Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Characterization of *cis*-elements and *trans*-factors that are involved in RNAi-mediated genome defense in *Drosophila melanogaster*

Stefan Kunzelmann

2017

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Characterization of *cis*-elements and *trans*-factors that are involved in RNAi-mediated genome defense in *Drosophila melanogaster*

Stefan Bernhard Kunzelmann

aus

Coburg, Deutschland

2017

Erklärung

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Klaus Förstemann betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, 21.06.2017

Stefan Kunzelmann

Dissertation eingereicht am 3.7.2017 ....

1. Gutachter: Prof. Dr. Klaus Förstemann

2. Gutachter: PD Dr. Dietmar Martin

Mündliche Prüfung am 26.7.2017 ....

Abstract

Small RNAs are key regulators of eukaryotic gene expression and essential for genome maintenance and integrity. In somatic cells of the fruit fly *Drosophila melanogaster*, small interfering RNAs (siRNAs) are crucial for the repression of transposable elements (TEs), which represent a threat to the genomic stability. Fly siRNAs are generated by the RNase III enzyme Dicer-2 (Dcr-2) and loaded onto Argonaute 2 (Ago2) to fulfill posttranscriptional gene silencing.

Convergent transcription or transcription of inverted repeats can lead to endogenously derived dsRNA precursors, which are a substrate of Dcr-2. However, not all transposons match these criteria. How TEs are recognized to trigger antisense transcription is elusive. Using a GFP-based reporter system for *Drosophila* cells to reconstruct TE recognition and silencing, I tried to understand the prerequisites *in cis*, i.e. within the targeted locus, that trigger siRNA generation. I was able to show that there is a clear copy number dependence of siRNA generation. Neither the artificial combination of genetic elements *in cis* (such as an intron and the histone stem loop termination signal) nor impaired splicing reactions (as described in the fungus *Cryptococcus neoformans*) can overcome this requirement. Thus, the initiation of siRNA generation seems to be strictly regulated in order to prevent self-targeting siRNAs or superfluous efforts.

With CRISPR/Cas9-mediated genome engineering having evolved as a standard technique to modify DNA sequences, novel and artificial insertions may pose a challenge for genome integrity comparable to the activity of TEs. By exploiting a GFP-based reporter assay and sRNA-seq, I could demonstrate that *bona fide* siRNAs are generated upon insertion of homologous recombination donors that contain a selection cassette. These siRNAs target predominantly the inserted sequence but also spread to adjacent transcribed regions. Importantly, this RNAi response vanishes upon removal of the selection cassette and marker free tagging circumvents siRNA production altogether. Yet, this underlines that genome editing can trigger endogenous cellular defense mechanisms against the manipulation of the genome.

In addition to canonical RNAi biogenesis and effector factors such as Dcr-2, Ago2 and Loquacious, further proteins are involved in efficient RNA interference in flies. Blanks is a dsRNA binding protein, which is expressed predominantly in testes and in Schneider cells. Its depletion results in derepression of TE reporter constructs and transposon transcripts in cultured cells. However, Blanks is not involved in the processing and effector function of dsRNA derived from external sources. In addition, it has neither a strong influence on the abundance of endo-siRNAs mapping to TEs nor does it affect their loading onto Ago2. However, Blanks is crucial for the production of siRNAs derived from at least 12 distinct loci in the genome. Some of the loci are in proximity to annotated INE-1 element insertions, indicating that Blanks may be part of the defense against TEs. Due to its distinct expression pattern in flies that is limited to testes, Blanks may be necessary for the silencing of TE that are insufficiently repressed by the rather weak piRNA pathway in the male germ line. Interaction studies by Co-IP and MS analysis revealed nuclear import and export factors as potential interactors. This suggests that Blanks functions as a dsRNA export factor for selected substrates. Upon blocking of the nuclear re-import, Blanks accumulates in the cytoplasm, consistent with the hypothesis. Altogether, Blanks might be an RNAi factor that links the nuclear and cytoplasmic phases in genome defense.

Table of Contents

1 Introduction to RNA interference in *Drosophila*

1.1 RNA interference mediates cellular regulation and ensures genome integrity

Cells are faced with complex, environmental challenges that threaten the integrity of cellular function and genomic DNA sequence. Therefore, adaptive and reliable mechanisms are required to regulate gene expression, to fight invasion by selfish genetic elements.

Beside regulation mechanisms that are based on proteins, small RNAs are involved in such processes. One class of small RNAs, microRNAs, were originally described by (Lee et al., 1993) in the nematode *Caenorhabditis elegans*. The authors found that *lin-14* mRNA translation is regulated by a small, 22 nt long antisense RNA. The phenomenon turned out to be conserved in many eukaryotes (Fire et al., 1998).

Two protein-families proved to be key players in RNA interference (RNAi): Dicer-proteins and Argonaute-proteins. The Ribonucleases III (RNase III)Dicer is involved in the biogenesis of functional small RNAs by cleaving longer double-stranded RNA precursors (dsRNA) but it participates also in other processes such as Toll immune signaling (Wang et al., 2015b). RNase III enzymes are endoribonucleases that cleave dsRNA molecules and consist of nuclease domains, dsRNA binding domains, helicase domains and PAZ domains (Lamontagne et al., 2001). However, the effector function of the small RNAs to targets with complementary sequence is mediated by Argonaute proteins (Azlan et al., 2016; Ghildiyal and Zamore, 2009; Meister, 2013; Wilson and Doudna, 2013). Argonaute proteins are crucial for RNAi-mediated gene silencing but are also involved in other mechanisms such as transcriptional regulation and alternative splicing (Huang and Li, 2014). They can be loaded with small RNAs as well as interact with binding partners (e.g. GW proteins) to fulfill their effector functions (Meister, 2013).

The biogenesis of small RNAs and thus the RNAi mechanism has been intensively studied in the model organism *Drosophila melanogaster*. Three classes of small RNAs contribute to genetic regulation mechanisms: small-interfering RNAs (endo- and exo-siRNAs), microRNAs (miRNAs) and PIWI-interacting RNAs (piRNAs).

piRNAs are the guardians of germ cell genome stability. The 26-31 nt long piRNAs derive from heterochromatic regions that consist of multiple and varying transposon fragments which are called piRNA clusters. Despite their heterochromatic nature, these regions give rise to long, single-stranded RNA transcripts that eventually give rise to piRNAs. They may be amplified in a reaction loop called the ping-pong cycle. The Argonaute family proteins Ago3 and Aub are the key players of the amplification loop. Loaded in PIWI, another Argonaute family member, piRNAs are able to silence the activation and translocation of transposons (Hartig et al., 2007; Khurana and Theurkauf, 2010; Siomi et al., 2010; Siomi et al., 2011; Wang et al., 2015a). The coding capacity for piRNAs is stored in the above-mentioned master loci, heritable and changing "databases" of sequences that have to be repressed (Yamanaka et al., 2014).

1.2 miRNAs are involved in gene expression regulation

miRNAs derive from genomic loci and are predominantly transcribed by RNA polymerase II (Bartel, 2004). The resulting transcript is folded into a hairpin (pri-miRNA) and processed by the RNase III enzyme Drosha together with the dsRNA binding protein (dsRBP) Pasha into the shorter pre-miRNA (Denli et al., 2004), as depicted in Figure 1—1. The pre-miRNA is exported from the nucleus via Exportin-5 and the Ran gradient (Yi et al., 2003). In addition, introns of protein coding genes can give rise to pre-miRNAs. The so called mirtrons are debranched and serve as substrate for the following processing steps (Okamura et al., 2007).

In the cytoplasm, a complex of the RNase III Dicer-1 (Dcr-1) and dsRBP LoqsPB binds to the pre-miRNA and generates an approximately 22nt long duplex. LoqsPB is a splice variant of the *loquacious* gene and contains three dsRBDs (Forstemann et al., 2005). The miRNA/miRNA*-duplex is preferentially loaded onto Ago1 to build the RNA-induced silencing complex (RISC) (Forstemann et al., 2007). RISC binds to the 3'UTR of complementary cellular target mRNAs and inhibits translation initiation, destabilizes the transcript by deadenylation and thus induces its degradation (Bartel, 2009; Fukaya and Tomari, 2012). These functions are mediated by the GW-family proteins (Eulalio et al., 2009).

miRNAs are involved in several cellular processes such as development and cell fate decisions (Chawla and Sokol, 2011; Choi et al., 2013). Moreover, they participate in many housekeeping functions, regulate gene expression after environmental stress (Ghildiyal and Zamore, 2009) and take part in cell signaling (Luhur et al., 2013).

1.3 siRNAs fight viral infection and block transposable element activity

Contrary to miRNAs, exogenous long double-stranded RNAs (dsRNA), which appear upon viral infection, can be the substrate for the siRNA biogenesis pathway (Sabin et al., 2013). The resulting small RNAs are known as exo-siRNAs.

Dcr-2 together with the dsRBP R2D2 processes dsRNA precursors in a highly processive manner into 21 nt long siRNA-duplexes with the following characteristics: a 19 nt long perfect complementarity, a two nucleotide overhang at the 3′-end and a 5′ phosphate (Kandasamy and Fukunaga, 2016; Kim et al., 2006; Kim et al., 2009; Patel et al., 2006; van Rij and Berezikov, 2009). The duplex is loaded by the RISC-loading complex (RLC) consisting of Dcr-2 and R2D2 and by the chaperone Hsp70/90 into the RISC comprising of Ago2 and the siRNA guide strand (Tomari et al., 2004). Thereby, R2D2 determines the fate of guide and passenger strand by sensing the duplex formation energy of either end. The endonuclease C3PO facilitates separation of the strands of the duplex by cutting the passenger strand endonucleolyticly. The guide strand remains within RISC and fulfills its effector function by guiding it to cognate mRNAs, which are endonucleolyticly cleaved (Meister, 2013). The remaining fragments of the mRNA are degraded by the exosome and Xrn1.

Another class of siRNAs is represented by the endo-siRNAs whose function is mainly to suppress the harmful effects of transposable elements (TEs). According to current knowledge, dsRNA precursors are generated upon transcription from structured loci or pseudogenes, convergent or bidirectional transcription events and read-through transcription of antisense oriented transposons. They can be processed into endo-siRNAs by Dcr-2 (Ghildiyal and Zamore, 2009; Okamura et al., 2008a; Okamura et al., 2008b; Okamura and Lai, 2008; van Rij and Berezikov, 2009). The production of endo-siRNA is comparable to exo-siRNAs. However, Dcr-2 interacts with Loqs-PD, another isoform of the *loquacious* gene (Hartig et al., 2009; Hartig and Forstemann, 2011). The small RNAs are loaded onto the Ago2-RISC by Dcr-2 and R2D2 within the D2 bodies in the cytoplasm of the cells (Nishida et al., 2013). Additional experiments then showed that R2D2 and Loqs-PD are partially redundant and that RISC-loading still works – albeit at lower levels – in the absence of R2D2 (Mirkovic-Hosle and Forstemann, 2014).

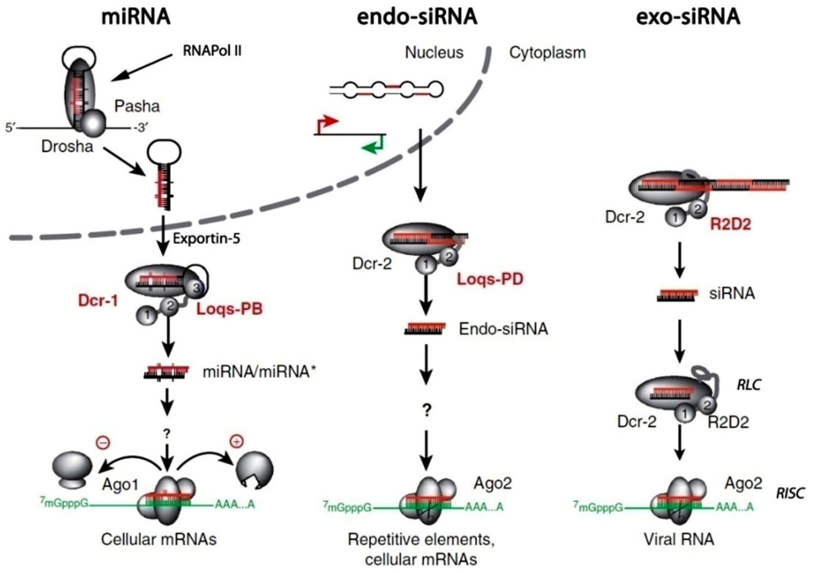

Figure 1—1 – Introduction to the biogenesis and effector pathways of miRNAs and siRNAs. miRNAs and endo-siRNAs come from endogenous sources, whereas exo-siRNAs are produced from exogenous dsRNA precursors introduced via viral infection. The long dsRNA precursors are processed by RNase III enzymes (Drosha, Dcr-1 and Dcr-2) that interact with a dsRBP (e.g. R2D2 or Loquacious). The regulative function of the RISC is mediated via complementarity of the small RNA to the target RNA and the endonuclease activity of the Argonaute proteins (Ago1 and Ago2). Figure modified from (Hartig et al., 2009).

2 Specific Aims

The goal of this thesis was to gain a deeper understanding of the RNA interference pathways in *Drosophila melanogaster*. On the one hand, I investigated if and how genetic elements *in cis* can stimulate the generation of siRNAs. On the other hand, I tried to characterize the non-canonical dsRBP and RNAi factor Blanks and how it is involved in RNAi *in trans*. In parallel, I participated in the development and establishment of methods such as genome editing or interactomics in order to study the above mentioned biological processes.

In general, *Drosophila* is well suited for the study of RNAi due to its separated biogenesis pathways of miRNAs and siRNAs, whose generation is often intertwined in other species. In flies, the biogenesis of siRNAs can be studied by depleting Dcr-2 and other siRNA biogenesis factors without affecting the miRNA pathway. Thus, no global deregulation of gene expression occurs.

The following projects were conducted and are described in this thesis:

- First, I present my contributions to the CRISPR/Cas9-mediated genome editing protocol that was developed in our lab. I established reagents to generate shut down cell lines that can be used to generate conditional knockdowns of specific genes; I also generated a number of such cell lines.

- In the next chapter, so far unknown side effects of genome engineering are examined in the context of siRNAs that target the modified chromosomal locus. Fortunately, I was able to show that the selection cassette is responsible for triggering the RNAi response and that this effect can be reverted by removing the marker.

- To understand transposon recognition, I investigated the prerequisites *in cis* (i.e. on the local genomic sequence level) that are able and/or necessary to stimulate a siRNA response. To this end, I used GFP-based reporter assays and deep sequencing.

- Moreover, I developed and robustly established a protocol that allows the identification of the interactome of epitope tagged proteins. *In vivo* cross-linking, immunoprecipitation and mass spectrometry based readout were optimized to provide a tool for functional analysis of RNAi factors and their associated protein partners.

- Finally, I applied several of the developed methods to answer the question how the recently discovered RNAi factor Blanks mechanistically contributed to the RNAi pathway. I was able to show that Blanks is a potential dsRNA export factor and might be part of a so far uncharacterized mechanism that is essential for the generation of siRNAs from distinct genomic loci.

Parts of these projects are already published; this is annotated at the relevant positions.

3 Expanding the CRISPR/Cas9-mediated genome editing protocol for cultured *Drosophila* cells

Parts of this chapter are published as:

> Kunzelmann et al. "A Comprehensive Toolbox for Genome Editing in Cultured Drosophila Cells" *G3 (Bethesda)* (2016) 6:1777-1785.

3.1 Introduction

The CRISPR/*cas* system has evolved as a bacterial anti-viral defense system and consists of the Cas proteins and the CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) locus. The Cas proteins are the mediators of bacterial immunity against phages. Their CRISPR loci comprise an array of spacers that contain the information about the target sequences, separated by repeats that usually can fold into hairpins and fulfill a structural role for interaction with the CRISPR associated (Cas) proteins. During a process that is called acquisition, parts of the viral genomes are integrated into the CRISPR locus to trigger an inheritable defense mechanism against the corresponding phage.

The long transcript that derives from this CRISPR array (pre-crRNA) is cleaved into shorter crRNAs which are incorporated into a Cas protein family nuclease (Figure 3—1A). The resulting complex is then directed via base-pairing with the complementary sequence to its target locus, the DNA of phages that infect the cell. The nuclease introduces a DNA double-strand break (DSB) which results in destruction of the target DNA (Bhaya et al., 2011). Thus, evolution has provided us with an RNA-programmable nuclease that induces DNA double-strand breaks.

In recent years, genome editing with the help of the CRISPR/*cas* system has become an indispensable tool for molecular biology. The CRISPR-"revolution" has shifted our attention from cloning work in the context of plasmids or bacterial artificial chromosomes to the modification of genes in their chromosomal context. Tagging at the genomic locus offers several advantages compared to the transient expression of tagged proteins: The expression levels of the epitope tagged proteins are similar to the endogenous protein and the risk of overexpression artifacts is limited. Moreover, the expression of the untagged protein is reduced or, if all alleles are modified, absent. This results in less competition of tagged and untagged proteins for incorporation into the relevant complexes. Finally, this approach allows for stable and long-term expression of the modified proteins.

Especially the Cas9 nuclease that derives from *S. pyogenes* has been engineered into a widely used enzyme to generate defined DSBs in eukaryotic organisms, the first step of genome editing (Jiang and Doudna, 2017). In the case of the Cas9 enzyme, the RNA component consists originally of two distinct molecules, the crRNA and an additional tracrRNA. For experimental convenience, a fusion of both, the so-called sgRNA, is used to program the Cas9 nuclease. The sgRNA codes for the target sequence (CRISPR sequence) which is followed by a stemloop structure that is necessary for the incorporation into the Cas9 enzyme.

As depicted in Figure 3—1B and C, the protein component of the enzyme recognizes the so called PAM (protospacer adjacent motif) sequence (NGG) on the target DNA. Two distinct arginines (R1333, R1335) interact with the guanosines of the DNA in a sgRNA independent manner and position the nuclease on the DNA strand(Anders et al., 2014). In a second step, an interaction between the CRISPR sequence and the nucleotides upstream of the PAM occurs when the sequences are complementary to each other, followed by the cleavage of the dsDNA between the third and fourth nucleotide upstream of the PAM(Sternberg et al., 2015). This two-step recognition mechanism prevents the Cas9 enzyme from targeting the bacterial CRISPR locus, which of course also contains a perfect match to the CRISPR sequence. Because no PAM is encoded within the CRISPR repeat sequence, an interaction between Cas9 and the bacterial DNA is prevented.

A sequence example is depicted in Figure 3—1B to illustrate the features of a targeted locus. The CRISPR sequence (black box) that targets the C-terminal end of the *act5C* locus is adjacent to the PAM (aGG), which is located 2nt downstream of the stop codon (TAA).

The introduced DSB can be repaired either by an error-prone pathway (end-joining activities), potentially leading to targeted mutagenesis, or by homology directed repair (Figure 3—1D). The latter is normally free of errors (when the sister chromatid serves as donor) but offers the possibility to introduce an experimentally provided, custom-modified homologous recombination (HR) donor into the desired locus (Doudna and Charpentier, 2014). Various methods that combine programmable cleavage of DNA by CRISPR/Cas9with artificial homologous recombination donors in cell culture have been described, e.g. (Bassett et al., 2014; Byrne et al., 2014; Böttcher et al., 2014; Fetter et al., 2015; Fu et al., 2014; Li et al., 2014; Wyvekens et al., 2015). These strategies can be broadly grouped according to the particular type of HR donor material employed: Cloned homology arms, single-stranded synthetic oligonucleotides or PCR-products with flanking homology regions.

In the Förstemann group, we have previously presented a protocol for cultured *Drosophila* cells that applies PCR to generate both, an expression cassette for the Cas9-programming sgRNA (CRISPR construct, 19 nt sequence homology to the target locus) and HR donors for selectable genome modification (Figure 3—2). The PCR products are transfected into *Drosophila* S2 cells which stably express the Cas9 enzyme. The CRISPR construct can be transcribed into the sgRNA by RNA-Pol III and is then loaded into Cas9 to introduce the cut. The HR donor contains the sequence that codes e.g. for an epitope tag and the Blasticidin resistance as a selection marker. This cassette is flanked by 60nt long homology regions of the desired locus to mediate integration of the HR donor via homologous recombination repair of the DSB. After enrichment of cells that have successfully integrated the HR donor(via antibiotic-containing medium), the resistance marker can be removed via the Flp/FRT system (Böttcher et al., 2014).

Here, I present the introduction of a second resistance cassette and the application of the N-terminal tagging approach to generate conditional knockdown cell lines.

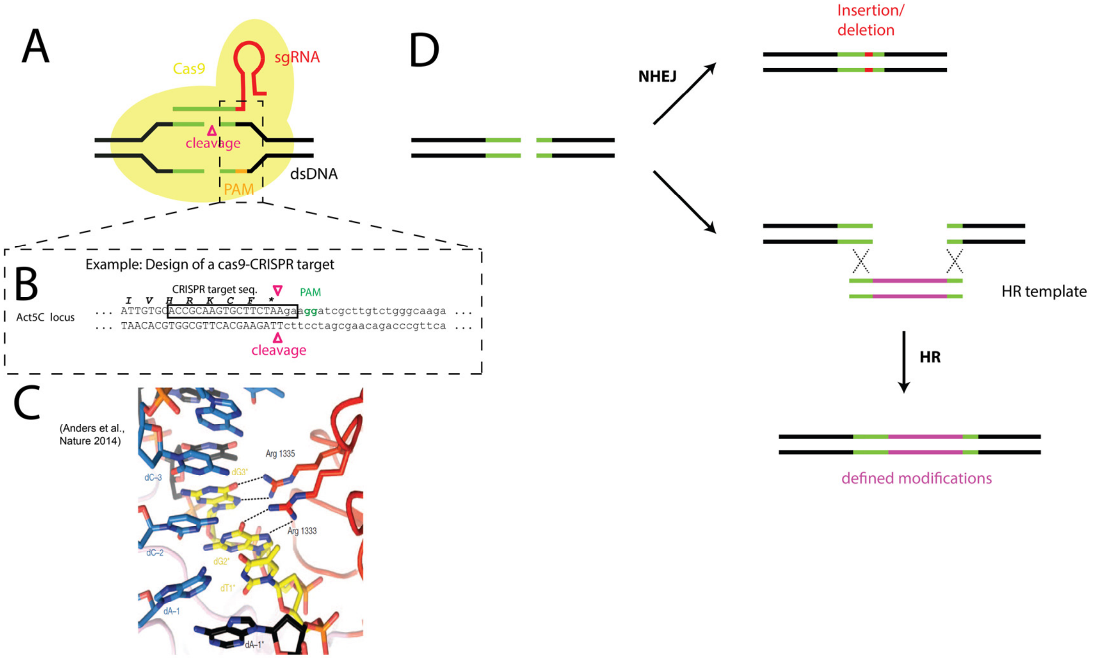

Figure 3—1: Schematic representation of the Cas9-mediated introduction of the DSB. (A) The complex of Cas9 nuclease (yellow) and sgRNA targets the dsRNA via the PAM (orange) interaction followed by base pairing between the CRISPR sequence (green) and the complementary target sequence. Cleavage of the dsDNA occurs. (B) As an example, the Cas9-CRISPR target of the *act5C* locus is depicted. A PAM is located close to the stop codon. The CRISPR sequence is marked by a black box. The position of the cleavage is indicated by red arrowheads. (C) Illustration of the structural features underlying the PAM recognition process. R1333 and R1335 establish hydrogen bonds with the two guanosines of the PAM and position the enzyme on the dsDNA. (D) After the introduction of the DSB, the DNA lesion can be repaired either by the potentially error-prone end-joining pathway (e.g. NHEJ) or via homologous recombination. The latter pathway enables the integration of heterologous sequences at the locus.

