mitochondrial poly(A) polymerase results in the stabilization of histone mRNAs (Mullen and Marzluff, 2008). However, this effect could not be observed by other investigators (Schmidt and Norbury, 2010). Instead, the knockdown of the uridyl transferase ZCCHC11, a different member of the human Cid1-like proteins, was shown to strongly reduce the oligouridylation of histone mRNAs, leading to their stabilization (Schmidt and Norbury, 2010). It is therefore possible that PAPD5, depending on the RNA substrate or the presence of different protein cofactors, can catalyze the polyadenylation as well as the oligouridylation of RNAs. To better understand the function of PAPD5, we set out to investigate the activity of the protein in vitro and to explore its function in mammalian cells through cross-linking and immunoprecipitation (CLIP) (Hafner et al., 2010).

3.3 Results

3.3.1 PAPD5 is a poly(A) polymerase in vitro

PAPD5 (UniProt accession number Q8NDF8) is involved in the polyadenylation of aberrant ribosomal RNA precursors in mouse (Shcherbik et al., 2010). In addition, it may also be responsible for the uridylation-mediated degradation of histone mRNAs (Mullen and Marzluff, 2008). However, this function of PAPD5 is controversial (Schmidt and Norbury, 2010). To test the activity of PAPD5 in vitro, the coding sequence was amplified from cDNA of HeLa cells. The protein was expressed in HEK293 cells as a fusion protein with an N-terminal Flag tag and a His8 tag. After cell lysis, the protein was bound to anti-Flag agarose, washed, and eluted with Flag peptide. The protein was further purified via nickel affinity chromatography (Figure. 1A).

To test the nucleotidyltransferase activity of PAPD5, the oligoribonucleotide A15 was incubated with the protein either in the presence of ATP, CTP, GTP, or UTP, or in a mixture of all four dNTPs (Figure. 2A). When PAPD5 and A15 were incubated in the presence of ATP, the incorporation of several nucleotides into the RNA substrate was observed. In contrast, the incorporation of the other three ribonucleotides was low and limited to single residues, showing a strong preference of PAPD5 for ATP. In addition, PAPD5 is able to discriminate against dNTPs, as these were not used as substrates.

3.3.2 PAPD5 does not need a cofactor for polyadenylation in vitro

To further analyze the properties of PAPD5, the protein was expressed in Escherichia coli. To prevent contamination with bacterial poly(A) polymerase, an expression strain with a disruption in the gene for poly(A) polymerase was used. Two PAPD5 expression constructs were tested: a His6 fusion and a combination of a C-terminal His6 tag and an N-terminal fusion of a GB1 tag that was shown to increase the solubility of recombinant proteins (Zhou et al., 2001). Because the activity of both wild-type fusion proteins was the same but the yield of the GB1-tagged protein was higher, the latter construct was used for further experiments. The purified protein is shown in Figure 1B.

Trf4p, the poly(A) polymerase of the TRAMP4 complex in baker's yeast, was previously shown to be inactive on its own. Trf4p contains no obvious RNA-binding domain and is believed to rely on the zinc-knuckle proteins Air1p or Air2p as the substrate-binding domain of the complex. Indeed, only the combination of Trf4p with either Air1p or Air2p has poly(A) polymerase activity (LaCava et al., 2005, Vanacova et al., 2005).

Surprisingly, the bacterially expressed PAPD5 (Figure. 1B) showed the same activity as the protein purified from HEK293 cells. When the ribo-oligonucleotide A15 was incubated with PAPD5 in the presence of different nucleotides, efficient incorporation could only be observed with ATP (Figure. 2A). A PAPD5 variant in which two aspartate residues (D177, D179) in the catalytic domain were exchanged to alanine could not catalyze the addition of nucleotides to the RNA substrate (PAPD5 DADA)

(Figure. 2B). Therefore, the observed polyadenylation activity is an intrinsic activity of PAPD5, and a protein cofactor is not needed for activity.

The nucleotide incorporation experiment was repeated with other RNA substrates: Ribo-oligonucleotide G2U15 was used to test whether the last nucleotide of the RNA primer influences the specificity for the nucleotide substrate (Figure. 2C). In addition, in vitro–synthesized tRNAiMet (a known substrate for yeast Trf4p) and an in vitro–synthesized fragment of the 39 UTR of mouse histone H2a mRNA comprising the stem–loop structure were tested as substrates (results not shown). Again, only AMP incorporation could be observed. Therefore, the specificity for ATP incorporation does not depend on the last nucleotide of the RNA primer but results from nucleotide specificity of the protein. The specificity for ATP incorporation was also observed when a reaction buffer containing manganese instead of magnesium was used (results not shown).

As reported for yeast Trf4p (LaCava et al., 2005), the polyadenylation reaction catalyzed by PAPD5 is not processive: In all cases, the reaction products were heterogeneous in size, the smallest product usually corresponding to the incorporation of a single nucleotide. When DNA oligonucleotides were used as primers, none of the nucleotide substrates was incorporated (results not shown), showing that the reaction is RNA-specific.

3.3.3 A C-terminal stretch of basic amino acids is involved in binding of the RNA primer

The amino acid sequence of PAPD5 does not contain any motif known to bind nucleic acids. However, the poly- adenylation activity of PAPD5 observed in vitro and in the absence of other protein cofactors argues for the presence of a domain in the protein that is involved in binding of the RNA substrate. To analyze the RNA binding of PAPD5, electrophoretic mobility shift experiments were performed. PAPD5 was incubated with in vitro– synthesized tRNA, and the reaction mixture was resolved on a nondenaturing polyacrylamide gel. Indeed, a shift of the RNA band to a band of lower mobility could be observed, showing that PAPD5 is able to bind RNA. To determine which part of the protein is responsible for RNA binding, protein variants comprising only the N terminus including the catalytic domain (amino acids 1–111, PAPD5 N), the central domain (amino acids 112–368, PAPD5 M), or the C terminus of PAPD5 (amino acids 369–572, PAPD5 C1) were analyzed separately. Of these, only the C-terminal part, PAPD5 C1, could shift the tRNA (Figure. 3), indicating the presence of an RNA-binding function in the C terminus of PAPD5.

Inspection of the amino acid sequence showed the C terminus of human PAPD5 to contain a stretch of basic amino acids (amino acids 557–563), a motif that is conserved among higher eukaryotes, but not in the yeast protein Trf4p (Figure. 4). To test the influence of the basic motif on the RNA-binding ability of PAPD5, a deletion variant of the C-terminal part lacking this motif was used (amino acids 369–551, PAPD5 C2). Whereas the C terminus binds to RNA (Figure. 3, PAPD5 C1), the deletion variant showed no shift of the RNA in EMSA experiments, even when used in higher concentrations than the original C-terminal fragment (Figure. 3, PAPD5 C2).

Next, we wanted to analyze the influence of this basic amino acid stretch on the catalytic activity of PAPD5. Therefore, a protein variant PAPD5DC, comprising amino acids 1–551, i.e., lacking only the basic stretch, was constructed and tested for polyadenylation activity. In addition, a point mutant of PAPD5 was constructed in which a lysine residue in the basic domain was exchanged to glutamate (PAPD5 K560E). The variants were expressed in E. coli and purified as the wild-type protein. For both variants, the yield of soluble protein was comparable to the wild type. In comparison to the wild-type protein, both variants showed a reduced activity when incubated with the A15 substrate in the presence of ATP, resulting in fewer incorporated nucleotides and therefore shorter polyadenylation products (Figure. 5).

To further examine the influence of the basic stretch of amino acids on the polyadenylation activity of PAPD5, the wild-type protein and the variants were incubated with different concentrations of either ATP or A15 substrate to determine the kinetic parameters of the reaction. To this end, radioactive ATP was used as nucleotide substrate, and the radioactively labeled reaction products were separated from nonincorporated nucleotides with DE81 filters. The polyadenylation activity was measured as total incorporation of nucleotides into the A15 substrate, independent of the length of the resulting product.

When the concentration of the RNA substrate A15 was varied, the maximal velocities for the wild-type protein and the variants were similar (data not shown). However, the KM of the proteins with changes in the C-terminal basic motif was increased by a factor of z3 to 5. In contrast, there was no significant difference in the apparent affinity for ATP. The concentration of ATP during A15 titration was in much higher excess than the concentration of the RNA primer when ATP was titrated; therefore, the kcat values determined by titration of ATP are lower than the values observed for the titration of A15. This result con- firms the participation of the stretch of basic amino acids in the binding of the RNA substrate by PAPD5.

3.3.4 Specificity for RNA structure

The yeast TRAMP complex was shown to specifically polyadenylate unmodified initiator tRNA, but not the native and therefore fully modified tRNA (Vanacova et al., 2005). This is thought to be a result of a more compact tertiary structure of the native tRNA. It was shown recently that a core complex consisting of the central part of Trf4p and two out of the five zinc knuckles of Air2p is sufficient to discriminate between the two tRNA substrates (Hamill et al., 2010). To test whether PAPD5 shows a similar ability, PAPD5 purified from HEK293 cells as well as bacterially expressed PAPD5 were incubated with either in vitro–synthesized yeast tRNAiMet or the native tRNA purified from yeast cells in the presence of ATP. Interestingly, both enzyme preparations were able to polyadenylate the in vitro transcript but not the native tRNA, whereas at the same time E. coli poly(A) polymerase polyadenylated both tRNA substrates (Figure. 6; results not shown). Therefore, similarly to the yeast Trf4p, PAPD5 seems to be sensitive to the structure of the RNA substrate.

3.3.5 Subcellular localization of PAPD5

In yeast, Trf4p is part of a RNA surveillance machinery residing in the nucleus. To determine the localization of PAPD5, the protein was detected in cells expressing Flag- tagged PAPD5 with anti-Flag antibody followed by a secondary antibody conjugated to a fluorescent dye (Figure. 7A). In addition, PAPD5 fusion proteins with green fluorescent protein (GFP) fused to the N or C terminus were expressed transiently (Figure. 7B and 7C). All fusion proteins were detected in the nucleus as shown by colocalization of the fluorescence signals with the DAPI stain, no staining of the cytoplasm could be observed. The basic stretch of amino acids at the C terminus of PAPD5 that is involved in RNA binding does not play a role for localization of PAPD5, because a Flag- tagged deletion variant corresponding to PAPD5DC, i.e., lacking the basic stretch, is still exclusively located in the nucleus (Figure. 7D).

3.3.6 Ribosomal RNAs are in vivo targets of PAPD5

PAPD5 was able to polyadenylate a variety of RNA substrates in vitro, ranging from oligo(A) and oligo(U) to different tRNAs as well as the 39 UTR of histone mRNAs. To determine the potential substrate spectrum of PAPD5 in vivo, RNAs bound to the protein in HEK293 cells were probed by the PAR-CLIP technique (Hafner et al., 2010). Briefly, HEK293 cells expressing Flag-tagged PAPD5 were grown in medium containing 4-thio-uridine and RNA– protein complexes were UV-cross-linked and immunoprecipitated. The covalently bound RNAs were extracted, adapter oligonucleotides were ligated to the 59 and 39 ends, and the RNA was reverse-transcribed. The cDNA was amplified and the sequences determined by deep sequencing. We then compared the results obtained in two separate PAPD5 PAR-CLIP experiments and those previously obtained for the insulin growth factor 2 binding protein 1 (IGF2BP1) (Hafner et al., 2010) to identify genomic regions from which PAPD5 targets derive. The IGF2BP1 PAR-CLIP sample contained 24,914,594 reads and the PAPD5 PAR- CLIP samples 4,949,000 and 5,563,677 reads, respectively. The proportion of reads mapping with at most one error to the human genome was 45.64% in the case of IGF2BP1 and 32.77% and 27.13% for PAPD5 replicates. We normalized the count of mapped reads across samples and calculated the number of reads originating in each 100-nt- long window of the genome, and the variance of these counts in the PAPD5 PAR-CLIP replicates with the method described by Balwierz et al. (Balwierz et al., 2009). We then computed the Z-score indicating the significance of the enrichment or depletion in reads in a PAPD5 sample compared to the IGF2BP1 sample. The distributions of the Z-scores across all genomic regions (Figure. 8) show a much higher similarity of PADP5 replicates relative to the IGF2BP1 sample. With cut-off values of -5 and 5, at which only a very small number of genomic regions will be considered differentially represented between PAPD5 replicates, we obtained 988 regions for the first PAPD5 replicate and 923 for the second (data not shown) that were differentially expressed relative to IGF2BP1. Four hundred thirty-four regions were differentially expressed between both PAPD5 replicates and the IGF2BP1 sample. Of these, 358 regions were enriched, whereas 76 were depleted in PAPD5 samples. One hundred ninety-two out of the 358 enriched windows corresponded to repeat elements, 61 of which were rRNA. These latter windows were located in genomic regions annotated as rRNA-like in the Genome Browser of the University of California Santa Cruz. They contained reads that we also annotated as rRNA-derived, based on the fact that they mapped with at most one error to the rRNA subunits. In contrast, only two of the 76 regions depleted in PAPD5 binding had rRNA annotations (Fisher's exact test, $p < 0.001$), indicating that rRNAs could be the primary targets of PAPD5 in human cells.

3.4 Discussion

PAPD5 belongs to the family of Cid1-like proteins or noncanonical PAPs. In this family of nucleotidyltransferases, several members do not contain a recognizable RNA-binding domain. For the noncanonical PAP Trf4p from yeast, it was shown that nucleotidyltransferase activity can only be observed for the complex of Trf4p and one of the two zinc-knuckle proteins Air1p or Air2p, but not for Trf4p alone (LaCava et al., 2005, Vanacova et al., 2005). Therefore, it is thought that Trf4p is the catalytic subunit of this bipartite PAP, whereas the zinc-knuckle proteins confer RNA- binding specificity to the complex. Such a modular organization of a catalytic subunit that could even be combined with different RNA-binding subunits to result in an active nucleotidyltransferase may allow the targeting of different classes of substrates, depending on the RNA-binding subunit used. For the TRAMP4 complex in yeast, it was shown that two out of the five zinc knuckles of Air2p are necessary and sufficient for the polyadenylation of a short RNA oligonucleotide. How- ever, efficient polyadenylation of a tRNA required the most N-terminal zinc knuckle in addition. Although the other two zinc knuckles seem to be dispensable for the tRNA substrate, they may be important for the recognition of other TRAMP substrates (Hamill et al., 2010).

Surprisingly, PAPD5, a human Trf4p homolog lacking a recognizable RNA-binding domain, is able to catalyze the addition of nucleotides to RNA substrates in the absence of a protein cofactor. Analysis of the amino acid sequence revealed a stretch of basic residues in the C terminus of PAPD5 that is highly conserved among higher eukaryotes. When this motif was deleted, the nucleotidyltransferase activity was strongly reduced, and both electrophoretic mobility shift experiments and the analysis of the activity of a point mutant in this part of the protein showed the stretch of basic amino acids to take part in binding of the RNA substrate. PAPD5 was able to incor- porate nucleotides to different RNA substrates [oligo(A), oligo(U), tRNA, histone mRNA 39 end], therefore the RNA-binding function is very likely not sequence-specific, resulting in polyadenylation of all RNAs with a free 39-hydroxyl end in vitro. However, like the yeast TRAMP complex, PAPD5 discriminates between in vitro–synthesized, unmodified and native, fully modified tRNAs. Very likely, this is due to a difference in the stability of the RNA structure, resulting in the 39 end of the in vitro transcript being more accessible to the poly(A) polymerase. PAPD5 has been shown to be involved in the polyadenylation of aberrant ribosomal RNAs: Aberrant pre-rRNAs were found to be partially polyadenylated in a murine cell line by reverse transcription with an oligo(dT)-adapter primer followed by PCR with a forward primer specific for the 59-ETS region of pre-rRNA. After knockdown of PAPD5, the amount of polyadenylated pre-rRNAs was decreased (Shcherbik et al., 2010). In agreement with this finding, the analysis of RNAs cross-linked to PAPD5 in vivo revealed rRNAs to be potential substrates of PAPD5. Ribosomal RNAs were represented in the potential RNA substrates in a high percentage, and they were specifically enriched in the PAPD5 samples compared to a control IGF2BP1 sample. The same was shown for Trf4p-associated RNAs in yeast: 50% of the sequences identified by an in vivo cross-linking approach were mapped to ribosomal sequences (Wlotzka et al., 2011). In addition, other classes of noncoding RNAs as well as pre-mRNAs and mRNAs were found to be associated with PAPD5. Although in Trf4-deletion yeast strains as well as in the cross-linking approach the same classes of RNAs were found to be enriched, the relative proportions of these transcript types were different from that found here for PAPD5 (San Paolo et al., 2009, Wlotzka et al., 2011), and in the PAPD5 samples such targets were not robustly enriched. Whereas in yeast only two noncanonical poly(A) polymerases are present, human cells contain seven members of this protein family. Therefore, other noncanonical poly(A) polymerases may be involved in the polyadenylation of other aberrant RNAs.

The catalytic parameters observed for the wild-type PAPD5 protein (data not shown) are similar to the values determined for bacterial poly(A) polymerase, an enzyme also involved in RNA degradation (Li et al., 1998, Li et al., 2002). In contrast, canonical poly(A) polymerases show higher maximal velocity even under distributive reaction conditions, reflecting the function of the two different types of polyadenylation. Whereas the canonical poly(A) polymerases add long tails that are immediately bound by poly(A)-binding proteins and allow protection of the mRNA from degradation, polyadenylation-mediated degradation is characterized by the addition of short tails as well as distributive polyadenylation to allow access of the degradation machinery.

In the cell, noncoding RNAs as well as mRNAs are bound by proteins during processing and reside in RNPs to perform their function. In contrast, aberrant RNAs or processing intermediates that—due to misfolding or misprocessing—cannot form the correct interactions to protein partners would be accessible for polyadenylation and degradation. Therefore, PAPD5 could act as a kind of scavenger enzyme, similar to the TRAMP4 complex in yeast, to polyadenylate RNA 39 ends and thereby tag the corresponding RNA for degradation. Like yeast Trf4p, PAPD5 is located in the nucleus of the cell and may therefore be part of a nuclear surveillance machinery similar to the TRAMP4–exosome system of yeast. In contrast to Trf4p, PAPD5 is able to bind different RNAs, and when His-Flag-tagged PAPD5 was purified from mammalian cells, no copurifying RNA-binding proteins could be identified by mass spectrometry. However, it still remains to be determined whether protein cofactors are necessary for the efficient binding of PAPD5 to substrates, as has been shown for TRAMP4.

When the levels of several of the putative substrates identified by the CLIP experiments were analyzed in cells treated with siRNAs directed against PAPD5 or control siRNA, no significant change could be observed (results not shown). However, mammalian cells contain genes for seven proteins of the Cid1-like family. PAPD7, one of these homologs, is very similar to PAPD5. If more than one of these non- canonical PAPs are able to target the same RNAs, depletion of the cell of a single PAP may not lead to a detectable change in the RNA level. It has been observed for poly(A) and A-rich tails on human rRNA that the knockdown of a single noncanonical PAP does not abolish tail addition (Slomovic et al., 2010). Therefore, the RNA surveillance system is most likely redundant, allowing potentially deleterious aberrant RNAs to be degraded with high efficiency.

3.5 Materials and Methods

3.5.1 Expression and purification of PAPD5

The coding sequence of PAPD5 was amplified from HeLa cell cDNA. The ORF was cloned into pcDNA5/FRT for expression as a fusion protein with N-terminal His8 and Flag tags in mammalian cells. For expression, an HEK293 cell line, Flp-In-293 (Invitrogen), was used. Cells were cultivated in DMEM medium containing 10% fetal calf serum, and cell lines expressing PAPD5 were selected according to the manufacturer's instructions. For analysis of the subcellular localization, PAPD5 was transiently expressed as an N- or C-terminal GFP fusion protein or with a Flag tag. For detection of Flag-tagged protein, cells were permeabilized, and the fusion protein was stained with mouse anti-Flag antibody (Sigma) and Cy3-conjugated donkey anti-mouse antibody (Dianova).

For expression in E. coli, the ORF of PAPD5 was cloned into pET30 for expression as a fusion protein with an N-terminal GB1 tag and a C-terminal His6 tag (Zhou et al.) or into pETDuet (Novagen) for expression as a C-terminal His6-tag fusion protein. A BL21 (DE3) strain deficient for E. coli poly(A) polymerase was constructed with the TargeTron Gene Knockout System (Sigma-Aldrich). The strain was transformed with the corresponding expression plasmids, grown in LB medium at 37°C to an OD600 of 0.8, and expression of PAPD5 was induced with 1 mM IPTG for 3 h.

For purification of the protein from mammalian cells, the cells were lysed by sonication in buffer containing 50 mM Tris (pH 7.5), 100 mM KCl, 10% glycerol, 0,05% NP-40, 1 mM PMSF, 1 mg/L Pepstatin, and 1 mg/L Leupeptin. The cleared lysate was incubated with M2 anti-Flag agarose (Sigma-Aldrich). The matrix was washed with lysis buffer containing 300 mM KCl, and the protein was eluted with lysis buffer containing 5 mM Flag peptide. The eluate was adjusted to 5 mM imidazole, bound to Ni-NTA agarose

(Sigma- Aldrich), washed with buffer containing 25 mM imidazole, and the protein was eluted with buffer containing 250 mM imidazole.

The protein expressed in E. coli was purified via metal chelating chromatography like the protein expressed in mammalian cells, but the cell lysate was adjusted to 25 mM imidazole, and the washing buffer contained 500 mM KCl and 50 mM imidazole.

3.5.2 RNA substrates

A plasmid ptRNAMet for the in vitro transcription of yeast tRNAiMet was obtained from Dr. Bruno Senger (Institut de Biologie Moleculaire et Cellulaire Strasbourg). The plasmid was linearized with BstNI and transcribed with T7 RNA polymerase. The native yeast tRNAMet was a gift of Dr. Gerard Keith (Institut de Biologie Mole ́culaire et Cellulaire Strasbourg). The plasmid pUC19-H2a for in vitro transcription of the mouse histone H2a mRNA 39 end with T7 RNA polymerase was obtained from Dr. Sophie Jaeger (Biozentrum, University of Basel). In vitro– transcribed RNAs were purified on 8.3 M urea/10% polyacrylamide gels. For 59-end labeling, RNAs were treated with alkaline phosphatase and labeled with [g-32P]ATP (GE Healthcare) and T4 polynucleotide kinase. Labeled RNAs were gel-purified. To allow folding of the structured RNAs, pellets were dissolved in RNase-free water and incubated for 1 min at 65°C, followed by incubation for 10 min on ice after the addition of an equal volume of twice-concentrated assay buffer containing 10 mM MgCl2.

3.5.3 Polyadenylation assays

Polyadenylation assays were carried out in 20-mL reaction mixtures containing 0.07–0.7 pmol of affinity-purified protein, 1 pmol of 59-end-labeled RNA, 1 mM ATP, 5 mM MgCl2, 25 mM Tris-HCl (pH 7.5), 50 mM KCl, 0.01 mM EDTA, 0.1 mg/mL BSA, 1 mM DTT, and 0.02% NP-40. Reactions were incubated for 30 min at 37°C or the times indicated and stopped by the addition of 25 mM EDTA. The RNA was precipitated by addition of 0.1 volume of 3 M ammonium acetate and three volumes of ethanol. Pellets were resuspended in formamide loading buffer and separated on denaturing poly-acrylamide gels. Radioactivity was scanned with a PhosphorImager, and results were analyzed with ImageQuant software (Molecular Dynamics).

For determination of the kinetic parameters, the activities of PAPD5 variants were measured in 20-mL reaction mixtures containing buffer as above with 0.2 mCi of [a-32P]ATP (3000 Ci/mmol) and 0.7–3.5 pmol (0.035–0.17 mM) recombinant protein. The concentration of ATP was varied be- tween 0.01 mM and 1 mM (the concentra- tion of A15 was kept constant at 2.5 mM), and the concentration of unlabeled RNA A15 was varied between 50 nM and 2.5 mM (ATP was kept constant at 1 mM). The reactions were incubated for 20 min at 37°C and stopped by spotting onto DE-81 paper. The filters were washed three times for 10 min in 0.3 M ammonium formate/10 mM Na-pyrophosphate, and the incorporated radioactivity was measured in a scintillation counter.

3.5.4 EMSA

Electrophoretic mobility shift assays (EMSAs) were carried out in 20-mL reaction mixtures containing 0.5–20 nM of affinity-purified protein, 1 nM 59-end-labeled RNA, 10% glycerol, and 1 mg/L total yeast RNA in buffer as above. Reactions were incubated for 10 min at room temperature and separated on a native 6% polyacryl- amide gel in 0.53 TBE. Radioactivity was scanned with a PhosphorImager, and results were analyzed with ImageQuant software (Molecular Dynamics).

3.5.5 PAR-CLIP

RNAs bound to PAPD5 in vivo were analyzed by the PAR-CLIP technique essentially as described in Hafner et al. (Hafner et al.). Flp-In-293 cells expressing His8-

Flag-PAPD5 were grown in medium containing 100 mM 4-thiouridine, RNA–protein complexes were cross- linked by exposure of the cells to UV light (366 nm, 150 mJ/cm2), and lysed in buffer containing 50 mM HEPES (pH 7.4), 150 mM KCl, 2 mM EDTA, 1 mM NaF, 0.5% NP-40, 0.5 mM DTT, 1 mM PMSF, 1 mg/L Pepstatin, and 1 mg/L Leupeptin. After removal of cell debris by centrifugation, the cleared lysate was incubated with RNase T1 to allow bound RNAs to be partially digested to an average size of 40 to 60 nt.

RNA–protein complexes were isolated with M2 anti-Flag an- tibody (Sigma-Aldrich) coupled to protein G Dynabeads (Invitrogen). The RNA–protein complexes were gel-purified, the RNA was isolated, and linkers were ligated to the 59 and 39 ends of the RNA (Hafner et al.). After reverse transcription and amplification, the resulting DNA library was sequenced by Solexa sequencing. Mapping of short reads to the genome (hg18 assembly version from the University of California, Santa Cruz obtained from hgdownload.cse.ucsc.edu) following adaptor removal was performed as described (Hafner et al.).

3.5.6 Normalization and multiplicative error model in PAR-CLIP

To compare the expression of genomic regions across samples, the number of reads needs to be normalized across samples. With several types of high-throughput expression profiling approaches such as SAGE (serial analysis of gene expression), CAGE (cap analysis of gene expression), and chromatin immunoprecipitation (ChIP), it has been reported that the number of reads originating from a genomic position or region has a power-law distribution (Ueda et al., Zhang et al., Balwierz et al.). We found that the same holds for our PAR-CLIP data, which we then analyzed as follows. We split the genome into 100-nt-long windows and calculated the number of reads originating in each of these regions in the IGF2BP1 CLIP samples obtained by Hafner et al. (Hafner et al.) and in our two PAPD5 replicates. The reverse cumulative distribution of the read count in a window (showing the number of windows with a read count at least as high as a given value, which is indicated on the x-axis) was power-law-distributed (data not shown). We standardized these distributions by transforming them to a reference power law distribution with slope -1.25 and a total copy of 1 million reads, as described (Balwierz et al.). The resulting normalized reverse cumulative distributions are shown in Supplemental Figure S1B. Based on the normalized read counts, we estimated the variance in read count per window between the PAPD5 replicates to be 0.41. In contrast, the variances between first and second PAPD5 replicates and IGF2BP1 were estimated to be 0.91 and 1.03, respectively.

3.5.7 Enrichment and depletion of reads in genomic regions in PAR-CLIP

Using the estimated variance above, we calculated the Z-scores for the difference in expression between PAPD5 reads and IGF2BP1 reads in 100-nt-long overlapping genomic windows as described in Balwierz et al. (Balwierz et al.). Inspecting the pairwise Z-score distributions, we chose a cut-off value of 5 and -5 for regions that are enriched and depleted, respectively, in the PAPD5 relative to the IGF2BP1 samples.

3.5.8 Annotation of the sequences

To determine whether the enriched or depleted regions over- lapped known elements of the genome, we intersected the genome coordinates of the regions with those of annotated functional elements (UCSC RepeatMasker, mRNA and noncoding RNA tracks for hg18 assembly version) (Smit et al., Pruitt et al., Lestrade and Weber, Griffiths-Jones et al.). If the overlap was at least 20 nt, we transferred the annotation of the genomic element to the region that was enriched/depleted in CLIP.

3.6 Acknowledgments

We are grateful to Markus Landthaler and Thomas Tuschl for their generous help with the CLIP experiments and to Georges Martin and Elmar Wahle for discussions and critical reading of the manuscript. We also thank Bruno Senger and Gerard Keith for clones and native tRNA samples. The work was supported by the University of Basel, the Louis-Jeantet Foundation for Medicine, the Swiss National Science Foundation (grant Nr. 3100A0-102132/2 to W.K.), and the Deutsche Forschungsgemeinschaft (grant Nr. RA 1892/1-1 to C.R.). B.B. is supported by a PhD fellowship from the Werner-Siemens-Foundation.

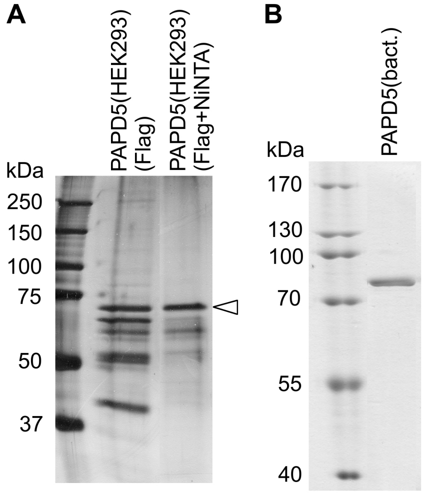


Figure 1. PAPD5 preparations used in the analysis. (A) PAPD5 was expressed as fusion protein with N-terminal His8 and Flag tag in HEK293 cells and purified via anti-Flag agarose (Flag), followed by metal chelating affinity chromatography (Flag + NiNTA). PAPD5 is marked by an arrowhead. (B) Bacterially expressed PAPD5 after metal chelating affinity chromatography (for details, see Materials and Methods).

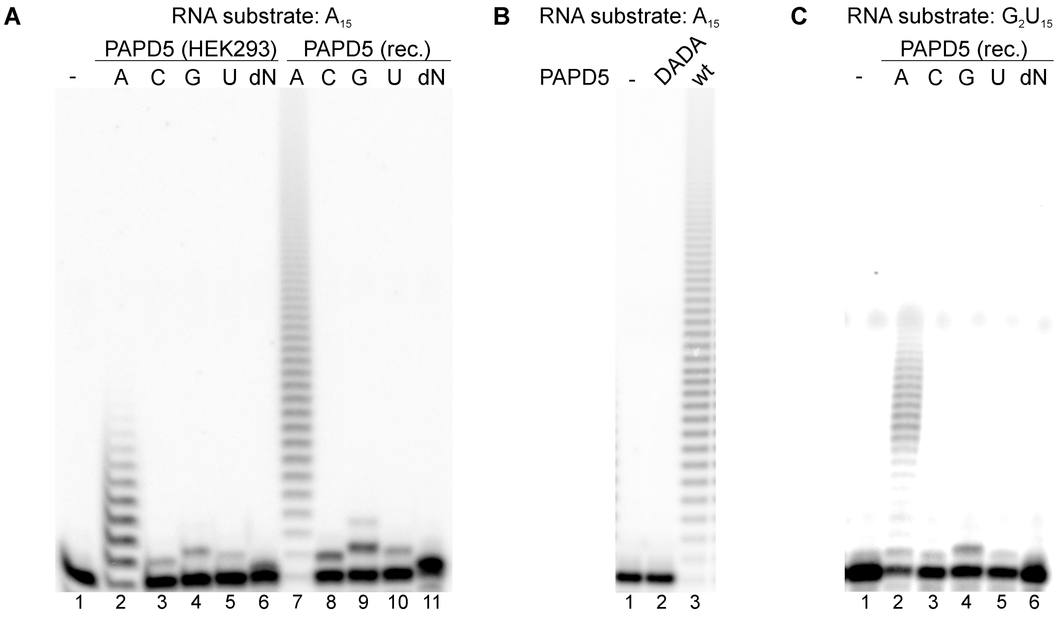


Figure 2. PAPD5 preferentially adds AMP residues to RNA substrates. (A) PAPD5 preparations from HEK293 cells (HEK293) or recombinant protein expressed in E. coli (rec.) were incubated with ribo-oligonucleotide A15 in the presence of ATP (A), CTP (C), GTP (G), UTP (U), or a mixture of all four dNTPs (dN), respectively. (B) PAPD5 wild-type protein (wt) or catalytic site mutant (DADA) expressed in E. coli were incubated with ribo- oligonucleotide A15 in the presence of ATP. (C) PAPD5 expressed in E. coli was incubated with ribo-oligonucleotide G2U15 in the presence of ATP (A), CTP (C), GTP (G), UTP (U), or a mixture of all four dNTPs (dN), respectively. Lane 1 always shows the RNA substrate incubated in the absence of protein.

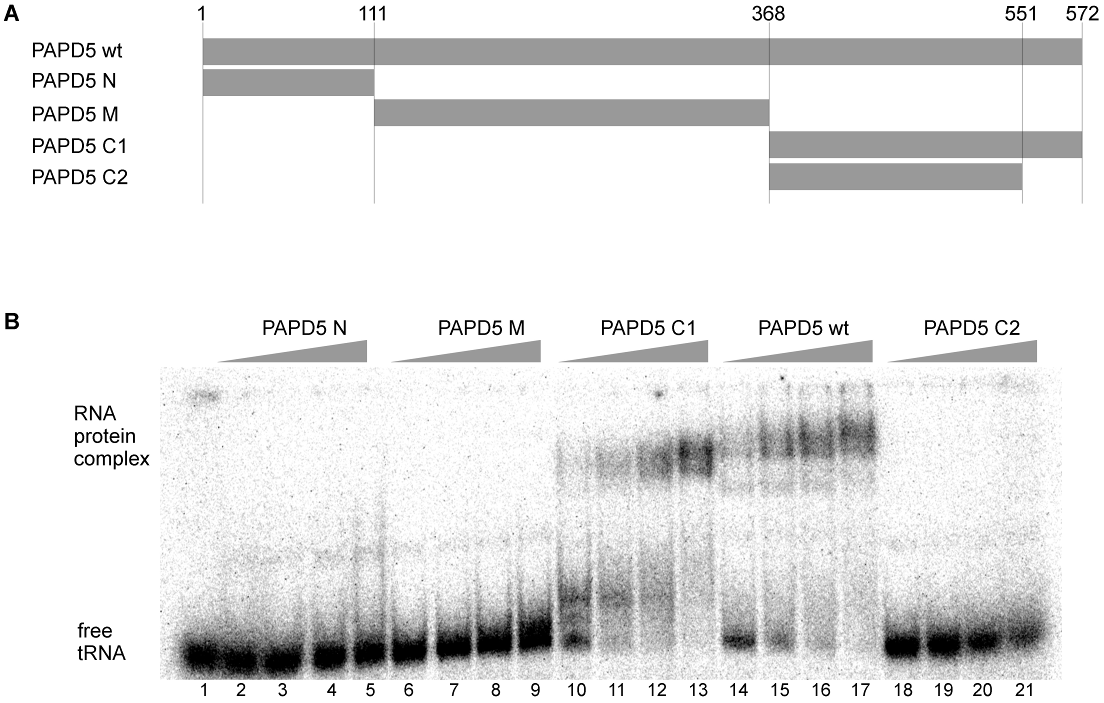
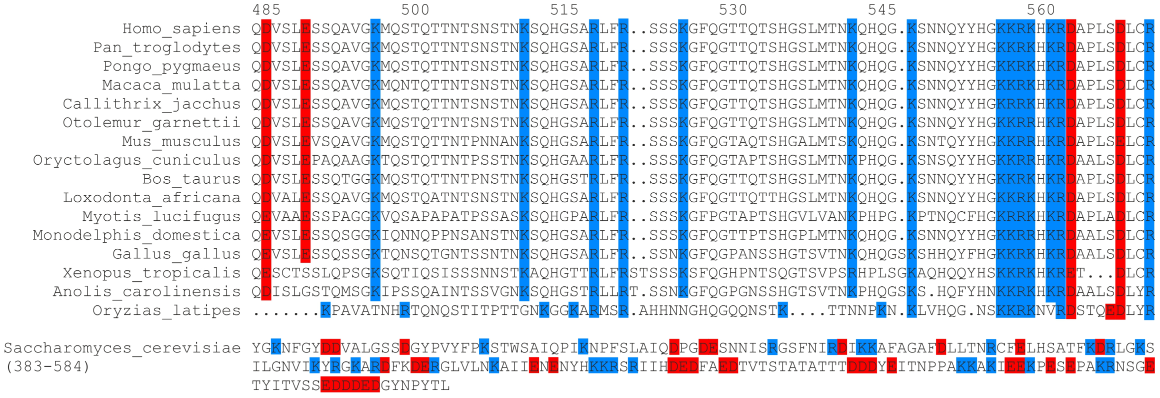


Figure 3. The C terminus of PAPD5 is involved in RNA binding. (A) Schematic representation of the truncation variants of PAPD5 tested for RNA binding. (B) Electropho- retic mobility shift assay of PAPD5 variants. Increasing amounts of PAPD5 full-length protein or fragments comprising the N terminus, central part, or C terminus of the protein were incubated with radioactively labeled in vitro–synthesized human tRNAiMet, and the reactions were resolved on a native polyacrylamide gel. Protein concentrations between 0.5 and 5 nM were tested for the full-length protein and the protein PAPD5 C2, 1–20 nM for the other variants. Lane 1 shows the tRNA after incubation in the absence of protein.



```
                           485            500            515            530            545            560
            Homo_sapiens QDVSLESSQAVGKMQSTQTTNTSNSTNKSQHGSARLFR..SSSKGFQGTTQTSHGSLMTNKQHQG.KSNNQYYHGKKRKHKRDAPLSDLCR
         Pan_troglodytes QDVSLESSQAVGKMQSTQTTNTSNSTNKSQHGSARLFR..SSSKGFQGTTQTSHGSLMTNKQHQG.KSNNQYYHGKKRKHKRDAPLSDLCR
          Pongo_pygmaeus QDVSLESSQAVGKMQSTQTTNTSNSTNKSQHGSARLFR..SSSKGFQGTTQTSHGSLMTNKQHQG.KSNNQYYHGKKRKHKRDAPLSDLCR
          Macaca_mulatta QDVSLESSQAVGKMQNTQTTNTSNSTNKSQHGSARLFR..SSSKGFQGTTQTSHGSLMTNKQHQG.KSNNQYYHGKKRKHKRDAPLSDLCR
     Callithrix_jacchus QDVSLESSQAVGKMQSTQTTNTSNSTNKSQHGSARLFR..SSSKGFQGTTQTSHGSLMTNKQHQG.KSNNQYYHGKRRKHKRDAPLSDLCR
     Otolemur_garnettii QDVSLESSQAVGKMQSTQTTNTSNSTNKSQHGSARLFR..SSSKGFQGTTQTSHGSLMTNKQHQG.KSNNQYYHGKKRRHKRDAPLSDLCR
            Mus_musculus QDVSLEVSQAVGKMQSTQTTNTPNNANKSQHGSARLFR..SSSKGFQGTAQTSHGALMTSKQHQG.KSNTQYYHGKKRRHKRDAPLSELCR
  Oryctolagus_cuniculus QDVSLEPAQAAGKTQSTQTTNTPSSTNKSQHGAARLFR..SSSKGFQGTAPTSHGSLMTNKPHQG.KSNSQYYHGKKRRHKRDAALSDLCR
              Bos_taurus QDVSLESSQTGGKMQSTQTTNTPNSTNKSQHGSTRLFR..SSSKGFQGTTQTSHGSLMTNKQHQG.KSNNQYYHGKKRKHKRDAPLSDLCR
     Loxodonta_africana QDVALESSQAVGKMQSTQATNTSNSTNKSQHGSARLFR..SSSKGFQGTTQTTHGSLMTNKQHQG.KSNNQYYHGKKRKHKRDAPLSDLCR
       Myotis_lucifugus QEVAAESSPAGGKVQSAPAPATPSSASKSQHGPARLFR..SSSKGFPGTAPTSHGVLVANKPHPG.KPTNQCFHGKRRKHKRDAPLADLCR
    Monodelphis_domestica QEVSLESSQSGGKIQNNQPPNSANSTNKSQHGSARLFR..SSSKGFQGTTPTSHGPLMTNKQHQG.KSNNQYYHGKKRRHKRDAALSDLCR
           Gallus_gallus QEVSLESSQSSGKTQNSQTGNTSSNTNKSQHGSARLFR..SSNKGFQGPANSSHGTSVTNKQHQGSKSHHQYFHGKKRKHKRDAALSDLCR
     Xenopus_tropicalis QESCTSSLQPSGKSQTIQSISSSNNSTKAQHGTTRLFRSTSSSKSFQGHPNTSQGTSVPSRHPLSGKAQHQQYHSKKRKHKRET...DLCR
    Anolis_carolinensis QDISLGSTQMSGKIPSSQAINTSSVGNKSQHGSTRLLRT.SSNKGFQGPGNSSHGTSVTNKPHQGSKS.HQFYHNKKRKHKRDAALSDLYR
         Oryzias_latipes .......KPAVATNHRTQNQSTITPTTGNKGGKARMSR.AHHNNGHQGQQNSTK....TTNNPKN.KLVHQG.NSKKRKNVRDSTQEDLYR

Saccharomyces_cerevisiae YGKNFGYDDVALGSSDGYPVYFPKSTWSAIQPIKNPFSLAIQDPGDESNNISRGSFNIRDIKKAFAGAFDLLTNRCFELHSATFKDRLGKS
(383-584)                ILGNVIKYRGKARDFKDERGLVLNKAIIENENYHKKRSRIIHDEDFAEDTVTSTATATTTDDDYEITNPPAKKAKIEEKPESEPAKRNSGE
                         TYITVSSEDDDEDGYNPYTL
```

Figure 4. The basic motif in the C terminus of PAPD5 is conserved among higher eukaryotes. Amino acid sequences of PAPD5 homologs (Ensembl) were aligned with ClustalW (Chenna et al.) and refined manually (upper panel). The amino acid sequence of the yeast protein Trf4p is shown for comparison (lower panel). Basic amino acids are labeled in blue, acidic amino acids in red.

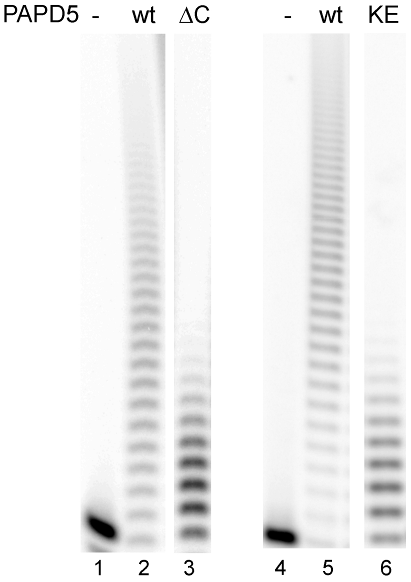


Figure 5. Mutations in the basic motif of PAPD5 lead to loss of activity. PAPD5 wild-type protein (wt), C-terminal deletion mutant (DC, amino acids 1–551), or a point mutant in the basic motif (KE, position 560) expressed in E. coli were incubated with ribo-oligonu- cleotide A15 in the presence of ATP.

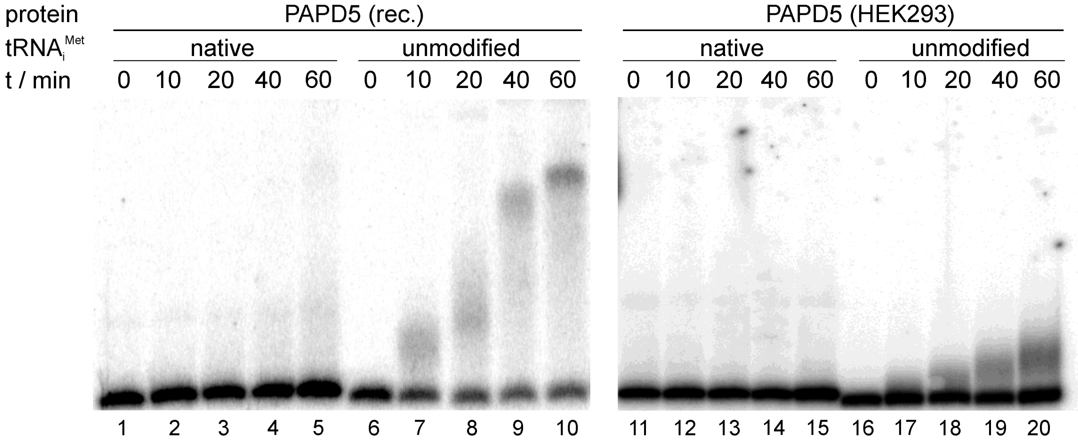


Figure 6. PAPD5 adds poly(A) tails to the unmodified in vitro transcript of S. cerevisiae tRNAiMet, but not to the native tRNA isolated from yeast. PAPD5 protein expressed in E. coli (PAPD5 rec.) or in HEK293 cells was incubated with the different RNA substrates in the presence of ATP, and the reactions were stopped after the reaction times indicated (0, 10, 20, 40, 60 min).

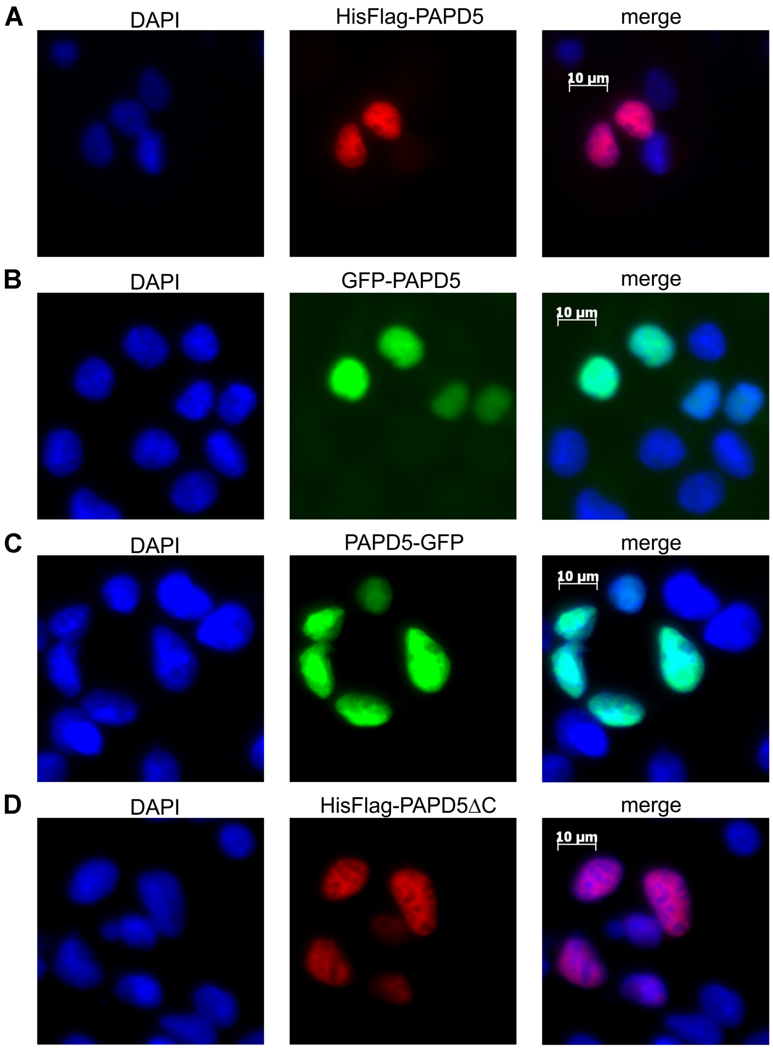


Figure 7. PAPD5 is located in the nucleus. (A) PAPD5 was expressed as a fusion protein with N-terminal His8-Flag tag in HEK293 cells. The protein was detected with anti-Flag antibody and Cy3-conjugated secondary antibody. Nuclei stained with DAPI (left panel); the tagged protein signal (middle); the merged picture (right panel). (B,C) PAPD5 was expressed as a fusion protein with green fluorescent protein fused to the N terminus (B) or the C terminus (C). Panels as in A. (D) PAPD5DC (amino acids 1–551) was expressed as a fusion protein with N-terminal His8-Flag tag in HEK293 cells. The protein was detected with anti-Flag antibody and Cy3-conjugated secondary antibody. Panels as in A. DAPI stain of the nucleus (blue), Cy3 staining (red), and GFP signal (Smit et al.).

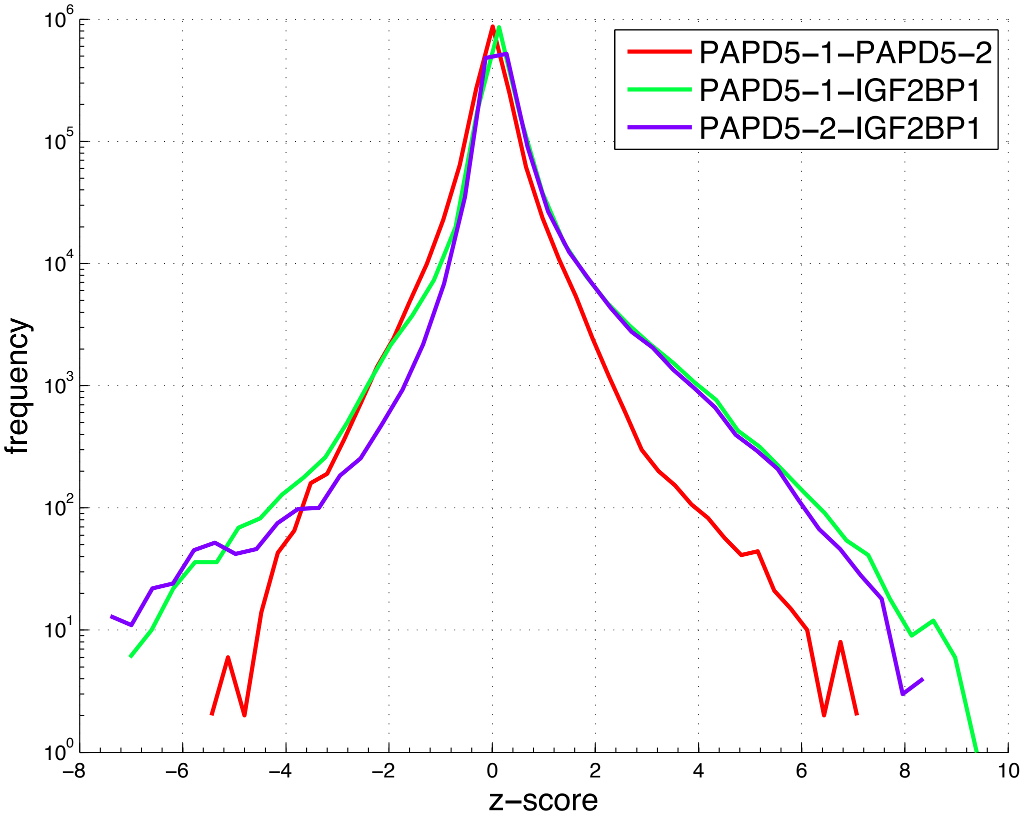


Figure 8. Distribution of pairwise Z-scores computed for 100-nt-long windows in pairs of samples. The reads obtained in two separate PAPD5 PAR-CLIP experiments as well as those previously obtained for the insulin growth factor 2 binding protein 1 (IGF2BP1) (Hafner et al.) were mapped to the genome. After normalization, the Z-scores, giving a measure of the differential expression of individual 100-nt-long windows in each pair of experiments, were calculated.

3.7 References

BALWIERZ, P. J., CARNINCI, P., DAUB, C. O., KAWAI, J., HAYASHIZAKI, Y., VAN BELLE, W., BEISEL, C. & VAN NIMWEGEN, E. 2009. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol,* 10**,** R79.

BUHLER, M., SPIES, N., BARTEL, D. P. & MOAZED, D. 2008. TRAMP-mediated RNA surveillance prevents spurious entry of RNAs into the Schizosaccharomyces pombe siRNA pathway. *Nat Struct Mol Biol,* 15**,** 1015-23.

CALLAHAN, K. P. & BUTLER, J. S. 2010. TRAMP complex enhances RNA degradation by the nuclear exosome component Rrp6. *J Biol Chem,* 285**,** 3540-7.

CARNEIRO, T., CARVALHO, C., BRAGA, J., RINO, J., MILLIGAN, L., TOLLERVEY, D. & CARMO-FONSECA, M. 2007. Depletion of the yeast nuclear exosome subunit Rrp6 results in accumulation of polyadenylated RNAs in a discrete domain within the nucleolus. *Mol Cell Biol,* 27**,** 4157-65.

CHENNA, R., SUGAWARA, H., KOIKE, T., LOPEZ, R., GIBSON, T. J., HIGGINS, D. G. & THOMPSON, J. D. 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res,* 31**,** 3497-500.

DAVIS, C. A. & ARES, M., JR. 2006. Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A,* 103**,** 3262-7.

DEZ, C., HOUSELEY, J. & TOLLERVEY, D. 2006. Surveillance of nuclear-restricted pre-ribosomes within a subnucleolar region of Saccharomyces cerevisiae. *EMBO J,* 25**,** 1534-46.

EGECIOGLU, D. E., HENRAS, A. K. & CHANFREAU, G. F. 2006. Contributions of Trf4p- and Trf5p-dependent polyadenylation to the processing and degradative functions of the yeast nuclear exosome. *RNA,* 12**,** 26-32.

GRIFFITHS-JONES, S., SAINI, H. K., VAN DONGEN, S. & ENRIGHT, A. J. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res,* 36**,** D154-8.

HAFNER, M., LANDGRAF, P., LUDWIG, J., RICE, A., OJO, T., LIN, C., HOLOCH, D., LIM, C. & TUSCHL, T. 2008. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods,* 44**,** 3-12.

HAFNER, M., LANDTHALER, M., BURGER, L., KHORSHID, M., HAUSSER, J., BERNINGER, P., ROTHBALLER, A., ASCANO, M., JR., JUNGKAMP, A. C., MUNSCHAUER, M., ULRICH, A., WARDLE, G. S., DEWELL, S., ZAVOLAN, M. & TUSCHL, T. 2010. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell,* 141**,** 129-41.

HAMILL, S., WOLIN, S. L. & REINISCH, K. M. 2010. Structure and function of the polymerase core of TRAMP, a RNA surveillance complex. *Proc Natl Acad Sci U S A,* 107**,** 15045-50.

HOUSELEY, J., KOTOVIC, K., EL HAGE, A. & TOLLERVEY, D. 2007. Trf4 targets ncRNAs from telomeric and rDNA spacer regions and functions in rDNA copy number control. *EMBO J,* 26**,** 4996-5006.

KADABA, S., KRUEGER, A., TRICE, T., KRECIC, A. M., HINNEBUSCH, A. G. & ANDERSON, J. 2004. Nuclear surveillance and degradation of hypomodified initiator tRNAMet in S. cerevisiae. *Genes Dev,* 18**,** 1227-40.

KADABA, S., WANG, X. & ANDERSON, J. T. 2006. Nuclear RNA surveillance in Saccharomyces cerevisiae: Trf4p-dependent polyadenylation of nascent hypomethylated tRNA and an aberrant form of 5S rRNA. *RNA,* 12**,** 508-21.

LACAVA, J., HOUSELEY, J., SAVEANU, C., PETFALSKI, E., THOMPSON, E., JACQUIER, A. & TOLLERVEY, D. 2005. RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. *Cell,* 121**,** 713-24.

LESTRADE, L. & WEBER, M. J. 2006. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res,* 34**,** D158-62.

LI, Z., PANDIT, S. & DEUTSCHER, M. P. 1998. Polyadenylation of stable RNA precursors in vivo. *Proc Natl Acad Sci U S A,* 95**,** 12158-62.

LI, Z., REIMERS, S., PANDIT, S. & DEUTSCHER, M. P. 2002. RNA quality control: degradation of defective transfer RNA. *EMBO J,* 21**,** 1132-8.

MARTIN, G. & KELLER, W. 2007. RNA-specific ribonucleotidyl transferases. *RNA,* 13**,** 1834-49.

MULLEN, T. E. & MARZLUFF, W. F. 2008. Degradation of histone mRNA requires oligouridylation followed by decapping and simultaneous degradation of the mRNA both 5' to 3' and 3' to 5'. *Genes Dev,* 22**,** 50-65.

NAKAMURA, R., TAKEUCHI, R., TAKATA, K., SHIMANOUCHI, K., ABE, Y., KANAI, Y., RUIKE, T., IHARA, A. & SAKAGUCHI, K. 2008. TRF4 is involved in polyadenylation of snRNAs in Drosophila melanogaster. *Mol Cell Biol,* 28**,** 6620-31.

PRUITT, K. D., TATUSOVA, T. & MAGLOTT, D. R. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res,* 33**,** D501-4.

SAN PAOLO, S., VANACOVA, S., SCHENK, L., SCHERRER, T., BLANK, D., KELLER, W. & GERBER, A. P. 2009. Distinct roles of non-canonical poly(A) polymerases in RNA metabolism. *PLoS Genet,* 5**,** e1000555.

SCHMIDT, M. J. & NORBURY, C. J. 2010. Polyadenylation and beyond: emerging roles for noncanonical poly(A) polymerases. *Wiley Interdiscip Rev RNA,* 1**,** 142-51.

SHCHERBIK, N., WANG, M., LAPIK, Y. R., SRIVASTAVA, L. & PESTOV, D. G. 2010. Polyadenylation and degradation of incomplete RNA polymerase I transcripts in mammalian cells. *EMBO Rep,* 11**,** 106-11.

SLOMOVIC, S., FREMDER, E., STAALS, R. H., PRUIJN, G. J. & SCHUSTER, G. 2010. Addition of poly(A) and poly(A)-rich tails during RNA degradation in the cytoplasm of human cells. *Proc Natl Acad Sci U S A,* 107**,** 7407-12.

SMIT, A., HUBLEY, R. & GREEN, P. 1996-2010. RepeatMasker Open-3.0.

UEDA, H. R., HAYASHI, S., MATSUYAMA, S., YOMO, T., HASHIMOTO, S., KAY, S. A., HOGENESCH, J. B. & IINO, M. 2004. Universality and flexibility in gene expression from bacteria to human. *Proc Natl Acad Sci U S A,* 101**,** 3765-9.

VANACOVA, S., WOLF, J., MARTIN, G., BLANK, D., DETTWILER, S., FRIEDLEIN, A., LANGEN, H., KEITH, G. & KELLER, W. 2005. A new yeast poly(A) polymerase complex involved in RNA quality control. *PLoS Biol,* 3**,** e189.

WIN, T. Z., DRAPER, S., READ, R. L., PEARCE, J., NORBURY, C. J. & WANG, S. W. 2006. Requirement of fission yeast Cid14 in polyadenylation of rRNAs. *Mol Cell Biol,* 26**,** 1710-21.

WLOTZKA, W., KUDLA, G., GRANNEMAN, S. & TOLLERVEY, D. 2011. The nuclear RNA polymerase II surveillance system targets polymerase III transcripts. *EMBO J,* 30**,** 1790-803.

WYERS, F., ROUGEMAILLE, M., BADIS, G., ROUSSELLE, J. C., DUFOUR, M. E., BOULAY, J., REGNAULT, B., DEVAUX, F., NAMANE, A., SERAPHIN, B., LIBRI, D. & JACQUIER, A. 2005. Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell,* 121**,** 725-37.

ZHANG, Z. D., ROZOWSKY, J., SNYDER, M., CHANG, J. & GERSTEIN, M. 2008. Modeling ChIP sequencing in silico with applications. *PLoS Comput Biol,* 4**,** e1000158.

ZHOU, P., LUGOVSKOY, A. A. & WAGNER, G. 2001. A solubility-enhancement tag (SET) for NMR studies of poorly behaving proteins. *J Biomol NMR,* 20**,** 11-4.