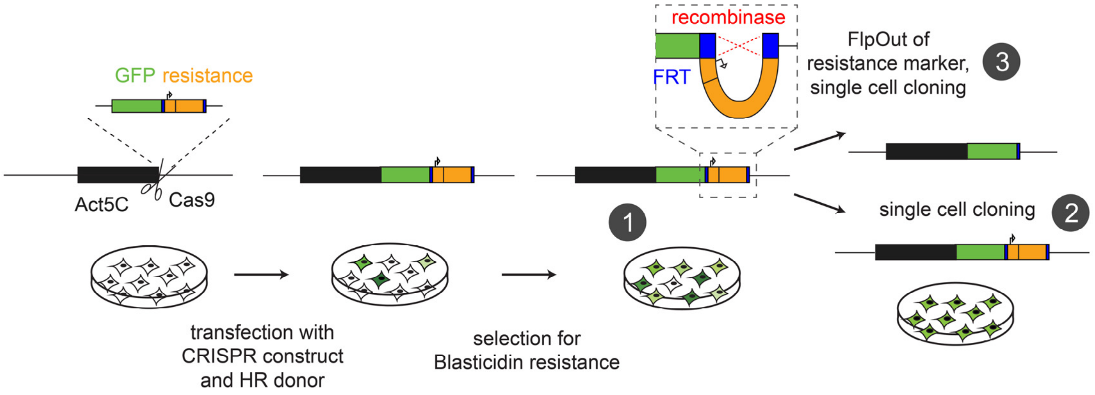
GFP resistance
Act5C
Cas9
recombinase
FRT
FlpOut of resistance marker, single cell cloning
3
1
single cell cloning
2
transfection with CRISPR construct and HR donor
selection for Blasticidin resistance

Figure 3—2: Workflow of the genomic tagging process. Cells stably expressing Cas9 are transfected with the sgRNA template and the HR donor. The Cas9 enzyme introduces a DNA double strand break at the desired locus, where HR-mediated repair can introduce the HR donor which contains the tag (GFP) and a selection cassette. Positive clones are enriched using the selection marker which results in a heterogeneous cell population (1). By single cell cloning, a clonal cell line can be generated (2). When transfecting the cells with a plasmid coding for the Flp recombinase, the selection cassette can be removed; subsequent single cell cloning can generate clonal cell lines that contain the genomic tag but no selection marker (3).

3.2 Results and discussion

3.2.1 The use of the puromycin resistance as an alternative selection marker

To enable straightforward introduction of a second epitope tag (e.g. for co-immunoprecipitation studies), we developed an independent selection cassette based on the puromycin acetyl transferase gene. Since the commonly used coding sequence of this gene proved to be rather refractory to PCR amplification (likely due to a high GC content), I created a *copia*-Puro resistance cassette as a synthetic gene. This element could be readily amplified by PCR and we have generated alternative versions of most template vectors, which are used during PCR for the generation of the HR donor, by exchanging the *copia*-Blast cassette with the *copia*-Puro marker. This allows the introduction of FLAG, V5, GFP or Strep epitope-tags either at the C-terminus or the N-terminus of proteins by selection for puromycin resistance.

I observed that our S2cell line is quite sensitive to puromycin. Selection works well at a concentration of 0.5 µg/ml of puromycin in Schneider's medium (Figure 3—3A). I quantified both the amount of dying cells and the density of the cells in the culture dish in order to determine the optimal selection concentration of the drug. The fraction of dying cells can be roughly estimated using flow cytometry measurements. While the forward scatter (FSC) describes the size of the cells, the side scatter (SSC) represents the granularity. Cells that do not withstand the selection pressure either due to overly concentrated puromycin or the absence of the resistance gene, will die. Dead cells have increased FSC and SSC values and are separated well in a FSC-SSC scatter plot from healthy cells. During analysis, the amount of cells whose FSC-SSC values in the scatter plot differ from the untreated control can be quantified by creating appropriate analysis regions. Moreover, the density of the cells in the culture disc is measured indirectly during flow cytometry as cell count events per second. Exploiting both methods, it becomes evident that 0.5 µg/ml puromycin is sufficient to kill naïve S2 cells and also blasticidin resistant cells, while no dying cells can be detected for the puromycin resistant cell line (dashed lines). The resistant cells are proliferating well, while the other cell lines have dramatically lower cell densities (survival plot, solid lines). These findings were validated by an independent readout, where the optical density was determined by eye and used as a proxy for selection success (Figure 3—3B). Again, 0.5 µg/ml is sufficient for successful selection.

Integration of a puromycin-construct in an already blasticidin-resistant cell line (or vice-versa) is efficient and does not require any changes to the protocol. I was not able to detect any cross resistance between the two markers.

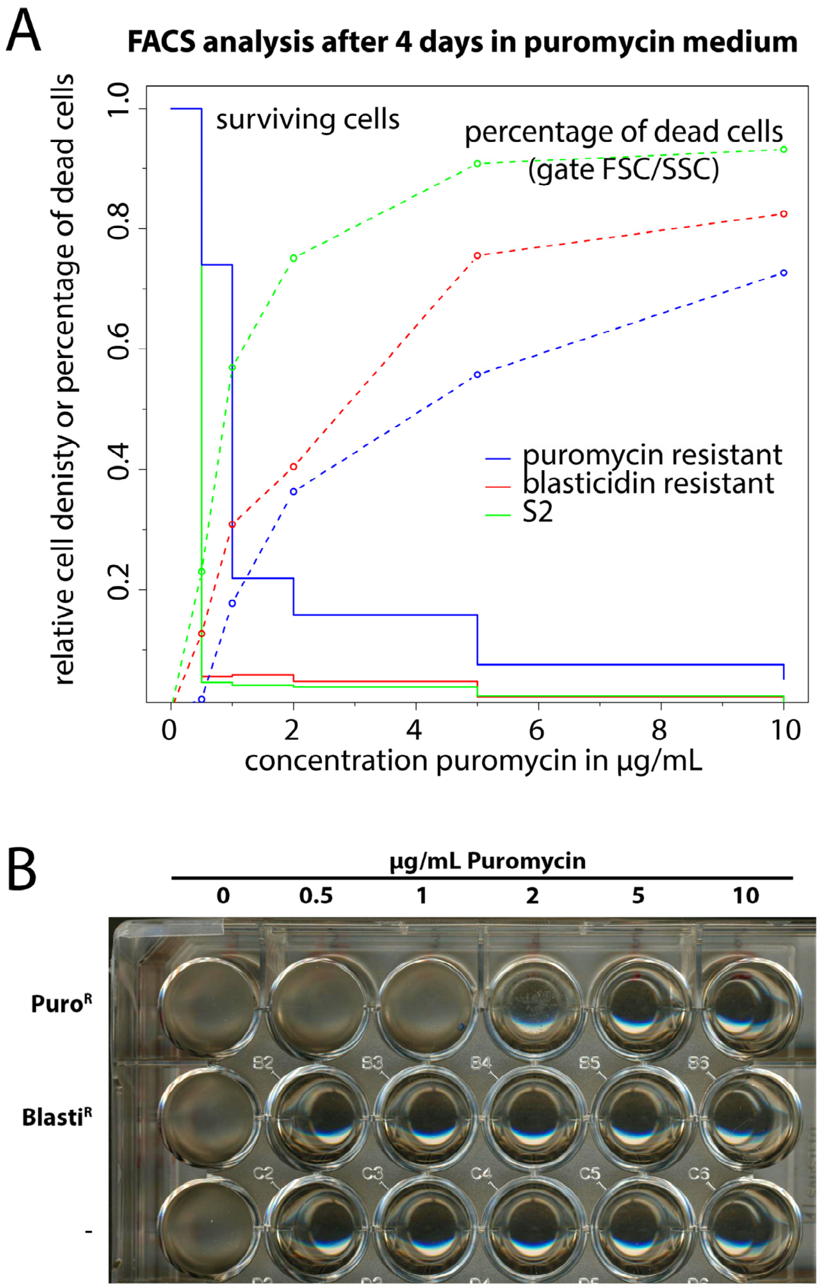

Figure 3—3: Assays to determine the optimal puromycin concentration to select successfully modified cells. (A) Flow cytometry data for puromycin resistant, blasticidin resistant or naïve cells after addition of various amounts of puromycin to the cell culture medium. FSC-SSC scatter plots of the samples were used for the quantification which is depicted. The percentage of dead cells (dashed lines) anti-correlates with the cell density of the surviving cells (survival graphs; solid lines). (B) Macroscopic readout of the cell density of puromycin resistant, blasticidin resistant or naïve cells (-) after addition of various amounts of puromycin to the cell culture medium.

3.2.2 The N-terminal tagging approach allows for inducible expression from the modified locus

Because our tagging protocol allows robust tagging of C-termini of proteins, we wanted to expand it by tagging N-termini of proteins via the same general principle. To this end, a new template vector design had to be developed that also contains a promoter for heterologous expression of the tagged gene (Figure 3—4A). This is necessary because in the case of an N-terminal tag, the selection cassette separates the endogenous promoter from the gene body. We chose the inducible, bi-directional *mtnDE* (metallothionein) promoter for this purpose. Induction is possible by adding e.g. $CuSO_4$ to the growth medium and the promoter can drive expression of the selection marker and the tagged protein concomitantly. As for the C-terminal approach, our vector templates contain constant regions for annealing of the homology-containing targeting primers during PCR. All N-terminal tags and selection constructs can thus be amplified with a single set of homology-containing primers. We have developed N-terminal vector templates for various epitope tags. Since the selection cassette and the *mtnDE* promoter are flanked by FRT-sites in all vectors, it can be removed with FLP recombinase to restore expression control via the endogenous promoter.

If all alleles for a given gene in the S2 cell genome have been modified, the introduced *mtnDE* promoter allows $CuSO_4$-dosage dependent heterologous control over the expression of the targeted gene. As an example, we derived cell lines with N-terminally FLAG-tagged Blanks protein. After clonal selection, I could readily identify cells that carried only modified *blanks* alleles using PCR reactions. This cell line allowed me to tune the expression level of Blanks; analysis by RT-qPCR demonstrated that both transcription shutdown and overexpression situations can be obtained (Figure 3—4B and C). A potential caveat is that the *mtnDE* promoter in the non-induced state appears to be partially leaky. The remaining transcript levels (12 % of wt levels) are comparable to those of an efficient knockdown of the gene (7 % of wt levels). This leakiness likely varies according to the genomic integration site due to local epigenetic marks, nearby enhancers or promoter strength of the endogenous gene.

By titrating copper ions to the cell culture medium the expression of *blanks* can be stimulated continuously. Directly after addition of the inducer, the mRNA levels of *blanks* increase dramatically up to 15-fold above the wt levels and decrease to steady-state levels after approximately one day. The first burst in transcription, however, may be due to a global up-regulation of transcription as a response to the challenge with heavy metal ions, since the transcription of endogenous *blanks* in the control cells increases slightly as well.

Besides the Blanks shutdown cell line (SD cell line, FLAG-Blanks A2), we have generated a Dcr-2 SD cell line (GFP-Dcr-2 #1) in the lab that is also used in further projects of this study.

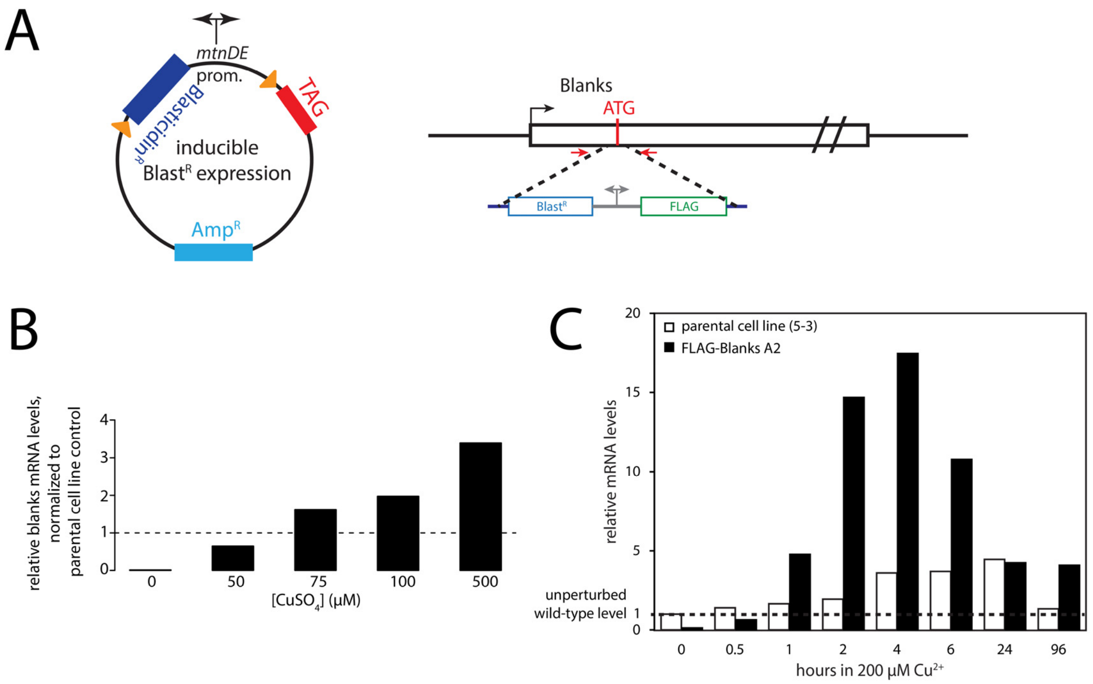

Figure 3—4: The N-terminal tagging approach can be used to generate shutdown cell lines if all alleles of a locus are modified. (A) The HR template vector design of the N-terminal constructs contains an inducible, bidirectional *mtnDE* promoter which is flanked by the resistance cassette and the epitope tag. This region can be amplified by PCR with primers that contain homology regions to target locus, see right panel. (B) Titration of copper and its effect on the induction of *blanks* in FLAG-Blanks A2 cells. Cells were harvested 5 days after induction. Transcript levels were determined by RT-qPCR. Data was normalized to the parental cell line using the Δct-method. (C) Induction kinetics of FLAG-Blanks A2 cells. Cells were cultured for two weeks in medium without copper and *blanks* expression was induced by the addition of 200 µM copper. After one hour the mRNA levels increase. Transcript levels were determined by RT-qPCR. Data was normalized to the house-keeping gene *rp49* using the ΔΔct-method.

3.3 Conclusions

During my studies, I was able to participate in simplifying the genome editing protocol and to introduce a second selection marker for more flexibility during the tagging process. Moreover, I characterized the features of the N-terminal tagging approach and was able to show that SD cell lines with tunable expression can be generated if all alleles are modified. The presented data demonstrate the usefulness of these SD cell lines and the versatility of the N-terminal tagging approach. For simple loss-of-function studies, RNAi is by far easier to apply. However, the *mtnDE* promoter “alleles” may present an interesting tool to study genetic interaction in combination with RNAi of a second factor. In particular, they may be convenient to create hypomorphic expression levels of essential genes in order to make them amenable for synthetic genetic screens. So, the SD cell lines are a powerful tool for genomic loss-of-function studies when RNAi cannot be used, e.g. to avoid circular arguments when knocking down factors of the RNAi pathway or to avoid altering the endogenous siRNA composition of the cells.

The genomic tagging system now offers significantly increased functionality, including the possibility to verify protein-protein interactions via co-immunoprecipitation. Both the N- and C-terminal template vectors can easily be modified using e.g. restriction enzyme based cloning to harbor other tags, fluorescent proteins or elements for genome functionalization. For example, I generated template vectors that allow the introduction of an *attP* target site based on the C-terminal epitope tag template series of vectors.

We have not quantified the tagging efficiencies with puromycin resistance based constructs in a manner analogous to the experiments described in our publications (Böttcher et al., 2014; Kunzelmann et al., 2016). The tagging success rates clearly depend on optimal sgRNA length and the extent of homology arms in the HR donor PCR product as published for the blasticidin constructs; since these elements are independent of the chosen marker, we do not expect major quantitative differences between puromycin and blasticidin based selections.

In principle, the two constructs could also be integrated in parallel rather than sequentially. If one begins with an inducible blasticidin resistance construct for an N-terminal tag and then continues with constitutive puromycin and blasticidin resistance cassettes, up to three epitope tags can be combined without the need to FLP out the marker in between.

In general, the knockout of genes should also easily be possible by exchanging at least parts of the CDS by a knock-in of the resistance cassette. For convenient detection of successful cassette exchange, GFP could be used as a second marker beside the resistance gene. Blasticidin and puromycin resistance can be concomitantly used to increase the number of targeted alleles. All in all, this shows that our PCR-based tagging approach is highly versatile and can be expanded on for further applications.

We estimate that it should be straightforward to extend our strategy to other *Drosophila* cell culture systems, potentially even to cultured cells from other insect species. Related PCR-based approaches have been described for use in cultured vertebrate cells (Li et al., 2014; Stewart-Ornstein and Lahav, 2016). We expect that our vector templates can be modified for use beyond insect cells by exchanging the *copia*-promoter that drives the expression of the selection gene and/or the inducible *mtnDE* promoter with sequences of corresponding functionality in e.g. vertebrate cells. Perhaps even more importantly, it may be possible to transfer the conclusions from our optimization efforts to other cell culture systems as well.

4 Reversible perturbations of gene regulation after genome editing in *Drosophila* cells

Parts of this chapter are published as:

> Kunzelmann and Förstemann "Reversible perturbations of gene regulation after genome editing in Drosophila cells" *PLOS One* (2017), *in press*

4.1 Introduction

As already mentioned, the CRISPR/*cas*-system has become an indispensable method to manipulate genomes with little effort and few side effects. It allows for the generation of mutant chromosomal loci as well as epitope tag knock-ins. Concerns were raised about possible off-target effects and their consequences on experimental results (Lin and Potter, 2016; Zhang et al., 2015). In contrast, we know little about how organisms deal with the on-target manipulation once it is in place. Do the cells "recognize" inserted sequences and respond to these foreign elements?

The artificial manipulations bear similarities with transposable elements (TEs), which are naturally occurring insertion events that threaten genomic stability. TEs code for enzymes that mobilize and re-insert them in new genomic locations. Cells have developed several defense strategies to suppress transposition (Levin and Moran, 2011). As detailed in the general introduction, the RNA interference (RNAi) pathway is responsible for the posttranscriptional silencing of TEs in somatic cells of *Drosophila melanogaster*.

The aim of this study was to address the question how the cells deal with genetic manipulations introduced via the CRISPR/Cas9-mediated genome editing approach. To this end, I used *Drosophila* S2 cells as a model system and the epitope tag knock-in protocol of the Förstemann group as described (Böttcher et al., 2014; Kunzelmann et al., 2016).

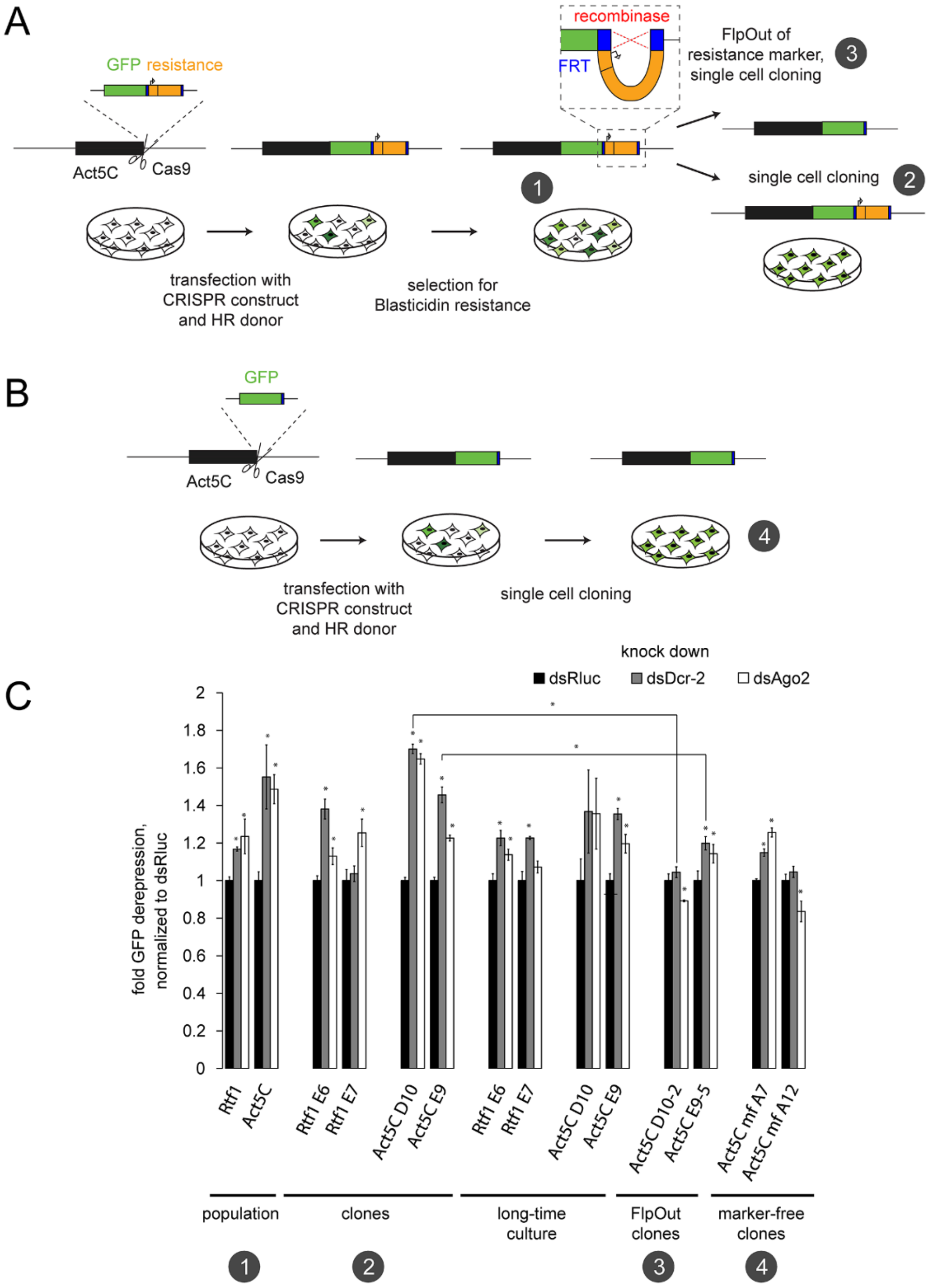

Figure 4—1: GFP-based reporter assay can detect siRNA mediated repression after genome engineering. (A) PCR-based tagging workflow using CRISPR/Cas9 in *Drosophila* Schneider cells. After introducing a DSB at the *act5C* locus by the Cas9 enzyme, the HR template (consisting of homology regions, the GFP coding sequence and the resistance cassette) integrates and GFP-positive cells can be eriched by drug selection [1] and cloned [2]. The FLP recombinase mediates the FlpOut of the resistance cassette and subsequent single cell cloning results in FlpOut clones [3]. (B) Marker-free tagging of the *act5C* locus with GFP. Similar to (A), the *act5C* locus can be tagged without an selection marker. Single cell cloning resulted in homogeneous cell lines [4]. (C) GFP-based reporter assay detecting the presence of functional siRNAs in several cell lines. Knockdown of Dcr-2 and Ago2 as key players of the RNAi pathway leads to derepression of the GFP fluoresence in the twoAct5C-GFP and Rtf1-GFP cell lines. Fluorescence levels (FL1 channel) were normalized to control knockdown (Rluc). Error bars represent standard deviation (n = 3). Significant differences were determined by applying upaired t-test (unequal variance) on the data (* p < 0.05).