Chapter 4: Human DIS3L2 exonuclease is involved in the processing of tRNA-derived small RNAs

Dmytro Ustianenko[1,3], Biter Bilen[2,3], Katerina Chalupnikova[1], Zuzana Feketova[1], Georges Martin[2], Kristyna Hrazdilova[1], Mihaela Zavolan[2], Stepanka Vanacova[1,*]

[1]CEITEC-Central European Institute of Technology, Masaryk University, Kamenice 5, 625 00, Brno, Czech Republic

[2]Biozentrum, University of Basel, Klingelbergstrasse 50 / 70 CH - 4056 Basel, Switzerland

[3]Equal contribution

[*]Corresponding author (E-mail: vanacova@chemi.muni.cz)

Running title: DIS3L2 is a processing factor of tRFs

Keywords: DIS3L2, exosome, tRNA-derived fragments, tRF, small RNAs, CLIP

4.1 Summary

Processing of noncoding RNAs in eukaryotes involves the 3' to 5' exoribonucleolytic activity of type II ribonucleases that are part of the exosome complex. While in yeast it is known that the catalytic activity resides in Rrp44/Dis3 subunit, the functions of the three human Dis3 homologs have not been well characterized. In this work we aimed to uncover the physiological RNA targets of DIS3L2, the DIS3 homolog whose mutation is at the origin of the Perlman syndrome. We show that DIS3L2 is an exosome-independent cytoplasmic exoribonuclease, and through crosslinking and immunoprecipitation followed by RNA sequencing, we demonstrate that DIS3L2 is involved in the formation of tRNA-derived small RNAs (tRFs). A fraction of DIS3L2 along with its tRFs associates with ribosomes and DIS3L2 overexpression alters the polysome/monosome ratio, suggesting that DIS3L2 is involved in translational

regulation. Most importantly, we found that the DIS3L2-dependent tRNA-derived fragments associate with Argonaute2 (AGO2) pointing to a crosstalk between the DIS3L2-dependent processing of tRNAs and AGO2-dependent regulation of gene expression the human cells.

4.2 INTRODUCTION

Every minute, living cells produces millions of RNA molecules. This rapid RNA production is balanced by equally rapid trimming and degradation, and it is this interplay that maintains a proper RNA homeostasis and allows cells to respond to new stimuli. The amount of protein that is produced from an mRNA is proportional to the lifetime of the mRNA, which is directly related to its turnover rate. Thus, mRNA degradation is a key step in gene expression that is subjected to extensive regulation, particularly through Argonaute protein-bound miRNAs (reviewed in (Czech and Hannon, 2011, Lee et al., 1993, Wightman et al., 1993)). miRNAs are only one class of small RNAs that have been uncovered in studies employing deep sequencing. Other small RNAs derive from well-known non-coding RNAs (ncRNAs) such as tRNAs, snoRNAs and/or rRNAs (Cole et al., 2009, Lee et al., 2009, Li et al., 2008, Li et al., 2012)

These small RNAs do not appear to be exclusively unstable degradation intermediates. tRNAs in particular, give rise to fragments that are stably present in various cell types, and this lead to the emergence of the paradigm of tRNA-derived small RNAs having additional biological functions (Jochl et al., 2008, Thompson et al., 2008, Thompson and Parker, 2009, Yamasaki et al., 2009, Heyer et al., 2012). In various studies and depending on the processing pattern, tRNA fragments have acquired different names; the shorter, 20-30 nucleotide (nt) long stable forms originating from 5' or 3' tRNA ends or pre-tRNA trailers were called tRFs, or tsRNAs (Lee et al., 2009, Haussecker et al., 2010, Heyer et al., 2012), while the longer, 30-50 nt stress-induced fragments were called tiRNAs (Yamasaki et al., 2009, Emara et al., 2010), sitRNAs (Li et al., 2008) and tRNA halves (Jochl et al., 2008, Thompson et al., 2008, Fu et al., 2009). Little is known however about the physiological significance of this diverse set of tRNA-derived fragments. The tiRNAs appear able to displace translation initiation factors eIF4G/eIF4A, eIF4F from capped and uncapped mRNAs (Yamasaki et al., 2009, Ivanov et al., 2011), thereby inhibiting translation. Some tRNA (but also other non-coding RNA)-derived fragments were found to associate with Argonaute 2 (AGO2), the effector protein of the RNA-induced Silencing Complex (RISC), indicating that they could have miRNA-like or siRNA-like functions (Haussecker et al., 2010, Li et al., 2012).

Mechanisms responsible for generating short RNA fragments have also started to emerge. For example, Angiogenin has been identified as the key enzyme responsible for the cleavage of tRNAs (Saxena et al., 1992, Fu et al., 2009) upon the exposure of cells to various stress conditions (Thompson et al., 2008, Fu et al., 2009, Li et al., 2012). Enzymes such as DICER (Cole et al., 2009), the cytoplasmic form of RNase Z (Takaku et al., 2003, Elbarbary et al., 2009a, Elbarbary et al., 2009b, Lee et al., 2009), and RNase A (Li et al., 2012) have also been linked to the production of short tRNA-derived fragments. Through endonucleolytic cleavage, these enzymes often generate extended products that need to be further digested exonucleolytically to yield the tRFs that are typically detected through RNA sequencing.

Processing of non-coding RNAs in eukaryotes involves 3' to 5' exoribonucleolytic trimming catalyzed by the exosome complex that contains a type II nuclease (RNase II) (Mitchell et al., 1997). In yeast, the exosome consists of ten core subunits, all of which are essential. Nine of the subunits are catalytically inactive (Vanacova and Stefl, 2007)) and form a barrel-shaped structure, while the tenth subunit, Rrp44 (also known as Dis3), carries the nuclease activity (Liu et al., 2006, Dziembowski et al., 2007, Schneider et al., 2007). Exosomes operate in both nucleus and cytoplasm, being involved in RNA interference, processing, maturation and turnover as well as in RNA surveillance pathways that recognize and degrade aberrant RNAs (Houseley and Tollervey, 2009)). In the cytoplasm, the exosome is responsible for degradation of unstable mRNAs containing AU-rich sequence elements and of defective mRNAs that are detected by quality control pathways such as nonsense- or nonstop-mediated mRNA decay (Houseley et al., 2006, Houseley and Tollervey, 2009)). The nuclear exosomes of *S. cerevisiae* additionally contain a non-essential 3' to 5' exonuclease, termed Rrp6 (Vanacova and Stefl, 2007).

The human genome encodes three Rrp44/Dis3 homologs: DIS3, DIS3L and DIS3L2, which slightly differ in domain composition (Figure 1A). The DIS3 and DIS3L proteins associate with the nuclear and cytoplasmic exosomes, respectively (Staals et al., 2010, Tomecki et al., 2010) and have been implicated in nuclear rRNA processing, degradation of promoter upstream transcripts (PROMPTS) and mRNAs such as *c-MYC* and *c-FOS* (Tomecki et al., 2010). DIS3L2 is the most distant Rrp44 homolog. Lacking the PilT N-terminus (PIN) domain that mediates DIS3 association with the exosome core in yeast (Lebreton et al., 2008, Schneider et al., 2009), DIS3L2 may not associate with the exosome, but rather act independently or in association with other proteins. Recent works have linked DIS3L2 mutations with the Perlman syndrome, a genetic overgrowth disorder appearing at birth (Perlman, 1986), and revealed that DIS3L2 plays a role in cell cycle regulation and cell division (Neumann et al., 2010, Astuti et al., 2012). Interestingly, disruptions of the *DIS3L2* gene lead to aneuploidy, mitotic errors and changes in expression of mitosis-related proteins such as Aurora-B kinase, Cyclin B1 and p21 (Astuti et al., 2012). However, the molecular mechanism of DIS3L2 function remained elusive.

In this work we aimed to uncover the biochemical properties and physiological RNA targets of DIS3L2. By using *in vivo* crosslinking and immunoprecipitation followed by next generation sequencing (CLIP-seq) we show that DIS3L2 is bound to relatively short (18-34 nt) 5' tRNA-derived fragments (5' tRFs). Biochemical analyses indicate that DIS3L2 is involved in the production of 5' tRFs, and that the generation of stable tRNA-derived fragments requires both endo- as well as exo-ribonucleolytic activities. Together with 5' tRFs, a fraction of DIS3L2 associates with ribosomes and DIS3L2 overexpression shifts the polysome/monosome ratio. These results suggest an involvement of DIS3L2 in translational regulation. Importantly, we found that some tRFs associate with AGO2 and have significant reverse complementarity to mRNAs. These findings point to a role of tRFs and DIS3L2 in the regulation of gene expression in human cells via an RNAi-like pathway.

4.3 Results

4.3.1 DIS3L2 is a cytoplasmic exonuclease that does not interact with the exosome core

DIS3L2 is a member of the RNase II family of enzymes. In contrast to DIS3 and DIS3L, DIS3L2 lacks the PIN domain that in yeast mediates the interaction with the core exosome (Lebreton et al., 2008, Schneider et al., 2009). To determine whether the absence of the PIN domain results in DIS3L2 acting independently of the core exosome, we analyzed immunoprecipitated (IPed) FLAG-DIS3L2 samples by western blot with antibodies against two core exosome subunits RRP40 and RRP41 (Figure 1B). Each of the two positive controls, FLAG-DIS3 and FLAG-DIS3L, coimmunoprecipitated both factors, whereas no signal over background was observed in FLAG-DIS3L2 IP samples (Figure 1B). We concluded that absence of the PIN domain renders DIS3L2 unable to interact with exosomes *in vivo*. The two exosome-interacting homologs show distinct subcellular localization, with DIS3 being primarily nuclear and DIS3L mainly cytoplasmic (Staals et al., 2010, Tomecki et al., 2010). To determine the specific localization of DIS3L2, we transiently expressed N- and C-terminal EGFP fusion proteins (EGFP-DIS3L2 (data not shown) and DIS3L2-EGFP, Figure 1C and 1D), and performed immunolocalization with DIS3L2 specific antibodies (Figure 1C). Both approaches revealed that DIS3L2 is a cytoplasmic protein (Figure 1C), in agreement with a recent report (Astuti et al., 2012).

To test whether DIS3L2 has 3' to 5' exoribonucleolytic activity, we expressed and purified recombinant DIS3L2 from *E. coli* and performed *in vitro* RNA degradation assays with unstructured RNA as a substrate. The recombinant protein showed processive 3' to 5' exonucleolytic activity on U30 RNA whereas a D391N mutation of the predicted catalytic residue abolished this activity (Figure 1E). Similar activity was observed with FLAG-DIS3L2 purified from HEK293T cells on three different unstructured RNA substrates. Furthermore, similarly to Rrp44 (Liu et al., 2006), DIS3L2 is also able to digest highly structured RNAs such as tRNA. Finally, akin to yeast and mammalian DIS3 proteins and to the *E. coli* RNase II (Frazao et al., 2006, Dziembowski et al., 2007), DIS3L2 requires bivalent metal cations for catalysis and its typical end product of digestion is 2-4 nucleotides in length.

Because DIS3L2 does not associate with the cofactors that provide RNA-binding specificity and affinity to DIS3 and DIS3L, we tested its ability to directly bind RNA in an electromobility shift assay (EMSA). The assay was performed in absence of Mg2++ ions, thus abolishing the enzymatic activity but without affecting the RNA binding of DIS3L2 (Schneider et al., 2007). Interestingly, DIS3L2 displayed rather high affinity for U30 RNA, in the nanomolar range (estimated KD 100-125 nM) (Figure 1F).

4.3.2 Identification of DIS3L2 RNA targets by PAR-CLIP

Having proven that DIS3L2 is a *bona fide* 3' to 5' cytoplasmic exoribonuclease, we applied the PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) method combined with high-throughput Illumina-based sequencing (Ule et al., 2005, Wang et al., 2009, Martin et al., 2012) in a HEK293 cell line overexpressing FLAG-tagged DIS3L2 (Figure 2A), to identify its *in vivo* physiological targets. We performed two independent experiments and deposited the sequence data in the Gene Expression Omnibus (Edgar et al., 2002) with accession number (GSE42311). The experiments resulted in 18,113,339 and 16,151,628 good quality reads, respectively, which we mapped to the human genome assembly (hg19 from the http://genome.cse.ucsc.edu) (Kent et al., 2002, Rosenbloom et al., 2012) using the CLIPZ server (http://www.clipz.unibas.ch/, (Khorshid et al., 2011)). 51% and 21% of the reads, respectively, mapped to the genome and the 3,632,533 (20%) and 1,192,858 (6%) reads that mapped uniquely to the genome were used for our subsequent analyses. For both replicates, more than half of the uniquely mapping reads originated from mRNAs (Figure 2B). Repeat elements and tRNAs together represented approximately 30% of the reads, whereas rRNAs, miscRNA (various other known ncRNAs), and reads mapping to un-annotated regions accounted for approximately 15% of the reads (Figure 2B). To identify the sites that were most significantly crosslinked to DIS3L2, we used a method that takes advantage of the T-to-C mutations that are introduced during PAR-CLIP sample preparation (Jaskiewicz et al., 2012). We then associated the genomic locations corresponding to the most significantly crosslinked nucleotides with genes based on the various gene tracks available from the genome browser of the University of California Santa Cruz (see Methods). Our two PAR-CLIP samples showed substantial reproducibility at the level of crosslinked genes, especially when considering a limited number (1,000) of top most significantly crosslinked positions. These positions mapped to 225 and 160 distinct genes, with 106 of the genes being identified in both replicates (Fisher's Exact test p-value=1.3e-160, odds-ratio=225). Furthermore, around half of the 1,000 most significantly crosslinked positions from one sample were within 10 nucleotides distance from at least one of the 1,000 most significantly crosslinked nucleotides in the second sample further showing the reproducibility of the crosslinking sites across the replicates. To identify the main *in vivo* substrates bound by DIS3L2, we tested whether specific categories of genes were over-represented among the most significantly crosslinked 1,000 positions compared to crosslinked positions recovered with much more permissive crosslink score cut-offs (30,318 and 25,811 positions, respectively, taking half of the positions that we would consider significantly crosslinked if they were supported by only one read that had a T-to-C mutation). We found that tRNAs were the most enriched category of genes represented among the most significantly crosslinked positions: DIS3L2 bound 109 and 107 distinct tRNA genes in the first and second replicates, respectively, 81 being common (Figure 2C). The number of significantly crosslinked positions on individual tRNA genes that were crosslinked in both replicates had a modest linear correlation (R=0.47) further supporting the reproducibility of DIS3L2 preference for these substrates. The most crosslinked tRNAs were Leu(CAA), Val(TAC), Arg(CCG), Ser(GCT), Lys(TTT), Val(AAC), SeC(e)(TCA), Arg(TCT), and Ala(AGC). Although more than half of the uniquely mapping reads were annotated as mRNAs, only a small number of mRNAs (N=7) contained positions with a crosslink score in the top 1,000 in both libraries. These transcripts corresponded to *HSPA1B, MAZ, BIRC5, ACTG1, USP22,* and *HIST1H4B*. Gene Ontology enrichment analysis with the Ontologizer software (Bauer et al., 2008), see also Methods) of an extended set of 89 mRNA genes that gave rise to significantly crosslinked sites in either of the replicates revealed endoplasmic reticulum (adjusted p-value=0.001, N=21) and the endomembrane system (adjusted p-value=0.004, N=24) cellular components. The abundantly and ubiquitously expressed (Landgraf et al., 2007) mature miRNAs from the miR-17-92 locus (miR-17, mir-18a, mir-20a) were the only miRNAs that were crosslinked to DIS3L2. Among rRNA, we obtained crosslinks to *RN5S9* and *RN5S215* 5S rRNA transcripts, which are part of the large ribosomal subunit. Although snoRNAs are predominantly nuclear and DIS3L2 predominantly cytoplasmic, we identified 35 snoRNAs, among which *U60, U14A, U8, mgU6-77, HBII-99B, U33,* and *U56* were reproducibly identified in both replicates. With the exception of *U1* and *RNU12,* almost all crosslinked snRNAs *(RNU2-2, RNU2-6P, RNU4-2, RNU5E-1, U2,* and *RNU11)* were identified in both replicates.

To validate the CLIP results, we performed RNA Immunoprecipitation (RIP) from the FLAG-DIS3L2 Flp-In HEK293T stable cell line followed by reverse transcription with oligo(Landgraf et al., 2007) primer and semi-quantitative PCR with oligonucleotides specific for genes encoding crosslinked mRNAs from both replicates (Figure 2D). FLAG-DIS3L2 samples showed significant enrichment of most selected target mRNAs *(ACTG1, BIRC5, HSPA1B, MAZ)* over the unspecific binding to FLAG-matrix detected with extracts from empty HEK293 Flp-In cells (Figure 2D). To test the specific binding to tRNAs, we used northern blot analysis of the RIP samples and DNA probes specific to the 5' and 3' part of tRNAVal, respectively (Figure 2E). Although the mature tRNAVal was detected also in the control sample, the FLAG-DIS3L2 specifically immunoprecipitated pre-tRNA (slower migrating band in Figure 2E) and shorter forms migrating as 50 nt and 30 nt fragments when probed with a 5'-end probe and migrating as 50 nt and 20 nt fragments when probed with a 3'-end tRNA probe (Figure 2E).

4.3.3 DIS3L2 predominantly crosslinks to tRNA 5' ends

To further investigate the relevance of DIS3L2 binding to tRNA-derived fragments (Figure 2E), we used 44 tRNAs that are precisely 72-nucleotides-long and were CLIPed with DIS3L2 to compute the profile of the average crosslink score along the tRNAs. We found that the highest binding scores were in the 5' region of the tRNAs (Figure 3A and 3B). Importantly, a similar profile was obtained for the set of 39 73-nucleotide-long tRNAs. This positional preference was not due to preferential loss by our processing pipeline of 3' end reads that end in the non-templated CCA tri-nucleotides, and was also not dependent on whether we used uniquely or multi-mapping reads in our analysis. Because the DIS3L2-crosslinked tRNA fragments resemble most the 5' tRFs (Lee et al., 2009, Haussecker et al., 2010), we decided to use this designation for the DIS3L2-bound 5' tRNA fragments.

The specific binding of DIS3L2 was further verified on tRNALeu- and tRNAVal-derived tRFs (Figure 3C) by semi-quantitative RT-PCR applied to RNA immunoprecipitated (RIP) with DIS3L2 (see above). To specifically amplify the 5' tRFs and not the full length tRNAs, we applied the method of stem-loop RT-PCR in which stem-loop RT primer is hybridized to a 5' tRF molecule and then reverse transcribed with a reverse transcriptase. For the cDNA synthesis we used primers aligning specifically to the last 6 nucleotides of tRFs identified in CLIP (see methods). Next, the RT products are amplified with the combination of a forward tRF-specific primer and a reverse stem-loop-RT-primer-specific oligonucleotide (Chen et al., 2005, Varkonyi-Gasic et al., 2007). We have observed specific coimmunoprecipitation of 5' tRFs with FLAG-DIS3L2 protein (Figure 3C). The identity of amplified 5' tRFs was further validated by DNA sequencing Unfortunately, the same approach could not be used for the specific amplification of 3' tRFs, because it does not allow for selective amplification at certain 5'-end position of a 3' tRF. In summary, we conclude that DIS3L2 binds the 5' ends of a subset of tRNAs.

4.3.4 DIS3L2 is a processing factor of tRNA-derived small RNAs that may have a miRNA-like function in mRNA gene silencing

Having shown that DIS3L2 binds 5' tRNA regions, we next investigated whether it is responsible for 5' tRFs turnover or processing. As certain stress conditions are responsible for the generation of 5' tRNA halves (Thompson et al., 2008, Fu et al., 2009), we first checked, whether DIS3L2 is involved in this process or in the degradation of the cleaved fragments. We monitored tRNA cleavage during oxidative shock in HeLa cells treated with 5 mM H2O2. Whereas the cleavage of mature tRNAs proceeded with similar dynamics in the DIS3L2 and control siRNA-treated cells, the cleaved tRNA fragments showed a slightly stronger accumulation in the DIS3L2 knockdown. This could indicate that DIS3L2 targets Angiogenin-cleaved tRNAs. Reasoning that DIS3L2 may process longer forms of tRNA fragments to tRFs observed in this and other studies, we monitored the steady state levels of the 22 and 23 nt long 5' tRFLeu(CAA) in response to modified levels of DIS3L2. Whereas DIS3L2 overexpression resulted in 1.5 to 2.5-fold increase of the tRFs analyzed (Figure 4A), DIS3L2 downregulation in HeLa cells led to 60% and 80% reduction of 5' tRFLeu(Neumann et al., 2010) and 5' tRFLeu(CAA), respectively (Figure 4C). To find out whether the catalytic activity of DIS3L2 was responsible for the changes in tRF expression, we used a stable cell line with inducible expression of the catalytically inactive form of FLAG-DIS3L2 (D391N, Figure 4B). Overexpression of the D391N mutant resulted in 20 - 40% reduction in 5' tRF levels (Figure 4A), a phenotype similar to DIS3L2 down-regulation (Figure 4C). This implies that DIS3L2 catalytic activity is responsible for tRF production and that D391N overexpression acts in a dominant negative manner. Interestingly, we observed similar changes in the levels of mature miR-30b in HEK293T cells and miR-30b and let-7a in HeLa cells (Figure 4A and 4C), but the miRNAs did not change as profoundly as the 5' tRFs. Cole et al. 2009 implicated Dicer1 (DCR1) in the production of certain tRNA fragments (Cole et al., 2009). To rule out the possibility that DIS3L2 regulates DCR1 expression by targeting *DCR1* mRNA (which would in turn lead to defects in tRF processing) we investigated whether altered levels of DIS3L2 lead to changes in the expression of DCR1 or to defects in miRNA processing. Neither over-expression, nor down-regulation of WT DIS3L2 modified DCR1 expression and no significant changes were detected in the levels of mature or pre-miR-16 or pre-let-7 miRNAs in HEK293T cells. In contrast, ectopic expression of LIN28A, a known negative regulator of specific miRNA processing, caused pre-miRNA accumulation in control samples. Altogether, we conclude that DIS3L2 is directly responsible for the processing of tRNAs into 5' tRFs. To gain insight into the possible function of the 5' tRFs, we investigated whether they could function within the cytoplasmic RNA silencing machinery. Other categories of non-coding RNAs have recently been found to be loaded into AGO and to function in mRNA gene silencing (Ender et al., 2008, Li et al., 2012). To this end, we applied our processing pipeline to an AGO2 PAR-CLIP library (accession GSM714646 in the GEO database of NCBI) from Kishore et al. (Kishore et al., 2011) and used the top 1,000 crosslinked sites as for the DIS3L2 libraries. Although most of the significantly crosslinked AGO2-binding sites were located in miRNA genes, we also obtained 10 significantly crosslinked tRNAs (Val(CAC), Val(AAC), Val(TAC), Leu(CAA), Pro(CGG), Ala(AGC), Asn(GTT), Ile(TAT), Ala(CGC)). Interestingly, all of these tRNAs were also crosslinked in the DIS3L2 replicates (Fisher's Exact test, p-value=1.7e-07). In contrast to the reads derived from miRNAs, that have a sharp peak at 23 nucleotides in both AGO2 and DIS3L2 CLIP samples, tRNA-derived fragments have a much broader length distribution (Figure 4D).

Moreover, the tRNA-derived fragments that were found crosslinked to AGO2 had a length distribution with a wide plateau, starting at 23 nt (Figure 4D). To determine which mRNAs these tRNA-derived fragments could target, we applied a method for predicting miRNA-binding sites in CLIP data that we recently developed (Khorshid et al., 2013). For comparison, we predicted sites for *bona fide* miRNAs as well as for mono- and di-nucleotide conserving shuffled variants of both tRNA and miRNA. We found that although the best hybrids that were predicted for the tRNA-derived small RNAs had on average smaller scores compared to hybrids involving *bona fide* miRNAs, none of the randomized tRNA or miRNA fragments was predicted to target one of the mRNA fragments obtained in the AGO2 PAR-CLIP. This suggests that tRNA-derived small RNAs may load into AGO2 and function in mRNA gene silencing in a way similar to miRNAs (Figure 4E).

Since tRNAs are not the main AGO2-bound RNAs, it is possible that a more permissive crosslink scoring cut-off for AGO2 would reveal more tRNA-derived fragments processed by DIS3L2. Indeed, among the top 2,000 crosslinked sites of AGO2 we obtained 51 tRNA genes out of which 37 were represented among top 1,000 most significantly crosslinked DIS3L2 positions (which covered a total of 135 tRNAs, Fisher's Exact Test p-value=3.2e-16, odds-ratio=13). Importantly, the tRNA-derived fragments identified through small RNA sequencing (Mayr and Bartel, 2009) and in the AGO2 PAR-CLIP samples also preferentially originated from the tRNA 5' ends (Cole et al., 2009), just as in the DIS3L2 PAR-CLIP samples. This further supports the hypothesis that DIS3L2 is involved in the processing of the majority of tRFs in cytoplasm under standard growth conditions. Among the targets that we predicted for tRFs that did not match any miRNA with a comparable score, and whose hybrid score was higher than the median score of the miRNA-mRNA hybrids (median score = 147.6, see Methods) were mRNAs encoding *ACVR1B, ATF7, DNAJB12, HMGA2, HNRNPA2B1, MTHFD2, PPIL1, PSMB1, RPA2, SLC19A2, SLC38A2, SRSF6* and *TUBB* which could be targeted by small RNAs derived from tRNAAsn(GTT), tRNAVal(TAC), tRNAVal(AAC), tRNALeu(CAA), tRNAAla(CGC) and tRNAPro(CGG).

4.3.5 DIS3L2 and 5' tRFs associate with active polysomes

Although we tested whether DIS3L2 directly regulates the stability of top CLIPed mRNA targets, we did not observed significant changes neither upon DIS3L2 knockdown, nor over-expression (data not shown). tRNAs are a fundamental component of the translation machinery and tRFs themselves have been shown to interact with proteins with a function in translation regulation, such as YB1, TIA1, PABP1 and AGO2 (Ivanov et al., 2011). We thus reasoned that DIS3L2 and its 5' tRFs could regulate the translation rather than the stability of mRNAs. To investigate this hypothesis we sought to determine whether DIS3L2 associates with translating ribosomes, and we performed polysomal profiling of DIS3L2-overexpressing and control HEK293T cells combined with western blot analysis of individual sucrose gradient fractions (Figure 5A). The endogenous as well as overexpressed DIS3L2 signal was present along the whole sucrose gradient indicating the presence of DIS3L2 in monosomes as well as polysomal fractions (Figure 5B). The polysomal profiles showed an increased signal for 80S monosomes in FLAG-DIS3L2 overexpressing samples. Comparing the relative abundance of polysome and monosome peaks of control and FLAG-DIS3L2 overexpressing cells, we found that the monosomal fraction increased three times upon DIS3L2 overexpression (Figure 5C). Furthermore, overexpression of the D391N mutant did not result in a similarly strong shift towards monosomes indicating that the nuclease activity of DIS3L2 is responsible for this phenotype (Figure 5C, last column). This was not due to different levels of expression of the tagged proteins. The changes in translation activity may be one of the reasons for growth defects associated with DIS3L2 overexpression (Astuti et al., 2012) In agreement with our polysomal profiling results, the overexpression of catalytically inactive DIS3L2 did not result in growth retardation.

Because the yeast Dis3 and the associated exosome are essential for ribosomal RNA (rRNA) processing and degradation, we checked whether DIS3L2 overexpression leads to changes in the rRNA profile. We did not find however, any accumulation of aberrant rRNA species or an imbalance in SSU/LSU rRNA levels in cells with altered DIS3L2 expression. Moreover, Astuti et al. ruled out the involvement of DIS3L2 in 5.8S rRNA maturation (Astuti et al., 2012).

Reasoning that 5' tRFs mediate at least some of DIS3L2 activity on ribosomes, we isolated RNA from the individual fractions of the sucrose gradient and used it for stem-loop RT-PCR analysis (see above) with specific primers for two 5' Leu tRFs. We found that the tRF signal paralleled DIS3L2 distribution along the whole sucrose gradient, with a stronger signal over the 60 and 80S monosomes (Figure 5D), indicating that they may have a role during translation initiation.

4.4 DISCUSSION

The high-throughput sequencing of various RNA size fractions have led in the recent years to the identification of new types of small and long RNAs, some of which have functions in the regulation of gene expression. In this work we have established DIS3L2 as a factor responsible for the generation of tRNA-derived fragments with regulatory potential. DIS3L2 is a human homolog of Dis3, the key exonuclease of yeast exosomes. In contrast to its DIS3 yeast and human homologs however, DIS3L2 does not associate with exosomes *in vivo*.

To uncover the DIS3L2-specific RNA targets *in vivo,* we applied photo-crosslinking followed by DIS3L2 immunoprecipitation and next generation sequencing, (DIS3L2 PAR-CLIP) in HEK293 Flp-In cells. Analysis of CLIP data combined with further experimental validation uncovered a strong enrichment in binding of DIS3L2 to tRNA molecules, particularly to fragments originating from tRNA 5'-ends. We designated these molecules tRFs, akin to the tRNA-derived fragments recently reported by Haussecker et al. (Haussecker et al., 2010).

A number of studies reported that tRNA halves are produced in response to certain stress stimuli, but also exist stably in cells (Fu et al., 2009, Lee et al., 2009, Li et al., 2012). To date, several endonucleases such as DICER1, Angiogenin or RNase Z, have been proposed to generate different types of tRNA fragments (Takaku et al., 2003, Cole et al., 2009, Lee et al., 2009). In many cases, such cleavage would generate fragments longer than those identified by RNA sequencing, which presumably captures mature/functional fragments. This suggests that an additional exonuclease activity is involved in the processing of tRFs. Because DIS3L2 knockdown leads to a reduction in 5' tRF levels and DIS3L2 upregulation increases 5' tRF levels, we propose that DIS3L2 provides this so-far-unknown activity.