4.2 Results and discussion

4.2.1 Functional siRNAs target integrated epitope tag cassettes

The previously developed CRISPR/Cas9-mediated genome editing workflow for *Drosophila* cell culture that allows the introduction of epitope tags adjacent to the coding sequences of genes at their chromosomal loci is depicted in Figure 4—1A and B. After enrichment of positive cells by antibiotic selection, the resistance marker can be removed via Flp/FRT. In order to study the potential of modified loci to trigger siRNA generation, we introduced a C-terminal GFP-tag at the *act5C* and *rtf1* loci in S2 cells. If these foreign sequences are targeted by siRNAs, then the GFP-fusion proteins should be de-repressed upon inactivation of the RNAi pathway. I thus monitored GFP expression with flow cytometry. Knockdown of the siRNA biogenesis enzyme Dcr-2 as well as the effector protein Ago2 resulted in derepression of the Act5C-GFP and Rtf1-GFP fusion proteins. This effect was less than two-fold, already visible in the cell population after one split into selective medium and remained after clonal selection. Even after prolonged cultivation of these cell lines (approximately 12 weeks) without selection pressure, the effect did not vanish (Figure 4—1C, stages 1 and 2). This argues for a stable situation that is not transiently triggered by the induced DNA double-strand break.

I sequenced the small RNA profile of the genome-engineered cell lines and mapped the reads back to the modified loci. This provided direct evidence for the presence of small RNAs in sense and antisense orientation targeting the *act5C* locus (Figure 4—2) or the *rtf1* locus in cells of the drug-selected population as well as single cell clones. I first examined the size distribution of the reads that were mapped to the locus. They showed a clear peak of 21 nt long reads in sense and antisense orientation (Figure 4—3). Together with their Dcr-2 and Ago2 dependent activity, this argues for *bona fide* siRNAs.

Since sense matching reads can also be mRNA degradation products, we quantified the strength of the siRNA response by summing up only antisense reads mapping to either the HR integrate (= the HR donor after integration), the upstream sequence or the downstream sequence of this locus (Figure 4—4 and Figure 4—5). The majority of siRNAs derived from the HR integrate, but reads also mapped upstream of the integration site. In particular, we found reads in antisense orientation that span the junction between the HR integrate and the *act5C* host gene (Figure 4—2D). This suggested that the dsRNA precursor of the siRNAs extends beyond the inserted sequence and excludes off-target integration events being the sole source of those siRNAs. The strength of the siRNA response decreased after clonal selection compared with the initial drug-selected population after genome editing. Nevertheless, the measurement of GFP fusion protein levels proved the potential of the remaining siRNAs to act as repressors (Figure 4—1C). It depends on the particular situation if these small changes in expression levels can interfere with experimental results and introduce biases to studies. Nevertheless, they may be an indicator that further epigenetic changes may have occurred at the modified locus.

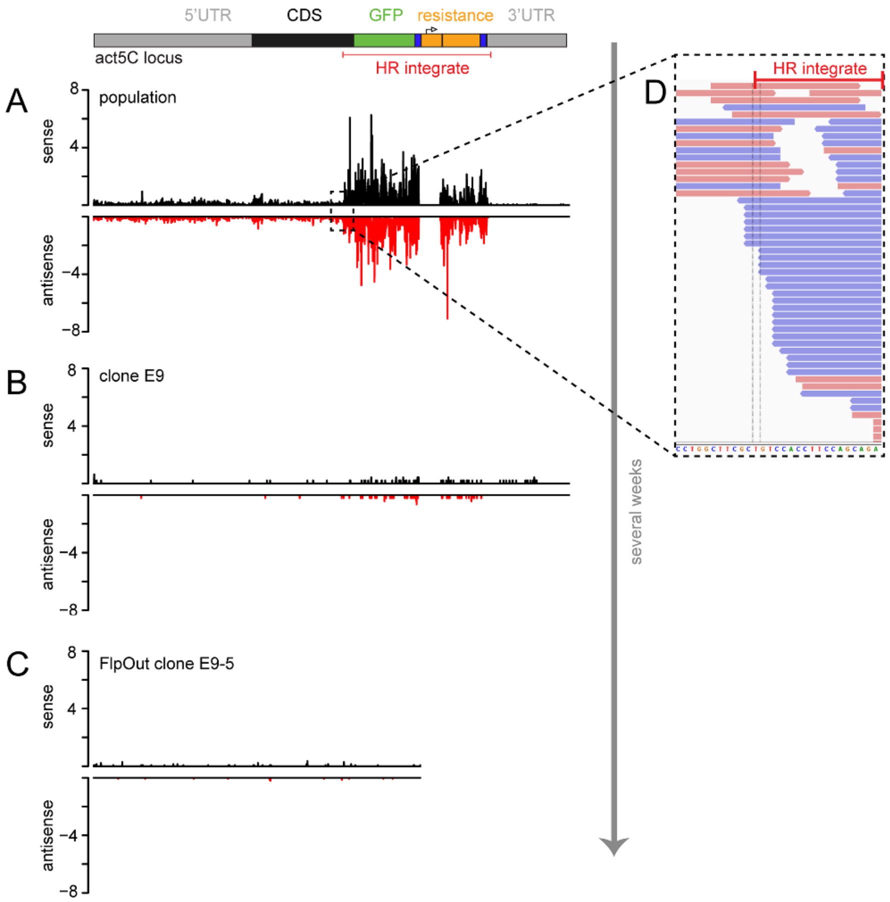

Figure 4—2: Profiling of siRNAs after genome editing by deep sequencing at the *act5C* locus. The siRNA distribution along the modified *act5C* locus was determined by binning into 1 nt intervals and normalized to the number of genome-matching reads in each library. The graphs depict the sense (black) and antisense (red) matching reads as reads per million of genome matching 19-25 nt reads in the respective library. Shown are the sequencing traces for the initial drug-selected population (A), the single cell clone E9 (B) and the respective FlpOut clone E9-5 (C) as representive examples. At the top, a scheme depicts the functional regions of the locus (drawn to scale); the HR donor is annotated in red. Reads derived from the *copia* promotor sequence are removed prior to mapping the remaining reads to the construct. Thus, the corresponding region seems to be masked. The box (D) shows the magnification of the transition between the endogenous sequence and the HR integrate (annotated red bar). Spanning siRNA reads in sense (red) and antisense (blue) orientation can be detected.

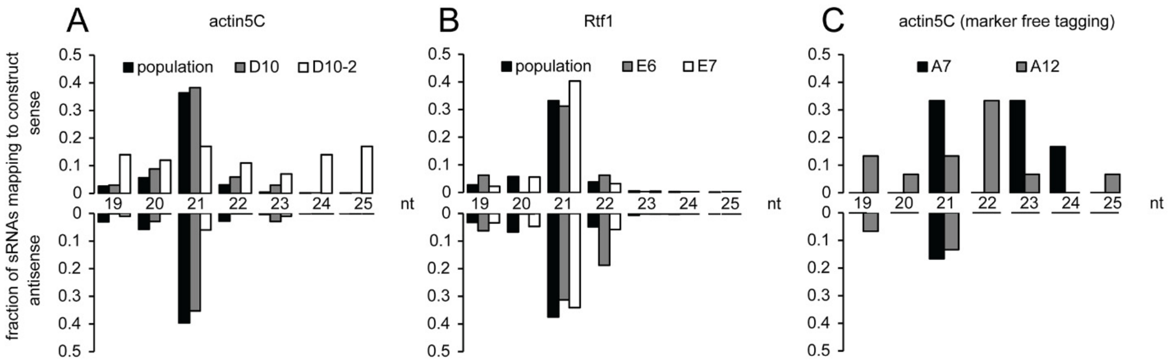

Figure 4—3: Read length distribution of act5C (A, C) and rtf1 (B) locus matching reads in sense and antisense orientation of representative cell lines. Data is presented as fraction of total siRNAs mapping to the construct. (Actin5C D10 = clone, D10-2 = FlpOut clone; Rtf1E6 and E7 = clones; Actin5C A7 and A12 = marker-free tagged clones)

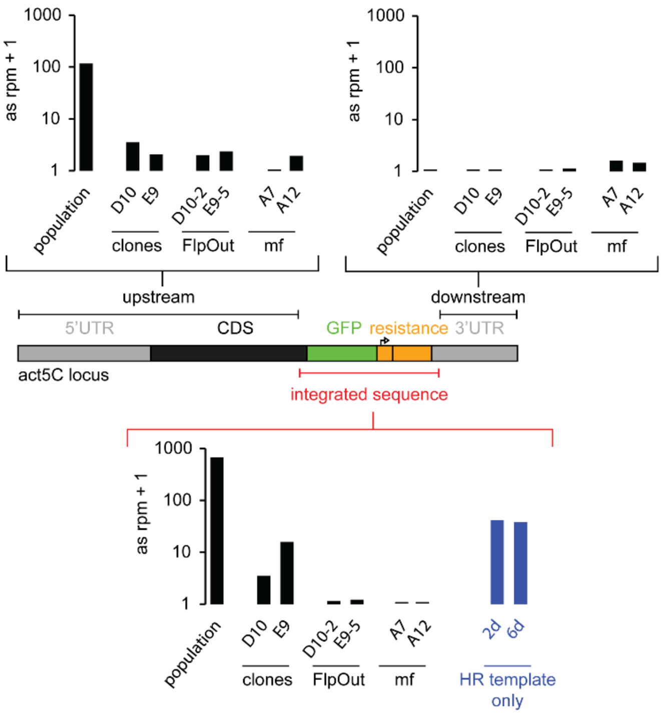

Figure 4—4: Quantification of the siRNA strength at the *act5C* locus for different cell lines. Sequenced siRNAs were mapped to the modified loci and antisense reads (only) mapping either to the upstream or downstrem region of the integrated sequence or the HR donor were summed up and normalized to genome matching reads and length of the sequence region. (mf = marker-free tagged cell lines, FlpOut = FlpOut cell lines)

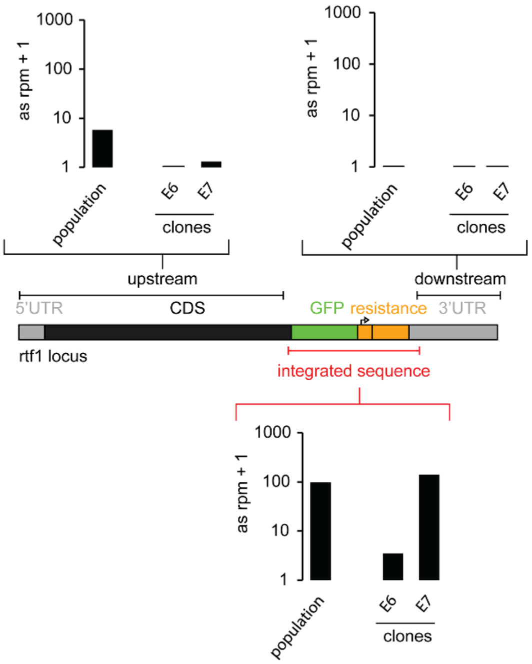

Figure 4—5: Quantification of the siRNA strength at the *rtf1* locus for different cell lines.Sequenced siRNAs were mapped to the modified loci and antisense reads mapping either to the upstream or downstrem region of the integrated sequence or the HR donor were summed up and normalized to genome matching reads and length of the sequence region.

4.2.2 Excision of the selection cassettes removes the siRNA trigger

I then tested whether specific parts of the introduced sequence were responsible for triggering the siRNA generation. To this end, I used Flp recombinase to remove the FRT-flanked selection cassette, which consists of the *copia* promoter and the blasticidin resistance gene (see Figure 4—1A, stage 3). The "Flp-out" of the selection cassette resulted in loss of small RNAs repressing the fusion protein, observed both in the GFP-based expression assay and by small RNA sequencing (Figure 4—1C,Figure 4—4). The remaining small RNAs were predominantly sense oriented and did not show an accumulation of 21 nt long reads. Most likely, they represent mRNA degradation products (Figure 4—3A, clone D10-2).

To further validate the hypothesis that the resistance cassette is the trigger for siRNA biogenesis, I generated GFP-tagged *act5C* clones after marker-free genome editing (Figure 4—1B). I employed the same template plasmid but generated HR donor PCR products that only contained the GFP coding sequence and the homology arms. After transfecting our Cas9-expressing cell line with the sgRNA expression construct and the HR donor, I established Act5C-GFP positive cell lines by single cell cloning and brute-force screening. From the initial 93 hand-picked clones, two lines had the desired *act5C*-GFP modification. The fusion protein neither showed Dcr-2 and Ago2 dependent repression (Figure 4—1C, stage 4), nor did I detect any corresponding siRNA reads by small RNA sequencing (Figure 4—4). Thus, it is not the tagging process *per se* that is responsible for the siRNA response, but rather the selection cassette comprising a promoter and resistance gene. The *copia* promoter, which drives expression of the Blasticidin resistance in our cassette, has sequence identity with an endogenous transposable element that is constitutively targeted by siRNAs (note that I excluded this region in the siRNA sequencing analysis). It is conceivable that these siRNAs serve to nucleate a response that then spreads into the surrounding sequence analogous to siRNA-directed heterochromatin formation in fission yeast (Halic and Moazed, 2010; Verdel and Moazed, 2005).

However, since the cassette excision completely reverts the siRNA generation, we favor the hypothesis that a low-level of antisense transcription activity of the *copia* promoter causes convergent transcription with the host gene and thus the generation of dsRNA at the site of integration in this case. Whatever the precise molecular mechanism may be, we recommend implementing strategies for removal of selection cassettes where possible.

4.2.3 Integration of the HR donor is a prerequisite for the generation of siRNAs

In higher eukaryotes, defense mechanisms target linear dsDNA in a context of DNA virus infection (Barber, 2011; Rathinam and Fitzgerald, 2011) and RNA polymerase III can serve as a sensor for cytoplasmic DNA (Chiu et al., 2009). I thus tested if the introduction of a linear PCR product, the HR donor used for GFP-tagging at the *act5C* locus, without a corresponding Cas9-mediated cut in the DNA is sufficient to trigger the generation of siRNAs. Small RNAs were sequenced two and six days after transfection and the sense and antisense reads mapping to the PCR product were quantified. In contrast to the robust response I detected at a comparable time point for the productively genome modified Act5C-GFP cell population (~680 reads per million genome matching sequences, rpm), the response was approximately 15-fold weaker (40 rpm) when the HR-stimulating site-specific DNA cut was omitted (Figure 4—4). Together with our observation that siRNAs repress the targeted locus even after prolonged culture, when all non-replicated sequences have been lost, this argues against a major contribution of episomal linear DNA to the siRNA pool.

4.2.4 The strength of the siRNA response depends on transcription levels of the gene locus

The N-terminal tagging approach of our lab uses the bidirectional and inducible *mtnDE* promoter to drive the concomitant expression of the tagged protein as well as the resistance gene. No TE-derived sequences are used in this setting. A tagging cassette consisting of the resistance gene, the promoter and the epitope tag is introduced between the endogenous promoter and the start codon of the gene. As described above, I was able to generate a conditional shutdown cell line for the *blanks* locus. I again performed small RNA-seq of cells that have no, medium or high expression of the tagged protein and measured the amount of siRNAs targeting the locus. Here, I observed a correlation between transcription levels and strength of the siRNA response. The higher the transcription activity, the more siRNAs are generated (Figure 4—6). Moreover, I validated the presence of siRNA reads that span the endogenous sequence and the integrated template to exclude that off-target integration of the HR template in antisense orientation within a transcribed locus is the source of siRNAs (see Figure 4—6D).

Consistent with our explanation for the origin of the siRNA at the C-terminally tagged loci - where the *copia* promoter generates some antisense transcripts -transcripts that were initiated at the endogenous promoter and antisense transcripts from the *mtnDE* promoter are the likely source of dsRNA and siRNAs respectively. However, many siRNAs can be detected downstream of the *mtnDE* promoter, where no obvious antisense transcription occurs. In this context, the generation of these siRNAs seems to be due to the modification of the locus and shows clear transcription level dependence. This transcription dependent phenomenon might argue for the presence of epigenetic marks at the modified locus. The marks may interact with RNA polymerases or other factors in order to initiate e.g. antisense transcription. Furthermore, by coupling transcription rate to dsRNA precursor generation, cells avoid unnecessary defense strategies at loci that are not transcribed.

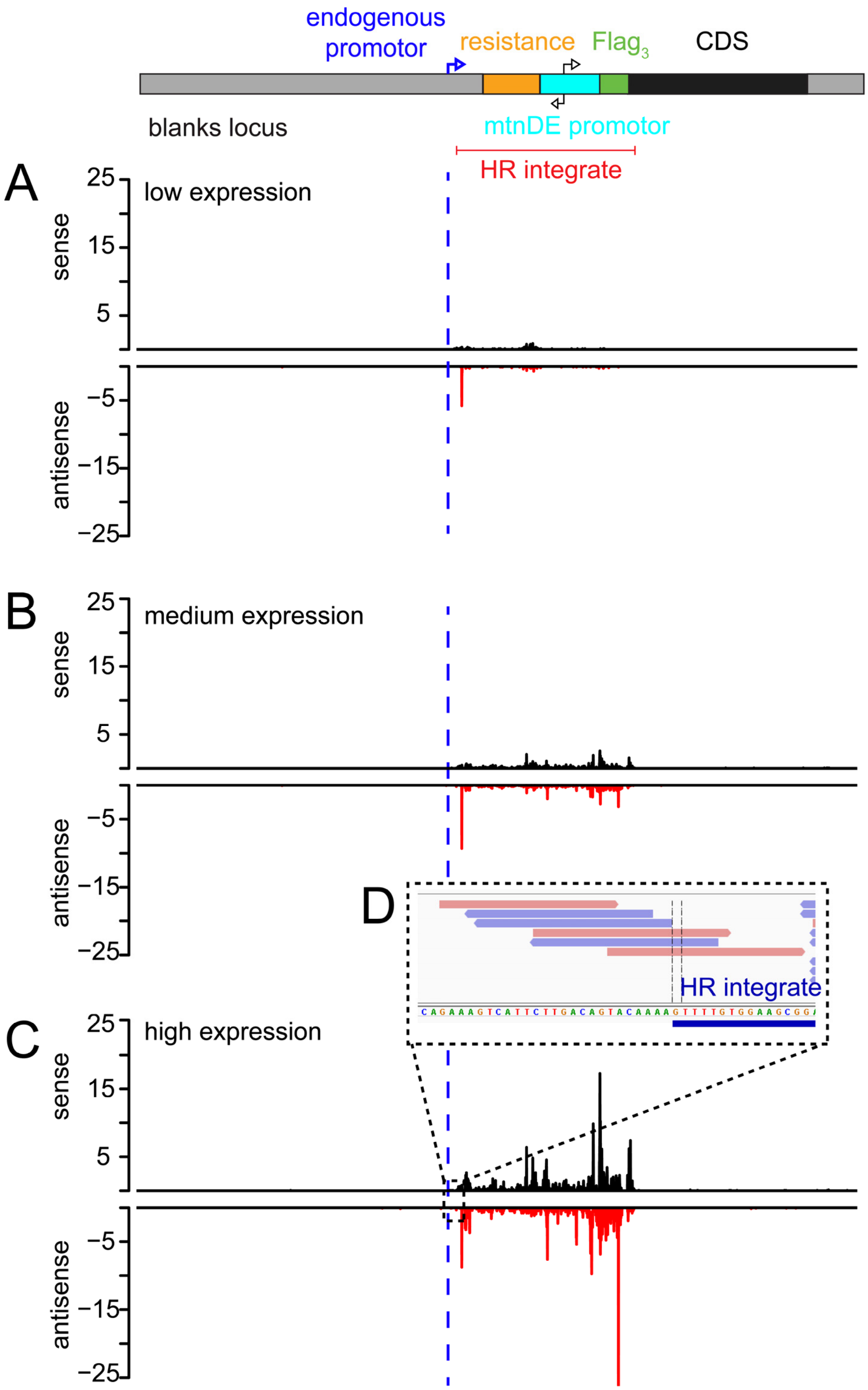

Figure 4—6: siRNA distribution in dependence of transcription levels at the *blanks* locus. The distribution of the siRNAs along themodified *blanks* locus was determined by binning into 1 nt intervals and normalized to the number of genome-matching reads in each library. The graphs depict the sense (black) and antisense (red) matching reads as reads per million of genome matching 19-25 nt long reads in the respective library. Shown are the sequencing traces for the N-terminally tagged Flag-Blanks cells for no (A), medium (B) and high (C) expression levels. The induction was performed with eiter 0 µM, 60 µM or 200 µM copper ions in the cell culture medium. The representation at the top depicts the functional regions of the locus; the HR template is annotated in red. The box (D) shows the magnification of the transition between the endogenous sequence and the HR integrate. Spanning siRNA reads in sense and antisense orientation can be detected.

4.3 Conclusions

Above, I described the induction of an siRNA response after genome editing in cultured *Drosophila* cells. This response is elicited by the presence of a selection cassette, which serves to enrich for cells with the desired modification. Fortunately, removal of the FRT-flanked cassette with FLP recombinase abolishes this response. The same result was obtained when genome editing was performed without selectable markers. Our measurements of GFP-fusion protein levels show that the quantitative extent of siRNA-mediated repression is less than two-fold. This is comparable to the effect of a heterozygous, recessive loss-of-function mutation. There are several reasons why removal of the selection cassette is recommended if the least invasive genome modification is to be achieved. The finding that in cultured cells even the epigenetic phenomenon of RNA interference can be reversed is an encouraging observation. It may be possible to benefit from the advantages of marker selection without inducing irreversible changes in gene expression. Nonetheless, perturbations of the targeted protein’s stability and/or functionality caused by the appended epitope tag remain a concern that should be experimentally addressed.

Our observations have led us to review possible drawbacks of current genome editing strategies in general: First, the introduction of epitope tags at proteins may hinder their function, e.g. by interfering with binding partners or localization mechanisms. Second, by integrating the epitope tag at the C-terminus of the protein, the endogenous 3′ UTR of the mRNA is disrupted and replaced by a short artificial sequence. Since the 3′ UTR is an important platform for posttranscriptional gene regulation such as miRNA binding or mRNA localization processes, the physiological function of the gene may be impaired. Third, off-target editing events are to be expected when the HR donor integrates elsewhere in the genome. Finally, epigenetic changes and influences on transcription regulation of the modified loci take place upon genome engineering.