At the moment, it is unclear whether DIS3L2 targets tRNA halves that result from an initial endonucleolytic cleavage or whether it targets full length tRNAs. The fact that both 5' and 3' tRNA cleavage products became stabilized in DIS3L2 knockdown cells upon H2O2 stress, suggested that DIS3L2 may either be responsible for their decay or further processing. On the other hand, the infrequent occurrence of crosslinks to fragments derived from tRNA 3' ends under non-stressed conditions and our finding that DIS3L2 processively digests both unstructured and structured RNAs *in vitro* (Figure 1, Figure 4) suggest that rapid decay is responsible for the depletion of 3' ends of tRNA fragments in our CLIP data. The fact that some 5' tRFs escape complete degradation could indicate that an additional factor controls DIS3L2 activity and sets the boundaries of mature 5' tRFs. Previous studies have demonstrated AGO2 binding to tRNA-derived fragments in mammals (Cole et al., 2009, Haussecker et al., 2010, Li et al., 2012), *Tetrahymena* (Couvillion et al., 2012), and fission yeast (Buhler et al., 2008). Interestingly, Xue et al. have recently reported that Dis3 and the exosome can produce, together with Argonaute, mature miRNAs in *Neurospora crassa* (Xue et al., 2012). Our analysis indicated a significant overlap between AGO2 and DIS3L2 binding to tRFs. We propose that AGO2 binding to the 5' region of tRFs acts as a molecular ruler (Wang et al., 2008, Haussecker et al., 2010) controlling DIS3L2 digestion (Figure 6B). 5' tRFs are thus protected by AGO2, leaving the 3' ends of the tRNAs or fragments accessible to DIS3L2 or other nucleases. We have observed that DIS3L2 also copurified extended tRNA species (Figure 2E). These may represent pre-tRNAs that for some reason escaped nuclear processing and were exported to cytoplasm. It is possible, that pre-tRNAs that fail RNaseP and/or RNaseZ processing and therefore also the subsequent aminoacylation serve as the cytoplasmic pool for tRF production. This hypothesis could also explain the earlier finding of another type of tRNA-derived fragments originating from 3'-end tRNA trailers that are found in cytoplasmic fractions (Li et al., 2012). Our results rule out the possibility that DIS3L2 affects the stability of the mRNAs corresponding to the endonucleases that are already known to be responsible for tRF formation (Babiarz and Blelloch, 2008). Particularly, we show that DICER1 levels and miRNA processing are not regulated by DIS3L2. Furthermore, Angiogenin-mediated cleavage is not altered upon DIS3L2 knockdown, and finally, our northern analyses did not show any accumulation of pre-tRNA molecules that would indicate defects in RNAse Z activity. It is currently unclear how DIS3L2 selects its targets. Apart from 5' tRFs, DIS3L2 CLIP identified other small RNAs originating from RNA Polymerase III transcripts, such as 5S rRNA and a miR-1975 originating from Y-RNA (Canella et al., 2010). We have shown that DIS3L2 binds U30 RNA with nanomolar affinity. RNA Pol III transcription termination often involves a U-rich stretch that could form a landing platform for initial DIS3L2 binding. On the other hand, it is possible that DIS3L2 associates with other RNA-binding cofactors that provide RNA substrate specificity.

The key question is whether DIS3L2-linked tRFs possess any biological function. Mostly due to tRF association with Argonaute proteins, some recent studies focused on the function of tRFs in the RNAi pathway. It was shown that certain 3' tRFs in association with AGO2 are able to induce cleavage of a complementary RNA *in vitro* (Li et al., 2012), yet *in vivo* their trans-silencing potential was rather small (Haussecker et al., 2010). On the other hand, in *Tetrahymena,* nuclear tRNA derived 3' RNAs associate with a Piwi-like protein and, in complex with the nuclear 5' to 3' exonuclease Xrn2, are responsible for the processing of the ribosomal RNAs (Couvillion et al., 2012). In fission yeast, tRNAGlu-derived small 5' tRFGlu represent a highly abundant class of sRNAs that have been proposed to interfere with the RNAi pathway (Buhler et al., 2008). Last, but not least, in mammalian cells, tRNA fragments have been shown to induce translational arrest (Yamasaki et al., 2009). In this context, we show that a fraction of DIS3L2 and the respective 5' tRFs associate with ribosomes and that DIS3L2 overexpression changes ribosomal profiles towards monosomes, which is generally understood as a sign of translational inhibition. The majority of DIS3L2 is not bound to ribosomes. However, this also seems to be the case for other proteins that are known to associate with ribosomes, e.g. CIRP2 (Matsumoto et al., 2000), S6 kinase 1 (Volarevic and Thomas, 2001), PRMT3 (Bachand and Silver, 2004) or Gemin3/4 (Nelson et al., 2004). Because we detect also the endogenous DIS3L2 in the individual sucrose fractions, it is unlikely that we detect FLAG-DIS3L2 only due to its overexpression. Although it is possible that DIS3L2 co-sediments with polysomes due to its association with another high molecular complex, the effect of DIS3L2 overexpression on polysome/monosome ratio further supports its association with translation machinery. Even though we do not know the stoichiometry of DIS3L2 loading to ribosomes we speculate that DIS3L2 associates with ribosomes under distinct conditions and dependent on specific subsets of mRNAs. This would be consistent with the fact that a few specific types of tRNAs (tRNALeu and tRNAVal) were especially enriched in our CLIP data. Interestingly, a recent study in yeast pointed to a correlation between the stress-induced general tRNA cleavage, protection of tRNALeu(CAA) and translation of mRNAs with a high frequency of Leu(CAA) codon occurrence (Chan et al., 2012). Upon stress, yeast cells upregulate Trm4-mediated tRNALeu(CAA) methylation (Trm4 is a yeast homolog of the mammalian DNA methyltransferase DNMT2), thereby protecting it from endonucleolytic cleavage. A subset of ribosomal proteins encoded by mRNAs with a high frequency of Leu(CAA) codons is then preferentially translated under stress conditions (Chan et al., 2012). The set of human genes with a very high frequency of Leu codons is strongly enriched in Gene Ontology categories that reflect membrane biosynthetic processes (protein lipidation, phospholipid/lipoprotein/glycerolipid biosynthetic process). Whether these or other Leu(CAA)-rich factors are regulated in a DIS3L2-dependent manner remains to be determined.

On the other hand, expression of the DIS3L2 may primarily affect protein synthesis via small regulatory RNAs. Fragments derived from tRNAs have been previously linked to translational inhibition through the association with a translation initiation factor (Ivanov et al., 2011). The fact that we detected slightly higher 5' tRF LeuCAA levels at 60S and 80S ribosomal fractions, could indicate that they function in translational initiation. However, the overall 5' tRFs distribution over the whole polysomal profile points to a more specific role in translational repression. Because it is well established that microRNAs inhibit mRNA translation, we sought evidence for DIS3L2-linked 5' tRFs acting in a miRNA-like pathway. A computational search for putative mRNA targets revealed several mRNAs whose disregulation could explain the slow growth and division phenotype observed upon DIS3L2 knockdown or mutation (Astuti et al., 2012)), such as *DNAJB12, RPA2,* or others. Nevertheless, future studies will be necessary for clarifying whether 5' tRFs act in a miRNA-like manner on these predicted targets. Apart from the predicted mRNA targets for 5' tRFs, our DIS3L2-CLIP analyses identified crosslinks to several mRNAs. Interestingly, the most significantly crosslinked mRNAs encode proteins linked to cell cycle *(HIST1H4B, USP22)* and mitosis *(BIRC5)*. In agreement with the proposed indirect role of DIS3L2 on gene expression regulation, DIS3L2 knockdown or overexpression did not cause stabilization or down-regulation of these mRNAs, respectively (data not shown). This resembles results of Astuti et al. who reported on differences in expression of several proteins involved in mitosis upon the knock down of DIS3L2 in HeLa cells, without the change of their corresponding mRNA levels (except for one, the *TTK)* (Astuti et al., 2012). Thus, DIS3L2 may regulate these genes only at the level of translation. In summary, here we have uncovered the role of a cytoplasmic exonuclease that mediates production of potentially functional tRNA-derived fragments. We propose that DIS3L2 and tRFs cooperatively regulate gene expression. As different types of tRNA-derived fragments exist stably in various cell types, future studies will be needed to assess tRF specific functions *in vivo*.

4.5 Material and Methods

4.5.1 Mammalian cell culture and transfection

Mammalian cells (HEK Flip In, HEK Flip In FLAG-DIS3L2, HEK293T, HeLa) were maintained in Dulbecco's Modified Eagle's Medium (DMEM), supplemented with 10% fetal calf serum at 37°C in the presence of 5% CO2. Inducible conditions: 10-50ng/ml of doxycycline. For transient expressions, cells were grown to 70% confluency, plasmid DNA was transfected using TURBOFECT (Fermentas) following the manufacturer manual. Preparation of stable cell lines

Plasmids (pcDNA5 FRT/TO FLAG DIS3L2, pcDNA5/FRT FLAG DIS3L2, pcDNA5/FRT/TO DIS312 D391N mutant) were transfected to Flp-InTM or to Flp-InTM T-RExTM (Invitrogen) cell line with TURBOFECT reagent. The stable inducible cell lines overexpressing different versions hDIS3L2 were generated with the use of the HEK293 Flp-In™ and Flp-In T-REx™ system (Invitrogen) according to the protocol of the manufacturer.

4.5.2 Polysomal profiling

The polysomal profiling was performed with minor modifications according to (Damgaard and Lykke-Andersen, 2011). For details see supplementary material.

4.5.3 RNA analysis by RT-PCR.

Total RNA was isolated with TRIzol (Invitrogen) according to manufacturer instructions followed by RNase-free DNase (TURBO DNase, Fermentas) treatment. 1 ug of purified RNA was reverse transcribed using random hexamers or oligo dT (as indicated) and Superscript III reverse transcriptase (Invitrogen) according to manufacturer instructions. Real-time quantitative PCR (RT-qPCR) was performed using

FastStart Universal SYBR Green Master (Roche) and gene-specific primer pairs on Q-PCR Light Cycler 7500 (Applied Biosystems). For details see Supplementary material.

4.5.4 PAR-CLIP

PAR-CLIP was performed essentially as described (Bauer et al., 2008). For details see supplementary material.

4.5.5 In silico analysis of high-throughput sequencing data

We extensively used custom Perl (version 5.8.1) and R (version 2.15.1) scripts as well as the BEDTools (version 2.16.1) package (Quinlan and Hall, 2010) for many steps of the analysis. Details are indicated in supplementary material.

4.6 Acknowledgements

We thank Prof. Ger Pruijn for antibodies against hRRP41 and hRRP40, Prof. Torben Jensen for HEK293T-Rex cell lines expressing FLAG-DIS3 and FLAG-DIS3L, Prof. Narry Kim for pri-miRNA DNA constructs and Dr. Petr Svoboda for DICER1 antibodies. We are grateful to Tomas Masek, Martin Pospisek and Leos Valasek for invaluable help with the polysomal profiling analyses, Svitlana Kurinna for technical advice and Leona Svajdova for excellent technical support. This work was supported by the Wellcome Trust [084316/Z/07/Z to S.V.], EMBO Installation Grant [1642 to S.V.], Czech Science Foundation [305/11/1095 to S.V, P305/12/G034 to S.V.], and CEITEC - Central European Institute of Technology [CZ.1.05/1.1.00/02.0068] from European Regional Development Fund. B.B. was supported by a Werner-Siemens fellowship.

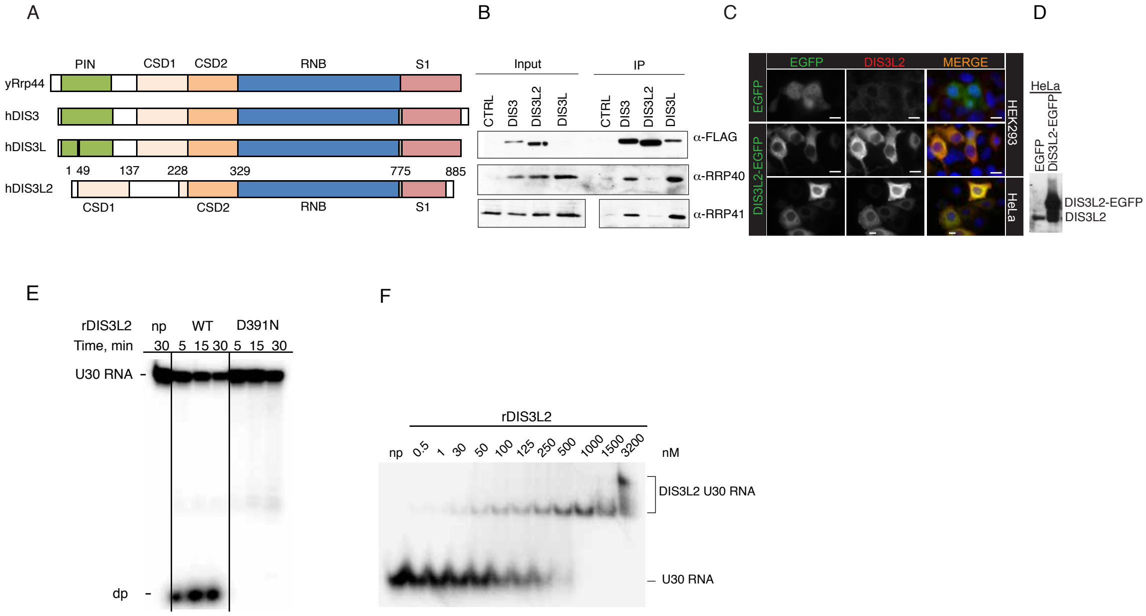


Figure 1. DIS3L2 is a cytoplasmic 3' to 5' exoribonuclease that does not associate with the exosome. A. Schematic representation of the domain organization of DIS3L2 homologues from *S. cerevisiae* and *H. sapiens*. The PIN domain is in green, the CSD1 and CSD2 RNA binding domains are in orange, the RNB ribonuclease domain in blue, and the S1 domain in pink. B. DIS3L2 does not interact with core exosome components. FLAG-tagged DIS3, DIS3L and DIS3L2 were immunoprecipitated from stable cell lines with inducible overexpression of the individual proteins. The composition of IP samples was analyzed by western blot analysis with antibodies indicated. C. DIS3L2 is a cytoplasmic protein. DIS3L2 C-terminally fused to an EGFP tag was transiently expressed in HeLa and HEK293T cells, respectively. Plasmid expressing EGFP alone was used for background control. The scale bar corresponding to 10 microM is shown in white. D. Relative abundance of endogenous and EGFP-tagged DIS3L2 in HeLa cells expressing either only EGFP or EGFP-tagged DIS3L2 were detected with antibodies specific to DIS3L2. E. *In vitro* degradation assays with recombinant DIS3L2 and its catalytical mutant (D391N) with U30 RNA as a substrate. The migration of input RNA and degradation products are indicated. F. DIS3L2 binds U30 RNA with nanomolar affinity. Electromobility shift assay with recombinant DIS3L2 (concentrations indicated on top) and 5' end 32P labeled U30 RNA. Migration of free RNA and of protein-RNA complexes, respectively are indicated.

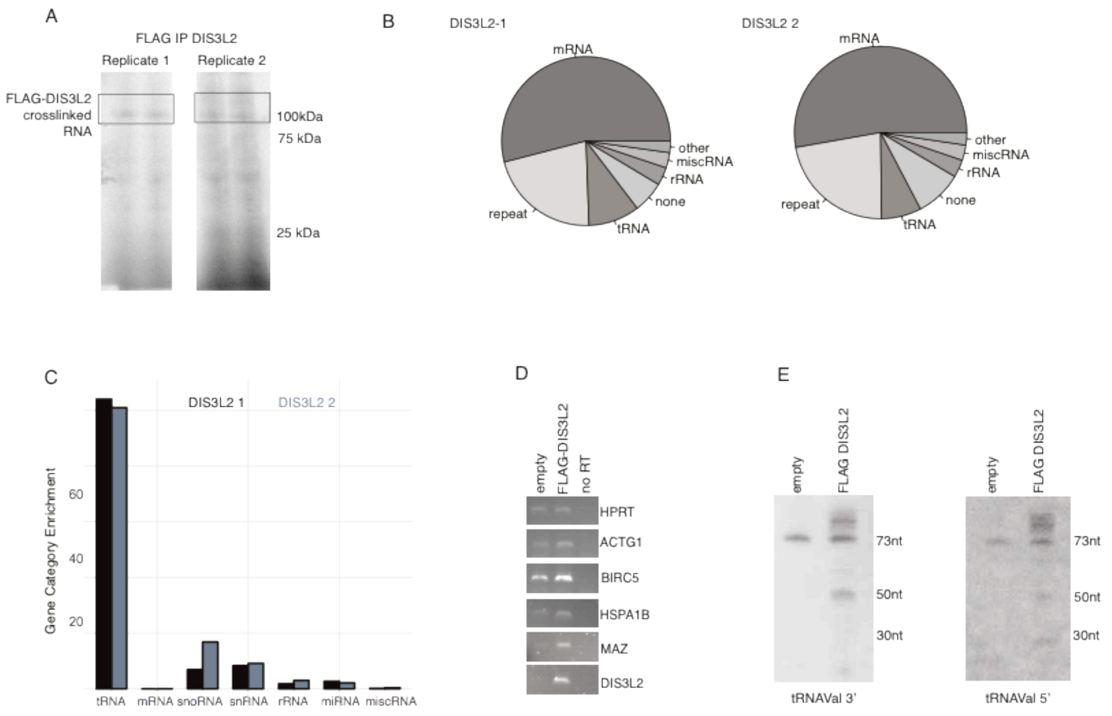


Figure 2. tRNAs are the main *in vivo* substrates of DIS3L2. A. Autoradiography of y- P-labeled RNAs crosslinked to DIS3L2. Each CLIP was split to two halves and loaded into two lanes. The squares designate area that was cut from the gel and used for subsequent analyses. B. Functional annotation of reads uniquely mapping to the genome in two DIS3L2 PAR-CLIP replicates. Categories that accumulate less than 1% of the reads are not represented individually, but are grouped in the "other" category and reads mapping to genomic regions without a functional annotation are represented as none category. C. tRNA genes are enriched among the 1,000 most significantly crosslinked positions. Categories shown on the x-axis are in descending order of the number of distinct genes represented in the set of the 1,000 most significantly crosslinked positions. Y-axis is p-value (-log10) of Fisher's Exact test. D. RNA immunoprecipitation with FLAG-DIS3L2 followed by oligo(Landgraf et al., 2007) cDNA synthesis and semi-quantitative PCR with gene specific primers. Empty HEK 293T-Rex cells (empty) were used as a negative control for unspecific coimmunoprecipitation of RNAs on FLAG matrix. No RT is a control where no reverse transcriptase was used. E. Northern blot analysis of tRNA ValAAC immunoprecipitated with FLAG-DIS3L2. RNAs obtained by RIP as in C were separated on 10 % denaturing polyacrylamide gel, transferred to nylon membrane and hybridized with 5'-end labeled DNA oligos corresponding to 3' and 5' part of *Homo sapiens* chr6.trna136-ValAAC (27,648,885-27,648,957), respectively.

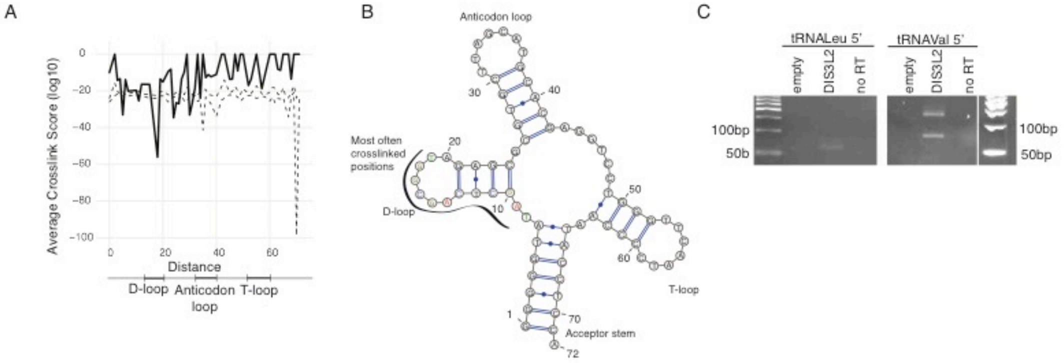


Figure 3. DIS3L2 preferentially binds 5' halves of tRNAs. A. Profile of average crosslink score (log of the probability that crosslink-diagnostic T-to-C mutations are more frequent than expected by chance, see Methods) along the 72-nucleotide long tRNAs that were crosslinked in the two replicate samples. The empirical 90% confidence intervals of the average crosslink scores obtained by shuffling the crosslinked positions along the T positions of the crosslinked tRNA species 2,000 times is shown with dashed lines. The crosslink position on specific regions of tRNAs are indicated below the x-axis. B. Secondary structure of tRNA AGC as a representative tRNA species of length 72 (predicted by tRNAscan-SE (Lowe and Eddy, 1997) and rendered by VARNA (Darty et al., 2009)). C. 5' fragments of tRNALys and tRNAVal are specifically bound by DIS3L2. RIP-RT-PCR analysis of DIS3L2 associated RNAs. The PCR product corresponding to 5' tRFs migrates as 60 - 70 bp band. The band migrating as 150 bp fragment corresponds to tRNAVal. D. Profile of read lengths of tRNA and miRNA annotated tags from the DIS3L2, AGO2, and small RNA-seq libraries. The left panel shows read length of miRNAs, where the number of reads at ~22 is highest for AGO2, sRNA and weakly also for DIS3L2. The right panel represents read length of tRNAs.

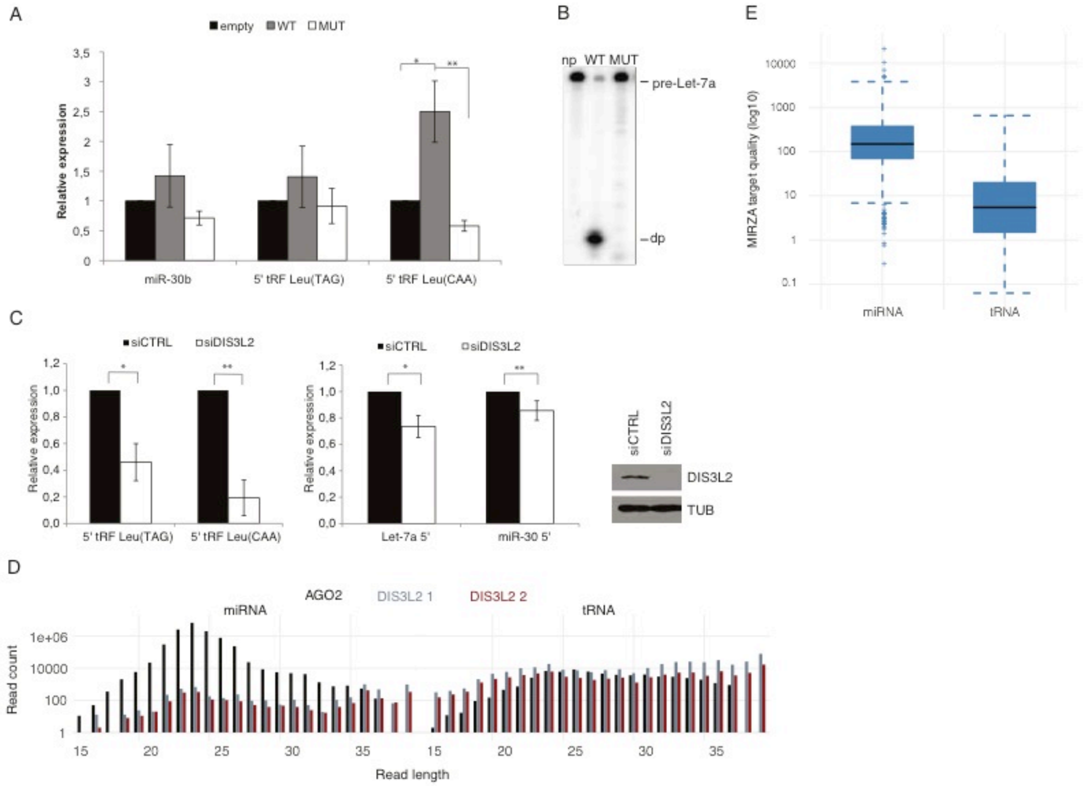


Figure 4. DIS3L2 generates 5'-derived tRNA fragments (tRFs) that are bound by AGO2 and may have miRNA-like function. A. DIS3L2 overexpression increases 5' tRFs levels. RT-PCR of mature miRNA30b, 5' tRFLeu TAG and 5' tRFLeu CAA in HEK293T-Rex cell lines with or without (empty) overexpression of wild type (WT) and D391N (MUT) FLAG-DIS3L2. Results are derived from three independent biological replicates, *) p-value = 0.03, n=3. **) p-value < 0.015, n=3. B. FLAG-D391N DIS3L2 mutant shows no exonucleolytic activity *in vitro*. Wild type (WT) and D391N (MUT) were immunoprecipitated from HEK 293T-Rex cells and incubated with 5'-end labeled pre-let-7a RNA for 30 minutes. np is a reaction where no protein was added. Reactions were stopped with formamide buffer and RNAs were resolved on 20% denaturing polyacrylamide gel, dp are degradation products. C. DIS3L2 downregulation decreases 5' levels of tRFLeu TAG and tRFLeu CAA in HeLa cells. Q-PCR of total RNA purified from control and siRNA treated cells. Results are derived from three independent biological replicates, *) p-value = 0.021, n=3. **) p-value < 0.009, n=3. In A and C the levels of tRFs and miRNAs were normalized to the expression of RNU44 snoRNA. P-values estimated by Student's paired t-test, number of replicates is indicated. Error bars represent standard deviation. The level of DIS3L2 knockdown (right panel) monitored with DIS3L2 specific antibodies. Tubulin is a loading control. D. Some hybrids predicted to form between tRNA-derived small RNAs and (AGO2-CLIPed) mRNA sites have scores that are comparable to those of miRNA-mRNA hybrids. Black lines indicate the median, the box edges the 25th and 75th percentiles, and plus symbols the outliers.

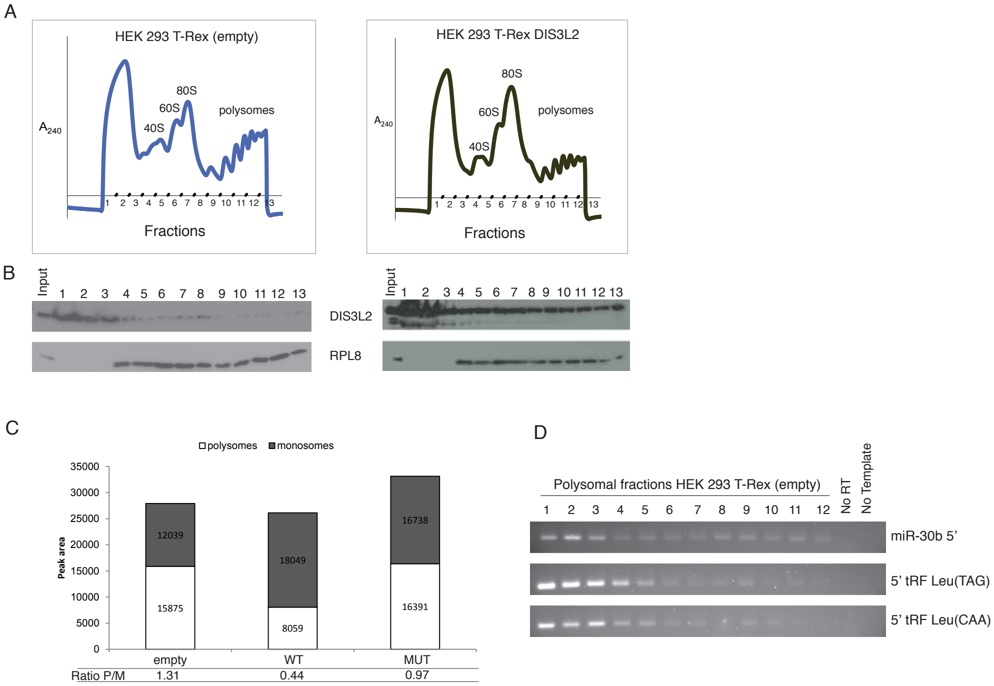


Figure 5. DIS3L2 and DIS3L2-derived 5' tRFs associate with translating ribosomes. A. Polysomal profiles of sucrose gradients from HEK293T-Rex cells. A polysome-enriched fraction of HEK293T was centrifuged through a linear 10-50% (wt/vol) sucrose gradient. The top of the gradient is on the left. The peaks corresponding to particular ribosomal fractions are indicated above. B. A fraction of DIS3L2 copurifies with HEK293T ribosomes on sucrose gradients. The distribution of DIS3L2 and ribosomal protein L8 (RPL8) in the fractions of the gradient were analyzed by immunoblotting using anti-DIS3L2 and anti-RPL8 antibodies, respectively. C. Ratio of the polysomes versus monosomes extracted from the graphs shown above (panel A). D. 5' tRFs and miR-30 associate with ribosomal fractions. PCR analysis of cDNAs prepared from RNAs associated with individual fractions from the sucrose gradient shown in (A). The RT-PCR protocol is designed to amplify 5' tRFs, not full length tRNAs (see methods). No RT is a control where no reverse transcriptase was used.

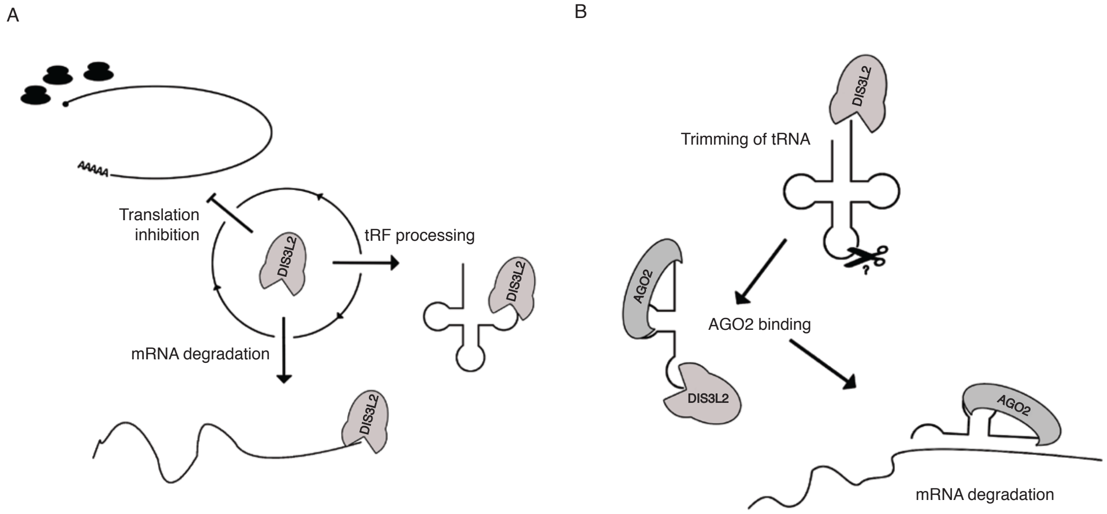


Figure 6. Proposed model of DIS3L2 and 5' tRF processing and function. A. Different roles of DIS3L2 in mRNA degradation, tRF production and translationalcontrol. B. Proposed mechanism of 5' tRF processing.