The CRISPR/*cas* system provides a method for generating cell lines that express e.g. proteins with epitope tags at their endogenous levels. In general, this would offer a possibility to conduct experiments that resemble native conditions. For example, pull downs could be performed without overexpression of the bait protein. Considering the described results, one has to be aware that the expression levels of the fusion protein may be different than expected and the chromatin state of the locus might be altered compared to unmodified alleles, which has to be investigated in further studies. Nevertheless, the addition of a protein tag (e.g. GFP) might have a higher stabilizing effect on the fusion protein than the silencing effect by the siRNAs that are generated against the modified locus.

Since the strength of the RNAi response varies between clones of the same modified locus, one could suggest two different chromatin states. The cells could either adapt to the modification and restore the epigenetic environment to the native state or keep on struggling with the manipulation and generate siRNAs. This process would show parallels to the evolutionary contribution of TEs to gene expression regulation, which can become incorporated into the genomic context and evolve into regulatory elements (Sundaram et al., 2014).

All experiments shown were performed in *D. melanogaster* cells. It is therefore an open question whether mammalian genetic engineering systems show similar side effects. Though mammals generate no or very low levels of siRNAs to silence TEs post-transcriptionally; rather, they methylate cytosines to generate repressive heterochromatin in order to suppress selfish genetic elements. A study addressing the question if CpG methylation occurs at genome-modified loci in mammalian systems would be beneficial for further research. Therefore, I recommend to evaluate each genome editing protocols and check for those phenomena prior to assuming a native chromatin state at the modified regions and endogenous protein levels.

Moreover, the transcription dependence of the siRNA generation demonstrates that the generation of the dsRNA precursor is coupled to the production of the sense directed transcript of the locus. Nevertheless, it is still elusive how antisense transcription is initiated and it remains unknown whether epigenetic marks are set prior to generation of siRNAs or vice versa.

All in all, I demonstrated that the CRISPR/*cas* system is an extremely powerful tool that facilitates many cloning processes and genetic studies. Still, this precise manipulation of the genome harbors the capacity to alter the genomic environment and thereby to introduce unknown and unwanted biases into experiments.

5 Transposable element recognition in *Drosophila* depends on copy number and *cis*-structure

5.1 Introduction

5.1.1 Transposable elements and their corresponding silencing mechanisms to ensure genome integrity

Transposable elements (TEs) are components of all organisms' genomes. They constitute a large fraction of the genomic DNA, display a huge diversity and are characterized and subdivided by their mobilization mechanism (Kazazian, 2004; Rebollo et al., 2012; Wicker et al., 2007). DNA transposons are "cut-and-paste" TEs coding for enzymes that excise the TE and catalyze the insertion at another site in the genome. Retrotransposons, however, transpose via an RNA intermediate that is reverse transcribed and inserted at the target site by an integrase. Retrotransposons can either be flanked by long terminal repeats (LTR) or lack these LTRs (Levin and Moran, 2011).

During evolution, TEs contributed to the development of genomes and regulatory networks (Feschotte, 2008; Kazazian, 2004). However, they can integrate into the ORF and *cis*-regulatory sequences of genes during transposition and thereby disrupt essential genes. Moreover, they provide the basis for ectopic recombination events during meiosis, which can result in harmful chromosomal rearrangements and deletions. Finally, the energy spend on replication, transcription and translation of a large number of TEs challenges the cell (Werren, 2011).

Therefore, TEs represent a serious threat to the integrity of a cells' genome. Especially germline cells need reliable mechanisms to inhibit transposition. In *Drosophila* and mice, a class of small Argonaute protein associated RNAs, the piRNAs, play a leading role in repressing TE activity. Further silencing mechanisms exist in somatic cells: In vertebrates, cytosines are methylated – a feature that represses TEs on a transcriptional level (Ooi et al., 2009). *Drosophila,* however, exploits the RNAi pathway and endo-siRNAs for transposon silencing (Levin and Moran, 2011). There are no universal transposon features that could be exploited for recognition by the cell. dsRNA precursors are – according to current knowledge – crucial for the initiation of the siRNA biogenesis, as only dsRNA molecules can be processed by the RNase III enzyme Dicer. Yet it remains elusive how TE genes or mRNAs are recognized and which additional mechanisms lead to production of dsRNA molecules.

In *C. elegans*, plants and some fungi, a RNA dependent RNA Polymerase (RdRP) can synthesize the complementary strand to the TE transcript, which can anneal to the transposon mRNA. Flies lack a functional RdRP and must thus rely on other mechanisms for dsRNA formation (Ahlquist, 2002; Ghildiyal and Zamore, 2009). To enable biogenesis of siRNAs that can target TE mRNA in flies, antisense transcription of DNA must be essential to generate dsRNA.

In *Drosophila,* about 19 % of all genes show a detectable amount of antisense transcription(Sun et al., 2006). In human cell culture, transcriptome analysis revealed hotspots of antisense transcription at first exon and the 5′ end of the first intron. The underlying sequence resembles bidirectional promoter sequences for RNApol II (Finocchiaro et al., 2007). Moreover, in *S. cerevisiae* global antisense transcription occurs at the promoters of almost all genes (Schulz et al., 2013). Additionally, stalled spliceosomes are known to be the trigger for siRNA production in fungi and plants (Dumesic and Madhani, 2013; Dumesic et al., 2013). In *Neurospora crassa*, repetitive DNA sequences together with DNA double stand breaks are sufficient for triggering a RNAi response (Yang et al., 2015b).

A recent publication claims that most TE families in *Drosophila* cells are characterized by both sense and antisense transcription due to the existence of internal antisense transcription starting sites. These sites resemble canonical RNA Pol II promoters. This could mean that the trigger for the antisense transcription seems to come from the TE itself. Moreover, it was shown that the antisense transcripts are retained in the nucleus as they are less efficiently polyadenylated in comparison to the sense transcript. This results in the formation of a nuclear dsRNA precursor that can be further processed by the RNAi machinery to silence the TE(Russo et al., 2016).

5.1.2 Starting point for my studies

Stable integration of a reporter gene in high copy numbers into the genome of somatic *Drosophila* S2 cells triggers an endo-siRNA response that can be abolished by the depletion of the RNAi factors Dcr-2, Loqs-PD or Ago2.This knowledge was used to generate a reporter system to investigate proteins which are involved in the generation and effector function of small RNAs (Forstemann et al., 2007; Hartig et al., 2009).

Katharina Elmer, a former member of the Förstemann lab, discovered that low level antisense transcription of the reporter gene is the trigger for dsRNA generation (Elmer, 2013). Moreover, the level of siRNA generation seems to correlate with the copy number of the reporter (Figure 5—1A and B).

Consequently, it was reasoned that the copy number stimulates the level of response. However, histone genes that are encoded in high copies (comparable to TEs) and do not have introns do not give rise to endo-siRNAs. The question is if the histone stem loop termination signal that is unique for histone genes protects them from generation of dsRNA precursors. Reporter cell lines were generated that contain this non-canonical termination signal and investigated the effect on siRNA generation. Strikingly, siRNAs in high abundance and in a reproducible manner accumulated at the intronic region of the *ubi* promoter (Figure 5—1C). Thus, a reasonable hypothesis was that the artificial combination of an intron – and thus splicing – with histone stem loop mediated termination may trigger the generation of siRNA. The combination of these processes does not occur in a physiological context in the cells.

So, we hypothesized that artificial, non-physiological structures when combined *in cis* might trigger the recognition of foreign genetic elements by the cell. We aimed to understand more about the transposon recognition by studying the prerequisites *in cis* that lead to the generation of siRNAs.

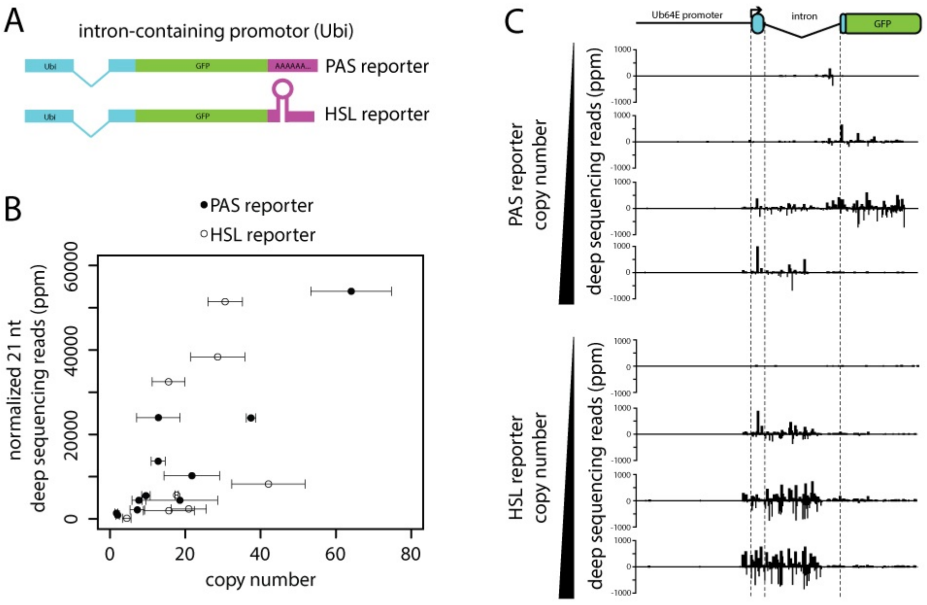

Figure 5—1: *Cis*-structure of the reporter gene has an influence on the quality and quantity of the corresponding RNAi response. (A) Schematic representation of the polyA (PAS) and histone stem loop (HSL) reporter cassette. The GFP coding sequence is flanked by the intron-containing *ubi-p63E*promoter and either the SV40 polyadenylation or histone stem loop in the 3'UTR. (B) Small RNAs of selected PAS (solid circle) and HSL (open circle) reporter cell lines were isolated and analyzed by deep sequencing. 21 nt long reads were mapped to the reporter sequence, quantified and normalized to genome matching reads. Copy number was determined by qPCR, error bars represent standard deviation. (C) Sequencing traces showing the siRNAs mapping to the reporter sequence. The schematic on the top shows the cis-structure of the gene. Deep sequencing reads were normalized to genome matching reads. All data of this figure was generated by Katharina Elmer.

5.2 Results and discussion

5.2.1 The *cis*-structure of a gene cannot overcome the copy number dependence for stimulating the siRNA generation

The reporter cell lines that were used by Katharina Elmer consist of the intron-containing *ubi-p63E* promoter, the GFP coding sequence and either the SV40 polyA or histone stemloop (HSL) termination signal. To further analyze the effect of *cis*-elements on the recognition as RNAi targets, I modified the GFP reporter and generated several reporter plasmids. The GFP coding sequence was flanked by varying functional genetic elements: intron-less (*sucb*, *H2A*) or intron-containing (*ubi*) promoters and a polyA signal or HSL termination signal (Figure 5—2A). Cell lines which stably express the different reporter constructs were established using the ΦC31 integrase system. The ΦC31 integrase can be used to mediate site-specific recombination in a unidirectional manner. *attP* recognition sites, which were integrated into the genome of S2 cells before, and *attB* sites on the plasmid containing the reporter cassette and a resistance gene are recognized by the recombinase. In this configuration, only the integration is catalyzed (Smith et al., 2010). In the absence of the integrase, spontaneous integration of the reporter plasmid can take place but at random sites. In contrast, if the integrase is expressed, the integration of the plasmid is highly site-specific (Figure 5—2B). Since the parental acceptor cell line contains several *attP* landing sites, the plasmid can integrate multiple times in the genome. However, experimental data shows that not all sites are occupied in each cell and therefore the resulting population is heterogeneous. Nevertheless, the distribution of different cells, having integrated the plasmid at varying numbers of *attP* sites, seems to be equal. All biological replicates are significantly correlated (Figure 5—2B). Thus, the cell populations which were generated with different reporter constructs are comparable to each other and have comparable genetic backgrounds. This enables us to directly compare the results of the reporter assay with the varying genetic features.

The copy number of the GFP gene in the reporter cells was determined by qPCR. To this end, the signal was normalized to *rp49,* which presumably has four copies in the genome due to the tetraploid state of the S2 cells. Accordingly, all reporter cell populations contain between two and four copies of the reporter genes (Figure 5—2C). The depletion of several proteins essential for the RNAi pathway revealed a potential involvement of endo-siRNAs in silencing the GFP expression for all cell lines. The Dcr-2 knockdown resulted in derepression of the GFP signal as depicted in Figure 5—2D. However, no differences among the various reporter constructs could be observed.

In order to investigate the siRNA response generated against the reporter region further, small RNAs were isolated and sequenced (Figure 5—3B). In contrast to the high copy situation as described by Katharina Elmer (see Figure 5—1C), only a small fraction of siRNAs maps to the constructs – independent of the genetic structure. For the reporter genes that are expressed via the *ubi* promoter (pRB10, pSK13) more sense than antisense siRNAs could be detected. This might be due to the high expression strength of the promoter and an excess of degradation products of the mRNA. The expression of the GFP via the *sucb* (pSK5, pSK7) or *H2A* (pSK7, pSK11) promoter was much weaker. Moreover, siRNAs did not accumulate within the intronic region of the reporter cell line with the *ubi* promoter and the HSL termination signal (pSK13); this should resemble the construct described by Katharina Elmer.

Apparently, a licensing mechanism that "measures" the copy number prior to the biogenesis of small RNAs exists. Nevertheless, the *cis*-structure of the reporter can influence the abundance and profile of the generated siRNAs as described in section 5.1.2. However, it was not possible to address this in the low copy situation.

To gain insight into the threshold for licensing, I cloned a second identical GFP reporter cassette into the corresponding plasmids and generated stable cell lines as described above. The cells contained approximately twice as many GFP copies compared to the initially used cell lines, thus four to eight copies. Upon depletion of Dcr-2 in the reporter assay, neither a derepression of the GFP signal nor a difference between the diverse reporter constructs could be observed. This shows that four to eight copies are again too few to trigger the RNAi response. This is consistent with the data obtained by Katharina Elmer. Figure 5—1B shows clearly that the amount of siRNAs increases aboveten copies.

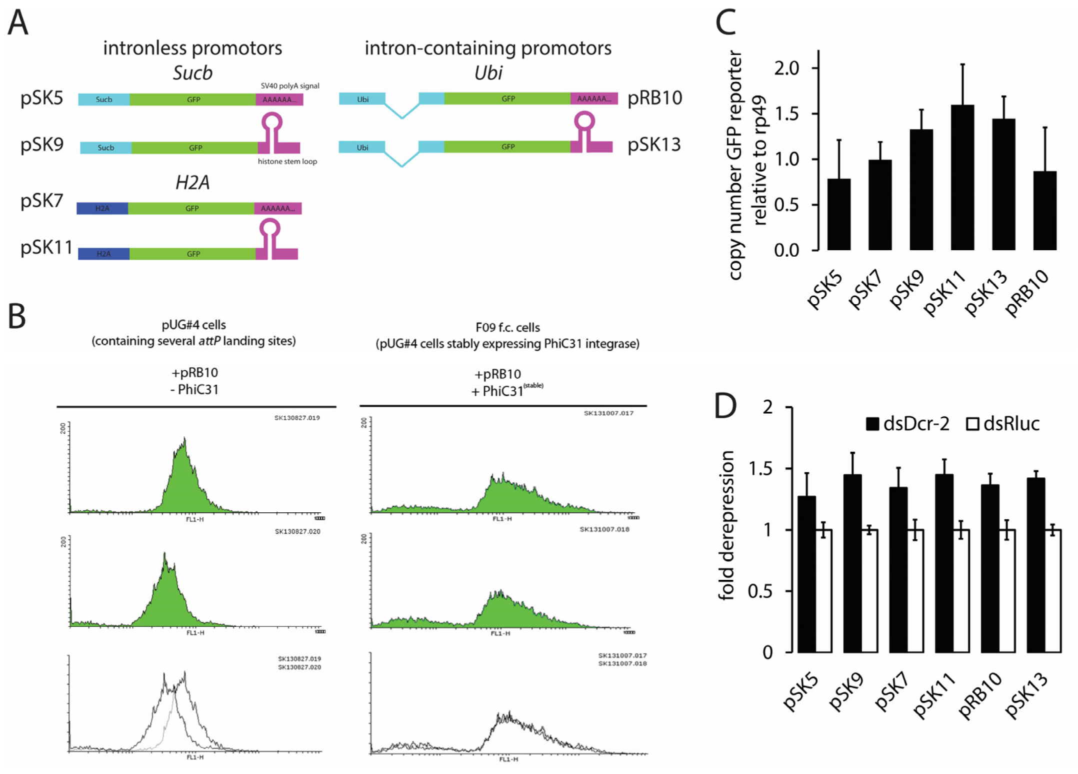

Figure 5—2: Characterization of stable low copy reporter cell lines. (A) Schematic representation of the gene showing the gene structure of the reporters. GFP expression was either driven by the intron-less *sucb* or *H2A*promoters or the intron-containing *ubi* promoter; transcription was terminated by the SV40 polyA or the histone stem loop. (B) The ΦC31 integrase system provides a method, which allows generation of highly reproducible stable cell lines with identical genetical background. Shown are flow cytometry histogram plots. The frequency of cells with distinct fluorescence intensity (Fl1-H channel) is depicted for each of the two independent biological replicates. The panel at the bottom represents the overlay histogram. The left column shows the classical approach without targeted integration of the plasmid – the replicates are distinct. Whereas the right column shows two replicates of cell lines generated with targeted integration. (C) qPCR on genomic DNA of stable reporter cell lines for quantification of integrated GFP reporters. Data was analyzed using the $2^{-\Delta\Delta ct}$ method with the *rp49* gene as negative control. Error bars represent standard deviation, n=3. (D) Dcr-2 knockdown in stable reporter cell lines leads to derepression of the GFP reporter as measured by flow cytometry. Depicted is the fold derepression (normalized to luciferase (Rluc) control knockdown) of GFP expression of different reporter cell lines. Error bars represent standard deviations, n=3.

5.2.2 Transiently transfected reporter plasmids can stimulate the RNAi response

The next goal was to investigate whether the episomal presence, i.e. not stably integrated into the chromosomes, of the reporter plasmids can trigger an siRNA response and if differences among the reporter constructs can be observed. To this end, I transiently transfected the reporter plasmids into *Drosophila* S2 cells and isolated the small RNAs. Deep sequencing libraries were generated and analyzed. Strikingly, a strong RNAi response against the whole reporter plasmid could be detected. Episomal DNA, which presumably has not assembled proper chromatin architecture, seems to trigger an RNAi response. The strength is comparable to reporter genes that were integrated at high copy number into the genome (Figure 5—3C). Moreover, the strength of the response seems to depend on the promoter of the reporter gene, whereas the termination signal has no influence. In contrast, no correlation between promoter strength and amount of generated siRNAs could be observed as the strongest promoter (*ubi*) triggered the weakest siRNA response.

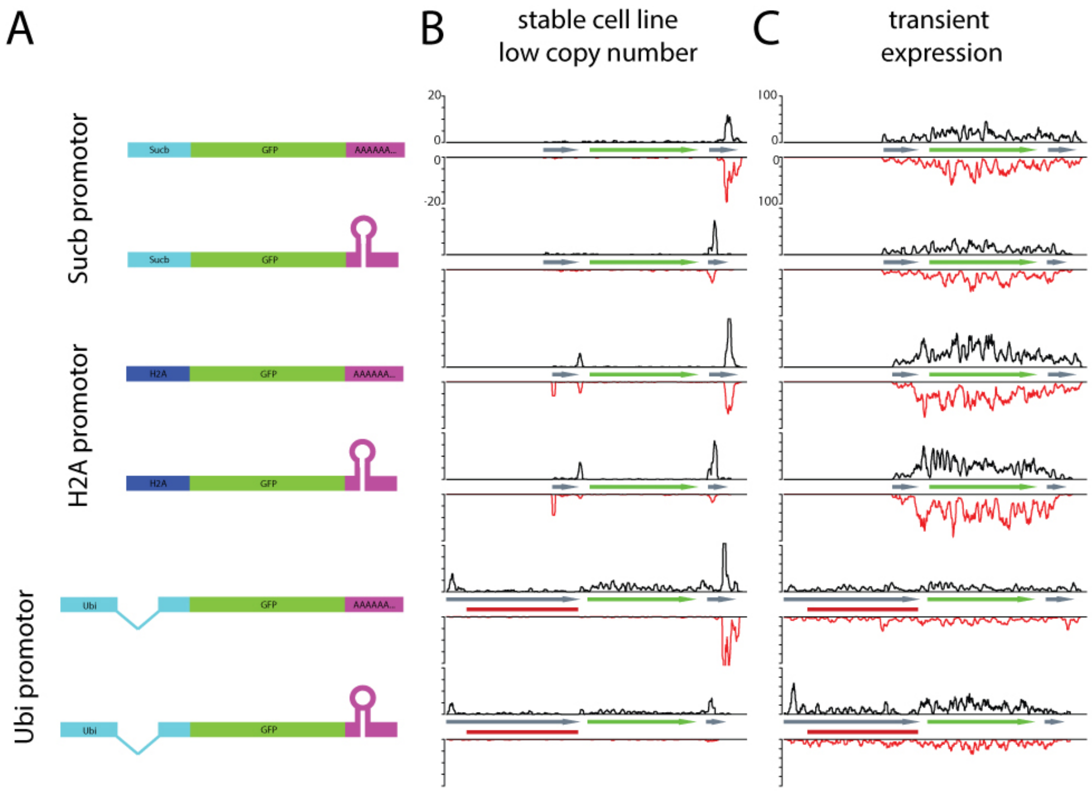

Figure 5—3: The *cis*-structure of stably integrated low copy or episomal reporter genes can only marginally affect the strength of the RNAi response. (A) Schematic representation of the gene showing the gene structure of the reporters. GFP expression was either driven by the intron-less *sucb* or *H2A* promoters or the intron-containing *ubi* promoter; transcription was terminated by the polyA or the histone stem loop. (B) Stable cell lines were generated using the ΦC31 recombinase system to insert the reporter plasmids targeted at specific positions in the genome. Approximately 2-4 copies of the reporter genes were introduced. Small RNAs were isolated and analyzed by deep sequencing. Shown are 19-25 nt long siRNAs mapping to the reporter genes (normalized to transposon and miRNA matching reads) in sense (black) and antisense (red) orientation. (C) 19-25 nt long siRNAs mapping to the reporter genes (normalized to transposon and miRNA matching reads) in sense (black) and antisense (red) orientation after transient transfection of S2 cells with the reporter plasmids.

5.2.3 Impaired splicing does not trigger the RNAi response against low copy reporters

In the yeast *Crypotococcus neoformans*, it was observed that stalled splicing can trigger the generation of siRNAs. This mechanism is used to target TE mRNAs and it depends on the activity of the intron-lariat debranching enzyme and an RNA-dependent RNA polymerase (RdRP) for the generation of the dsRNA precursor (Dumesic and Madhani, 2013; Dumesic et al., 2013). Although *Drosophila* lacks the RdRP enzyme and the described mechanism can thus not be completely conserved, the general principle of TE recognition might be shared among the species. Can imperfect or impaired splicing enhance the siRNA generation?

In order to investigate this, I designed GFP based reporter constructs that contain introns with different mutations. Strong and constitutive expression of GFP was driven by the short and intronless *tctp* promoter. The *mini-white* intron was then inserted between the promoter and the GFP coding sequence. The intronic sequence was flanked by either perfect splice signals (-AGgt … agGT-pSK20), a signal that leads to blocked 3' end recognition of the intron (-AGgt … **cc**GT- pSK21) and the corresponding sequences without the exonic consensus sequences (-gt … ag- and -gt … **cc** -, pSL3 and pSL4), see Figure 5—4A.

The plasmids were transiently transfected into *Drosophila* S2 cells and GFP expression was observed for all reporters by flow cytometry (Figure 5—4C). The plasmids pSK21, pSL3 and pSL4 showed lower levels of GFP signal. Since mRNAs derived from these plasmids are either imperfectly spliced or have extended 5'UTRs, they are likely substrates for maintenance processes in the cells such as the non-sense mediated decay (NMD) pathway. Factors of the NMD pathway degrade unproductive mRNAs and are hypothesized to be also involved in the degradation of transcripts with extended UTRs. RT-PCR proved that the majority of pSK20-derived transcripts are spliced, whereas pSK21 with the blocked 3' intronic splice signal as well as pSL3 and pSL4 lacking the exonic consensus sequences are not spliced efficiently. Yet, a small amount of unspliced RNA remains detectable also for pSK20. This is likely due to the lack of appropriate exonic splice enhancers that are present in normal fly genes and that influence the efficiency of the splicing reaction.

To study the influence of impaired splicing on siRNA production, I cloned the reporter genes into plasmids containing a selection marker and the *attB* site. Subsequently, I generated stable cell lines using the ΦC31 integrase system. The cells contained one to three copies of the reporter gene as measured by qPCR (Figure 5—5A) and did not respond on the depletion of Dcr-2 with derepression (Figure 5—5B); integration of pSK33 (-gt … ag-) appears to be an exception. This argues for only a low amount of siRNAs targeting the locus. Upon small RNA sequencing, only few siRNAs mapped to the reporter (Figure 5—6A). Since the sequence of the *mini-white* intron is also present on the plasmid backbone as well as at the endogenous *white* gene, it is difficult to draw conclusions from the amount of siRNA mapping to this region (Figure 5—6B). In contrast, the siRNAs mapping to the unique GFP region can be used for estimating the strength of the siRNA response. We did not observe any reliable difference between the reporter constructs. As a trend, there are more reads visible with the reporter pSK35 (-AGgt … **cc**GT-) that is characterized by impaired splicing. The amount of sense and antisense reads was equal, which argues for *bona* fide siRNAs derived from a dsRNA precursor. Also the pSK36 reporter (-gt … **cc**-) showed equal numbers of sense and antisense reads, whereas all other reporters have an excess of antisense reads. To conclude about potential biological processes, the experiments must be repeated with deeper coverage of the sequencing.

In summary, it seems that the inefficient splicing process by itself cannot trigger an siRNA response in the low copy situation. As mentioned before, the copy number might be the (or at least another) licensing step that works upstream of the potential recognition of impaired splicing. Thus, the constructs described should be studied in a high copy context.

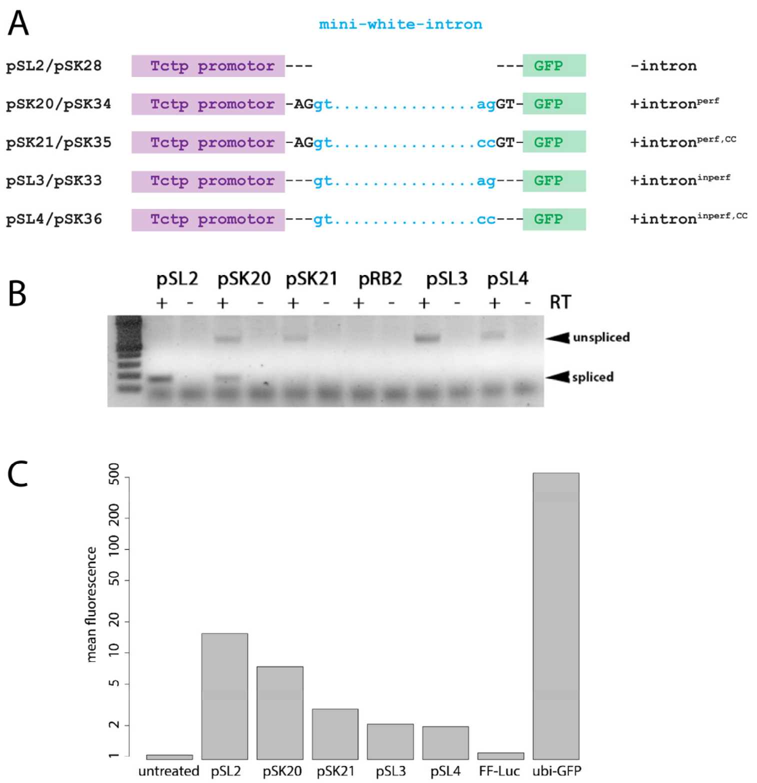

Figure 5—4: (A) Schematic representation of the splicing reporter constructs. The GFP expression was driven constitutively by the *tctp* promoter. The *mini-white* intron was integrated between the promoter and the GFP coding sequence and contains perfect or mutated exonic and/or intronic splice sites. (B) RT-PCR spanning the region of the intron to assess the splice efficiency of the different reporter constructs. The products are annotated. pRB2 codes for the Renilla luciferase and served as a negative control. (C) Flow cytometric analysis of the GFP levels that are generated from the different reporter plasmids. Equal amount of plasmids were transiently transfected into S2 cells. FF-Luc (=pRB2) and (ubi-GFP, pKF63) served as negative and positive controls respectively.

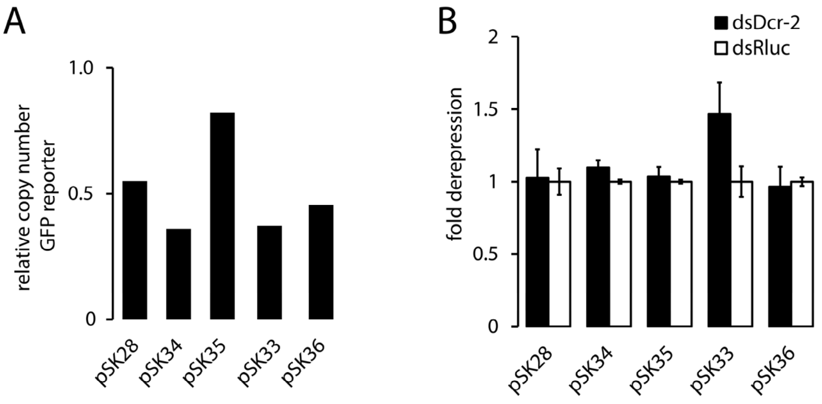

Figure 5—5: (A) Copy number of the splice reporter constructs in the stable F09 cell lines. Copy numbers were determined by qPCR on gDNA by normalization to *rp49*. Since the used S2 cells are probably tetraploid, four copies of *rp49* are present. (B) GFP levels were measured after Dcr-2 knockdown in stable reporter cell lines by flow cytometry. Depicted is the fold derepression (normalized to luciferase (Rluc) control knockdown) of GFP expression of different reporter cell lines. Error bars represent standard deviations, n=3.

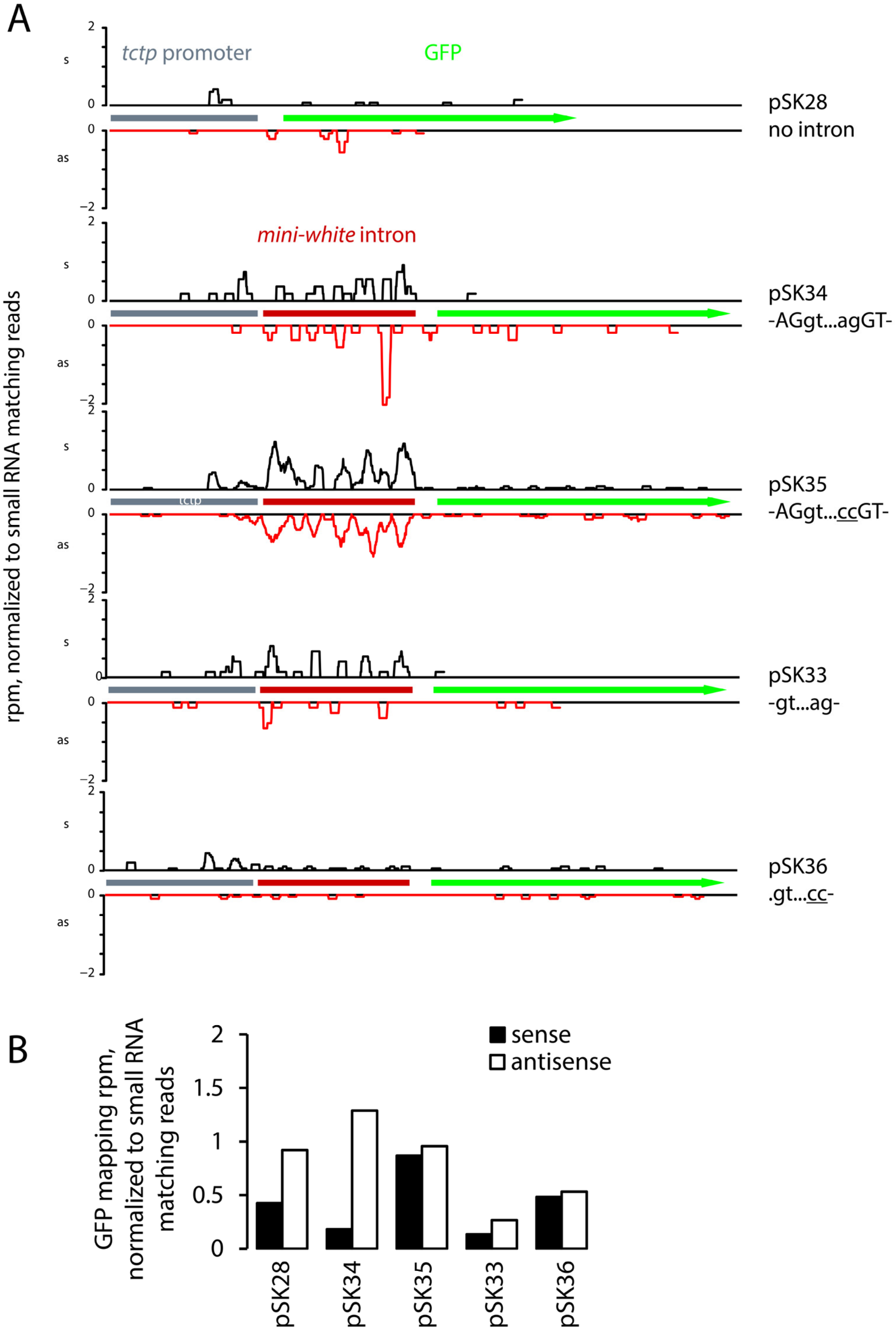

Figure 5—6: (A) Stable cell lines were generated using the ΦC31 recombinase system to insert the reporter plasmids targeted at specific positions in the genome. Approximately 1-3 copies of the reporter genes were introduced. Small RNAs were isolated and analyzed by deep sequencing. Shown are 19-25 nt long siRNAs mapping to the reporter genes (normalized to transposon and miRNA matching reads) in sense (black) and antisense (red) orientation. (B) Quantification of the GFP mapping reads in sense and antisense orientation, normalized to the small RNA mapping reads of the corresponding library.

5.3 Conclusions

Altogether, I was able to advance our understanding of how cells deal with foreign genetic elements such as reporter genes or TEs. With the help of GFP based reporter constructs I could show that siRNAs are essentially not produced when the copy number of the reporter gene is below a threshold of approximately 10 copies per cell – independent of the functional elements assembled *in cis*. This phenomenon might serve as a licensing step for the cell to ensure efficient focusing of the resources on harmful processes. This way, no energy is wasted on the silencing of elements that are inactive. When the copy number of these elements, however, passes the threshold due to transposition and multiplication, the RNAi response is triggered.

Interestingly, the episomal existence of DNA – here shown by the transient transfection of the reporter plasmids – seems to be sufficient to trigger a relatively strong RNAi response. This situation mimics the infection of cells by viral dsDNA that is targeted by RNAi as well. For many reporter assays and other approaches, however, the transient transfection of plasmids into cells is used in order to study various processes – potentially in a quantitative manner. Since these vectors are likely targets for RNAi, the expression from these plasmids is affected.

In addition, it is not possible to exclude the hypothesis that impaired splicing has an influence in the siRNA generation in *Drosophila*, because the described experiments were only conducted in the low copy situation. If the licensing processes acts indeed upstream, the influence of the *cis*-structure cannot be studied in the low copy context. Thus, it would be necessary to generate stable cell lines that contain more than 10 copies of all reporter constructs – both for the *cis*-element reporter or the splice reporters. One possibility would be to generate cell lines that contain several *attP* landing sites in their genome. Using the CRISPR/Cas9-mediated genome editing protocol described in the chapters before, this approach should be feasible. Cell lines generated this way could then be used for the fast and reproducible generation of the final reporter cell lines. Assuming that differences in the strength of the RNAi response are then detectable for the different reporters, genome wide RNAi screens with these cell lines could be performed in order to discover factors that are involved in triggering the silencing process. This can reveal mechanistic details of the copy-number sensing system.

Finally, it remains elusive whether the stably integrated reporter genes are an appropriate mimic of the TE silencing phenomenon. To bring this analysis forward, an investigation of histone marks could provide information about the chromatin state and the epigenetic pattern of the reporters in comparison to endogenous TEs.

In summary, I could demonstrate that changes in the reporter gene structure cannot overcome the copy number dependence for triggering an RNAi response.

6 Establishing a protocol for mass spectrometry-based identification of protein-protein interactions

6.1 Introduction

Many small molecules such as hormones, second messengers and ions as well as a large variety of macromolecules ensure cellular functionality and mediate the appropriate responsiveness that is crucial for cellular integrity. Beside nucleic acids, proteins are highly abundant and involved in most cellular processes. By interacting with other proteins or nucleic acids they engage in context-dependent activities. Consequently, the identification of binding partners of those proteins gives further insights into their function in the cellular context.

In order to study protein function, the complex structures of living cells have to be homogenized and proteins have to be isolated and separated from the insoluble debris. For eukaryotic cells, a huge number of different lysis protocols exist. By combining non-mechanical, chemical parameters such as detergents and salt concentrations with mechanical force, the membrane structures of cells and their organelles are disrupted and the soluble components become accessible. Depending on the stringency of the lysis conditions, nuclear and membrane-integrated proteins are isolated as well.

Detergents interfere with lipid-lipid interactions and solubilize proteins. Additives like urea denature proteins, nucleic acids and cellular structures. High or low salt concentrations generate osmotic pressure on organelles and membranes and therefore facilitate the lysis (Lottspeich and Engels, 2006).

In addition or as alternatives to those approaches, physical disruption methods can be used to lyse the cells: By using a coffee mill with dry ice or a blender with rotating blades the cells can be mechanically disturbed. Similarly, sonication of the sample results in shearing of membranes, organelles and nucleic acids to facilitate the extraction of proteins. Moreover, using a Dounce homogenizer can give comparable results: The cell suspension is pushed through a narrow gap, which shears the membranes. Finally, multiple freeze-and-thaw cycles using liquid nitrogen can mildly lyse the cells due to their “cyclic” swelling and the generation of ice crystals that cause the rupture of the cells (Islam et al., 2017).

However, the appropriate lysis method depends on the cell type and also on the physical and chemical properties of the protein of interest. Therefore, the protocol has to be optimized for each specific application.

After cell lysis, immunoprecipitation (IP) can be used to enrich the protein of interest with its interaction partners still bound. This method can be used to investigate protein-protein interactions. Epitope-tagged proteins can be pulled down with corresponding, well characterized antibodies that are immobilized to beads (Kaboord and Perr, 2008). Subsequent washing removes unspecific binding partners and the sample can be analyzed. If potential interactors are known, western blotting can be a fast and inexpensive readout. If, however, unknown interactors have to be identified, mass spectrometry is the method of choice to determine the associated proteins in an explorative fashion.

Similar to the lysis protocol, many relevant parameters such as incubation time, buffer composition or washing stringency influence the efficiency of the pull down and the background binding of proteins.

Weak or transient interactions can be stabilized prior to immunoprecipitation by the addition of reactive chemicals that crosslink the protein complexes *in vivo*. Formaldehyde is a cross-linker that generates covalent interactions. Its cross-links are quite short (2 Å distance) which ensures that only amino acids that are in very close proximity to each other are linked, i.e. tightly interacting proteins. It is highly cell permeable and reacts by linking nucleophilic amino acid side chains (e.g. Lys, Cys, Tyr or Arg) with each other. As an example, the amino group of a lysine and one formaldehyde molecule react to a Schiff base (imine). The imine can attack a nucleophile, e.g. another amino group from the same or a different protein, and thereby establish a covalent crosslink. Remaining unreacted formaldehyde can be quenched by adding the amino acid glycine to the reaction. The concentration of formaldehyde used roughly correlates with the number of introduced cross-links. Finally, the cross-links can be reverted by incubation at high temperature (70°C) in the presence of reducing agents (Sutherland et al., 2008; Vasilescu et al., 2004).

The aim of this project was to establish an optimized and robust lysis and IP protocol for *Drosophila* S2 cells in order to analyze the interactome of endogenous, epitope tagged proteins by mass spectrometry and thereby gain insights into the function of unknown interactor proteins.

6.2 Results and discussion

6.2.1 Optimizing the lysis conditions

In the beginning, the lysis conditions were optimized in order to isolate highly concentrated protein extract that contains not only the cytosolic but also the nuclear fraction of the proteome. Since mild *in vivo* cross-linking for stabilizing weak interactions was used in the final experiments, the lysis of cells that were previously treated with 0.1 % formaldehyde was also evaluated in this section.

In general, two lysis conditions were tested (see Table 1):

(a) Mild, physiological salt concentrations after mild *in vivo* cross-linking (0.1 % formaldehyde)
(b) Stringent conditions to gain maximal protein extraction efficiency using urea and SDS-containing buffers (Gao et al., 2014)

Table 1: Composition of the lysis buffers used for the disruption of *Drosophila* cells.

Mild (a)		Stringent (b)	
150 mM	KAc, pH 7.4	0.5 M	Urea
30 mM	HEPES, pH 7.4	0.01 %	SDS
5 mM	MgAc2	2.0 %	Tergitol (NP-40)
1 mM	DTT	1 tabl./10ml	Complete Protease Inhibitor, EDTA-free
15 %	Glycerol	in 1x PBS	
1.0 %	Tergitol (NP-40)		
1 tabl./10 ml	Complete Protease Inhibitor, EDTA-free		

In addition to the two different lysis buffers, the efficiency of four mechanical disruption methods – "douncing", multiple freeze-and-thaw cycles, sonication and coffee milling –were tested. SpnA, the *Drosophila* Rad51 homolog, was used as a model protein because of its known nuclear localization. Thus, the extraction efficiency of nuclear proteins could easily be tested by probing for SpnA. Cytosolic proteins are already extractable with very mild conditions.

Drosophila S2 cells expressing C-terminally FLAG-tagged SpnA were harvested and washed with 1x PBS. The pellet was resuspended in an appropriate volume of lysis buffer and the cell suspension was used for mechanical disruption. Prior to a centrifugation step, which is necessary to separate the insoluble debris and chromatin fraction from the soluble protein extract, a sample for analysis was taken (crude extract, CE). The supernatant (SN) that remains after centrifugation contains only the proteins which were successfully extracted from the cell using the corresponding method, whereas the CE also contains the fraction of proteins that were not extractable with the applied protocol. However, these proteins are detectable on a western blot by boiling the sample in SDS and DTT containing loading buffer prior to the SDS-polyacrylamide gel electrophoresis (SDS-PAGE). Comparing the intensity of the SpnA-FLAG signal for CE and SN gives information about the lysis and extraction efficiency of the corresponding method.

Figure 6—1shows the results for the different mechanical and non-mechanical extraction approaches. When using the stringent buffer conditions, no difference between the mechanical methods can be detected. However, after mild cross-linking in combination with more physiological and milder buffer conditions, douncing and sonifying the sample results in much higher extraction efficiency of the nuclear SpnA protein than thawing-freezing the suspension or using the coffee mill for destroying the cells.

In the following experiments, cells were lysed by using sonication as the mechanical disruption method in combination with either the stringent or mild buffer conditions. Besides its efficiency, sonication brings along the advantage that several samples can be treated in parallel since only little hands-on work is necessary during this step.

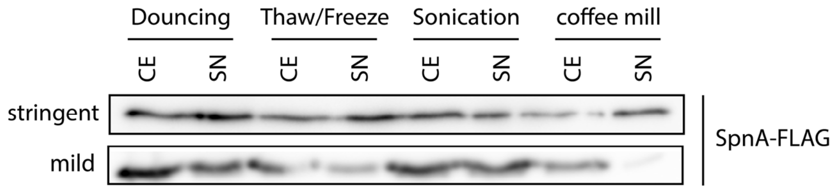

Figure 6—1: Western Blot that was probed for SpnA-FLAG using anti-FLAG-HRP antibody. The stringent and mild buffer conditions are compared as well as the different mechanical disruption methods. CE = crude extract prior to centrifugation, SN = supernatant after centrifugation, protein extract

6.2.2 Optimizing the immunoprecipitation protocol

The immunoprecipitation protocol contains many parameters that can influence the efficiency, recovery and purity of the immunoprecipitated bait protein and its interactors. The amount of antibody and beads as well as the incubation time and the order in which the bead-antibody-protein complex is assembled have to be considered.

Magnetic beads that are covalently coupled with protein G (Dynabeads Protein G, Invitrogen) were used. Protein G binds to the constant region of the antibodies that can interact with its antigen and thereby immobilize the protein complexes with its bound interactors. Dynabeads Protein G have a binding capacity of ~ 8µg human IgG per mg beads. During the experiments, I used 20 µl beads per reaction as a standard, which corresponds to an overall binding capacity of 4.8 µg antibody in one sample. The predominantly used antibody for the optimization process was the monoclonal anti-FLAG M2 antibody (Sigma) that has a concentration of 1 mg/ml. I used 2 µL= 2 µg of antibody per pull down, which results in a 2.4-fold excess of binding capacity on the beads. This ensures that a high fraction of antigen – antibody complexes can be recovered.

6.2.2.1 Assessing the incubation of Dynabeads Protein G with the antibody

First, I checked for the optimal condition to generate the complex consisting of beads, antibody and antigen (=bait protein).

In general, the beads can be pre-incubated with antibody, unbound antibody can be washed away and the Protein G-antibody interaction can be additionally stabilized by covalently cross-linking both with the bi-functional cross-linker DMP. The antibody is then covalently bound to the beads and cannot be eluted. For IP, the protein extract is then added to beads prepared in this way.

Two other possibilities are:

(a) the pre-incubation of the beads with the antibody and subsequent addition of the lysate or
(b) the pre-incubation of the protein extract with the antibody followed by the addition of the beads.

In order to test the efficiencies of all three methods, I generated protein lysate of Blanks-FLAG / Blanks-V5 cells and performed the immunoprecipitation with either anti-FLAG or anti-V5 antibodies. The pre-incubation steps were conducted in the lysis buffer and lasted for one hour, followed by another hour of incubation when all components were combined. The supernatant of the IPs was kept together with the IP fraction for analysis via SDS-PAGE and western blotting; see Figure 6—2A.