4.7 Supplementary material

4.7.1 Dis3L2 constructs

Constructs for bacterial expression: The coding sequence of human DIS3L2 protein (isoform 1, 885 aminoacids, NP_689596.4) was cloned into pET28b between NheI and BamHI sites allowing expression of N-terminaly fused 6xHis-Smt3 tag. DIS3L2 was amplified by PCR in two parts (because of XhoI internal cleavage site) using cDNA prepared from HEK293 RNA as a template. First, the 3′ part of CDS (1664 – 2658 nt) was amplified with primers (Forward 5′ AGCAGCGAGGAGGTACACCAG3′, Reverse 5′

ACCTCGAGTCAGCTGGTGCTTGAGTCCTCG 3′) and subcloned into pET28b using HindIII and XhoI sites, than the 5′ part (1 – 1663 nt) was amplified with primers (Forward 5′ CGGGATCCATGAGCCATCCTGACTACAG 3′, Reverse 5′ ATGACATCCTTGAGGCAATCC 3′) and ligated to the 3′ end via BamHI and HindIII sites. The sequence of the final clone was verified by sequencing.

Catalytical mutants D391 was designed based on sequence alignments with RNase II in which, D391 position in DIS3L2 corresponds to the D209 residue of E. coli RNase II that is critical for nuclease activity. D209N mutation within RNB motif allows RNA binding but prevents cleavage (Amblar and Arraiano, 2005). The mutation was introduced by site-directed mutagenesis using oligos: forward DIS3L2 D391N F, reverse DIS3L2 D391N Rv.

For the expression of the protein in human cells, coding sequence of DIS3L2 was cloned in pcDNA5/FRT and pcDNA/FRT/TO. FLAG DIS3L2 was prepared by amplification of DIS3L2 coding sequence using DIS3L2 pcDNA/FRT Not1 Rv and DIS3L2pcDNA/FRT Kpn1 Fw. pcDNA5/FRT/TO DIS3l2 was created by endonucleolytic cleavage of DIS3L2 coding sequence from pcDNA5/FRT FLAG DIS3L2 by KpnI and NotI endonucleases and ligating them to the pcDNA4 FRT/TO vector.

For immunofluorescence, the coding sequence of DIS3L2 was subcloned into pEGFP-N1 (GenBank Accession #U55762) and pEGFP-C3 vectors (GenBank Accession #: U57607) to obtain N-terminal and C-terminal EGFP fused tags, respectively.

4.7.2 Preparation of specific DIS3L2 antibodies

Part of coding sequence of DIS3L2 corresponding to aminoacids 135 – 298 was cloned into pET28b vector containing N-terminal fusion 6xHis-SMT3 tag (between NheI and BamHI sites) using primers (forward 5′ CGGGATCCGCTGCGTATGAATCAGATATC 3′, reverse 5′ TGAAGCTTTTAGGCATAATCTTTAGGCCGTGC 3′) standard cloning techniques. Protein fragment was expressed in E. coli BL21-CodonPlus (DE3)-RIPL strain; cells were lysed in 100 mM dihydrogen sodium phosphate, 10 mM Tris-HCl and 8 M urea, pH 8. Protein was purified by metal affinity chromatography on NiNTA beads (Qiagen) and released by low pH. Fractions containing eluted protein were collected together and dialysed for 72 hrs to 50 mM Tris, 300 mM NaCl, 5% glycerol, 0.02 % NP-40 and 1 mM β-mercaptoethanol. SMT3 tag was removed by Ulp1p cleavage at room temperature for 6 hrs. After concentration by VIVASPIN 6 MWCO10 000 (Sartorius Stedim Biotech S.A.), 3 mg of protein were separated from SMT3 tag on 18% SDS-PAGE, coomassie stained band corresponding to fragment of Dis3L2 protein was cut out from the gel and sent to Moravian-Biotechnology Ltd for rabbit immunization.

4.7.3 Expression and purification of recombinant wild type and mutant proteins

Recombinant DIS3L2 was expressed and purified from BL21-DE3 RIPL strain of E. coli. Bacterial cells were grown in at 37 °C, protein expression was induced at O.D. 0.7 with 0.5 M IPTG at 27 °C for 2 hours. Cells were harvested by centrifugation and lysed by sonication in buffer containing 50 mM Tris pH 7.9, 500 mM NaCl, 10% glycerol, 2 mM 2-mercaptoethanol, and 0.1 % NP-40. Lysate was cleared by centrifugation (14000 rpm, 30 minutes, 4 °C). Protein was purified by Ni-NTA chromatography. SMT3 tag was removed by proteolysis with Ulp1 protease. Resulting recombinant DIS3L2 (rDIS3L2) was further purified by gel filtration on Superdex 200 column (GE Healthcare) in buffer containing 20 mM Tris pH 7.9, 300 mM NaCl, 10% glycerol, 0.01 NP-40, 2 mM 2-mercaptoethanol, 10 mM imidazole.

4.7.4 In vitro degradation assay

In vitro degradation assays were performed in 10 ul reaction volumes containing 10 mM Tris pH 8.0, 50 mM KCl, 5 mM MgCl2, 10mM DTT (modified from (Staals et al., 2010, Tomecki et al., 2010, Lorentzen et al., 2008). Typically, 150 nM of purified protein and 20 pmol of 5′-end labeled RNA substrate, were incubated at 37 °C for the times indicated. Reactions were terminated with 1 volume of formamide loading buffer (80% formamide, 0.1% bromphenol blue, 0.1% xylene cyanol, 5mM EDTA). Reactions were resolved on denaturing 20% polyacrylamide gels containing 8 M urea. The radioactivity was exposed to phosphorimaging screen (FUJI) and scanned by phosphorimager FLA-9000 (FUJIFILM).

4.7.5 RNA binding assay

The binding reactions were performed in 10 μl volume. Recombinant DIS3L2 was incubated with 5′-end labeled U30 RNA for 20 minutes at room temperature in binding buffer (50 mM KCl, 10 mM Tris pH 8.0, 10 mM DTT, 10% glycerol, 0.1 μg BSA). Proteins and RNAs were resolved on native 8% polyacrylamide under 100V for approximately 8 hours. Visualized by exposing to phosphorimaging screen (Fuji) and scanned by FLA-9000 (FUJIFILM).

4.7.6 Northern blot analysis

Total RNA was resolved on 10% denaturing polyacrylamide gel and transferred to Hybond-N+ membrane (GE Healthcare) by electroblotting (BioRad). The hybridization with radioactively labeled oligonucleotides was performed in ULTRAhyb-oligo hybridizatioin buffer (Ambion) at 38 °C. Prior adding the labeled probe membrane was prehybridized at 42°C for 2 hours. The radioactive signal was monitored by phosphorimager FLA-9000 (FUJIFILM). Quantification of signals was done using Multi Gauge software v3.2 (FUJIFILM).

4.7.7 Purification of FLAG-DIS3L2 from human cells

Flag-DIS3L2 was purified from stable HEK293 FlipIn cell line expressing the fusion protein. Cells were resuspended in 4 ml of ice cold lysis buffer (50 mM Tris pH 8.0, 150 mM NaCl, 0.5% Triton X100 and Complete Protease Inhibitor Cocktail (Roche) and incubated rocking at 4 °C for 15 min. Lysate was cleared by centrifugation (14000 rpm, 30 minutes, 4 °C). For purification of FLAG DIS3L2, 100 μl of anti-FLAG M2 beads (Sigma-Aldrich) washed with lysis buffer were incubated with cell extract for 1 hour in a cold room rotating. Beads were extensively washed with 10 volumes of wash buffer (50 mM Tris pH 8.0, 300 mM NaCl, 0.1% Triton X100). Protein elution was done with 1 volume of 3x FLAG peptide (Sigma-Aldrich) resuspended in lysis buffer or by boiling with SDS loading buffer for 5 min.

4.7.8 PAR-CLIP

Human embryonic kidney cells (HEK293 FlpIn, Invitrogen) stably expressing human N-terminal FLAG-DIS3L2 fusion protein were grown overnight in a medium supplemented with 100 mM 4-thiouracyl (4-SU). Cells were washed with 1xPBS and exposed to 150 mJ of 365 nm UV light in a Stratalinker 2400 device (Stratagene). Cells were collected, frozen in liquid nitrogen and stored at -80°C. PAR-CLIP was performed essentially as described (Martin et al., 2012). Whole cell lysate was prepared in PNDS buffer (1x PBS, 0.5 NP-40, 0.25% deoxycholate, 0.05% SDS supplemented with 1 mM DTT, protease inhibitor cocktail (Roche) and RNase inhibitor RNAsin (Promega). Lysates were cleared by centrifugation and split in two aliquots. One aliquot was treated with 4 Units/ml and the other aliquot with 12 Units/ml of RNase I (Ambion, AM2294) for 10 min at 32°C and pooled on ice. FLAG-tagged DIS3L2 was immunoprecipitated using anti-FLAG M2 monoclonal antibody (Sigma) bound to Protein G Dynabeads (Invitrogen). Beads were treated with alkaline phosphatase (Fast-AP, Fermentas). 3′ adaptor (P-UCGUAUGCCGUCUUCUGCUUGU-Pur) was ligated to the bound RNA with T4 RNA ligase (Fermentas) in buffer containing 25% PEG 8000 at 16°C over night. The crosslinked RNAs were radiolabeled with polynucleotide kinase (T4 PNK, NEB) and g-32P ATP. Protein-RNA complexes were resolved on a 4-12% gradient SDS-PAGE (NuPAGE, Invitrogen), the band corresponding to tagged DIS3L2 was cut out from the gel and eluted with proteinase K containing elution buffer (50mM Tris pH 7.5, 50 mM NaCl, 10 mM EDTA, 2 M urea, 2 mg/ml proteinase K) at 50°C for 2 hrs. RNAs were then ligated to 5′ adaptor (OH-GUUCAGAGUUCUACAGUCCGACGAUC-OH). RNA was size fractionated on 8% polyacrylamide- 8 M urea gel and 70-110 nt long RNA fragmets were eluted. Reverse transcription was done with a 3′ primer (5′ CAAGCAGAAGACGGCATACGA 3′) and with Superscript III reverse transcriptase (Invitrogen). Resulting cDNAs were used as templates for PCR amplification (5′ Primer: 5′ AATGATACGGCGAC-CACCGACAGGTTCAG-AGTTCTACAGTCCGA 3′ and the lowest possible number of PCR cycles were used. The PCR products were sequenced on an Illumina Genome Analyzer IIx.

4.7.9 Growth analysis

Control HEK293T-Rex cells and cells overexpressing WT and D391N form of DIS3L2 were seeded into 24-well format plates at seeding density 0.5 x 105/ml per well. At times indicated, cells were collected, stained with 0.2% trypan blue solution (Sigma) and counted with LUNATM automated cell counter (Logos Biosystems) according to manufacturer instructions. Only live cells were included in the analysis. The resulting growth curves represent the average result of at least tree independent measurements and error bars indicate standard deviations.

4.7.10 siRNA-mediated knock down

The day before transfections 1,7 x 105 cells were seeded. siRNA were transfected using INTERFERin transfecting reagent (Polyplus transfections) following the manufacturer instructions. Briefly, 10 – 30 nM of siRNAs were taken, polysomes formed by addition the transfecting reagent (8 – 15 μl). After 24 hours cells were re-transfected and collected for further experiments the next day. siRNA used in this study were obtained: siDIS3L2 (Dharmacon), Non-Targeting Control (Sigma).

4.7.11 Immunofluorescence and image processing

Cells expressing GFP-fusion proteins were plated in dishes with coverslips coated with 0.2% gelatin. Paraformaldehyde fixed cells were permeabilized with 0.2% Triton-X100 in PBS in the presence of DAPI to stain the nucleus. Coverslips with FlouroMount reagent (Invitrogen) were mount to glass slides and fluorescent images were captured with a Leica DM 6000 B microscope. To visualize DAPI and GFP, 405 nm diode and Argon 488nm lasers were applied, respectively.

Transfected HeLa cells were fixed with 3.7% PFA in PBS for 30 min, permeabilized with 0.2% Triton X-100 in PBS for 30 min and blocked with 5% horse serum in PBS for 1 h. All steps were performed at RT. Samples were incubated with primary antibodies (rabbit anti-DIS3L2) diluted 1:200 in the blocking buffer overnight at 4°C. The next day, fixed cells were washed three times with 0.2% Triton X-100 in PBS (5 min) and incubated in the dark with the secondary antibody for 40 min at RT. Cy2-, Cy3-, or Cy5-conjugated donkey secondary antibodies (Jackson ImmunoResearch Laboratories, West Grove, PA) were used at dilutions of 1:200, 1:2,000, or 1:200 with

2.5% horse serum in PBS buffer, respectively. The cells were mounted in FluoroMount reagent (SouthernBiotech, Birmingham, AL) and pictures were captured with a confocal microscope Leica TSC SP5.

4.7.12 H_2O_2 treatment

Cells were treated with a standardized doze of hydrogen peroxide (5 mM) for times indicated and collected for the RNA isolation.

4.7.13 Sucrose gradient fractionation for polysomal profilling

Polysomes were isolated from one 15 cm dish of the cell lines HEK293T-Rex empty control, HEK293T-Rex Flag-DIS3L2 and HEK293T-Rex Flag-DIS3L2 D391N. Cells were cultured in DMEM supplemented with 10% fetal bovine serum without antibiotics. When indicated overproduction of both wt and mutant forms of the DIS3L2 protein was achieved by adding doxycycline (100 ng/ml) directly to media. Actively growing cells (confluence around 85-90%) were treated with 10 µg/ml cycloheximide for 5 minutes at 37°C, all subsequent steps were performed in a cold room. Cells were washed twice with cold PBS supplemented with 10 µg/ml cycloheximide and then lysed on plate using PB (100 mM KCl, 5 mM MgCl2, 10 mM HEPES pH 7, 10 µg/ml cycloheximide, 0.5 % NP-40, 1 mM DTT, 1 mM PMSF, 200 µg/ml Heparin, 100 U/ml RNAsin Plus, 2 mM vanadyl ribonucleoside complex, protease and phosphatase inhibitors). Lysates were incubated on ice for 15 min. and cleared by centrifugation at 15 000 g for 15 min. Supernatants (24 Units) were separated in 5-45% sucrose gradients by ultracentrifugation (35 000 rpm, 105 min. in Beckman SW41 rotor at 4°C). Polysome profiles were fractionated (1 ml fractions) using an ISCO UA-5 gradient analyzer connected to a Clarity data acquisition station (DataApex). Proteins from individual fractions were ethanol precipitated and separated on 9% SDS-PAGE gels and proteins were transferred and immobilized onto PVDF membranes. Proteins were detected by western blot analyses with antibodies indicated (α-DIS3L2 rabbit polyclonal serum 1:800, and α-RPL8 (Abcam) goat polyclonal serum 1:500, Abcam). Both α-goat (Sigma) and α-rabbit (Promega) HRP secondary antibodies were diluted 1:5000.

The polysome to monosome ratios were acquired using the Image SXM 193 software (NIH).

4.7.14 RNA immunoprecipitation

Empty HEK293T-REX cells and HEK293T-REX FLAG DIS3L2 were grown to 80% confluence. Washed with ice cold PBS and UV cross-linked (400mJ, 254nm). Cells lysed in buffer containing 150mM NaCl, 50mM Tris pH 7.6, 0.5% Triton X 100, supplemented with protease inhibitors (EDTA-free Complete Protease Inhibitor Cocktail, Roche), 0.5mM EDTA, 1mM DTT, RNase In (Promega). Lysate cleared by centrifugation. Supernatant applied on FLAG M2 Magnetic beads (Sigma) and incubated for 60 min. Washed two times with Lysis buffer, two times with Lysis buffer containing 300mM NaCl. RNA was eluted by treating whole beads with 2mg/ml Protease K (New England Biolabs) for 120 min at 37C. RNA was phenol/chlorophorm cleaned and ethanol precipitated. After DNase treatment (Turbo DNase, Fermentas) equal amount of RNA was taken for the cDNA synthesis by Superscript III (Invitrogen). Obtained cDNA was used for semi quantitative PCR. Reaction resolved on 2% agarose gel.

4.7.15 RNA isolation, cDNA synthesis and quantitative PCR

Total RNA was isolated with TRIzol (Invitrogen) according to manufacturer instructions followed by RNase-free DNase (TURBO DNase, Fermentas) treatment. The RNA concentration was measured in a Beckman Coulter DU 730. 1 ug of purified RNA was reverse transcribed using random hexamers or oligo dT (as indicated) and

Superscript III reverse transcriptase (Invitrogen) according to manufacturer instructions. Real-time quantitative PCR (RT-qPCR) was performed using FastStart Universal SYBR Green Master (Roche) and gene-specific primer pairs on Q-PCR Light Cycler 7500 (Applied Biosystems). Each experiment was performed in triplicate. Transcript abundance was calculated by the Delta Ct method. Data were normalized to an internal control of the housekeeping gene HPRT mRNA or RNU44 for small RNA detection. Results are expressed as means and standard errors of the mean. Statistical analyses were performed with a t test; P values < 0.05 were considered significant.

The RNA for in vitro tests were either prepared by in vitro transcription (tRNA Ala) from DNA templates containing T7 promotor and T7 RNA polymerase or purchased as synthetic molecules from Dharmacon or Sigma. For the degradation assays and EMSA, the RNAs were radioactively labeled at the tRNA 5′ end by T4 polynucleotide kinase (New England Biolabs) and 32P gamma-ATP (company, MGP-Zlin).

4.7.15 Read-to-genome mapping

The high-throughput sequencing resulted in an average of 4 x 107 reads per sample. After filtering out reads with nucleotides that could not be unambiguously called, we mapped the reads to the human genome 19 assembly downloaded from the University of California at Santa Cruz (hg19) on the CLIPZ server (www.clipz.unibas.ch and (Khorshid et al., 2011). The copy numbers of reads that mapped to multiple locations (multi-mappers) were distributed equally among the possible mapping locations. Reads that did not map in the first step were processed to remove the intact and truncated CCAs at the 3' end and then remapped with the CLIPZ server. We merged the results of these two mapping steps and calculated the genome-wide crosslink scores as described in (Jaskiewicz et al., 2012). In the calculation of the crosslink score we used only uniquely mapping reads that were not annotated by the CLIPZ server as being of bacterial, fungal, vector, marker/adaptor or viral origin.

4.7.16 Annotation of the genomic crosslinked sites

We used the RefSeq, snomir, and tRNA gene tracks (Dreszer et al., 2011) and the wgEncodeGencodeBasicV12 tracks for snRNA and rRNA genes (Rosenbloom et al., 2012) from University of California at Santa Cruz (http://genome.cse.ucsc.edu). We further annotated the RefSeq genes with their biotype information and used one representative transcript from long-noncoding RNAs or mRNAs. The selection of the representative transcripts of single-locus genes were done based on a hierarchy of several criteria ordered by biotype (mRNA > non-coding RNA), RefSeq gene status (Reviewed > Validated > Inferred > Provisional > Predicted), length of genomic locus, and length of the exonic region with a higher precedence for the longer loci and transcripts.

4.7.17 Estimation of expressed RefSeq mRNA genes and long non-coding RNA genes

We used the CLIPZ server (Khorshid et al., 2011) to map the reads from two RNA-seq libraries prepared from HEK293 FlpIn cells overexpressing Flag-tagged DIS3L2 to the human genome (assembly version hg19) and processed the expression levels of the representative RefSeq transcripts to which at least one read longer than 25 nt and annotated as either “mRNA” or without a know annotation mapped. We performed the expression estimation by fitting a two-component Gaussian Mixture model with unequal variance to the per-nucleotide read densities of the transcripts with the mclust R package (Fraley and Raftery, 2002). We used the union of genes belonging to the higher expression component of the mixture in the two libraries as the set of “expressed genes” and for the annotation of genomic crosslinked positions.

4.7.18 Gene set enrichment analysis

We used the Ontologizer (version 2.0) tool for calculating the enriched Gene Ontology (GO) terms considering the expressed RefSeq genes as the total population of genes with the following additional parameters: -n -c Parent-Child-Union -m Westfall-Young-Single-Step -d 0.1 -r 1000.

4.7.19 Identification of mRNA targets of tRNA derived fragments

To define the (tRNA/miRNA derived) guide sequences, we used uniquely mapping reads in AGO2 PAR-CLIP library that were longer than 20 nt. We identified those reads that mapped to tRNA/miRNA loci that contained one of the top 1,000 most significantly crosslinked genomic positions in the AGO2 CLIP library. We grouped these reads by their genomic start positions and summed up the read counts in each group. We then isolated 21-nucleotide-long regions from the start of each group that contained at least 100 reads and defined these as potential guide sequences. For each miRNA, we only used one guide, the one with the highest read count. For tRNAs, we used those putative guides that were at least over ½ of their length located in the 5' half of the tRNA. As potential target sequences, we took the top 3,000 mRNA-annotated, 51-nucleotides-long, crosslinked-centered regions. Finally, we used the MIRZA model for calculating the energy of hybrid structures formed between putative guide RNAs and target mRNAs (Khorshid et al., 2013) The MIRZA model is similar to those typically used for predicting RNA secondary structure and RNA-RNA interactions, but has an additional set of parameters corresponding to energy contributions of individual positions in the miRNA to the energy of miRNA-target interaction. Parameters of the model were inferred in Khorshid et al. (Khorshid et al., 2013) from the Ago2-CLIP site data of Kishore et al. (Kishore et al., 2011).

4.8 References

AMBLAR, M. & ARRAIANO, C. M. 2005. A single mutation in Escherichia coli ribonuclease II inactivates the enzyme without affecting RNA binding. *FEBS J,* 272**,** 363-74.

ASTUTI, D., MORRIS, M. R., COOPER, W. N., STAALS, R. H., WAKE, N. C., FEWS, G. A., GILL, H., GENTLE, D., SHUIB, S., RICKETTS, C. J., COLE, T., VAN ESSEN, A. J., VAN LINGEN, R. A., NERI, G., OPITZ, J. M., RUMP, P., STOLTE-DIJKSTRA, I., MULLER, F., PRUIJN, G. J., LATIF, F. & MAHER, E. R. 2012. Germline mutations in DIS3L2 cause the Perlman syndrome of overgrowth and Wilms tumor susceptibility. *Nat Genet,* 44**,** 277-84.

BABIARZ, J. E. & BLELLOCH, R. 2008. Small RNAs - their biogenesis, regulation and function in embryonic stem cells. *StemBook*. Cambridge (MA).

BACHAND, F. & SILVER, P. A. 2004. PRMT3 is a ribosomal protein methyltransferase that affects the cellular levels of ribosomal subunits. *EMBO J,* 23**,** 2641-50.

BAUER, S., GROSSMANN, S., VINGRON, M. & ROBINSON, P. N. 2008. Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics,* 24**,** 1650-1.

BUHLER, M., SPIES, N., BARTEL, D. P. & MOAZED, D. 2008. TRAMP-mediated RNA surveillance prevents spurious entry of RNAs into the Schizosaccharomyces pombe siRNA pathway. *Nat Struct Mol Biol,* 15**,** 1015-23.

CANELLA, D., PRAZ, V., REINA, J. H., COUSIN, P. & HERNANDEZ, N. 2010. Defining the RNA polymerase III transcriptome: Genome-wide localization of the RNA polymerase III transcription machinery in human cells. *Genome Res,* 20**,** 710-21.

CHAN, C. T., PANG, Y. L., DENG, W., BABU, I. R., DYAVAIAH, M., BEGLEY, T. J. & DEDON, P. C. 2012. Reprogramming of tRNA modifications controls the oxidative stress response by codon-biased translation of proteins. *Nat Commun,* 3**,** 937.

CHEN, C., RIDZON, D. A., BROOMER, A. J., ZHOU, Z., LEE, D. H., NGUYEN, J. T., BARBISIN, M., XU, N. L., MAHUVAKAR, V. R., ANDERSEN, M. R., LAO, K. Q., LIVAK, K. J. & GUEGLER, K. J. 2005. Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Res,* 33**,** e179.

COLE, C., SOBALA, A., LU, C., THATCHER, S. R., BOWMAN, A., BROWN, J. W., GREEN, P. J., BARTON, G. J. & HUTVAGNER, G. 2009. Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA,* 15**,** 2147-60.

COUVILLION, M. T., BOUNOVA, G., PURDOM, E., SPEED, T. P. & COLLINS, K. 2012. A Tetrahymena Piwi bound to mature tRNA 3' fragments activates the exonuclease Xrn2 for RNA processing in the nucleus. *Mol Cell,* 48**,** 509-20.

CZECH, B. & HANNON, G. J. 2011. Small RNA sorting: matchmaking for Argonautes. *Nat Rev Genet,* 12**,** 19-31.

DAMGAARD, C. K. & LYKKE-ANDERSEN, J. 2011. Translational coregulation of 5'TOP mRNAs by TIA-1 and TIAR. *Genes Dev,* 25**,** 2057-68.

DARTY, K., DENISE, A. & PONTY, Y. 2009. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics,* 25**,** 1974-5.

DRESZER, T. R., KAROLCHIK, D., ZWEIG, A. S., HINRICHS, A. S., RANEY, B. J., KUHN, R. M., MEYER, L. R., WONG, M., SLOAN, C. A., ROSENBLOOM, K. R., ROE, G., RHEAD, B., POHL, A., MALLADI, V. S., LI, C. H., LEARNED, K., KIRKUP, V., HSU, F., HARTE, R. A., GURUVADOO, L., GOLDMAN, M., GIARDINE, B. M., FUJITA, P. A., DIEKHANS, M., CLINE, M. S., CLAWSON, H., BARBER, G. P., HAUSSLER, D. & JAMES KENT, W. 2011. The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res,* 40**,** D918-23.

DZIEMBOWSKI, A., LORENTZEN, E., CONTI, E. & SERAPHIN, B. 2007. A single subunit, Dis3, is essentially responsible for yeast exosome core activity. *Nat Struct Mol Biol,* 14**,** 15-22.

EDGAR, R., DOMRACHEV, M. & LASH, A. E. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res,* 30**,** 207-10.

ELBARBARY, R. A., TAKAKU, H., UCHIUMI, N., TAMIYA, H., ABE, M., NISHIDA, H. & NASHIMOTO, M. 2009a. Human cytosolic tRNase ZL can downregulate gene expression through miRNA. *FEBS Lett,* 583**,** 3241-6.

ELBARBARY, R. A., TAKAKU, H., UCHIUMI, N., TAMIYA, H., ABE, M., TAKAHASHI, M., NISHIDA, H. & NASHIMOTO, M. 2009b. Modulation of gene expression by human cytosolic tRNase Z(L) through 5'-half-tRNA. *PLoS One,* 4**,** e5908.

EMARA, M. M., IVANOV, P., HICKMAN, T., DAWRA, N., TISDALE, S., KEDERSHA, N., HU, G. F. & ANDERSON, P. 2010. Angiogenin-induced tRNA-derived stress-induced RNAs promote stress-induced stress granule assembly. *J Biol Chem,* 285**,** 10959-68.

ENDER, C., KREK, A., FRIEDLANDER, M. R., BEITZINGER, M., WEINMANN, L., CHEN, W., PFEFFER, S., RAJEWSKY, N. & MEISTER, G. 2008. A human snoRNA with microRNA-like functions. *Mol Cell,* 32**,** 519-28.

FRALEY, C. & RAFTERY, A. E. 2002. Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.,* 97**,** 611–631.

FRAZAO, C., MCVEY, C. E., AMBLAR, M., BARBAS, A., VONRHEIN, C., ARRAIANO, C. M. & CARRONDO, M. A. 2006. Unravelling the dynamics of RNA degradation by ribonuclease II and its RNA-bound complex. *Nature,* 443**,** 110-4.

FU, H., FENG, J., LIU, Q., SUN, F., TIE, Y., ZHU, J., XING, R., SUN, Z. & ZHENG, X. 2009. Stress induces tRNA cleavage by angiogenin in mammalian cells. *FEBS Lett,* 583**,** 437-42.

HAUSSECKER, D., HUANG, Y., LAU, A., PARAMESWARAN, P., FIRE, A. Z. & KAY, M. A. 2010. Human tRNA-derived small RNAs in the global regulation of RNA silencing. *RNA,* 16**,** 673-95.

HEYER, R., DORR, M., JELLEN-RITTER, A., SPATH, B., BABSKI, J., JASCHINSKI, K., SOPPA, J. & MARCHFELDER, A. 2012. High throughput sequencing reveals a plethora of small RNAs including tRNA derived fragments in Haloferax volcanii. *RNA Biol,* 9**,** 1011-8.

HOUSELEY, J., LACAVA, J. & TOLLERVEY, D. 2006. RNA-quality control by the exosome. *Nat Rev Mol Cell Biol,* 7**,** 529-39.

HOUSELEY, J. & TOLLERVEY, D. 2009. The many pathways of RNA degradation. *Cell,* 136**,** 763-76.

IVANOV, P., EMARA, M. M., VILLEN, J., GYGI, S. P. & ANDERSON, P. 2011. Angiogenin-induced tRNA fragments inhibit translation initiation. *Mol Cell,* 43**,** 613-23.

JASKIEWICZ, L., BILEN, B., HAUSSER, J. & ZAVOLAN, M. 2012. Argonaute CLIP - A method to identify in vivo targets of miRNAs. *Methods*.

JOCHL, C., REDERSTORFF, M., HERTEL, J., STADLER, P. F., HOFACKER, I. L., SCHRETTL, M., HAAS, H. & HUTTENHOFER, A. 2008. Small ncRNA transcriptome analysis from Aspergillus fumigatus suggests a novel mechanism for regulation of protein synthesis. *Nucleic Acids Res,* 36**,** 2677-89.

KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002. The human genome browser at UCSC. *Genome Res,* 12**,** 996-1006.

KHORSHID, M., HAUSSER, J., ZAVOLAN, M. & VAN NIMWEGEN, E. 2013. A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nat Methods,* 10**,** 253-5.

KHORSHID, M., RODAK, C. & ZAVOLAN, M. 2011. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res,* 39**,** D245-52.

KISHORE, S., JASKIEWICZ, L., BURGER, L., HAUSSER, J., KHORSHID, M. & ZAVOLAN, M. 2011. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods,* 8**,** 559-64.

LANDGRAF, P., RUSU, M., SHERIDAN, R., SEWER, A., IOVINO, N., ARAVIN, A., PFEFFER, S., RICE, A., KAMPHORST, A. O., LANDTHALER, M., LIN, C., SOCCI, N. D., HERMIDA, L., FULCI, V., CHIARETTI, S., FOA, R., SCHLIWKA, J., FUCHS, U., NOVOSEL, A., MULLER, R. U., SCHERMER, B., BISSELS, U., INMAN, J., PHAN, Q., CHIEN, M., WEIR, D. B., CHOKSI, R., DE VITA, G., FREZZETTI, D., TROMPETER, H. I., HORNUNG, V., TENG, G., HARTMANN, G., PALKOVITS, M., DI LAURO, R., WERNET, P., MACINO, G., ROGLER, C. E., NAGLE, J. W., JU, J., PAPAVASILIOU, F. N., BENZING, T., LICHTER, P., TAM, W., BROWNSTEIN, M. J., BOSIO, A., BORKHARDT, A., RUSSO, J. J., SANDER, C., ZAVOLAN, M. & TUSCHL, T. 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell,* 129**,** 1401-14.

LEBRETON, A., TOMECKI, R., DZIEMBOWSKI, A. & SERAPHIN, B. 2008. Endonucleolytic RNA cleavage by a eukaryotic exosome. *Nature,* 456**,** 993-6.

LEE, R. C., FEINBAUM, R. L. & AMBROS, V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell,* 75**,** 843-54.

LEE, Y. S., SHIBATA, Y., MALHOTRA, A. & DUTTA, A. 2009. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev,* 23**,** 2639-49.

LI, Y., LUO, J., ZHOU, H., LIAO, J. Y., MA, L. M., CHEN, Y. Q. & QU, L. H. 2008. Stress-induced tRNA-derived RNAs: a novel class of small RNAs in the primitive eukaryote Giardia lamblia. *Nucleic Acids Res,* 36**,** 6048-55.

LI, Z., ENDER, C., MEISTER, G., MOORE, P. S., CHANG, Y. & JOHN, B. 2012. Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs. *Nucleic Acids Res,* 40**,** 6787-99.

LIU, Q., GREIMANN, J. C. & LIMA, C. D. 2006. Reconstitution, activities, and structure of the eukaryotic RNA exosome. *Cell,* 127**,** 1223-37.

LORENTZEN, E., BASQUIN, J., TOMECKI, R., DZIEMBOWSKI, A. & CONTI, E. 2008. Structure of the active subunit of the yeast exosome core, Rrp44: diverse modes of substrate recruitment in the RNase II nuclease family. *Mol Cell,* 29**,** 717-28.

LOWE, T. M. & EDDY, S. R. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res,* 25**,** 955-64.

MARTIN, G., GRUBER, A. R., KELLER, W. & 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**,** 753-63.

MATSUMOTO, K., AOKI, K., DOHMAE, N., TAKIO, K. & TSUJIMOTO, M. 2000. CIRP2, a major cytoplasmic RNA-binding protein in Xenopus oocytes. *Nucleic Acids Res,* 28**,** 4689-97.

MAYR, C. & BARTEL, D. P. 2009. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell,* 138**,** 673-84.

MITCHELL, P., PETFALSKI, E., SHEVCHENKO, A., MANN, M. & TOLLERVEY, D. 1997. The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'-->5' exoribonucleases. *Cell,* 91**,** 457-66.

NELSON, P. T., HATZIGEORGIOU, A. G. & MOURELATOS, Z. 2004. miRNP:mRNA association in polyribosomes in a human neuronal cell line. *RNA,* 10**,** 387-94.

NEUMANN, B., WALTER, T., HERICHE, J. K., BULKESCHER, J., ERFLE, H., CONRAD, C., ROGERS, P., POSER, I., HELD, M., LIEBEL, U., CETIN, C., SIECKMANN, F., PAU, G., KABBE, R., WUNSCHE, A., SATAGOPAM, V., SCHMITZ, M. H., CHAPUIS, C., GERLICH, D. W., SCHNEIDER, R., EILS, R., HUBER, W., PETERS, J. M., HYMAN, A. A., DURBIN, R., PEPPERKOK, R. & ELLENBERG, J. 2010. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature,* 464**,** 721-7.

PERLMAN, M. 1986. Perlman syndrome: familial renal dysplasia with Wilms tumor, fetal gigantism, and multiple congenital anomalies. *Am J Med Genet,* 25**,** 793-5.

QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26**,** 841-2.

ROSENBLOOM, K. R., DRESZER, T. R., LONG, J. C., MALLADI, V. S., SLOAN, C. A., RANEY, B. J., CLINE, M. S., KAROLCHIK, D., BARBER, G. P., CLAWSON, H., DIEKHANS, M., FUJITA, P. A., GOLDMAN, M., GRAVELL, R. C., HARTE, R. A., HINRICHS, A. S., KIRKUP, V. M., KUHN, R. M., LEARNED, K., MADDREN, M., MEYER, L. R., POHL, A., RHEAD, B., WONG, M. C., ZWEIG, A. S., HAUSSLER, D. & KENT, W. J. 2012. ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res,* 40**,** D912-7.

SAXENA, S. K., RYBAK, S. M., DAVEY, R. T., JR., YOULE, R. J. & ACKERMAN, E. J. 1992. Angiogenin is a cytotoxic, tRNA-specific ribonuclease in the RNase A superfamily. *J Biol Chem,* 267**,** 21982-6.

SCHNEIDER, C., ANDERSON, J. T. & TOLLERVEY, D. 2007. The exosome subunit Rrp44 plays a direct role in RNA substrate recognition. *Mol Cell,* 27**,** 324-31.

SCHNEIDER, C., LEUNG, E., BROWN, J. & TOLLERVEY, D. 2009. The N-terminal PIN domain of the exosome subunit Rrp44 harbors endonuclease activity and tethers Rrp44 to the yeast core exosome. *Nucleic Acids Res,* 37**,** 1127-40.

STAALS, R. H., BRONKHORST, A. W., SCHILDERS, G., SLOMOVIC, S., SCHUSTER, G., HECK, A. J., RAIJMAKERS, R. & PRUIJN, G. J. 2010. Dis3-like 1: a novel exoribonuclease associated with the human exosome. *EMBO J,* 29**,** 2358-67.

TAKAKU, H., MINAGAWA, A., TAKAGI, M. & NASHIMOTO, M. 2003. A candidate prostate cancer susceptibility gene encodes tRNA 3' processing endoribonuclease. *Nucleic Acids Res,* 31**,** 2272-8.

THOMPSON, D. M., LU, C., GREEN, P. J. & PARKER, R. 2008. tRNA cleavage is a conserved response to oxidative stress in eukaryotes. *RNA,* 14**,** 2095-103.

THOMPSON, D. M. & PARKER, R. 2009. The RNase Rny1p cleaves tRNAs and promotes cell death during oxidative stress in Saccharomyces cerevisiae. *J Cell Biol,* 185**,** 43-50.

TOMECKI, R., KRISTIANSEN, M. S., LYKKE-ANDERSEN, S., CHLEBOWSKI, A., LARSEN, K. M., SZCZESNY, R. J., DRAZKOWSKA, K., PASTULA, A., ANDERSEN, J. S., STEPIEN, P. P., DZIEMBOWSKI, A. & JENSEN, T. H. 2010. The human core exosome interacts with differentially localized processive RNases: hDIS3 and hDIS3L. *EMBO J,* 29**,** 2342-57.

ULE, J., JENSEN, K., MELE, A. & DARNELL, R. B. 2005. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods,* 37**,** 376-86.

VANACOVA, S. & STEFL, R. 2007. The exosome and RNA quality control in the nucleus. *EMBO Rep,* 8**,** 651-7.

VARKONYI-GASIC, E., WU, R., WOOD, M., WALTON, E. F. & HELLENS, R. P. 2007. Protocol: a highly sensitive RT-PCR method for detection and quantification of microRNAs. *Plant Methods,* 3**,** 12.

VOLAREVIC, S. & THOMAS, G. 2001. Role of S6 phosphorylation and S6 kinase in cell growth. *Prog Nucleic Acid Res Mol Biol,* 65**,** 101-27.

WANG, Y., JURANEK, S., LI, H., SHENG, G., TUSCHL, T. & PATEL, D. J. 2008. Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature,* 456**,** 921-6.

WANG, Z., TOLLERVEY, J., BRIESE, M., TURNER, D. & ULE, J. 2009. CLIP: construction of cDNA libraries for high-throughput sequencing from RNAs cross-linked to proteins in vivo. *Methods,* 48**,** 287-93.

WIGHTMAN, B., HA, I. & RUVKUN, G. 1993. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell,* 75**,** 855-62.

XUE, Z., YUAN, H., GUO, J. & LIU, Y. 2012. Reconstitution of an Argonaute-dependent small RNA biogenesis pathway reveals a handover mechanism involving the RNA exosome and the exonuclease QIP. *Mol Cell,* 46**,** 299-310.

YAMASAKI, S., IVANOV, P., HU, G. F. & ANDERSON, P. 2009. Angiogenin cleaves tRNA and promotes stress-induced translational repression. *J Cell Biol,* 185**,** 35-42.

Chapter 5: Ewing sarcoma breakpoint region 1 protein binds G-rich RNAs and prevents transcription-associated genome instability

Shivendra Kishore[1,3,4], Biter Bilen[1,4], Christophe Kunz[2], Nitish Mittal[1], Severin Berger[1], Friedel Wenzel[2], Georges Martin[1], Erik van Nimwegen[1], Primo Schär[2], Mihaela Zavolan[1,*]

[1]Biozentrum, University of Basel, Klingelbergstrasse 50-70, 4056 Basel, Switzerland

[2]Department of Biomedicine, University of Basel, Mattenstrasse 28, 4058 Basel, Switzerland

[3]Present address: Centogene AG, Schillingallee 68, 18057 Rostock, Germany

[4]Equal contribution

[*]Corresponding author (Email: mihaela.zavolan@unibas.ch, Phone: +41 612671577, Fax: +41 612671585)

Running title: EWS binding, transcription, and genome instability

Keywords: transcription-associated stress; RNA-binding; genome instability; Rloop; Ewing sarcoma; FET family; RNA polymerase II elongation; RNAPII; intron retention; DNA double strand break

5.1 SUMMARY

RNA-binding proteins (RBPs) are not only involved in RNA processing but also in the maintenance of genome stability by modulating RNA-DNA interactions. In this context, we studied the Ewing's sarcoma breakpoint region 1 (EWS/EWSR1) protein, whose RNA-binding domain is lost when diagnostic chromosomal translocations occur in Ewing sarcoma. With in vivo crosslinking and high-throughput sequencing, we found that EWS binds G-rich RNAs that originate from intrinsically fragile genomic regions such as sub-telomeres, introns including those from its own chromosomal translocation-

prone region, and predicted R-loops. EWS appears to promote the RNA polymerase II elongation through these loci, which also increases intron retention. EWS depletion increases the instability of its own locus and the frequency of spontaneous DNA double strand breaks (DSBs), but does not affect the level of induced DNA damage. These results highlight the importance of EWS in relieving transcription-associated stress and may link its loss to Ewing sarcoma tumorigenicity.

5.2 HIGHLIGHTS

- EWS binds G-rich RNAs from intrinsically fragile regions
- RNA polymerase II elongation is promoted at the loci of EWS' RNA targets
- EWS depletion increases the level of stochastic but not induced DNA damage
- EWS relieves transcription-associated stress

5.3 INTRODUCTION

R-loops are DNA-RNA hybrid structures that form between guanine (G)-rich nascent RNA and the template DNA strand leaving the non-template strand in a single stranded conformation. They play a role in many biological processes such as immunoglobulin class switching, telomere homeostasis, and protection of CpG island promoters from methylation (reviewed in (Wahba and Koshland, 2013)). However, their formation was linked to genomic instability mostly through the stalling of RNA polymerase II (RNAPII), which leads to collisions between the transcription and DNA replication machineries and DNA breaks. These in turn lead to mutations and chromosomal rearrangements (reviewed in (Aguilera and García-Muse, 2012)), whose accumulation predisposes to ageing and cancer. Various mechanisms have evolved to counteract such deleterious effects of R-loops. Co-transcriptional assembly of ribonucleoprotein complexes prevents the accumulation of naked nascent RNA and thereby the accumulation of R-loops in eukaryotes (reviewed in (Kim and Jinks-Robertson, 2012)). Consistently, mutations in RNAPII subunits and perturbation of RNA processing lead to an increase in the frequency and length of R-loops (Wahba et al., 2011, Li and Manley, 2005, Stirling et al., 2012, Domínguez-Sánchez et al., 2011). Moreover, a high-throughput genetic screen revealed that many RBPs, particularly those that are involved in splicing and RNA processing, reduce R-loop formation and thereby the level of DNA damage. This indicates that RBPs have a substantial impact on the maintenance of genome stability (Paulsen et al., 2009).

EWS is a ubiquitously expressed nuclear RBP and was initially identified as part of a fusion protein in Ewing sarcoma, a rare but malignant bone and soft tissue tumor. The fusion protein is the result of a chromosomal translocation that juxtaposes the EWS gene to a transcription factor (TF) gene with the concomitant loss of the RNA-binding domain of EWS (Delattre et al., 1992). Similarly, recurrent translocations are observed in distinct types of sarcomas and leukemias involving EWS' paralogs fused in sarcoma (FUS) and TATA-box binding protein-associated factor 15 (TAF15) (reviewed in (Kovar, 2011)), which together with EWS form the FET (FUS-EWS-TAF15) family of genes. The function of the wildtype EWS is not well understood. In vitro, the protein binds uridine(U)- and guanine(G)-rich RNAs (Ohno et al., 1994). In vivo, it interacts directly with components of RNAPII and with splicing factors (reviewed in (Kovar, 2011)). Consistently, EWS was found to regulate alternative splicing of genes involved in genotoxic stress signaling (Paronetto et al., 2011, Sanchez et al., 2008), which in turn would predict an indirect role of EWS in DNA damage response (DDR). The knockout of EWS impairs B-lymphocyte maturation and meiosis in mice (Li et al., 2007) and its knockdown leads to chromosome segregation defects during mitosis in zebrafish (Azuma et al., 2007). These findings suggest that EWS and specifically, its RNA-binding activity, plays a role in the maintenance of genomic integrity, and that the loss of EWS' RNA binding domain contributes to the pathogenesis of Ewing sarcoma.

To investigate this hypothesis we identified the transcriptome-wide targets of EWS with an in vivo crosslinking method. We showed that EWS binds G-rich RNAs that originate from intrinsically fragile genomic regions such as sub-telomeres and the introns of the FET family members where cancer-related translocations occur. Overall, EWS predominantly bound intronic RNAs, and the EWS-targeted regions had a propensity to form R-loops. The rate of RNAPII elongation appeared to be higher in introns that are bound by EWS compared to R-loop forming introns, and, consistently, depletion of EWS lead to decreased intron retention. Furthermore, EWS depletion increased the frequency of DNA double-strand breaks (DSBs) in untreated cells but did not affect the level of induced DSBs. Finally, depletion of EWS resulted in breaks at the EWS locus, breaks that occurred precisely in the region that is prone to genomic rearrangements and from which EWS-bound RNA fragments originate. Altogether, these results suggest that EWS' activity in binding G-rich RNAs may play a role in the prevention of transcription-associated genotoxic stress at intrinsically fragile genomic loci. The loss of this protective activity could, in turn, contribute to the development of Ewing sarcoma.

5.4 RESULTS

5.4.1 EWS targets G-rich intronic RNAs in vivo

To identify in vivo targets of endogenous EWS we carried out photo-reactive nucleoside-enhanced crosslinking and immuno-precipitation (PAR-CLIP) (Hafner et al., 2010, Kishore et al., 2011) in human embryonic kidney 293 (HEK293) cells. While EWS specific antibody recognized a ~80kDa protein, immuno-precipitation of EWS with the associated RNAs revealed two bands at ~90kDa (EWS A) and ~120kDa (EWS B)

(Figure 1A). Deep sequencing libraries generated from these bands yielded 21,274,257 and 32,468,723 reads respectively, which we mapped to the reference human genome and to the known transcripts with CLIPZ server (Khorshid et al., 2011). More than 80% of the uniquely mapping reads originated from protein coding and long non-coding RNA genes, repeats, or unannotated regions. We therefore determined EWS binding sites in RNAs from these functional categories with a method that takes advantage of crosslink-induced mutations (Jaskiewicz et al., 2012). Illustrative patterns of read coverage in the translocation prone loci of the FET genes are shown in Figure 1B. EWS bound relatively broad regions, but the read coverage from the two samples was very similar. To further assess the similarity of the samples, we computed the distances between crosslinked positions in EWS A and the closest crosslinked positions in EWS B. For comparison, we also computed the distances between the crosslinked positions in EWS A and the closest positions where another nuclear protein, Hu antigen R (HuR), crosslinked. We found that the top crosslinking positions identified from the EWS A sample were located closer to the crosslinked positions inferred from the EWS B sample than they were to positions that crosslinked to HuR, with very high significance (Figure 1C; Kolmogorov-Smirnov test p-value=0, statistic=0.81); half of the top crosslinked positions inferred from the EWS A sample were located less than 10 nucleotides (nts) away from a top crosslinked position from the EWS B sample. Hence, to obtain the genome-wide binding sites of EWS, we combined the reads from these two samples and defined a unique set of targets for further analysis (Jaskiewicz et al., 2012).

To determine the in vivo sequence specificity of EWS, we submitted 10,000 top-scoring EWS target regions to the MEME motif discovery tool (Bailey, 2002a). We obtained a significant G-rich motif that occurred 8 nts upstream of the crosslinked position in almost all crosslinked regions (Figure 1D; e-value=2.8e-644). This indicates that, similar to what was observed in vitro (Ohno et al., 1994), EWS binds G-rich sequences in vivo. Scanning an extended set of regions with the FIMO motif search tool (Grant et al., 2011) we observed a modest inverse relation between the log-likelihood of the motif and the crosslink score of the region. The motif likelihood dropped slightly after the top 20,000 crosslinks and we therefore used only these positions for further analysis.

We mapped the crosslinking positions onto gene structures defined by representative RefSeq transcripts (Meyer et al., 2013) to determine the positional preference of EWS within gene bodies. We found that EWS preferentially binds introns and 3' untranslated regions (UTRs) (Figure 1E inset); among 1,143 target transcripts, 698 (61%) were crosslinked at least once in introns, 298 (26%) in 3' untranslated regions (UTRs), and 35-220 (<20%) in other sub-regions. While most transcripts harbored only a few of the top crosslinking positions (median count of 3) (Figure 1E), 62 (5%) transcripts were extensively crosslinked at >50 positions, most of which were located in introns. These results suggest that EWS mainly targets intronic regions of transcripts. Functional annotation of EWS target transcripts based on the KEGG pathway gene sets (Kanehisa et al., 2014) revealed that EWS targets components of signaling pathways that are involved in splicing, metabolism, immune response and cancers (Figure 1F).

5.4.2 EWS promotes RNAPII elongation

To better understand the role of EWS in intronic regions, we asked whether the protein has a more fine-grained positional preference. Indeed, we found that EWS tends to target initial introns. This positional preference was stronger for EWS than for GU-rich element binding heterogeneous nuclear ribonucleoproteins (hnRNPs) such as hnRNPH, hnRNPM, hnRNPF, hnRNPA2B1, hnRNPU, and hnRNPA1 whose binding specificity was studied by Huelga et al. (Huelga et al., 2012) (Figure 2A); in particular, EWS bound initial introns of 325 transcripts and second introns of 120 transcripts, which gives a 2.7-fold ratio. Furthermore, EWS binding was enriched downstream of transcription start sites (TSSs) in the initial introns; 81 (25%) initial introns were bound in the 5'-most 10% of their length, which is 2.5-fold higher than that would be expected if the coverage of the initial intron were uniform (Figure 2B; Kolmogorov-Smirnov test p−value=5.383182e−11, statistic=0.19). These results suggest that EWS preferentially targets transcripts downstream of TSSs.

The sequence composition of EWS-bound regions and their location within transcripts is reminiscent of R-loops (Ginno et al., 2012). To investigate the hypothesis that EWS targets RNAs that are prone to form R-loops, we analyzed a dataset of genome-wide R-loop predictions from Ginno et al. (Ginno et al., 2012). We found that 508 (49%) of the 1,043 transcripts bound by EWS at genic regions also formed R-loops (Figure 2C inset; Fisher's exact test p-value= 4.4e-35, odds-ratio=2.3) and the positions where EWS crosslinked within these 508 co-targeted transcripts were closer to R-loops than expected by chance (Figure 2C; Kolmogorov-Smirnov test p−value=5.775462e−18, statistic=0.28). Even more specifically, 60 (41%) of the 148 transcripts that were bound by EWS downstream of TSSs - that is, in the 5' UTRs or in the first 10% of initial introns - also formed R-loops precisely in this region. These results suggest that the RNAs that are targeted by EWS are prone to form R-loops with the template DNA.

To gain insight into the role of EWS binding at R-loop-forming regions, we analyzed the rate of RNAPII elongation in these regions. From chromatin immuno-precipitation and sequencing (ChIP-seq) data obtained from HEK293 cells by the ENCODE consortium (Rosenbloom et al., 2013), we determined the read density of RNAPII in introns, normalized by the read density in the respective gene locus. We then analyzed the RNAPII occupancy in the following mutually exclusive categories of introns: introns exclusively bound by EWS, introns that are only predicted to form R-loops, introns that are both predicted to form R-loops and bound EWS and all other introns of crosslinked genes that are expressed in HEK293 cells. Consistent with a slow rate of RNAPII elongation through these regions, R-loop-forming introns had the highest read densities (Figure 2D). In contrast, the introns that were exclusively bound by EWS had much lower densities of RNAPII compared to those that form R-loops (Figure 2D, Kolmogorov-Smirnov test p-value=1.657989e−50, statistic=0.32). The introns that are predicted to form R-loops but were also bound by EWS showed an intermediate level of RNAPII density (Kolmogorov-Smirnov test p-value=1.086846e−08, statistic=0.2). These results suggest that EWS binding promotes RNAPII elongation along pre-mRNAs and counteracts R-loop formation at specific regions.

The rate of RNAPII elongation affects the recognition of splicing signals and thereby alternative splicing (reviewed in (Dujardin et al., 2013). One of the models that was proposed to explain the coupling between transcription and splicing is the “kinetic model”, which postulates that a slow elongating RNAPII allows more time for the recognition of weak splice signals, reducing the frequency of intron retention. Consistently, a recent genome-wide study of co-transcriptional splicing in fly found that intron retention is reduced in cells that express a slow-elongating RNAPII mutant (Khodor et al., 2011). To test whether the apparent role of EWS in promoting RNAPII elongation is reflected in intron retention in EWS target genes, we analyzed the polyadenylated transcriptome obtained from HeLa cells that we transfected with either EWS or control small interfering RNAs (siRNAs). We obtained 30-60 million sequence reads, which we mapped to the reference human genome and to known transcripts with the CLIPZ server (Khorshid et al., 2011). The EWS mRNA levels were approximately 4-

fold reduced in the EWS siRNA-treated sample compared to the control indicating that the knockdown was efficient. We found that the EWS knockdown preferentially reduced the retention of EWS target introns: ~70% of the 699 introns that were only bound by EWS, compared to ~60% of the 235 that were also predicted to form R-loops and to ~50% of the 2,758 that were only predicted to form R-loops had a reduced level of retention in EWS compared to control siRNA-treated cells (Figure 2E; Kolmogorov-Smirnov test in comparing the first two categories of introns with the latter p-value=1.138928e−16, statistic=0.18 and p-value=0.000179687, statistic=0.14, respectively). These results suggest that, consistent with a role in promoting RNAPII elongation, EWS leads to increased retention of its target introns.

5.4.3 EWS acts in the prevention of transcription-associated stress

Since co-transcriptional formation of R-loops has been associated with chromosomal instability, we sought to determine whether EWS protects against DNA damage. Phosphorylation of histone 2A variant X (H2AX) is an early event in the DDR (Rogakou et al., 1998) and H2AX has been shown to localize preferentially at endogenous DNA stress sites such as sub-telomeres and transcriptionally active loci in rapidly dividing cells (Seo et al., 2012). Re-analyzing H2AX ChIP-seq data obtained from Jurkat T cells by Seo et al. (Seo et al., 2012), we found that the loci of the RNA targets of EWS are strikingly similar to those bound by H2AX. EWS associated with RNAs originating from sub-telomeric regions, where H2AX also preferentially binds (Figure 3A; Kolmogorov-Smirnov test in comparing the distribution of EWS binding regions relative to a uniform distribution p-value=0, statistic=0.24); 52% (10,443) of 19,964 crosslinked positions on the assembled human chromosomes were located within the first or last 5% of the chromosomes' lengths. EWS and H2AX co-targeting was evident also at the level of individual genes (Figure 3B inset; Fisher's exact test p-value=9.1e-39, odds-ratio=4.6) and even at the level of binding regions; within the 140 co-targeted gene loci, EWS crosslinks occurred significantly closer to the H2AX peaks relative to what would be expected if the H2AX loci were randomly distributed within the genes (Figure 3B; Kolmogorov-Smirnov test p−value=5.775462e−18, statistic=0.28). These results suggest that EWS targets RNAs that are transcribed from regions that are prone to genomic instability.

Phosphorylated H2AX (γH2AX) and p53-binding protein 1 (53BP1) mark sites where the non-homologous end joining (NHEJ) system engages to repair DSBs (Nakamura et al., 2006). To further determine whether EWS depletion increases the frequency of DNA DSBs, we examined by immunofluorescence the frequency of γH2AX and 53BP1 co-staining nuclear foci upon short hairpin RNA (shRNA) mediated knockdown of EWS. Transfection of EWS shRNA markedly reduced EWS protein levels compared to control shRNA in both HeLa and osteosarcoma (U2OS) cells. We found that depletion of EWS significantly increased the frequency of γH2AX/53BP1 co-foci in both HeLa and U2OS cells (Figure 3C; Kruskal-Wallis test p-value<0.0001), similar to the treatment with camptothecin (CPT), a DNA damaging agent that acts as a topoisomerase 1 poison to cause transcriptional arrest and replication fork collapse (Liu et al., 2000). We observed a median number of 1, 7.5, 13, and 22 γH2AX/53BP1 co-foci in HeLa cells treated with control shRNA, EWS shRNA, sublethal-dose of CPT, and high-dose of CPT, respectively and 0, 5, 23.5 γH2AX/53BP1 co-foci in U2OS cells treated with control shRNA, EWS shRNA, and high-dose of CPT, respectively. EWS shRNA did not significantly increase the number of foci in cells in which DNA damage was induced with a high-dose of CPT: a median number of 23 compared to 22 foci were observed in HeLa cells and 27.5 compared to 23.5 foci in U2OS cells that were treated with EWS shRNA and control shRNA, respectively (Figure 3C; Kruskal-Wallis test p-value=non-significant (ns)). Thus, while the knockdown of EWS has a significant effect on the steady state levels of γH2AX/53BP1 co-foci in HeLa and U2OS cells, it does not affect the frequency of CPT-induced foci, suggesting that EWS functions in the prevention of DNA DSBs.

The binding of EWS precisely within the chromosomal translocation-prone regions of transcripts from the FET genes is consistent with EWS being necessary for the maintenance of genomic integrity. To further test this hypothesis, we carried out fluorescence in situ hybridization (FISH) with EWS break-apart probes which flank the EWS breakpoint region between intron 7 and intron 10 in EWS and control shRNA transfected HeLa cells. EWS depletion lead to a significant 3-fold increase in the frequency of cells with one or more split signals (Figure 3D; Fisher's exact test p-value < 0.0001); compared to 20 split signals in control shRNA-treated cells, we observed approximately 60 split signals in EWS shRNA-treated cells. This result suggests that depletion of EWS creates an environment that is permissive for chromosomal translocations in its own genomic locus, in the region where EWS binds its own RNA. Finally, to more generally evaluate the propensity of EWS-bound transcripts to originate in translocation prone-loci, we analyzed the fusion transcripts from the ChimerDB knowledgebase (Kim et al., 2010). From 841 genes that are involved in fusions and are expressed in HEK293 cells, 124 (12%) were also targeted by EWS (Figure 3E; Fisher's exact test p-value=2.4e-09, odds-ratio=1.9. This suggests that EWS is important for the maintenance of genomic integrity genome-wide.

5.5 DISCUSSION

In addition to DNA damaging agents that induce genomic instability, single stranded DNA intermediates that predispose to chromosome fragility occur during normal processes such as replication and transcription. Not surprisingly, a variety of surveillance mechanisms, which act in DNA damage prevention, signaling and repair, have evolved to maintain the stability of genome. While the processes of DNA damage response and repair are relatively well studied, the mechanisms of DNA damage avoidance only recently began to be uncovered. A recent siRNA screen for factors that influence the frequency of γH2AX foci revealed that RNA binding proteins play an important role in the maintenance of genome stability by preventing the co-transcriptional formation of stable RNA-DNA structures (Paulsen et al., 2009). Our study suggests that EWS is part of this preventive mechanism. Namely, EWS promotes RNAPII elongation, thereby relieving transcription-associated stress at intrinsically fragile genomic regions. That the landmark chromosomal translocations in Ewing sarcoma are accompanied by the loss of EWS' RNA-binding domain raises the intriguing hypothesis that reduced EWS expression may trigger a vicious circle that leads to genomic instability and chromosomal translocations.