The amount of remaining, non-depleted protein in the supernatant fraction can be used as a proxy for the efficiency of the IP: the less protein remains, the higher the IP efficiency. First, the IP efficiency of the anti-FLAG antibody is much higher than efficiency of the anti-V5 antibody. Second, the pre-incubation of the beads with the antibody results in the highest IP efficiency. Cross-linking of the beads and the antibody prior to the IP results in reduced pull down ability of the beads. It is very likely that intramolecular cross-links within the antigen binding region impair the antigen recognition and binding. Although the cross-linking has its clear benefits by reducing the amount of antibody that elutes with the IP fraction, it has to be considered that immunoprecipitation is clearly affected with respect to its efficiency.

In addition, I determined the optimal incubation time of the antibody with the beads (Figure 6—2B). I was not able to detect any difference in the IP efficiencies when experimenting with incubation times between 30 min and overnight. However, this offers the possibility to adjust the pre-incubation time to the needs of the protocol and the time available.

All in all, I could show that it is beneficial to first incubate the beads with the antibody before incubating with the lysate. The incubation can be restricted to 30 min without losing IP efficiency.

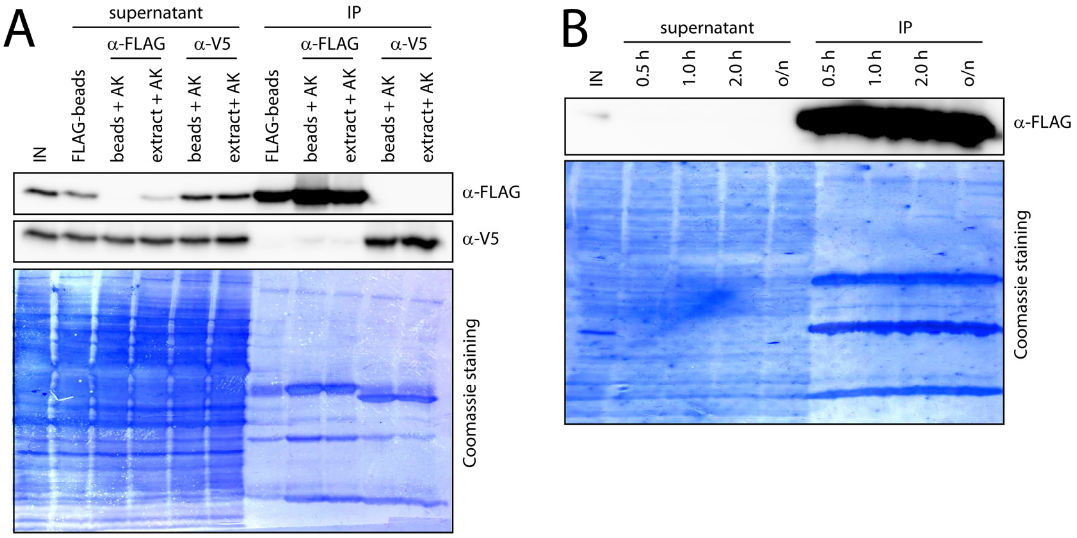

Figure 6—2: Western blots used to determine the IP efficiency of the samples. Blanks-FLAG / Blanks-V5 expressing cell lines were used. (A) The order of the addition of beads, antibody and protein extract was tested. “FLAG-beads” are Dynabeads Protein G that were pre-incubated with anti-FLAG antibody and covalently cross-linked with DMP. 5% input (IN) and supernatant of the IP reaction were loaded on the SDS-PAGE. The IP was either conducted with anti-FLAG or anti-V5 antibody. The Commassie staining of the membrane serves as a loading control. (B) The pre-incubation time of beads with anti-FLAG antibody was determined. 2.5% input and supernatant of the IP reaction were loaded. The Commassie blue staining of the membrane serves as a loading control.

6.2.2.2 Determining the optimal formaldehyde concentration for *in vivo* cross-linking to stabilize weak interactions

As already mentioned, weak and transient interactions can be stabilized by *in vivo* cross-linking of the protein complexes. In this study, formaldehyde was used as the reactive substance due to its cell permeability, short cross-link distance and simple handling. The *in vivo* cross-linking is conducted prior to cell lysis and immunoprecipitation. The specificity of the crosslink and thus the risk of conserving unspecific interactions is determined either by the amount of cross-linker or by the incubation time before the reaction is stopped by the addition of the quencher. I decided to keep the incubation time constant at 5 min at room temperature. The amount of formaldehyde was varied to modulate the cross-linking intensity.

For *in vivo* cross-linking in *Drosophila* S2 cells, the cells were washed twice with 1x PBS and resuspended in an appropriate volume of 1x PBS to reach a concentration of ~10^7 cells per ml. An appropriate volume of formaldehyde stock solution (37 %) was added to the cell suspension in order to create the desired concentration in the cell suspension. The sample was incubated for 5 min at room temperature on a rotating wheel and then the reaction was quenched with 100mM glycine. The cells were washed again with 1xPBS and the standard lysis and immunoprecipitation protocol was applied.

In order to determine the optimal formaldehyde concentration, I used Hrb27C-FLAG cells and titrated the formaldehyde concentration in different samples. Hrb27C is a highly abundant protein, which allows for the detection of small changes in IP efficiency. In general, two effects occur after cross-linking: First, the weak and transient interactions of the protein of interest are stabilized. Second, with increasing concentration of formaldehyde unspecific interactions are captured, which make both (a) the cell lysis more complicated due to the establishment of large macromolecular networks and (b) the interpretation of the results difficult. Consequently, I tried to find the lowest amount of formaldehyde that is necessary to obtain a stabilizing effect but still tried to prevent the linkage of unspecific interactions with the bait protein.

The amount of protein that could be successfully immunoprecipitated decreased with higher formaldehyde concentrations (Figure 6—3A). In contrast, the intensity of complexes of higher molarity that are visible on the Commassie staining of the membrane increases, starting at a concentration of 0.25 % formaldehyde. In general, no protein complexes should be visible on the western blot since the boiling of the sample in the SDS and DTT containing buffer reverts the cross-links. Nevertheless, the appearance of these bands and the smear may indicate that upon cross-linking very stable complexes are generated, which might also interfere with sample preparation for mass spectrometry.

An excess of cross-linker that was not efficiently quenched after the crosslink and therefore is still active during the IP reaction can be detected when the antibody complexes are not successfully reduced during the boiling process prior to the SDS-PAGE anymore. The reversion of the cross-links seems not to be possible for larger amounts of introduced formaldehyde. The band that corresponds to the unreduced IgG antibody (the one that was used for the pull down) can be detected at around 160 kDa. This band is visible due to the fact that a secondary antibody that detects the murine IgG was used during western blotting. Since these IgG complexes remain stable in spite of the boiling, it is likely that also additional complexes are excessively stabilized, which may interfere with further analysis.

Thus, I checked whether the lowest concentration of formaldehyde used (0.1 %) was already sufficient to enrich interactors. I compared the protein levels that co-purify with Hrb27C-FLAG with and without cross-linking by mass spectrometry (Figure 6—3B). An algorithm is implemented in the analysis tool MaxQuant that can quantify the amount of protein in a sample in a label-free manner. The resulting values were log-transformed and normalized (z-score of the sample). For proteins that were only detectable in one of the conditions, abundance values were imputed from the standard distribution of the corresponding sample values. Subsequently, the levels of each identified protein can be depicted in a scatter plot comparing both conditions. While no changes in the abundance of the bait protein (Hrb27C, red) and its known constitutive interactors (blue) could be observed, many other proteins are more abundant in the cross-linked sample. To discriminate which of these proteins are background (binding to beads) and unspecific (binding to proteins *per se*) binders or "real" interactors, negative controls have to be performed in parallel.

All in all, it seems as if 0.1 % formaldehyde is sufficient for the enrichment of weakly interacting proteins in the IP sample.

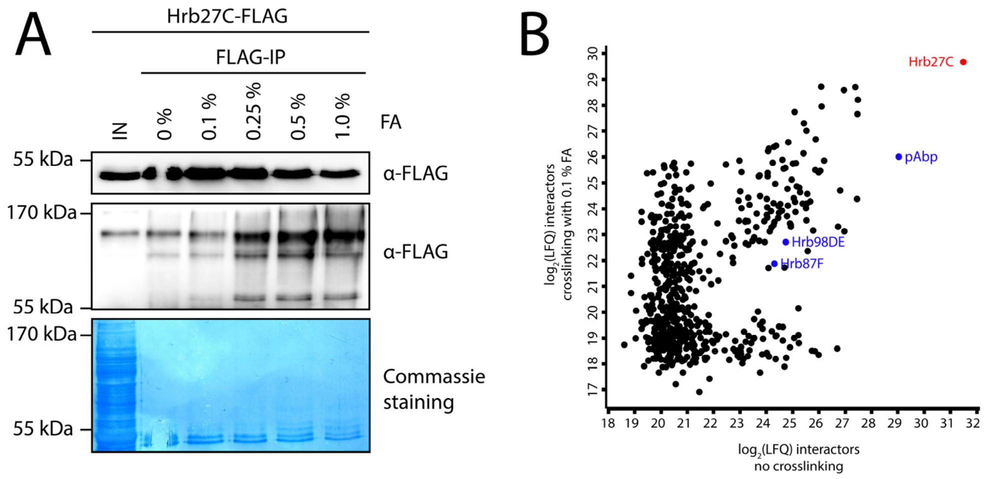

Figure 6—3: Assays to determine the appropriate formaldehyde concentration for *in vivo* cross-linking in *Drosophila* S2 cells. Hrb27C-FLAG expressing cell lines were used in the experiments. (A) Western blot showing the IP recovery and efficiency after cross-linking with different amounts of formaldehyde. After lysis, the same amount of protein was added to the beads. 5 % input (IN) was loaded on the SDS-PAGE. The Commassie blue staining of the membrane serves as a loading control. (B) Scatter plot of the protein levels of proteins that co-purify with Hrb27C either with in vivo cross-linking or without. Each data point represents one identified protein. Label-free quantification (LFQ) values are normalized to the z-score of the sample, missing values were imputed.

6.2.2.3 Optimizing the elution of the bait protein and its interactors

The easiest way to analyze the immunoprecipitated protein and its interactors on a western blot is to boil the washed beads in SDS and DTT containing loading buffer. However, this method results in rather “dirty” lanes since also all background binders that are associated with the matrix of the beads, the Protein G and the antibodies are eluted.

For some applications, it is necessary to elute the proteins more mildly. One possibility would be the use of glycine or arginine solutions at acidic pH values, which break up the interaction between the anti-FLAG M2 antibody and the FLAG epitope probably due to the protonation of the lysines within the FLAG peptide sequence. This change in the electrostatic properties interferes with the antigen-antibody interaction (Futatsumori-Sugai et al., 2009). However, this elution method does not work for the V5 tag.

When epitope tags are linked to the protein via a protease recognition site such as for the TEV or Prescission protease, the elution can be performed very mildly by incubating the beads with the corresponding protease. Native complexes can be thereby eluted without contamination of antibodies.

6.2.2.3.1 Arginine elutes FLAG-tagged proteins very efficiently

First, I checked the efficiency of 0.1 M glycine, pH 2.5 and 0.75 M arginine, pH 3.5 for elution of the FLAG-tagged Blanks protein from Dynabeads after immunoprecipitation with anti-FLAG antibody. The eluates, the proteins that are still bound to the beads after elution (= boiled beads after elution) and – as a control – the non-eluted beads were analyzed on a SDS-PAGE and stained with Commassie blue (Figure 6—4).

The elution with glycine does not seem to work properly since no band can be detected for the elution fraction and there is still Blanks-FLAG protein bound to the beads in a comparable amount as for the control. Arginine solution, however, can efficiently elute the protein form the beads (dashed box) and almost no protein remains on the beads (arrow).

In summary, 0.75M arginine, pH 3.5 can be used to elute FLAG-tagged proteins from beads that were incubated with anti-FLAG antibody. However, this method cannot be used for the other widely used epitope tag V5 which contains no lysines or other basic amino acids that can be protonated in order to break up the antigen-antibody interaction.

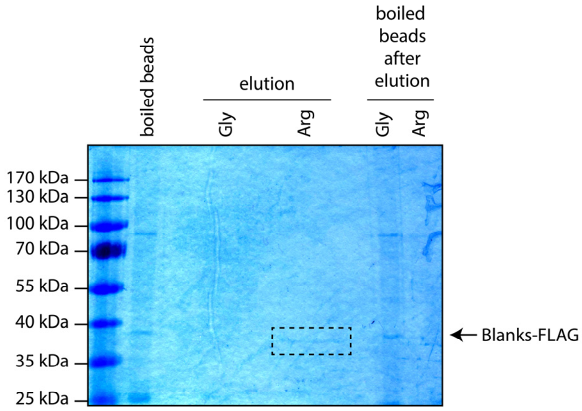

Figure 6—4: Commassie-stained SDS-PAGE that shows the immunoprecipitated proteins which were eluted by glycine or arginine solutions. The remaining proteins that were not eluted successfully are visible in the fraction of the boiled beads after elution. As a control, an equal amount of beads that were not subjected to elution are boiled. The dashed box highlights the eluted Blanks-FLAG protein.

6.2.2.3.2 Proteases elute native complexes

Next, I evaluated the use of site-specific proteases to elute the protein complexes from the beads. The V5 epitope tag used for the tagging of proteins at their chromosomal locus is separated from the protein by a TEV cleavage site. Moreover, we designed an FLAG epitope tag that contains a Prescission cleavage site (pFLAG). For the experiments, either Act5C-pFLAG or pAbp-V5 expressing cell lines were used.

First, I evaluated the optimal time for the cleavage reaction using the Prescission protease. Cell lysis and immunoprecipitation was performed as described before. The washed beads were resuspended in 30µl of the cleavage buffer that was provided by the manufacturer and 2 units of Prescission protease (Pierce, Thermo Scientific) were added. The reaction was incubated at 18°C from 15 min to two hours. The eluate and the proteins that remained on the beads were analyzed on a SDS-PAGE (silver stain); see Figure 6—5A. With increased incubation time, the band that corresponds to Act5C-pFLAG becomes fainter in the beads fraction, suggesting that prolonged elution is beneficial for the reaction. The eluted protein is not visible since it is mostly concealed by the band that corresponds to the Prescission protease.

Next, I tested which incubation temperature yields the best results. I incubated the cleavage mixture over night at either 4°C, 18°C, room temperature or 37°C. Although for all temperatures protein that is still bound to the beads can be detected, the highest cleavage efficiency could be observed for 4°C (Figure 6—5B).Consequently, I used incubation at 4°C over night for elution during the following experiments.

Since the clean and mild elution with the proteases should be preferentially used for subsequent analysis of the sample by mass spectrometry, it is undesirable to “contaminate” the sample with excessive amounts of protease whose peptides would decrease the detection sensitivity for other proteins due to their abundance. Therefore, I tried to deplete the protease after elution by utilizing the His_6-tag that is fused to both the TEV and Prescission protease. To this end, I incubated the eluate with magnetic Ni-NTA beads for 1h at 4°C and saved the supernatant, which should contain decreased levels of the protease. For both proteases, I was able to deplete most of the enzyme by the Ni-NTA beads and only a small fraction remained in the supernatant (Figure 6—5C).

Altogether, I established a mild elution method that enables the elution of native protein complexes from beads by proteases. Moreover, the enzymes could be removed by using an additional subtractive pull down step.

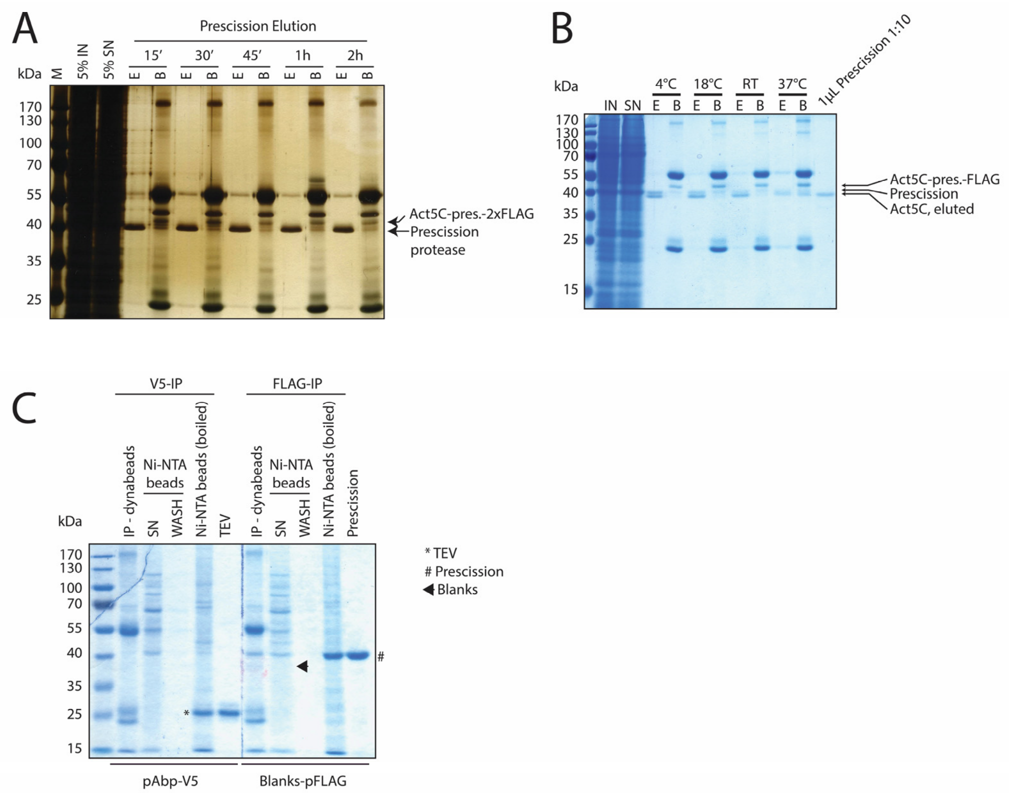

Figure 6—5: Elution of the immunoprecipitated proteins using Prescission or TEV proteases. (A) Act5C-pFLAG was immunoprecipitated and the bait protein was eluted from the beads by Prescission protease at 18°C for various incubation times. The eluate (=E) and the beads after elution (=B) were analyzed on a SDS-PAGE that was subsequently silver stained. Besides the heavy and light chain of the IgG the bound Act5C-pFLAG and the protease are also visible (annotated). (B) Act5C-pFLAG was immunoprecipitated and eluted from the beads overnight at different temperatures with Prescission protease. (C) Commassie stain of the subtractive Ni-NTA pulldown of the His tagged proteases which was conducted in order to deplete the eluateof protease. pAbp-V5 expressing and Blanks-pFLAG expressing cells were used. A sample of the IP fraction (Dynabeads) and the Ni-NTA beads treated supernatant and the wash of the Ni-NTA beads was analyzed as well as the Ni-NTA beads after the reaction. Some of the proteins are annotated.

6.2.3 Comparing the parameters of sample preparation for mass spectrometry-based analysis of the IP sample

Prior to analysis of the immunoprecipitated sample by mass spectrometry, the proteins have to be digested by trypsin into smaller peptides that can be analyzed via a mass spectrometer using the LC-MS/MS method (Mann et al., 2001; ten Have et al., 2011).

Two general approaches are possible:

(a) The protein sample can be eluted from the beads and the eluate is used for the tryptic digestion.
(b) The tryptic digestion reaction can be used to "elute" the proteins from the beads. In this case, beads are resuspended in a trypsin containing buffer and incubated for 30 min, so that the enzyme can cleave the proteins into shorter peptides, which remain in the supernatant. The supernatant can then be used for further sample preparation.

Traditionally, the digested and peptide containing samples were further purified in a reverse phase chromatography step and desalted using C18 stage tips. The peptides bind to the matrix while salt and other residual contaminants can be washed away in order to gain an ultra-pure peptide sample. For some mass spectrometers, it is not necessary to perform this step which has an inherent risk to lose peptides and thereby experimental data. Those mass spectrometers perform an additional purification step during the loading of the sample onto the HPLC column, which is upstream of the actual mass spectrometer and ESI unit.

In this section, I compared these two different sample preparation methods:

(a) No elution of the protein complexes from the beads but rather tryptic digestion of the proteins on the beads or
(b) Prescission protease mediated elution and subsequent tryptic digestion using the FASP protocol (Wisniewski et al., 2009).

Moreover, the peptides generated as described above were either further purified and the sample desalted or this step was omitted. Blanks-pFLAG expressing cell lines were used to evaluate the different methods. Therefore, the IP sample was split into two equal fractions, one of which was subjected to tryptic digestion on the beads and Prescission protease was added the other. The protease in the supernatant was then depleted with Ni-NTA beads and tryptic digestion was performed *in solution*. Again, the samples were equally split and one fraction was desalted. The final samples were lyophilized, resuspended in an equal volume and analyzed by mass spectrometry.

As summarized in Figure 6—6 and Table 2, most proteins (52) were identified in the non-desalted sample which was tryptically digested directly on the beads. 22 proteins of those were also recovered in the desalted sample. Dramatically fewer proteins (5 or 6) were identified after the elution by Prescission protease. Additionally, the bait protein Blanks is not the most abundant protein in the Prescission eluted fraction. Inefficient elution can explain this phenomenon where potential interactors or background binders are enriched. Since no negative control was prepared in parallel for this analysis, it is not possible to discriminate between those two possibilities.

Moreover, it became clear that during the desalting process many peptides were lost, which results in fewer identified proteins in the sample. For the method "digestion on the beads" 30 additional proteins were identified compared to the desalted fraction.

Consequently, to achieve the most sensitive analysis of co-immunoprecipitating proteins the sample should be digested directly on the beads and no further desalting steps are necessary. Since this protocol is highly sensitive, it is crucial to work with appropriate negative controls in parallel to identify potential unspecific binders.

Table 2: Overview of the identified proteins resulting from the different sample preparation methods. The proteins are ranked according to their abundance in the sample. The bold written proteins in the first column are also identified in the corresponding non-desalted sample.

digestion on beads		elution	
desalted	not desalted	desalted	not desalted
blanks	blanks	eIF-4B	eIF-4B
CG18501	CG18501	blanks	blanks
Map205	Map205	Hsc70-4	CG13090
eIF-4B	eIF-4B	CG7911	Hsc70-4
Hsc70-4	Hsc70-4		Hsc70-1
CG15415	CG15415		
Rcd1	Ir76a		
ncd	Rcd1		
mip120-RA	Hex-A		
Act5C	ncd		
CG13151	Act5C		
CG8478	mip120-RA		
Ef1alpha48D	CG13151		
Hsc70-3	CG8478		
dre4	Ef1alpha48D		
RpL28	CG13367		
CG2129	Hsc70-3		
RpL30	dre4		
pnr	CG2129		
CG1633	RpL28		
Ssrp	Ssrp		
RpL32	RpL30		
RpS15	pnr		
	RpLP1		
	ebi		
	RpS17		
	BcDNA.LD23876		
	RpL11		
	RpL23A		
	CG1633		
	RpS10b		
	sta		
	RpL14		
	RpS3		
	RpL13		
	RpS3A		
	Ef1gamma		
	CG3662		
	RpS8		
	RpL22		
	RpL18		
	RpS6		
	zip		
	RpL8		
	RpL32		
	CG10984		
	RpL31		
	CG6686		
	B52		
	RpS21		
	Rack1		
	RpL24		

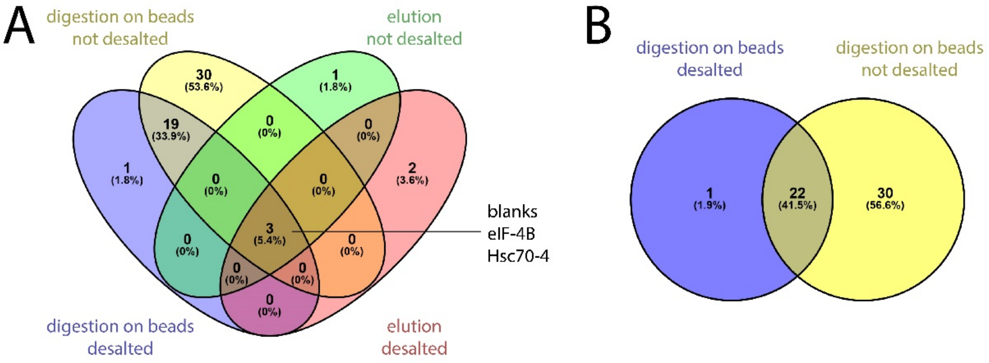

Figure 6—6: Venn diagrams representing the amount of identified proteins after the different sample preparation methods.

6.3 Conclusions

All in all, I was able to establish an optimized cell lysis and immunoprecipitation protocol for *Drosophila* S2 cells that is fast and robust. *In vivo* cross-linking can be used to stabilize weak interactions prior to immunoprecipitation. The analysis of the interactome can be performed by mass spectrometry. Therefore, I established several protocols for sample preparation and for the analysis of the data. The optimized lysis and IP protocol is described in section9.2.2.

In general, it is important to conduct the experiments with appropriate negative controls. Either naïve S2 cells with no epitope tagged proteins can be used for immunoprecipitation or the pull down can be performed with an antibody of the same isotype but different specificity. Both controls are useful to detect proteins that bind preferentially to the bead matrix or the Protein G-antibody complex. However, immunoprecipitation of different proteins carrying the same tag such as heterologously expressed luciferase or the abundant Act5C can serve as negative controls. Each method comes with its own advantages and disadvantages; consequently the best control has to be chosen for each experimental setup. During my optimization attempts, I found that there is no universally applicable optimal control. I therefore recommend including several different negative controls in order to evaluate the data to its full potential.

Moreover, it might be beneficial to combine the analysis of *in vivo* cross-linked samples with native, un-cross-linked ones. This may help to discriminate between real interactors and contaminants but also can give insights into transient and weak binders. Finally, the interactors resulting from the mass spectrometry should be validated by the traditional co-immunoprecipitation approach followed by western blotting. Due to the existence of the CRISPR/Cas9-mediated genome engineering protocol I described before, it is fast and convenient to generate a cell line that expresses an epitope tagged bait protein and the potential interactor with a different tag.

Despite the aforementioned pitfalls, the described and established method is an extremely powerful tool that – especially in combination with CRISPR/Cas9-mediated tagging of proteins at their chromosomal loci – can help to elucidate the interactome and function of so far unknown proteins.

7 Biochemical role of the double-stranded RNA binding protein Blanks for endo-siRNA biogenesis

7.1 Introduction

7.1.1 The discovery of Blanks

In 2008, Zhou and colleagues conducted a genome-wide screen in *Drosophila* cells to identify factors involved in the small RNA pathways and identified Blanks (CG10630) as a positive regulator of siRNA function (Zhou et al., 2008). Three years later, in 2011, two papers were published that further characterized Blanks and proposed mechanistic details on its function. Sanders and Smith provided evidence that Blanks (also called lump) is required for male fertility but not for efficient siRNA biogenesis and function (Sanders and Smith, 2011). Gerbasi et al., however, postulated that Blanks interacts with Rm62, a RNA helicase, CG6133, the predicted homolog of human Nsun2, and Xrn2, a 5'-3' exonuclease. They argued that this complex functions as a novel RISC (Figure 7—1). Gerbasi and colleagues already described the highly specific expression pattern in testes and *Drosophila* Schneider cells and the predominantly nuclear localization of Blanks. Their findings are consistent with the report of Sanders and Smith that male flies depleted of Blanks are infertile and they were able to show that the spermatogenesis is impaired due to the asynchronous individualization of the spermatids. Moreover, they were able to detect an upregulation of genes that are important for innate immunity or stress responses upon depletion of Blanks (Gerbasi et al., 2011).

Together, these data show that Blanks is involved in the RNAi pathway under specific conditions such as during sperm maturation. Due to their re-activation of blanks expression, S2 cells represent an attractive experimental model to study its biochemical function with respect to the RNAi pathway.

In the Förstemann lab, we conducted a genome-wide screen in S2 cells to identify proteins involved in RNA interference triggered by a DNA double-strand break. The significant candidates were re-screened in order to the involvement in efficient silencing of reporter genes that are integrated in high copy numbers in the genome and give rise to natural siRNAs repressing the reporter gene. This situation mimics TEs. For both scenarios, Blanks came up as a highly reproducible and positive regulator of RNAi emphasizing the physiological role of Blanks (Merk et al., 2017, *PLoS Genetics, in press*).

In addition, Blanks was identified as a specific interactor of HP1a, the heterochromatin mark that is amongst others necessary for the stable repression of TEs, and Blanks' functional involvement in position effect variegation of a reporter gene. Both for S2 and Kc167 cells a co-localization to heterochromatic regions within the genome was observed consistent with the interaction data (Swenson et al., 2016).

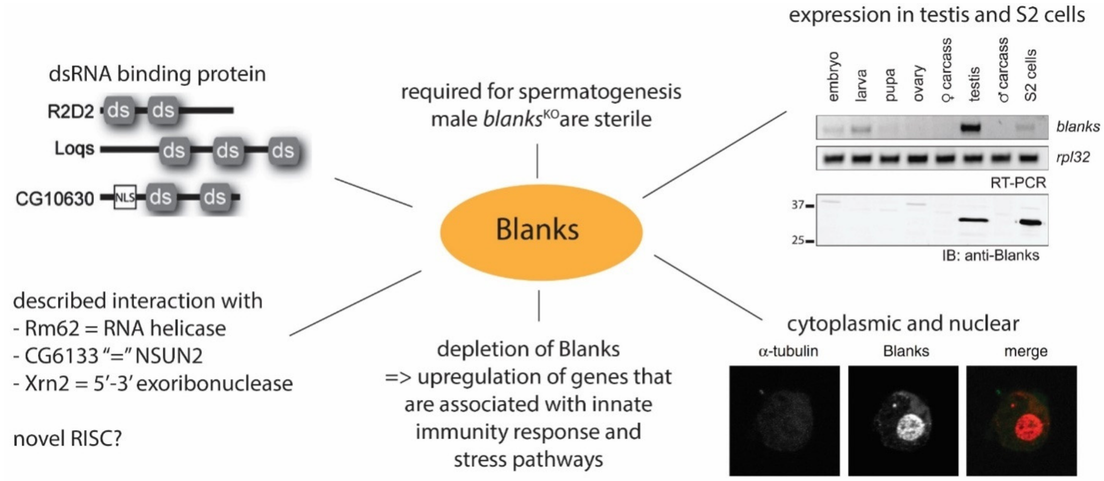

Figure 7—1: Graphical summary of the results by Gerbasi et al, 2011. CG10630 is the former gene name of Blanks.

7.1.2 Bioinformatic analysis

Bioinformatic analysis of Blanks revealed its homology to other dsRNA binding proteins (dsRBP) of *Drosophila* such as R2D2 or Loquacious. Blanks consists of an N-terminal part (103aa), followed by two dsRNA binding domains (dsRBDs) of 74aa and 73aa in length, which are separated by a linker region of 55aa's. A nuclear localization sequence (NLS, NGRKKQKKNKKAKIR) is placed within its N-terminal region as predicted by several web-based tools such as NLS mapper or NucPred.

The topology of a canonical dsRBD is highly conserved: A β-sheet with three strands is flanked by two α-helices (αβββα).Three regions are important for the binding of the protein to dsRNA as annotated in Figure 7—2B. Region 1 lies within α1 and consists of the amino acids glutamine and glutamate which interact with the 2'OH of the ribose ring via hydrogen bonds. Region 2 is located between β1 and β2 and also forms hydrogen bonds with the 2'OH's of the ribose via a conserved GxPH motif. Three conserved lysine residues (KKxAK) within region 3 are binding to the phosphodiester backbone of the RNA. Region 1 and 2 bind to the minor groove of the A-form helix of the dsRNA, region 3, however, to the major groove. Due to the fixed distances of the binding regions, dsRBDs are able to discriminate between the shape of A-formed dsRNA and B-formed dsDNA, which ensures the substrate specificity of the dsRBDs. Consequently, the dsRBPs do not recognize their substrates in a sequence specific manner but rather recognize the shape of the A-form dsRNA.

Blanks shows also homology to dsRBDs that are involved in RNA interference from other organisms, for example the human dsRBPs TRBP and PACT. Both proteins are known interactors of Dicer and functionally resemble *Drosophila* Loqs and R2D2 in small RNA biogenesis.

However, not all dsRBDs of the mentioned proteins Loqs, TRBP and PACT are as conserved that they easily match the consensus dsRBD sequence. For TRBP, the human ortholog to *Drosophila* Loqs, the third dsRBD is degenerate and unable to bind dsRNA; rather, it mediates the interaction with Dicer (Wilson and Doudna, 2013). The third binding domain of human PACT, also a Loqs ortholog, mediates protein-protein interaction and does not participate in the binding of dsRNA. Similarly, the third dsRBD of the PB isoform of Loqs (Loqs-PB) is not involved in RNA binding but rather necessary for the homodimerization of two Loqs-PB molecules or the binding of one monomer to Dcr-1 (Jakob et al., 2016). This difference is also visible in the underlying amino acid sequence, since all three domains lack the highly conserved KKxAK motif that is – based on the structural analysis – crucial for the interaction with the dsRNA (Figure 7—2C).

When comparing the amino acid sequence of both dsRBDs of Blanks with known representative domains of other dsRBPs, the phylogenic analysis revealed that dsRBD1 of Blanks is more closely related to the protein interaction domains of Loqs, TRBP and PACT, whereas the dsRBD2 is related to the classical dsRBDs that are able to bind dsRNA (Figure 7—2A). However, it looks as if the conservation of both dsRBDs is less pronounced than for R2D2 or Loqs, which fits well with the distinct expression pattern of Blanks and the fact that most tissues are RNAi-proficient without Blanks. Therefore, Blanks may be required for specific processes in the RNAi and is not essential for the standard RNA interference pathway.

Moreover, the classical KKxAK motif in the dsRBD2 is changed into a KKxAR pattern similar to the dsRBD2 in R2D2 (Figure 7—2B). Additionally, the surrounding sequence is less conserved than for the other known factors. The QE motif in region 1 and the GxPH motif of region 2 are degenerated as well. These changes, however, might offer the possibility for Blanks to gain substrate specificity to distinctly modified dsRNA that may differ in its secondary structure from the canonical A-form helix due to bulges or mismatches.

When submitting the amino acid sequence of Blanks to homology-based structure prediction tools (HHpred), the two dsRBDs are identified with high fidelity. Moreover, a region upstream of the dsRBD1 bears homology to the murine NF90 protein. NF90, also known as nuclear factors associated with dsRNA (NFAR), belongs to the dsRBP family as well. There is experimental evidence that NF90 proteins are involved in the host defense against viruses and regulate selectively mRNA levels and the export of RNAs from the nucleus, which are normally exported by the interaction with exportin. Furthermore, NF90 proteins seem to bind to nuclear export factors, especially to exportin-5 which is also responsible for the export of pre-miRNAs (Barber, 2009).

Using the homology-prediction data and the online-tool Modeller 9.14, a three-dimensional protein structure can be generated which is depicted in Figure 7—3. Since no homologous protein could be detected for the N-terminal region, this part is unstructured in the model.

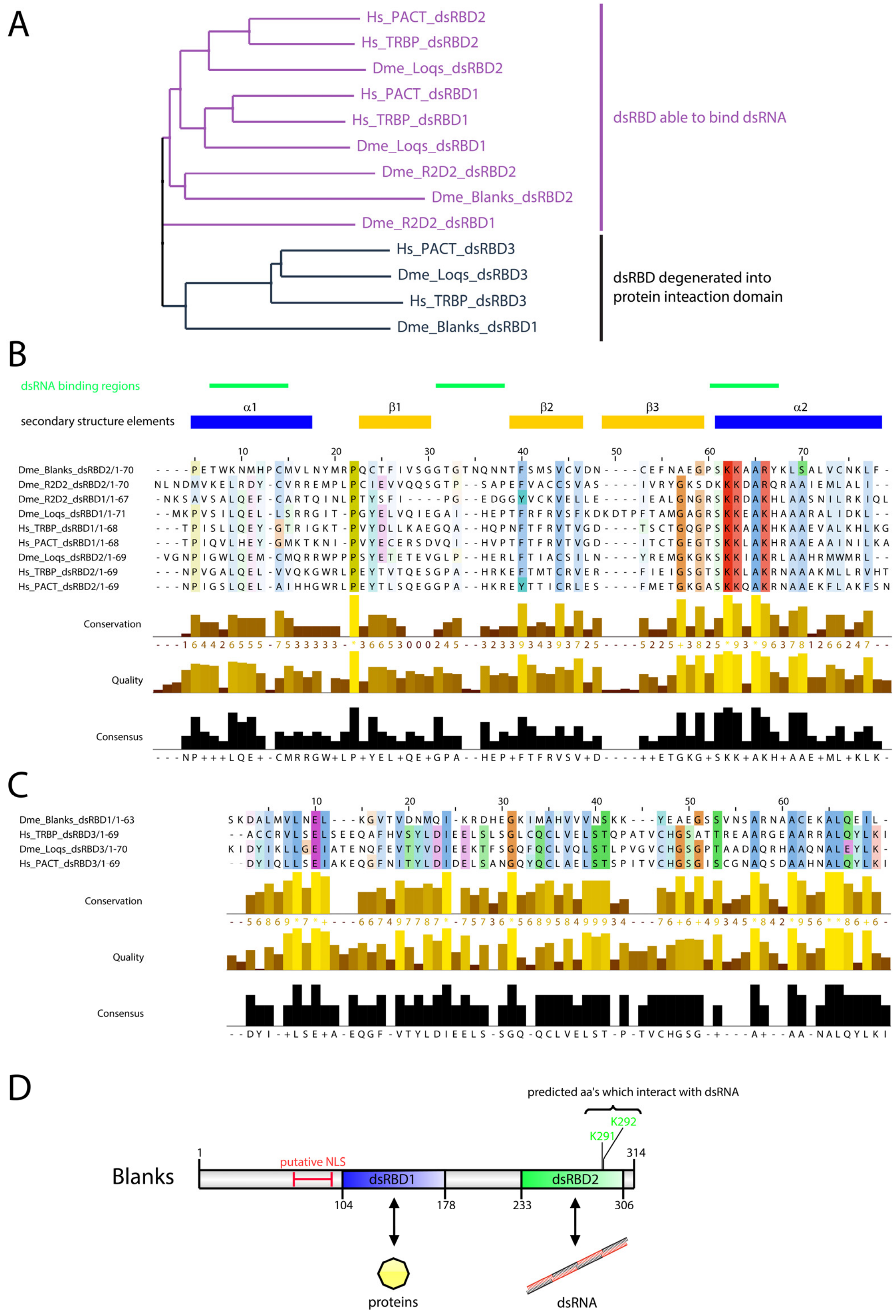
A
Hs_PACT_dsRBD2
Hs_TRBP_dsRBD2
Dme_Loqs_dsRBD2
Hs_PACT_dsRBD1
Hs_TRBP_dsRBD1
Dme_Loqs_dsRBD1
Dme_R2D2_dsRBD2
Dme_Blanks_dsRBD2
Dme_R2D2_dsRBD1
Hs_PACT_dsRBD3
Dme_Loqs_dsRBD3
Hs_TRBP_dsRBD3
Dme_Blanks_dsRBD1
dsRBD able to bind dsRNA
dsRBD degenerated into
protein inteaction domain
B
dsRNA binding regions
secondary structure elements
α1
β1
β2
β3
α2
10
20
30
40
50
60
70
Dme_Blanks_dsRBD2/1-70 ----PETWKNMHPCMVLNYMRPQCTFIVSGGTGTNQNNTFSMSVCVDN----CEFNAEGPSKKAARYKLSALVCNKLF-
Dme_R2D2_dsRBD2/1-70 NLNDMVKELRDY-CVRREMPLPCIEVVQQSGTP--SAPEFVACCSVAS----IVRYGKSDKKDARQRAAIEMLALI--
Dme_R2D2_dsRBD1/1-67 -NKSAVSALQEF-CARTQINLPTYSFI----PG--EDGGYVCKVELLE----IEALGNGRSKRDAKHLAASNILRKIQL
Dme_Loqs_dsRBD1/1-71 --MKPVSILQEL-LSRRGIT-PGYELVQIEGAI--HEPTFRFRVSFKDKDTPFTAMGAGRSKKEAKHAAARALIDKL--
Hs_TRBP_dsRBD1/1-68 ---TPISLLQEY-GTRIGKT-PVYDLLKAEGQA--HQPNFTFRVTVGD----TSCTGQGPSKKAAKHKAAEVALKHLKG
Hs_PACT_dsRBD1/1-68 ---TPIQVLHEY-GMKTKNI-PVYECERSDVQI--HVPTFTFRVTVGD----ITCTGEGTSKKLAKHRAAEAAINILKA
Dme_Loqs_dsRBD2/1-69 -VGNPIGWLQEM-CMQRRWPPSYETETEVGLP--HERLFTIACSILN----YREMGKGKSKKIAKRLAAHRMWMRL--
Hs_TRBP_dsRBD2/1-69 ---NPVGALQEL-VVQKGWRLPEYTVTQESGPA--HRKEFTMTCRVER----FIEIGSGTSKKLAKRNAAAKMLLRVHT
Hs_PACT_dsRBD2/1-69 ---NPIGSLQEL-AIHHGWRLPEYTLSQEGGPA--HKREYTTICRLES----FMETGKGASKKQAKRNAAEKFLAKFSN
Conservation
Quality
Consensus
---NP+++LQE+-CMRRGW+LP+YEL+QE+GPA--HEP+FTFRVSV+D----++ETGKG+SKK+AKH+AAE+ML+KLK-
C
10
20
30
40
50
60
Dme_Blanks_dsRBD1/1-63 SKDALMVLNEL---KGVTVDNMQI-KRDHEGKIMAHVVVNSKK---YEAEGSSVNSARNAACEKALQEIL-
Hs_TRBP_dsRBD3/1-69 --ACCRVLSELSEEQAFHVSYLDIEELSLSGLCQCLVELSTQPATVCHGSATTREAARGEAARRALQYLKI
Dme_Loqs_dsRBD3/1-70 KIDYIKLLGEIATENQFEVTYVDIEEKTFSGQFQCLVQLSTLPVGVCHGSGPTAADAQRHAAQNALEYLK-
Hs_PACT_dsRBD3/1-69 --DYIQLLSEIAKEQGFNITYLDIDELSANGQYQCLAELSTSPITVCHGSGISCGNAQSDAAHNALQYLKI
Conservation
Quality
Consensus
--DYI-+LSE+A-EQGF-VTYLDIEELS-SGQ-QCLVELST-P-TVCHGSG-+---A+--AA-NALQYLKI
D
predicted aa's which interact with dsRNA
K292
K291
1
314
putative NLS
Blanks
dsRBD1
dsRBD2
104
178
233
306
proteins
dsRNA

Figure 7—2: Bioinformatic analysis of Blanks. (A) Neighbor-joining tree without distance corrections of both dsRBDs of Blanks with other known dsRBPs. dsRBD1 clusters with dsRBD3 of Loqs, TRBP and PACT and is therefore very likely involved in protein-protein interactions. dsRBD2 is highly related to the canonical dsRBDs of R2D2, Loqs and the human orthologs. Hs, human; Dme, *Drosophila melanogaster*. Sequence alignment of canonical dsRBDs (B) respective dsRBD2 of Blanks and protein-protein interaction domains are shown in (C). The alignments were performed using MUSLE. (D) Predicted domain structure of Blanks, in which dsRBD1 is probably incapable of dsRNA binding, whereas dsRBD2 can interact with dsRNA.

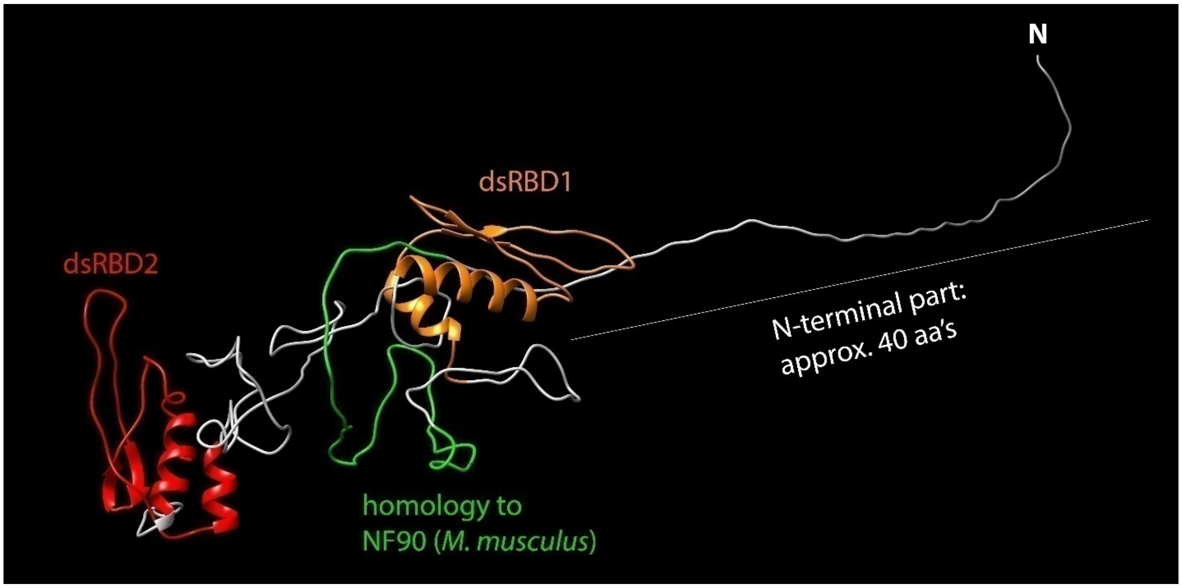

Figure 7—3: Predicted protein structure of Blanks using HHpred and conserved motif search. dsRBD1 (orange), dsRBD2 (red) and a region that is homologous to the murine NF90 protein (green) are highlighted. The model was generated using Modeller 9.14.

7.1.3 The aim of this project

Based on the results of the reporter assay as described in section 7.1, where Blanks was identified as a positive regulator of RNA interference and published observations, several possible roles of Blanks in the RNAi pathway can be postulated, as described in Figure 7—4.