In vitro, EWS was reported to bind G- and U-runs (Ohno et al., 1994). In our study, we determined the transcriptome-wide targets of EWS using PAR-CLIP (Hafner et al., 2010) and showed that it associates with G-rich sequences in vivo as well. Although crosslinking and immuno-precipitation is becoming a standard method to determine targets of RBPs in vivo, the protocols are quite involved and prone to biases at various levels (Kishore et al., 2011, Sugimoto et al., 2012), which may explain why CLIP studies of FET family proteins yielded somewhat divergent results. For instance, no G-rich motif emerged from a previous PAR-CLIP study of FET proteins, including EWS (Hoell et al., 2011). However, this could be due to the depletion of G-containing binding sites through T1 RNase digestion (Hoell et al., 2011, Kishore et al., 2011), because the same study identified an AU-rich stem-loop as the preferred binding motif of the FUS protein, while other CLIP studies of FUS found instead GU-rich elements (Lagier-Tourenne et al., 2012, Rogelj et al., 2012). Our results, obtained with mild T1 RNase digestion, are consistent with both the previous in vitro data on EWS and the in vivo data from multiple CLIP studies of the related FUS protein, and indicate that EWS indeed binds G-rich elements in vivo.

The location - downstream of TSSs - and composition - G-rich - of in vivo bound EWS targets are very reminiscent of regions that are prone to form R-loops (Ginno et al., 2012, Roy and Lieber, 2009), structures that impede RNAPII elongation during transcription (reviewed in (Aguilera and García-Muse, 2012)). While EWS binds transcripts that are predicted to form R-loops, our analysis of RNAPII ChIP-seq data and mRNA-seq data indicates that EWS counteracts the effect of R-loops. Namely, EWS promotes RNAPII elongation and intron retention. Recent reports provide extensive evidence that the lack of coordination between RNA processing steps leads to the formation of R-loop structures and transcription-associated deleterious effects on the genome (reviewed in (Montecucco and Biamonti, 2013)). The fact that EWS depletion alters the steady state levels of DNA DSBs but does not affect the frequency of DSBs that occur in response to the topoisomerase inhibitor CPT suggests that EWS binding has a role in the prevention rather than the repair of breaks. These results point to a direct link between EWS and the prevention of transcription-associated genomic instability.

The effect of EWS on intron retention is consistent with previously reported links between EWS and alternative splicing (Paronetto et al., 2011, Sanchez et al., 2008). EWS was found to interact directly with splicing regulatory proteins (reviewed in (Kovar, 2011)) and we here found that it binds preferentially transcripts of splicing factors. Finally, the relationship that we uncovered between the level of EWS and the level of intron retention is consistent with the kinetic model of coupling between RNAPII elongation and splicing (reviewed in (Dujardin et al., 2013)).

Other FET family members appear to share EWS' functions. For example, in vivo crosslinking studies of FUS also indicated binding at the rearrangement-prone loci of FET genes (Lagier-Tourenne et al., 2012, Rogelj et al., 2012). Interestingly, a recent study found that the FUS protein regulates the phosphorylation of RNAPII at serine 2 and thereby the rate of RNAPII elongation (Schwartz et al., 2012). In the RNAPII pull-downs presented in this study EWS was the second most abundant protein, but its interaction with RNAPII was not investigated further. Our data suggests however, that EWS promotes RNAPII elongation in regions that are prone to R-loop formation and transcription-associated stress, an activity that is probably common to all FET family members and perhaps other RBPs.

Ewing sarcoma is a rare tumor whose very distinctive characteristic is the presence of chromosomal translocation-induced fusion proteins. In about 85% of Ewing sarcoma cases the EWS locus is translocated to the FLI1 (friend leukemia virus integration 1) locus, which leads to the production of an EWS-FLI1 fusion transcription factor that lacks the RNA binding domain of EWS (Ohno et al., 1994). How these translocations take place is unclear. However, our FISH analysis suggests that limiting levels of EWS lead to DNA breaks in its own translocation-prone genomic locus, which could in turn lead to chromosomal translocations. This translocation per se may not lead to tumorigenicity, which requires additionally that cells escape senescence and apoptosis and proliferate with an increased rate (Hanahan and Weinberg, 2011). However, these features may be acquired as a result of the expression of EWS fusion proteins (reviewed in (Riggi et al., 2007)) or of DNA DSBs at other loci where the EWS RNA-binding activity is lost. Genome-wide translocation sequencing (Chiarle et al., 2011, Klein et al., 2011) following EWS knockdown could be an initial approach to identify such recurrent translocation-prone regions. Collectively, our analysis points to a direct role of EWS' RNA-binding activity in promoting RNAPII elongation and relief of transcription-associated stress. The fact that this activity is lost in Ewing sarcoma suggests a novel mechanism contributing to the pathogenesis of Ewing sarcoma.

5.6 Experimental Procedures

We used custom Perl (version 5.8.1) and R (version 2.15.1) scripts and BEDTools packages (version 2.17.0) (Quinlan and Hall, 2010) for in silico analysis and IGV (version 2.2) (Thorvaldsdóttir et al., 2012) for visualization of features. Detailed experimental and analysis methods can be found in the Extended Procedures.

5.6.1 PAR-CLIP experiment and analysis

PAR-CLIP of EWS in HEK293 cells was performed as described in (Hafner et al., 2010) and sequencing libraries were analyzed as described in (Jaskiewicz et al., 2012).

5.6.2 mRNA-seq experiment and analysis

The Illumina directional mRNA-seq protocol with minor modifications was applied to samples extracted from control or EWS siRNA transfected HeLa cells and untreated HEK293 cells.

5.6.3 H2AX ChIP-seq analysis

H2AX immunoprecipitation (SRR074195 and SRR074196) and input DNA (SRR074203) samples obtained from Jurkat T cells by Seo et al. (Seo et al., 2012) were downloaded from www.ncbi.nlm.nih.gov/sra and analyzed as described in (Arnold et al., 2012a) with small modifications.

5.6.4 Immunofluorescence

HeLa and U2OS cells transfected with control or EWS shRNA constructs on glass coverslips were enriched by puromycin selection, fixed in ice-cold methanol, stained with antibodies and imaged with a Leica DMI 4000 microscope.

5.6.5 Fluorescent In Situ Hybridization (FISH)

The status of EWS locus was evaluated with the Poseidon Repeat Free EWS (22q12) break-apart probes (KBI-10708; KREATECH Diagnostics, Netherlands) in HeLa cells transfected with control or EWS shRNA constructs. Hybridizations, washing and evaluation were done according to the manufacturer's instructions.

5.7 Accession Numbers

PAR-CLIP and RNA-seq libraries have been deposited in the NCBI GEO under accession number GSE54689.

5.8 Acknowledgments

We thank Manuel Stucki (University of Zürich) for his assistance in the evaluation of the 53BP1 foci, S. Piscuoglio, M. Kovac, K. Heinimann and lab members (Department of Biomedicine, University of Basel), and Zavolan lab members for helpful discussions. Work in the Zavolan laboratory was partly supported by SystemsX, the systems biology initiative in Switzerland. S.K. was the recipient of the EMBO long-term fellowship 184-2009, and B.B. of a Werner-Siemens Stiftung fellowship at Biozentrum.

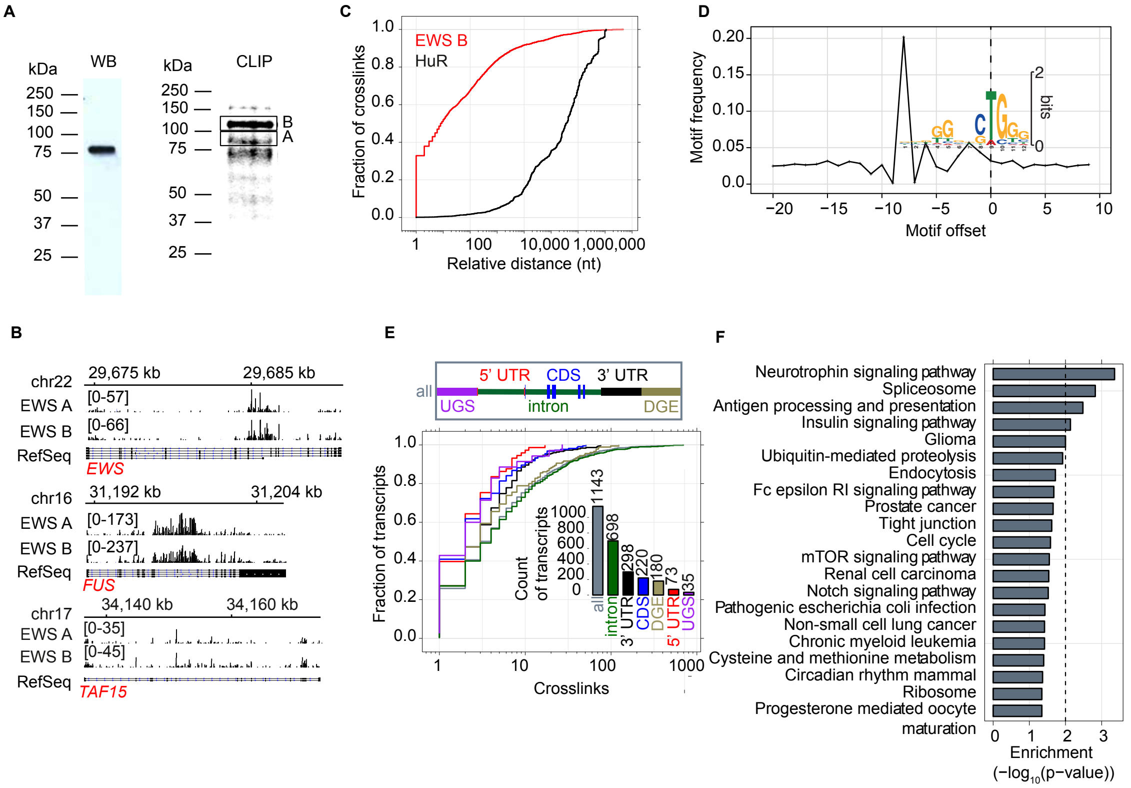


Figure 1. PAR-CLIP reveals that EWS targets intronic G-rich RNAs in vivo. A. The EWS antibody recognizes a ~80 kDa protein in western blot (WB) from HEK293 cells. Boxed ~90 kDa (A) and ~120kDa (B) bands in CLIP show the regions from which samples were prepared for sequencing. B. Genome browser screenshots of the coverage of FET gene loci by CLIP reads. Range of raw read counts are indicated on each respective track. (kb: kilobase) C. Cumulative distribution of the relative distance from each of the top 10,000 crosslink in EWS A sample to the closest crosslink among the top 10,000 crosslinks in EWS B and HuR samples. D. Frequency of the sequence motif identified as over-represented in EWS crosslinked sites as a function of position relative to the crosslink, which is indicated by the vertical dashed line at 0. E. Cumulative distribution of the number of crosslinks in various transcript regions, sketched at the top of the panel (UGS: upstream of gene start, DGE: downstream of gene end). Inset shows number of transcripts with crosslinks in different types of regions. F. KEGG pathways enriched in EWS target transcripts. P-values are calculated from Fisher's exact test taking the genes expressed in HEK293 as the universe. Vertical dashed line represents p-value=0.01.

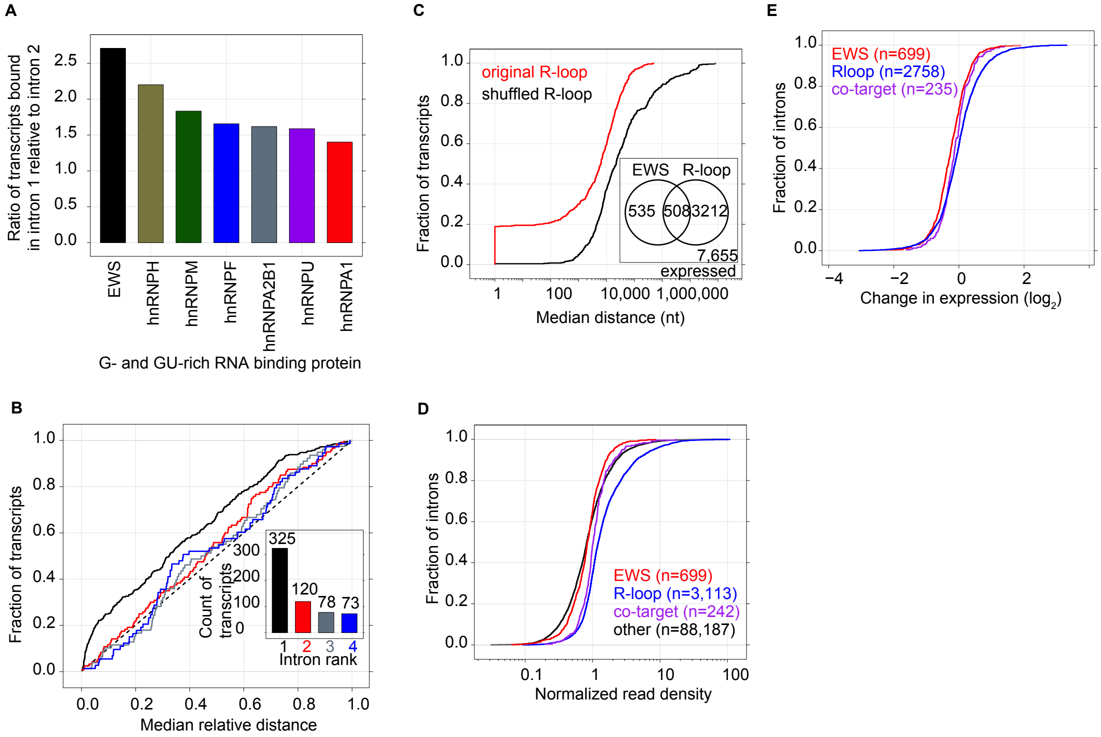


Figure 2. EWS associates with R-loops and promotes RNAPII elongation. A. Targeting of initial introns by EWS and GU-rich element-binding RBPs. B. Cumulative distribution of median displacements of EWS crosslinks from starts of introns. The dashed line indicates no positional bias within introns. Introns are ranked from transcript start. Inset shows the number of transcripts crosslinked in individual introns. C. Cumulative distribution of the median distance between an EWS crosslink and the closest R-loop within individual genes in the EWS and R–loop co-targeted gene loci. The background distribution was computed by randomizing the positions of R-loops within the co-targeted gene loci. Inset shows Venn diagram of co-targeting of genes expressed in HEK293 cells. D. Cumulative distribution of relative RNAPII ChIP-seq read densities in introns compared to the entire gene bodies in HEK293 cells. E. Cumulative distribution of normalized intron expression change between EWS and control siRNA-treated cells in HeLa cells. Normalized intron expression was computed as ratio of read densities in the intron and in the exons of the corresponding gene.

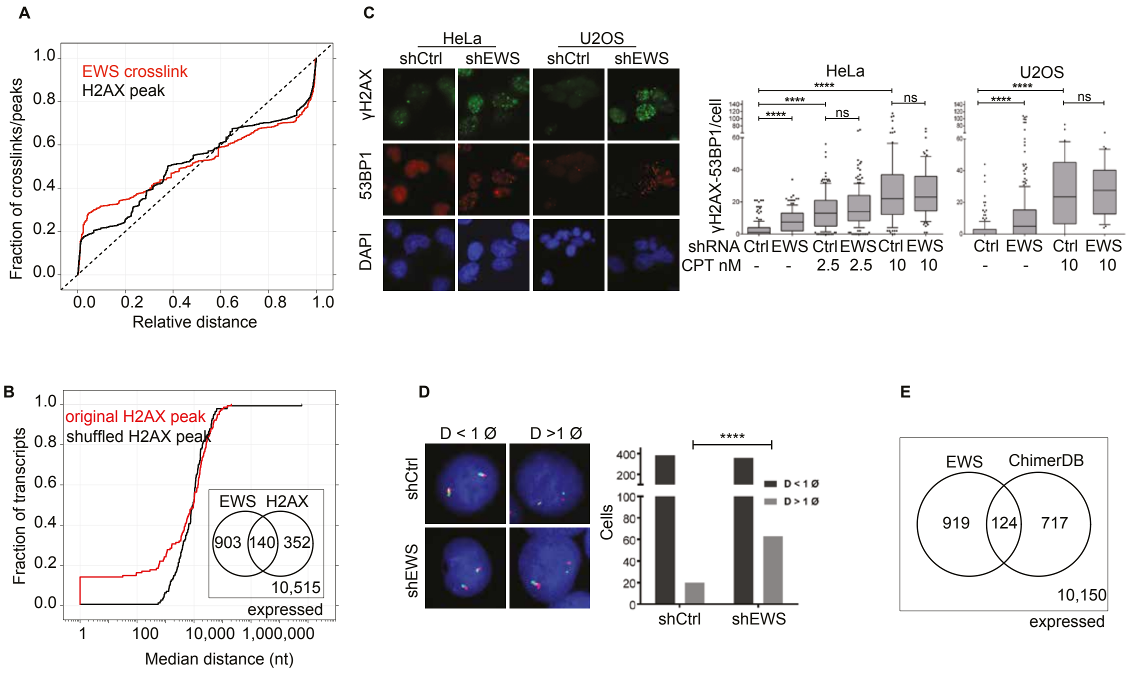


Figure 3. EWS binds RNAs derived from intrinsically fragile genomic sites and prevents transcription-associated stress. A. Cumulative distribution of the relative distance of EWS crosslinks and H2AX ChIP-seq peaks from chromosome starts. Dashed line corresponds to no positional bias. B. Cumulative distribution of the median distance between an EWS crosslink and the closest H2AX peak within individual genes in the EWS and H2AX co-targeted gene loci. The background distribution was computed by randomizing the positions of H2AX peaks within the co-targeted gene loci. Inset shows Venn diagram of co-targeting of genes expressed in HEK293. C. Representative immunofluorescence images and quantification of spontaneously formed nuclear γH2AX/53BP1 co-foci in HeLa and U2OS cells expressing EWS or control shRNA. CPT indicates camptothecin treatment. DNA was stained with DAPI. Box plots summarize the number of γH2AX/53BP1 co-foci per cell. ****: p < 0.0001 in the Kruskal-Wallis test; ns: not significant. D. Representative FISH images and the quantification of EWS break-apart probe signals showing intact (D < 1 ø) or split (D > 1 ø) patterns in the EWS locus in HeLa cells expressing EWS or control shRNA. ****: p < 0.0001 in Fisher's exact test. E. Venn diagram showing the overlap of EWS target genes with genes from ChimerDB. Only genes that are expressed in HEK293 cells were considered in this analysis.

5.9 EXTENDED PROCEDURES

5.9.1 Cell culture

HEK293 and HeLa cells were routinely cultured in DMEM medium supplemented with 10% FCS in the presence of pennicillin and streptomycin antibiotics until they reached 80-90% confluence.

5.9.2 PAR-CLIP experiment

HEK293 cells were grown on 15-cm cell culture plates. Approximately 30 plates, with cells at ~80% confluence were used. 4-thiouridine (Sigma) was added to the cells to a final concentration of 100 μM and cells were incubated for 12 hours before crosslinking. PAR-CLIP was carried out as described previously (Hafner et al., 2010, Kishore et al., 2011). The sc-48404 antibody (Santa Cruz Biotechnology) coupled to protein-G Dynabeads (Invitrogen) was used for immuno-precipitation. The RNA was partially digested with endonuclease T1 (Ambion) and then with micrococcal nuclease (MNase, from New England Biolabs) as previously described (Jaskiewicz et al., 2012). The RNA complexed to the protein was radiolabeled; the complexes were separated on SDS-PAGE gel followed by transfer onto nitrocellulose membranes to reduce the background of free RNAs (Konig et al., 2011). The two bands (90kDa and 120kDa) containing EWS were separately isolated and the protein was digested with proteinase K. The recovered RNA was ligated first to preadenylated 3' adaptor (5' TCGTATGCCGTCTTCTGCTTG 3') then size fractionated (20-50 nts) on denaturing acrylamide gels and ligated to 5' adaptor (5' GUUCAGAGUUCUACAGUCCGACGAUC 3'). The ligation mixture was once again size fractionated on denaturing acrylamide gels to remove unligated inserts along with free adaptors. The recovered RNA was reverse transcribed with Superscript III

(Invitrogen) reverse transcriptase using reverse complementary sequence of 3' adaptor as primer and treated with RNase H. The cDNA resulting from reverse transcription was PCR amplified (typically 14-16 cycles) using forward primer (5' AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA 3') and reverse primer (5' CAAGCAGAAGACGGCATACGA 3'). The PCR products were extracted from agarose gels with a Qiaex II Gel extraction kit (Qiagen). Libraries were then sequenced on an Illumina Genome Analyzer IIx instrument.

5.9.3 mRNA-seq

HeLa cells were grown in 6-well plates as described above. The siRNA-mediated EWS knockdown was done with siRNAs from Santa Cruz Biotechnology (sc-35347, a pool of 3 target-specific 19-25 nt siRNAs sc-35347A, sc-35347B and sc-35347C). Corresponding control siRNA-A (sc-37007) for control transfections was also obtained from the same manufacturer. siRNAs were transfected with RNAiMAX (Invitrogen) according to the manufacturers protocol. Two experiments were performed. In one, the sample was harvested at 48 hours (replicate 1) after transfection and in the other at 72 hours (replicate 2). In both cases, the samples were split into two. One aliquot was used for western blot assay to assess the down-regulation of EWS protein. The other aliquot was used for the preparation of an mRNA-seq sample. Poly(A)+ RNA was isolated directly from EWS and control siRNA treated HeLa cells with Dynabeads® mRNA DIRECT™ Kit from Invitrogen, following the manufacturer's protocol. After isolation, the mRNA was chemically fragmented by incubating mRNA solution with twice the volume of alkaline hydrolysis buffer (50 mM Na-CO3, 1 mM EDTA, pH 9.2, prepared by mixing 1 ml 0.1 M Na2CO3 with 9 ml 0.1 M NaHCO3, adding EDTA to 1 mM, adjusting the pH to 9.2 and the volume to 20 ml with H2O) at 95°C for 5 minutes to obtain fragments of approximately 200 nt. Fragmented mRNA was immediately purified with RNeasy MinElute Cleanup Kit (Qiagen) to stop the reaction and to remove small RNA fragments (<100 bases). Purified fragmented mRNA was then treated with phosphatase FastAP (Fermentas) at 37 °C for 30 minutes to remove 5' and 3' phosphate groups from the fragmented mRNA. Fragmented mRNA was further incubated with ATP and T4 polynucleotide kinase (Fermentas) at 37 °C for an hour to add phosphate at 5' end of fragmented mRNA (to define 5' and 3' end of fragmented mRNA). After phosphorylation fragmented mRNA was purified using above-mentioned kit to remove all the free ATP from phosphorylation reaction. Following purification, ligation of preadenylated 3' adaptor from Illumina (RA3, 5' TGGAATTCTCGGGTGCCAAGG 3') was performed with T4 RNA Ligase 2, truncated K227Q (New England Biolabs Inc) according to the Illumina protocol. Ligation was followed by another purification done as described above to remove un-ligated 3' adaptors. The 5' adaptor from Illumina (RA5, 5' GUUCAGAGUUCUACAGUCCGACGAUC 3') was ligated with T4 RNA ligase (Fermentas) according to the Illumina protocol followed by purification to remove un-ligated 5' adaptors. cDNA was synthesized using the sequence complementary to the 3' adaptor with SuperScript III (Invitrogen) as per Illumina protocol. Libraries were amplified for 14 cycles of PCR using forward primer and reverse primer with barcodes from Illumina. Barcodes were introduced in each library at the PCR step for multiplexing. Libraries were sequenced for 50 cycles on an Illumina HiSeq 2000 deep sequencer.

HEK293 cells were grown on 6-well plates. mRNA-seq sample preparation was carried out as described above for HeLa cells, except that for replicate 1, the adaptors that were used were those that were specified above for the PAR-CLIP sample preparation: 3' adaptor (5' TCGTATGCCGTCTTCTGCTTG 3') and 5' adaptor (5'

GUUCAGAGUUCUACAGUCCGACGAUC 3’), and the primers for PCR amplification were: forward primer (5' AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA 3’) and reverse primer (5' CAAGCAGAAGACGGCATACGA 3’).

5.9.4 Immunofluorescence

EWS shRNA constructs used were originally obtained from Origene (catalogue number TG313142, sequence shRNA-EWS_1 - GI352561: GAGCAGTTACTCTCAGCAGAACACCTATG) and contained the shEWS sequence cloned in a retroviral GFP vector. shEWS cassette was subsequently sub cloned into a retroviral pRS vector (Origene) without GFP using EcorI/HindIII site to eliminate interference of any GFP fluorescence. HeLa and U2OS cells grown on coverslips in 6-well plates were then transfected with EWS shRNA construct or respective control (sequence TR30013 GCACTACCAGAGCTAACTCAGATAGTACT) in pRS vector containing a puromycin-resistance cassette using the jetPEI transfection reagent (Polyplus-transfection SA, France) according to the manufacturers protocol. Medium was exchanged 12 hours after transfection. Successfully transfected cells were then enriched by the addition of 1 μg/ml puromycin to the culture medium. Additionally, for continuous CPT treatments, CPT was added to a final concentration of 10 nM. Four days after transfection, the cells were washed twice with PBS and fixed for 30 min in ice-cold methanol at 4°C. All following steps were performed at room temperature. After washing twice with PBS, cells were blocked with blocking buffer (5% FCS in PBS) for 30 min. Antibodies against γH2AX (clone JBW301; upstate, USA) and 53BP1 (H-300; Santa Cruz Biotechnology, USA) were hybridized in blocking buffer at dilutions of 1:1000 and 1:500, respectively. After one hour cells were washed three times with PBS for 5 min, and secondary antibodies (A11017 and A11012; Invitrogen, USA) hybridized in blocking buffer for 30 min. After washing 3 times with PBS for 5 min, coverslips were mounted using vectashield, containing DAPI (1 μg/ml). γH2AX and 53BP1 were evaluated and counted with a Leica DMI 4000 microscope. Statistical analysis of triplicate experiments was done using the Prism5 software (GraphPad, USA).

5.9.5 FISH

HeLa cells were grown and transfected with control or EWS shRNA constructs as mentioned above. Four days after transfection coverslips were washed twice with PBS and incubated in 2x SSC (pH 7.0) for 30 minutes at 37°C. Samples were then dehydrated by a series of two minutes incubations in 70%, 85% and 100% ice-cold ethanol and subsequently air dried. To avoid over-denaturation in the subsequent steps, samples were "aged" overnight at 4°C. Hybridization of the Poseidon Repeat Free EWSR1 (22q12) Break probe (KBI-10708; KREATECH Diagnostics, Netherlands) and subsequent steps were performed according to the manufacturer's instructions. Briefly, coverslips were covered with 10 μl of the probe, sealed with Fixogum, denatured at 80°C for 5 minutes and hybridized over night at 37°C in a humidified chamber. Post-hybridization washes included one time 2xSSC/0.1% Igepal for 2 minutes at room temperature, one time 0.4x SSC/0.3% Igepal for 2 minutes at 72°C and one time 2xSSC/0.1% Igepal for 2 minutes at room temperature, the latter two steps being performed without agitation. Samples were then dehydrated by a series of two minutes incubations in 70%, 85% and 100% ethanol, air dried and mounted using vectashield, containing DAPI (1 μg/ml). Intact and aberrant loci defined as red/green (yellow) fusion signals or split signals with a distance of more than one signal diameter, respectively, were evaluated and counted with a Leica DMI

4000 microscope. Statistical analysis of triplicate experiments was done using the Prism5 software (GraphPad, USA).

5.9.6 Read mapping

We mapped the reads from PAR-CLIP and mRNA-seq libraries to the human genome (assembly version hg19) downloaded from the University of California at Santa Cruz (genome.cse.ucsc.edu) (Meyer et al., 2013) and transcriptome with CLIPZ server (Khorshid et al., 2011).

5.9.7 Estimation of expressed RefSeq genes

We used 2 biological replicates of mRNA-seq libraries obtained from HEK293 cells for the estimation of expressed RefSeq genes. We used uniquely mapping >25 nts long reads, which were annotated as either “mRNA” or “none”, without a known annotation mapped, by CLIPZ server (Khorshid et al., 2011). We fitted a two-component Gaussian mixture model with unequal variance to the per-nucleotide read densities of the representative RefSeq transcripts with the mclust R package (Yeung et al., 2001). We used the set of genes belonging to the higher expression component of the mixture in the libraries as the set of “HEK293-expressed genes” and used for functional annotation.

5.9.8 PAR-CLIP analysis

We used reads that were uniquely mapped to the genome and annotated as “mRNA”, “repeat”, or “none”, without a known annotation mapped, by CLIPZ server (Khorshid et al., 2011) to calculate the genome-wide crosslink scores in two 90kDa and 120kDa EWS, their combination, and HuR (Martin et al., 2012) (sequencing data accessible from the Gene Expression Omnibus (GEO) accession GSM714639) PAR-CLIP libraries as previously described (Jaskiewicz et al., 2012).

5.9.9 H2AX ChIP-seq analysis

We used two H2AX ChIP-seq samples (SRR074195 and SRR074196) and one input DNA sample (SRR074203) obtained from Jurkat T cells by Seo et al. (Seo et al., 2012), downloaded from www.ncbi.nlm.nih.gov/sra to identify genome-wide H2AX binding loci with the method described Arnold et al. (Arnold et al., 2012b) but modeling the fluctuations on the read-counts as a convolution of log-normal and Poisson sampling noise (Balwierz et al., 2009). Briefly, raw sequencing reads were quality-filtered and adaptors were removed. Alignments to the human genome assembly hg19 were obtained with bowtie (Langmead, 2010) version 0.12.7 and parameters -v 2 -a -B 1 --quiet --best --strata -m 100. Further analysis was performed with aligned reads. Read lengths were estimated by finding the distance *d* that maximizes the cross-correlation between the numbers of reads starting at opposite DNA strands, and subsequently each read was represented by its estimated center position. i.e. shifted by half of the estimated read length. Read counts associated with a given genomic location were obtained by summing over all read centers within the region, weighing each read with the inverse of the number of loci to which the read mapped equally well. To detect H2AX binding peaks we compared read density in sliding windows of size 500 nt for immuno-precipitation samples (foreground) with the read density in 2,000 nt regions, centered at the same position, for the input DNA sample (background). We assumed that the differences in log read-densities within the windows derive from a mixture of background (a convolution of log-normal and Poisson sampling noise) and enriched regions (represented as a uniform distribution over the range spanned by the data). We fitted the parameters of this mixture model with expectation maximization, *i.e.* using maximum likelihood. Finally, we calculated a z-value for each window, corresponding to the number of estimated standard deviations that the log read-density is enriched relative to the background distribution.

The histogram of z-values showed a distinct shoulder at high z-scores and 8,298 regions with z-value > 4 were selected as H2AX binding peaks.

5.9.10 Functional annotation

We used the RefSeq gene track from University of California at Santa Cruz genome browser (genome.cse.ucsc.edu) (Kuhn et al., 2009) to map EWS and HuR crosslinks, and H2AX peaks, R-loops, and GU-rich RBP binding sites over genes. We selected a representative RefSeq transcript from long-noncoding RNAs and mRNAs. The selection of representative transcripts of single-locus genes was based on several criteria such as biotype (mRNA > non-coding RNA), RefSeq gene status (Reviewed > Validated > Inferred > Provisional > Predicted), length of genomic locus, and total length of the exons (with a higher precedence for the longer transcript locus and total exon-length). The 2,500 nt long regions upstream of representative transcripts were defined as UGS regions (Upstream of Gene Start) regions and the 2,500 nt-long regions downstream of the representative transcripts were defined as DGE (Downstream of Gene End) regions. Intra-genic regions were classified as 5' UTR, CDS, and 3' UTR on the basis of the representative transcript annotation. Finally genomic crosslinks or regions were functionally annotated with the following precedence rules: 3' UTR > 5' UTR > CDS > intron > DGE > UGS. Region annotations were prioritized based on the longer of the intersecting regions.

5.9.11 Motif discovery and search

Non-overlapping representative crosslink centered regions (41-nt-long regions centered on the crosslink) from combined EWS PAR-CLIP sample were used for the motif analysis with the MEME program (Bailey, 2002b). We used the same number of regions with the same lengths from the sequences flanking these sites that did not contain any crosslinked positions to compute a Markov model of order 1 which we used as background. Additional parameters were: -dna -bfile -mod zoops -nmotifs 1 -minw 5 -maxw 12. We used the position weight matrix obtained from the motif analysis as described above as well as the same background model to compute the motif likelihoods for an extended set of 41-nt-long CCRs with FIMO (Grant et al., 2011) and defined a cut-off for high-affinity EWS binding sites.

5.9.12 Gene set enrichment analysis

We used the KEGG Pathways dataset (Kanehisa et al., 2014) and checked their association with EWS crosslinks using Fisher's exact test.

5.9.13 GU-rich RNA binding data

Processed files for the genomic binding regions of heterogeneous nuclear ribonucleoproteins (hnRNPs) such as hnRNPH, hnRNPM, hnRNPF, hnRNPA2B1, hnRNPU, and hnRNPA1 by Huelga et al. (Huelga et al., 2012) was obtained from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE34993. These predictions were done on the hg18 version of the human genome assembly therefore we converted the coordinates to hg19 coordinates using liftover utility from UCSC (Hinrichs et al., 2006).

5.9.14 R-loop data

We obtained R-loop predictions data from Ginno et al. (Ginno et al., 2012) and extracted strand specific G-skewed regions. These predictions were done on the hg18 version of the human genome assembly and we therefore converted the coordinates to hg19 coordinates using liftover utility from UCSC (Hinrichs et al., 2006).

5.9.15 Fusion genes analysis

The genes corresponding to head and tail partners of fusion transcripts inferred from mRNA-seq data were downloaded from ChimerDB (Version 2.0) (Kim et al., 2010).

5.9.16 EWS association analysis

Association of EWS crosslinks with H2AX peaks, R-loops, and ChimerDB fusion genes was drawn based on the genes with which these features were annotated with HEK293 expressed transcript regions. The distance between a crosslink and H2AX peaks or R-loops on the co-targeted gene loci were further determined and used in statistical tests.

5.9.17 RNAPII ChIP-seq data and analysis

RNAPII ChIP-seq data obtained from HEK293 cells by ENCODE (ftp://hgdownload.cse.ucsc.edu/goldenPath/currentGenomes/Homo_sapiens/encodeDCC/wgEncodeSydhTfbs/wgEncodeSydhTfbsHek293Pol2StdAlnRep3.bam) (Rosenbloom et al., 2013) were used in the analysis of RNAPII elongation rate. Reads were mapped onto introns of genes expressed in HEK293 and containing at least one EWS crosslink and only those introns with more than one read were used in the statistical significance analysis.

5.9.18 Intron retention analysis

We used 2 replicates of EWS and control siRNA treated mRNA-seq samples obtained from HeLa cell lines for intron retention analysis. Reads were mapped onto introns of genes expressed in HEK293 and containing at least one EWS crosslink and only those introns with more than one read were used in the statistical significance analysis.

5.10 References

AGUILERA, A. & GARCÍA-MUSE, T. 2012. R loops: from transcription byproducts to threats to genome stability. *Molecular cell*.

ARNOLD, P., ERB, I., PACHKOV, M., MOLINA, N. & VAN NIMWEGEN, E. 2012a. MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics,* 28**,** 487-94.

ARNOLD, P., SCHÖLER, A., PACHKOV, M., BALWIERZ, P., JØRGENSEN, H., STADLER, M., VAN NIMWEGEN, E. & SCHÜBELER, D. 2012b. Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting. *Genome Res*.

AZUMA, M., EMBREE, L., SABAAWY, H. & HICKSTEIN, D. 2007. Ewing sarcoma protein ewsr1 maintains mitotic integrity and proneural cell survival in the zebrafish embryo. *PLoS ONE*.

BAILEY, T. 2002a. Discovering novel sequence motifs with MEME. *Curr Protoc Bioinformatics*.

BAILEY, T. L. 2002b. Discovering novel sequence motifs with MEME. *Curr Protoc Bioinformatics*.

BALWIERZ, P. J., CARNINCI, P., DAUB, C. O., KAWAI, J., HAYASHIZAKI, Y., VAN BELLE, W., BEISEL, C. & VAN NIMWEGEN, E. 2009. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol,* 10**,** R79.

CHIARLE, R., ZHANG, Y., FROCK, R., LEWIS, S., MOLINIE, B., HO, Y.-J., MYERS, D., CHOI, V., COMPAGNO, M., MALKIN, D., NEUBERG, D., MONTI, S., GIALLOURAKIS, C., GOSTISSA, M. & ALT, F. 2011. Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *CELL*.

DELATTRE, O., ZUCMAN, J., PLOUGASTEL, B., DESMAZE, C., MELOT, T., PETER, M., KOVAR, H., JOUBERT, I., DE JONG, P. & ROULEAU, G. 1992. Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours. *Nature*.

DOMÍNGUEZ-SÁNCHEZ, M., BARROSO, S., GÓMEZ-GONZÁLEZ, B., LUNA, R. & AGUILERA, A. 2011. Genome instability and transcription elongation impairment in human cells depleted of THO/TREX. *PLoS Genet*.

DUJARDIN, G., LAFAILLE, C., PETRILLO, E., BUGGIANO, V., GÓMEZ ACUÑA, L., FISZBEIN, A., GODOY HERZ, M., NIETO MORENO, N., MUÑOZ, M., ALLÓ, M., SCHOR, I. & KORNBLIHTT, A. 2013. Transcriptional elongation and alternative splicing. *Biochim Biophys Acta*.

GINNO, P., LOTT, P., CHRISTENSEN, H., KORF, I. & CHÉDIN, F. 2012. R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters. *Molecular cell*.

GRANT, C., BAILEY, T. & NOBLE, W. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics*.

HAFNER, M., LANDTHALER, M., BURGER, L., KHORSHID, M., HAUSSER, J., BERNINGER, P., ROTHBALLER, A., ASCANO, M., JUNGKAMP, A.-C., MUNSCHAUER, M., ULRICH, A., WARDLE, G., DEWELL, S., ZAVOLAN, M. & TUSCHL, T. 2010. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *CELL*.

HANAHAN, D. & WEINBERG, R. 2011. Hallmarks of cancer: the next generation. *CELL*.

HINRICHS, A. S., KAROLCHIK, D., BAERTSCH, R., BARBER, G. P., BEJERANO, G., CLAWSON, H., DIEKHANS, M., FUREY, T. S., HARTE, R. A., HSU, F., HILLMAN-JACKSON, J., KUHN, R. M., PEDERSEN, J. S., POHL, A., RANEY, B. J., ROSENBLOOM, K. R., SIEPEL, A., SMITH, K. E., SUGNET, C. W., SULTAN-QURRAIE, A., THOMAS, D. J., TRUMBOWER, H., WEBER, R. J., WEIRAUCH, M., ZWEIG, A. S., HAUSSLER, D. & KENT, W. J. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res,* 34**,** D590-8.

HOELL, J., LARSSON, E., RUNGE, S., NUSBAUM, J., DUGGIMPUDI, S., FARAZI, T., HAFNER, M., BORKHARDT, A., SANDER, C. & TUSCHL, T. 2011. RNA targets of wild-type and mutant FET family proteins. *Nat Struct Mol Biol*.

HUELGA, S., VU, A., ARNOLD, J., LIANG, T., LIU, P., YAN, B., DONOHUE, J., SHIUE, L., HOON, S., BRENNER, S., ARES, M. & YEO, G. 2012. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep*.

JASKIEWICZ, L., BILEN, B., HAUSSER, J. & ZAVOLAN, M. 2012. Argonaute CLIP - A method to identify in vivo targets of miRNAs. *Methods*.

KANEHISA, M., GOTO, S., SATO, Y., KAWASHIMA, M., FURUMICHI, M. & TANABE, M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research*.

KHODOR, Y., RODRIGUEZ, J., ABRUZZI, K., TANG, C.-H., MARR, M. & ROSBASH, M. 2011. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes & Development*.

KHORSHID, M., RODAK, C. & ZAVOLAN, M. 2011. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res*.

KIM, N. & JINKS-ROBERTSON, S. 2012. Transcription as a source of genome instability. *Nat Rev Genet*.

KIM, P., YOON, S., KIM, N., LEE, S., KO, M., LEE, H., KANG, H., KIM, J. & LEE, S. 2010. ChimerDB 2.0--a knowledgebase for fusion genes updated. *Nucleic acids research*.

KISHORE, S., JASKIEWICZ, L., BURGER, L., HAUSSER, J., KHORSHID, M. & ZAVOLAN, M. 2011. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Meth*.

KLEIN, ISAAC A., RESCH, W., JANKOVIC, M., OLIVEIRA, T., YAMANE, A., NAKAHASHI, H., DI VIRGILIO, M., BOTHMER, A., NUSSENZWEIG, A., ROBBIANI, DAVIDE F., CASELLAS, R. & NUSSENZWEIG, MICHEL C. 2011. Translocation-Capture Sequencing Reveals the Extent and Nature of Chromosomal Rearrangements in B Lymphocytes. *CELL*.

KONIG, J., ZARNACK, K., ROT, G., CURK, T., KAYIKCI, M., ZUPAN, B., TURNER, D. J., LUSCOMBE, N. M. & ULE, J. 2011. iCLIP--transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. *J Vis Exp*.

KOVAR, H. 2011. Dr. Jekyll and Mr. Hyde: The Two Faces of the FUS/EWS/TAF15 Protein Family. *Sarcoma*.

KUHN, R. M., KAROLCHIK, D., ZWEIG, A. S., WANG, T., SMITH, K. E., ROSENBLOOM, K. R., RHEAD, B., RANEY, B. J., POHL, A., PHEASANT, M., MEYER, L., HSU, F., HINRICHS, A. S., HARTE, R. A., GIARDINE, B., FUJITA, P., DIEKHANS, M., DRESZER, T., CLAWSON, H., BARBER, G. P., HAUSSLER, D. & KENT, W. J. 2009. The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res,* 37**,** D755-61.

LAGIER-TOURENNE, C., POLYMENIDOU, M., HUTT, K., VU, A., BAUGHN, M., HUELGA, S., CLUTARIO, K., LING, S.-C., LIANG, T., MAZUR, C., WANCEWICZ, E., KIM, A., WATT, A., FREIER, S., HICKS, G., DONOHUE, J., SHIUE, L., BENNETT, C., RAVITS, J., CLEVELAND, D. & YEO, G. 2012. Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. *Nat Neurosci*.

LANGMEAD, B. 2010. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics,* Chapter 11**,** Unit 11 7.

LI, H., WATFORD, W., LI, C., PARMELEE, A., BRYANT, M., DENG, C., O'SHEA, J. & LEE, S. 2007. Ewing sarcoma gene EWS is essential for meiosis and B lymphocyte development. *J Clin Invest*.

LI, X. & MANLEY, J. 2005. Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *CELL*.

LIU, L., DESAI, S., LI, T., MAO, Y., SUN, M. & SIM, S. 2000. Mechanism of action of camptothecin. *Ann N Y Acad Sci*.

MARTIN, G., GRUBER, A. R., KELLER, W. & 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**,** 753-63.

MEYER, L., ZWEIG, A., HINRICHS, A., KAROLCHIK, D., KUHN, R., WONG, M., SLOAN, C., ROSENBLOOM, K., ROE, G., RHEAD, B., RANEY, B., POHL, A., MALLADI, V., LI, C., LEE, B., LEARNED, K., KIRKUP, V., HSU, F., HEITNER, S., HARTE, R., HAEUSSLER, M., GURUVADOO, L., GOLDMAN, M., GIARDINE, B., FUJITA, P., DRESZER, T., DIEKHANS, M., CLINE, M., CLAWSON, H., BARBER, G., HAUSSLER, D. & KENT, W. 2013. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic acids research*.

MONTECUCCO, A. & BIAMONTI, G. 2013. Pre-mRNA processing factors meet the DNA damage response. *Front. Gene*.

NAKAMURA, K., SAKAI, W., KAWAMOTO, T., BREE, R., LOWNDES, N., TAKEDA, S. & TANIGUCHI, Y. 2006. Genetic dissection of vertebrate 53BP1: A major role in non-homologous end joining of DNA double strand breaks. *DNA Repair (Amst)*.

OHNO, T., OUCHIDA, M., LEE, L., GATALICA, Z., RAO, V. & REDDY, E. 1994. The EWS gene, involved in Ewing family of tumors, malignant melanoma of soft parts and desmoplastic small round cell tumors, codes for an RNA binding protein with novel regulatory domains. *Oncogene*.

PARONETTO, M., MIÑANA, B. & VALCÁRCEL, J. 2011. The ewing sarcoma protein regulates DNA damage-induced alternative splicing. *Mol Cell*.

PAULSEN, R., SONI, D., WOLLMAN, R., HAHN, A., YEE, M.-C., GUAN, A., HESLEY, J., MILLER, S., CROMWELL, E., SOLOW-CORDERO, D., MEYER, T. & CIMPRICH, K. 2009. A genome-wide siRNA screen reveals diverse cellular processes and pathways that mediate genome stability. *Molecular cell*.

QUINLAN, A. & HALL, I. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*.

RIGGI, N., CIRONI, L., SUVA, M. L. & STAMENKOVIC, I. 2007. Sarcomas: genetics, signalling, and cellular origins. Part 1: The fellowship of TET. *J Pathol,* 213**,** 4-20.

ROGAKOU, E., PILCH, D., ORR, A., IVANOVA, V. & BONNER, W. 1998. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *J Biol Chem*.

ROGELJ, B., EASTON, L., BOGU, G., STANTON, L., ROT, G., CURK, T., ZUPAN, B., SUGIMOTO, Y., MODIC, M., HABERMAN, N., TOLLERVEY, J., FUJII, R., TAKUMI, T., SHAW, C. & ULE, J. 2012. Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. *Sci Rep*.

ROSENBLOOM, K., SLOAN, C., MALLADI, V., DRESZER, T., LEARNED, K., KIRKUP, V., WONG, M., MADDREN, M., FANG, R., HEITNER, S., LEE, B., BARBER, G., HARTE, R., DIEKHANS, M., LONG, J., WILDER, S., ZWEIG, A., KAROLCHIK, D., KUHN, R., HAUSSLER, D. & KENT, W. 2013. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic acids research*.

ROY, D. & LIEBER, M. R. 2009. G clustering is important for the initiation of transcription-induced R-loops in vitro, whereas high G density without clustering is sufficient thereafter. *Mol Cell Biol,* 29**,** 3124-33.

SANCHEZ, G., BITTENCOURT, D., LAUD, K., BARBIER, J., DELATTRE, O., AUBOEUF, D. & DUTERTRE, M. 2008. Alteration of cyclin D1 transcript elongation by a mutated transcription factor up-regulates the oncogenic D1b splice isoform in cancer. *Proc Natl Acad Sci USA*.

SCHWARTZ, J., EBMEIER, C., PODELL, E., HEIMILLER, J., TAATJES, D. & CECH, T. 2012. FUS binds the CTD of RNA polymerase II and regulates its phosphorylation at Ser2. *Genes & Development*.

SEO, J., KIM, S., LEE, H.-S., KIM, J., SHON, H., SALLEH, N., DESAI, K., LEE, J., KANG, E.-S., KIM, J. & CHOI, J. 2012. Genome-wide profiles of H2AX and γ-H2AX differentiate endogenous and exogenous DNA damage hotspots in human cells. *Nucleic acids research*.

STIRLING, P., CHAN, Y., MINAKER, S., ARISTIZABAL, M., BARRETT, I., SIPAHIMALANI, P., KOBOR, M. & HIETER, P. 2012. R-loop-mediated genome instability in mRNA cleavage and polyadenylation mutants. *Genes & Development*.

SUGIMOTO, Y., KÖNIG, J., HUSSAIN, S., ZUPAN, B., CURK, T., FRYE, M. & ULE, J. 2012. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol*.

THORVALDSDÓTTIR, H., ROBINSON, J. & MESIROV, J. 2012. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinformatics*.

WAHBA, L., AMON, J. D., KOSHLAND, D. & VUICA-ROSS, M. 2011. RNase H and multiple RNA biogenesis factors cooperate to prevent RNA:DNA hybrids from generating genome instability. *Molecular cell*.

WAHBA, L. & KOSHLAND, D. 2013. The Rs of Biology: R-Loops and the Regulation of Regulators. *Mol Cell*.

YEUNG, K. Y., FRALEY, C., MURUA, A., RAFTERY, A. E. & RUZZO, W. L. 2001. Model-based clustering and data transformations for gene expression data. *Bioinformatics,* 17**,** 977-87.

Chapter 6: Discussion

In this thesis, we aimed to identify the targets and understand the function of three RNA-binding proteins that up to this point were only poorly characterized. An important approach towards this goal was the analysis of high-throughput, particularly next generation sequencing data. A second important aspect was the integration of multiple experimental data sets that were available in the public domain.

We presented in the 2nd chapter an experimental approach, AGO CLIP, for the identification of *in vivo* targets of Argonaute proteins, and two probabilistic methods for the analysis of the resulting data. The only step that is specific for Argonaute however is the immunoprecipitation with the specific antibody, so the method can be generally applied to other RNA-binding proteins. In AGO CLIP, cells are initially grown in the presence of a photo-reactive nucleoside and subjected to UV radiation. This step is not absolutely necessary, but it allows very precise pinpointing of the crosslink sites due to the introduction of crosslink-diagnostic, T-to-C mutations during reverse transcription. After crosslinking of proteins to their target RNAs, cells are lysed and RNA-protein complexes are pulled down with an antibody. To fragment the RNAs for sequencing, the extracts are treated with nucleases, labeled with radioactive phosphate, run on a radiograph for size selection and then the protein is digested and the RNA purified. Finally, adaptor-ligated RNA species are reverse transcribed, amplified and sent to deep sequencing. The computational analysis then begins by removing the adaptors and aligning the short reads to the reference genome assembly. Reads are then functionally annotated, and finally statistical analysis is performed to identify high affinity binding sites of proteins on the target RNAs. We presented two probabilistic methods that we developed for this purpose; one is based on the enrichment of sequence reads with respect to a background library (e.g. an RNAseq sample obtained from the same cell line), that can be applied to any CLIPed protein, and the other based on crosslink-diagnostic mutations. Noting that at bona fide crosslinked positions the mutation frequency is neither too low nor too high (in the latter case it would rather indicate polymorphisms), the method rests on the question of whether or not a genomic position retains the empirically estimated crosslinking rate estimated based on most of the data. Of course, a main advantage of the latter method is it does not need an independent background sample. An RNAseq library could still be utilized however to reduce false positives originating from de facto polymorphic sites observed with a comparable frequency to the estimated crosslinking rates. Existing SNP (single nucleotide polymorphism) data could also be used for the same purpose (Jaskiewicz et al., 2012). Although we developed this method to take advantage to the T-to-C mutations that are quite frequent in PAR-CLIP, work in our group established that crosslink-diagnostic mutations are introduced in the 254 nm-based crosslinking that does not use photo-reactive nucleosides. Thus, our method can be applied to CLIP data in general (Kishore et al., 2011).

In Chapter 3, we studied PAPD5, a non-canonical poly(A) polymerase whose homolog in yeast has been shown to have nucleotidyltransferase activity and to function in RNA surveillance pathway. Finding the targets of PAPD5 required a non-trivial amount of effort because, presumably due to the rapid processing of PAPD5 targets by the exosome complex, the CLIP replicates had a high variance. We applied very stringent methods and identified rRNA as the main substrates of the protein. An independent study recently arrived at the same conclusion (Shcherbik et al., 2010).

In Chapter 4, we studied DIS3L2, one of the three human homologs of the core exosome exonuclease Dis3 that is involved in the degradation of its targets in yeast. DIS3L2 is the least characterized of the three homologs; like PAPD5, DIS3L2 is also involved in RNA surveillance and thus target identification is again hampered by the very rapid degradation of the targets and the variation between the replicates experiments. However, the most strongly reproducible result was the enrichment of tRNA fragments in the DIS3L2 PAR-CLIP samples. Thus, we were able to identify tRNAs as the main *in vivo* substrates of DIS3L2, and experimental work in the laboratory of our collaborators confirmed this computational prediction. Additionally, by analyzing Ago2-CLIP data obtained in our laboratory, we found clues that the fragments that DIS3L2 generates from tRNAs are loaded into the RNA silencing complex and may function in gene regulation.

The last PAR-CLIP data set analyzed in this thesis corresponded to a protein of a very different nature, EWSR1. EWSR1 is very well known from the fusions that it forms with transcription factors. They occur in very specific pathologies, namely the Ewing's sarcoma. The function of the normal protein was addressed by much fewer studies. It has been determined for example that EWSR1 is essential for mitotic integrity as well as for meiosis (Azuma et al., 2007, Li et al., 2007) and the protein has been implicated in alternative splicing of genes involved in DNA repair induced upon genotoxis stress (Paronetto et al., 2011). Analysis of EWSR1 PAR-CLIP data generated in our lab showed that EWSR1 binds its own pre-mRNA exactly at positions where fusions are usually generated. Additional analyses of the sequence and positional binding specificity of EWSR1 along pre-mRNAs lead us to hypothesize that EWSR1 binds preferentially on RNAs originating from instability-prone DNA regions; we found a strong enrichment of EWSR1 at subtelomeric regions, which are known hotspots of genomic rearrangement. Furthermore, we found a very significant correspondence between the regions where EWSR1 binds and the regions where H2AX, a histone variant that associates with double-stranded DNA breaks binds. These results implicate EWSR1 in the prevention of double-strand breaks at genomic regions that are intrinsically unstable; however the exact mechanism still needs to be determined.

In summary, being experimentally tedious and complex, when integrated with proper computational analysis, CLIP methods are powerful in the identification of one important component of post-transcriptional gene regulation, targets of RNA-binding proteins. More sophisticated methods still need to be developed for combinatorial modeling of the RNA-protein and regulatory RNA networks.

6.1 References

AZUMA, M., EMBREE, L. J., SABAAWY, H. & HICKSTEIN, D. D. 2007. Ewing sarcoma protein ewsr1 maintains mitotic integrity and proneural cell survival in the zebrafish embryo. *PLoS ONE*.

JASKIEWICZ, L., BILEN, B., HAUSSER, J. & ZAVOLAN, M. 2012. Argonaute CLIP - A method to identify in vivo targets of miRNAs. *Methods*.

KISHORE, S., JASKIEWICZ, L., BURGER, L., HAUSSER, J., KHORSHID, M. & ZAVOLAN, M. 2011. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Meth*.

LI, H., WATFORD, W., LI, C., PARMELEE, A., BRYANT, M. A., DENG, C., O'SHEA, J. & LEE, S. B. 2007. Ewing sarcoma gene EWS is essential for meiosis and B lymphocyte development. *J Clin Invest*.

PARONETTO, M. P., MIÑANA, B. & VALCÁRCEL, J. 2011. The ewing sarcoma protein regulates DNA damage-induced alternative splicing. *Mol Cell*.

SHCHERBIK, N., WANG, M., LAPIK, Y. R., SRIVASTAVA, L. & PESTOV, D. G. 2010. Polyadenylation and degradation of incomplete RNA polymerase I transcripts in mammalian cells. *EMBO Rep*.

Vita

Biter Bilen

biterbilen@gmail.com

EDUCATION

University of Basel, Basel, Switzerland, *Dec 2012*, PhD Bioinformatics. Dissertation: "Development of Methods for the Analysis of Deep Sequencing Data; Applications to the Discovery of Targets of RNA-binding Proteins", Advisor: Dr. Mihaela Zavolan

Bilkent University, Ankara, Turkey, *Jan 2007*, MS Molecular Biology and Genetics. Thesis: "Analyses and Web Interfaces for Protein Subcellular Localization and Microarray Gene Expression Data", Advisor: Dr. Rengül Çetin-Atalay

Middle East Technical University, Ankara, Turkey, *Jun 2004*, BS Computer Engineering. Project: "A Computer Graphics Framework for the Design and Simulation of High School Level Physics Experiments"

RESEARCH EXPERIENCE

University of Basel, Basel, Switzerland, *Jan 2013 – Jan 2014*. Postdoctoral Research with Dr. Mihaela Zavolan

University of Basel, Basel, Switzerland, *Dec 2007 – Dec 2012*. Graduate Research with Dr. Mihaela Zavolan

SIB PhD Training Network, Switzerland, *May 2009 – Dec 2012*. Graduate Research

German Cancer Research Center, Heidelberg, Germany, *Oct 2007 – Nov 2007*. Research Intern with Dr. Michael Boutros

Bilkent University, Ankara, Turkey, *Sep 2004 – Jan 2007*. Graduate Research with Dr. Rengül Çetin-Atalay and Dr. Özlen Konu

PUBLICATIONS

Bilen B, Martin G, Kishore S, Rammelt C, Zavolan M: **Genome-wide prediction of alternative polyadenylation using ARE binding proteins**. *manuscript in progress.*

Bilen B*, Kishore S*, Kunz C, Mittal N, Berger S, Wenzel F, Martin G, van Nimwegen E, Schaer P, Zavolan M:
Ewing sarcoma breakpoint region 1 protein binds G-rich RNAs and prevents transcription-associated genome instability. *manuscript in review in Cell Reports, *equal contribution.*

Ustianenko D*, Bilen B*, Chalupnikova K, Feketova Z, Martin G, Hrazdilova K, Zavolan M, Vanacova S: **Human DIS3L2 exonuclease is involved in the processing of tRNA-derived small RNAs**. *manuscript in review in EMBO Journal, *equal contribution.*

Yildiz G, Arslan-Ergul A, Bagislar S, Konu O, Yuzugullu H, Gursoy-Yuzugullu O, Ozturk N, Ozen C, Ozdag H, Erdal E, Karademir S, Sagol O, Mizrak D, Bozkaya H, Ilk HG, Ilk O, Bilen B, Cetin-Atalay R, Akar N, Ozturk M: **Genome-Wide Transcriptional Reorganization Associated with Senescence-to-Immortality Switch during Human Hepatocellular Carcinogenesis**. *PLoS ONE* 2013, **8**(5):e64016.
Hausser J, Syed AP, Bilen B, Zavolan M: **Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation**. *Genome Res* 2013.
Jaskiewicz L, Bilen B, Hausser J, Zavolan M: **Argonaute CLIP - A method to identify in vivo targets of miRNAs**. *Methods* 2012.
Rammelt C, Bilen B, Zavolan M, Keller W: **PAPD5, a noncanonical poly(A) polymerase with an unusual RNA-binding motif**. *RNA* 2011, **17**(9):1737–46.
Arziman Z, Engert C, Bilen B, Nickles D, Pelte N, Boutros M: **miRseq: An R package to analyze multiplexed small RNA libraries**. *unpublished*.
Bilen B, Atalay V, Ozturk M, Cetin-Atalay R: **Localizome analysis of *Homo sapiens* and model organisms**. *unpublished*.

HONORS

Werner Siemens Foundation Fellowship, *Dec 2007 – Apr 2011*. **TUBITAK* Research Scholarship**, *Apr 2005 – Jan 2007*.
*Turkish Scientific and Technical Research Council
Bilkent University Full-Tuition Scholarship, *Sep 2004 – Jan 2007*.
Seckinler Private Teaching Institution Scholarship, *Sep 2000 – Jun 2001*.
Ranked in top-600* in OSS**, *Jul 2000*. *among over 1 million students, **University Entrance Examination of Turkey
Ranked 4th in biology in TUBITAK Scientific Projects Competition, *May 1999*.

SKILLS AND EXPERTISE

High-throughput Sequencing Data Analysis. PAR-CLIP, mRNA-seq, ChIP-seq
Molecular Biology. regulatory RNAs, genome instability, mRNA processing
Computing. bioinformatics, statistics, machine learning, pattern recognition, image processing, computer graphics, software engineering, database systems
Programming. Perl, Python, BASH, R, MATLAB, Octave, C, C++, Java, SQL, SGE, PHP, LATEX
Language. Turkish, English, German