First, it has to be checked whether Blanks is able to bind to dsRNA and interacts with Dcr-2 – similar to Loquacious and R2D2. Next, it has to be examined if Blanks plays a role in the processing and function of siRNAs derived from exogenous or endogenous dsRNA. Exogenous dsRNAs are present upon viral infection of cells or after application of dsRNA to the cell culture medium in order to induce a knockdown of specific genes. Although the dsRNA source of all previously cited experiments was nuclear, an involvement of Blanks in the exo-siRNA pathway would give hints for its precise role in the RNAi process (Zhou et al., 2008). If Blanks was involved in both the exo- and the endo-siRNA pathway, whose biogenesis routes converge at the processing of dsRNA by Dcr-2, the function of Blanks should be downstream of the dicing step.

If Blanks is involved in the endo-siRNA pathway selectively, it might mediate the shuttling of Dcr-2 or the export of dsRNAs which are generated from endo-siRNA loci such as TEs, cis-NAT loci or convergent transcripts. Furthermore, Blanks might be involved in the processing of siRNAs by Dcr-2 or in the loading of mature siRNAs onto Ago2. Finally, Blanks could facilitate the translational regulation of proteins via siRNAs, comparable to the function of miRNAs.

In this study, I aimed to characterize the mechanisms of Blanks' function in the RNAi pathway. To this end, I used S2 cells as a well-characterized model system for *Drosophila* biology.

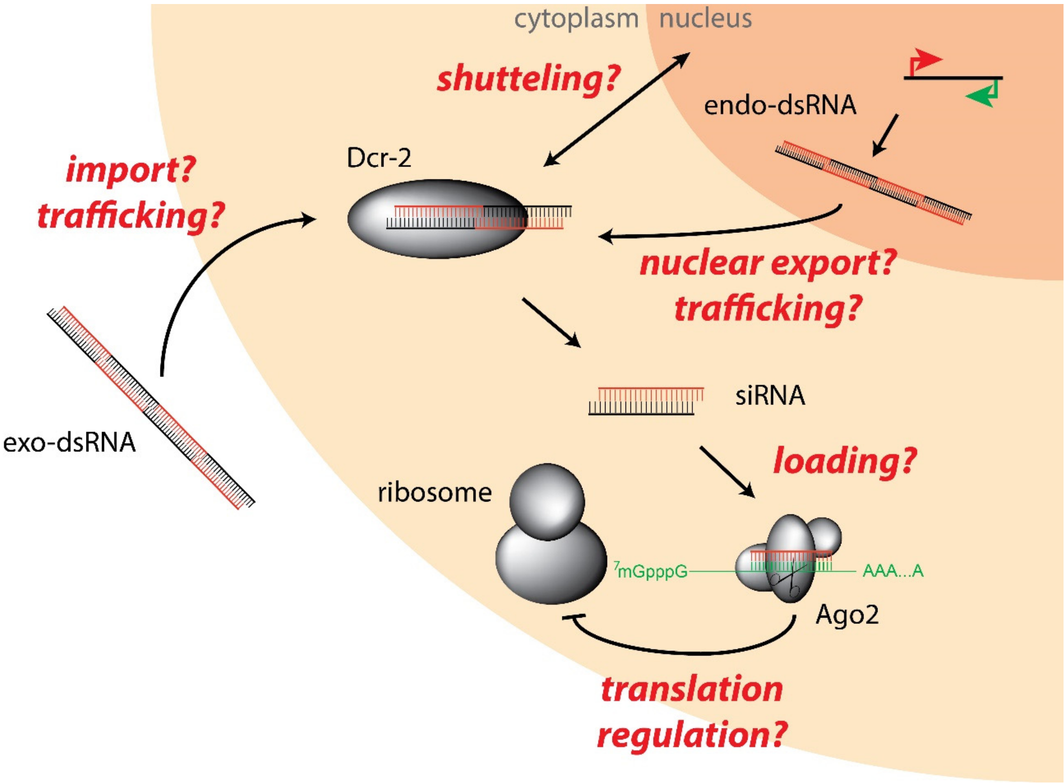

Figure 7—4: Model for the potential involvement of Blanks in the RNAi pathway. The siRNA pathway in *Drosophila* has two branches, and Blanks may participate in both: the exo-siRNA pathway that is triggered after viral infection or the endo-siRNA pathway that is fed by dsRNA which derives from TEs. In general, Blanks might be necessary for the proper processing or loading of siRNAs as well as for secondary effects such as shuttling of Dcr-2 or export of dsRNA from the nucleus.

7.2 Results and discussion

7.2.1 Blanks is a dsRNA-binding protein that does not interact with Dcr-2

In order to check if the predicted dsRBD2 of Blanks is capable of RNA binding or if its slightly modified KKxAK motif prevents the interaction with dsRNA, I recombinantly expressed and purified N-terminally GST-tagged Blanks protein. I measured the binding affinity (K_D value) to various RNA substrates such as siRNAs with perfect sequence complementarity of the sense and antisense strand, miRNA duplexes with mismatches, miRNA precursors (stemloop RNA) and ssRNA. GST-Blanks was titrated (1-2000 nM) to the fluorescently labeled RNA substrates (10 nM).After excitation of the fluorophore with linearly polarized light, the remaining polarization of the emitted light depends amongst others on the proper motion of the molecule. Unbound RNA rotates rapidly so that the emitted light is de-polarized, whereas RNA bound to the protein tumbles more slowly and thus a certain degree of polarization remains in the emitted light. Exploiting this, the binding of the RNA substrates to Blanks can be quantified (Figure 7—5).

Blanks shows a high binding affinity to siRNAs, miRNAs and the miRNA precursors (K_D = 175 – 258 nM), whereas ssRNA is bound in an erratic manner with clearly lower binding affinity (Figure 7—5C). Compared to the binding affinity of full length Loqs (K_D = ~ 50 nM), the binding of Blanks to dsRNA is approx. 4-fold lower (Fesser, 2013).

The Hill coefficient n describes the cooperativity of the binding to the substrate. If n > 1, the substrate binding occurs positively cooperative. For the double-stranded substrates the Hill coefficient is between 1.4 and 2.0 which suggests that the binding seems to happen in a cooperative manner. Based on the sequence homology search, however, Blanks likely contains only one functional dsRBD. Consequently, the values for the Hill coefficients cannot be explained with the model that upon binding of the first substrate to one dsRBD the binding of the second substrate to the other dsRBD is facilitated. These effects were observed for the dsRBD1 and dsRBD2 in Loqs (Fesser, 2013).

Yet, the measurements were conducted using GST-Blanks. It is well known that GST dimerizes with K_D(GST-GST) values between << 1nM and 5.1 µM (reviewed in (Fabrini et al., 2009)). Despite the inconsistent experimental results, it is likely that Blanks is present at least in part as dimeric complex formed via the GST-tag. Consequently, cooperativity of both components of the complex could be mediated by the GST-GST interaction.

Since Blanks binds to dsRNA and shares sequence and structural homology with R2D2 and Loqs – both known interactors of Dcr-2 –, I tested whether Blanks interacts with Dcr-2 as well. To this end, I generated cell lines with FLAG-epitope tags introduced at the genomic loci of Blanks, R2D2 and Act5C and a Strep-tag at the Dcr-2 locus via CRISPR/Cas9-mediated genome editing. After *in vivo* cross-linking (0.1 % formaldehyde) to stabilize weak interactions and whole cell lysis with mild conditions, immunoprecipitation with anti-FLAG antibody was performed and subsequently checked for Dcr-2 as an interactor (Figure 7—6). When pulling on R2D2, Dcr-2 co-purified as expected, whereas for the Blanks-IP no Dcr-2 could be detected, similar to the negative control Act5C. Several repeated experiments reproduced these results.

Since the interaction might be a very rare event *in vivo* and its detection might be prevented by the sensitivity limitation of western blotting, I performed an *in vitro* GST-pulldown assay. Recombinantly expressed GST-Blanks was immobilized on glutathione sepharose and incubated with highly concentrated (~ 10 mg/ml) cell extract of *Drosophila* S2 cells. The beads were washed twice and bound proteins were analyzed via western blotting. Empty glutathione sepharose beads were used as a negative control in order to detect background binding of proteins. Using an anti-Dcr-2 antibody, co-immunoprecipitated Dcr-2 could be detected in the GST-Blanks sample. However, the western blot was inconclusive since no Dcr-2 could be detected in the supernatant of the control sample.

All in all, Blanks binds with reasonably high affinity to dsRNA but is no constitutive interactor of Dcr-2 like R2D2 or Loqs. This argues for a novel, different function of Blanks in the RNAi pathway.

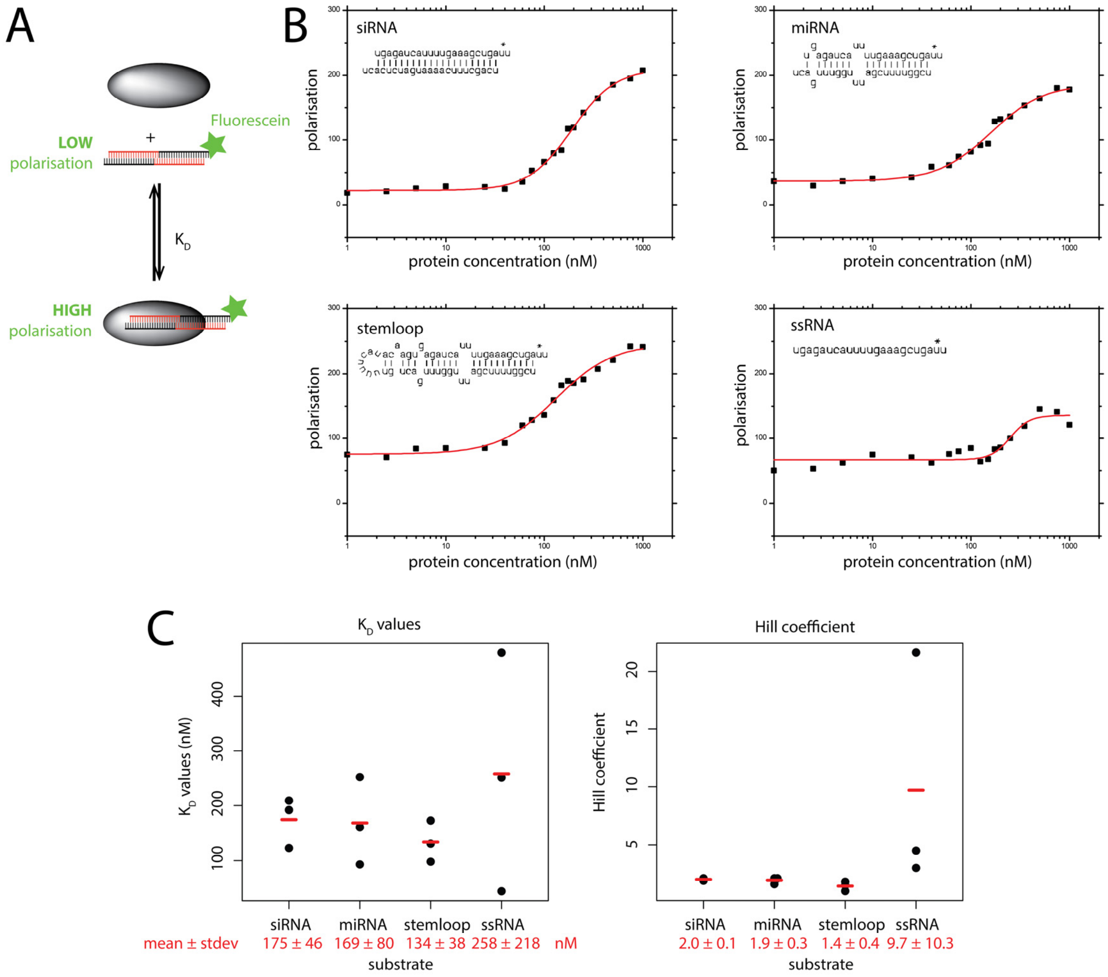

Figure 7—5: Blanks is a dsRNA-binding protein with high affinity to dsRNA.(A) Principle of anisotropy measurements to determine the KD values of Blanks with different RNA substrates. Recombinantly expressed GST-Blanks was incubated with fluorescein-labeled RNA. Emitted light of RNA bound to proteins remains highly polarized after excitation with polarized light, whereas unbound RNA emits less polarized light. These values can be used to calculate the KD value. (B) GST-Blanks was titrated to 10 nM fluorescently labeled RNA substrates and anisotropy measurements were performed. The sequence of siRNA, miRNA, miRNA-precursor (stemloop) and ssRNA derived from the bantam miRNA is depicted in the corresponding panels. Binding curves were fitted by applying the Hill1 model (y = START + (END – START) · x^n / (k^n + x^n) with k = KD and n = Hill coefficient) to the data. (C) KD values and Hill coefficients for GST-Blanks – RNA interaction. The values were obtained by fitting each experiment separately and calculating the average (depicted in red), n = 3.

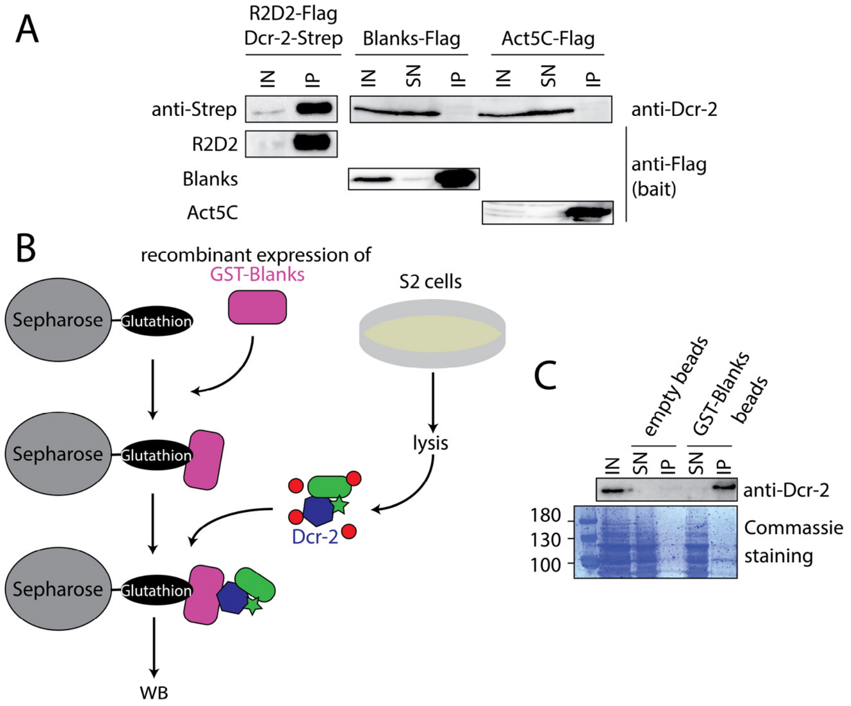

Figure 7—6: Blanks does not interact *in vivo* with Dcr-2. (A) Immunoprecipitation of R2D2-FLAG, Blanks-FLAG or Act5C-FLAG expressing cell lines using anti-FLAG antibody. Dcr-2 as potential interactor was detected using either anti-Dcr-2 or anti-Strep antibodies. (B) Schematic of GST-pulldown assay: Recombinantly expressed GST-Blanks was bound to glutathione sepharose and incubated with S2 cell extract. Proteins that bind to GST-Blanks were analyzed by immunoblotting. (C) Results of the GST-pulldown assay: Empty beads or GST-Blanks beads were incubated with S2 cell extracts and subsequently immunoblotting was performed to check for an interaction between Blanks and Dcr-2. The membrane was stained with Commassie as a loading control.

7.2.2 Blanks is not required for the processing of exo-dsRNA into siRNAs

First, I checked whether Blanks is necessary for the processing of long dsRNAs into siRNAs, when the dsRNA is derived from exogenous sources. The cells take up the dsRNA and produce functional siRNAs in a Dcr-2- and R2D2-dependent manner(Feinberg and Hunter, 2003; Zhou et al., 2014).

In order to test if Blanks is required for the generation of *bona fide* exo-siRNAs that are able to silence gene expression, I applied dsRNA which targets the GFP coding sequence to either Dcr-2 or Blanks shutdown (SD) cell lines (see chapter3.2.2) and transfected them subsequently with a plasmid that expresses GFP constitutively. In a Dcr-2 depleted background, an 8-fold derepression of the GFP fluorescence could be detected by flow cytometry, as expected. In contrast, the presence or absence of Blanks has no influence on the silencing of GFP (Figure 7—7A and B).

This result was validated by another, independent read-out (Figure 7—7C). In Blanks SD cells (FLAG-Blanks A2) which have an additional V5-tag at the C-terminus of the protein pAbp, the expression of Blanks was induced by the application of copper to the medium and a subsequent knockdown of pAbp by soaking of dsRNA targeting this gene was initiated. The protein levels of pAbp were detected by western blotting. No difference in the knockdown efficiencies could be noticed as a function of the expression of Blanks, which supports the finding of the first experiment.

In summary, these results demonstrate that Blanks is dispensable for the proper generation or function of siRNAs derived from exogenous dsRNA which is applied by soaking.

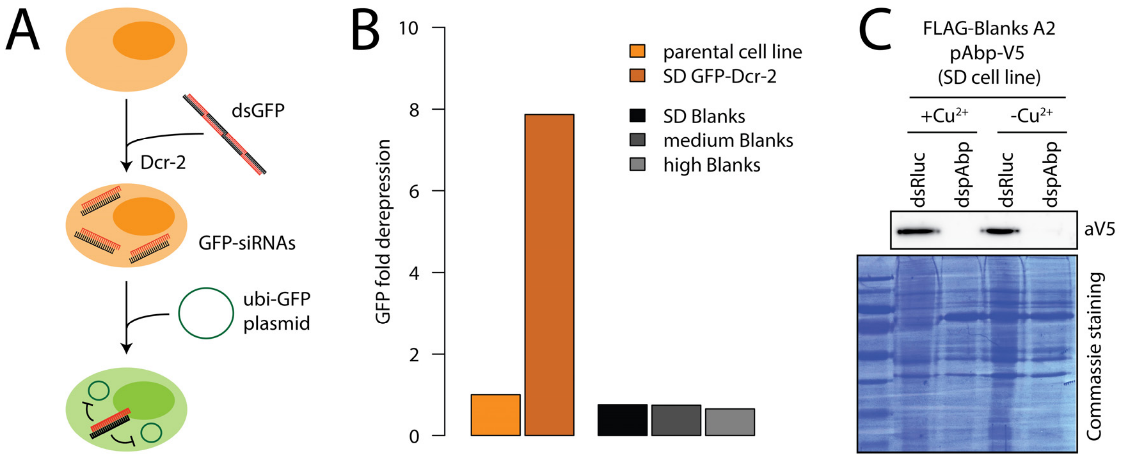

Figure 7—7: Blanks is dispensable for the exo-siRNA pathway. (A) Overview of the reporter assay to check for the role of Blanks in the exo-siRNA pathway. Cells are treated with dsRNA targeting GFP (dsGFP) by adding the dsRNA to the cell culture medium. After uptake of the dsRNA, it is processed into siRNAs in a Dcr-2 dependent manner. Subsequent transfection of a plasmid coding for GFP results in the expression of GFP which is, however, suppressed by functional siRNAs. Consequently, depletion of factors which are necessary for the processing of dsRNA into functional siRNAs that are also loaded into Ago2 to fulfill their suppressive function results in a derepression of the GFP signal. (B) Flow cytometric analysis of the reporter assay from (A) in cells that are depleted for Dcr-2 or Blanks. Dcr-2 shutdown (SD) leads to a derepression of GFP, Blanks protein levels have no influence on the GFP suppression. (C) Alternative assay to demonstrate that Blanks is not necessary for proper exo-siRNA function. Blanks expression in Blanks SD cells that constitutively express V5-tagged pAbp was either induced or not and dsRNA targeting pAbp was applied to the cell culture medium. The dsRNA was taken up by soaking. Western blotting shows that knockdown of pAbp works independently of the expression of Blanks. pAbp was detected by anti-V5; the membrane was stained with Commassie as a loading control.

7.2.3 Endogenous TEs are slightly de-repressed, TE-mapping siRNAs biogenesis is slightly impaired and the loading onto Ago2 is unaffected upon Blanks depletion

Since Blanks was identified as a positive regulator of RNAi using – amongst others – a reporter cell line that mimics TEs, I checked if endogenous TEs, the physiological targets of RNAi, are de-repressed in a similar way as the reporter gene upon depletion of Blanks.

After knockdown of either Blanks or Dcr-2 as a positive control, total RNA was extracted from the cells and reverse transcribed to cDNA using random hexamers for priming. The transcript levels of selected, representative endogenous TEs were measured by qPCR (Figure 7—8A).Roo, blood, mdg-1 and 297 are LTR-retrotransposons; F-element belongs to the class of LINE-like elements (non-LTR retrotransposons).

Similar to the Dcr-2 depleted situation, a slight increase in TE transcript levels could be detected upon Blanks knockdown arguing for a slightly less efficient silencing of the TEs. However, the effect is less pronounced than for Dcr-2 depletion. Moreover, no correlation between copy number of the TEs in the genome and the level of derepression could be observed. Taken together, the data presented is consistent with the notion that Blanks is required for efficient silencing of TEs *in vivo*.

The most straightforward explanation for the observed phenotype upon Blanks depletion could be an altered abundance of siRNAs targeting the TEs. In order to test this, the small RNA population was quantified after Blanks and Dcr-2 depletion using the corresponding SD cell lines. Small RNAs were isolated, sequenced and the 19-25nt long reads were mapped to the genome (Figure 7—8B). Upon Dcr-2 shutdown a clear decrease in transposon mapping reads was obvious, while – as expected – no change in miRNA mapping reads could be observed since they are produced in a Dcr-2 independent manner. The miRNA levels are also stable independently of the Blanks expression, whereas a marginal increase in TE mapping reads can be detected when comparing the siRNA levels in Blanks shutdown cells with cells after strong induction of the Blanks expression. This induced situation (200 µM copper) resembles a slight overexpression of the protein, approximately 2-fold. However, this minor effect cannot explain the detected elevated TE transcript levels completely. Additional processes must be involved.

Bona fide siRNAs are prominently 21nt long. The accuracy of the processing of siRNAs from longer dsRNA by Dcr-2 is guaranteed by Dcr-2 itself (Kandasamy and Fukunaga, 2016). However, the substrate specificity and processivity of the dicing process is modulated by its cofactors R2D2 and Loqs (Miyoshi et al., 2010a; Miyoshi et al., 2010b). Although no *in vivo* interaction of Dcr-2 with Blanks could be demonstrated, Blanks might have an indirect effect on Dcr-2 and the processing of the dsRNA into 21nt siRNAs. However, the length distribution of the small RNAs does not change with shutdown of the Blanks expression, see Figure 7—9.

Although the abundance of siRNAs mapping to TEs seems not to be strongly altered after Blanks depletion, the siRNAs can only function in silencing TEs there are properly loaded into Ago2, the effector protein that mediates the cleavage of the corresponding mRNA.

Small RNAs can be either sorted into Ago1 or Ago2 (Czech and Hannon, 2011; Czech et al., 2009). In *Drosophila*, Ago2 is the catalytically active protein that is loaded with siRNAs and cleaves endonucleolytically the target mRNA, which is then degraded(Okamura et al., 2004). Ago1 is loaded with miRNAs to fulfill its regulative function via translational repression and destabilization of the mRNA but not by direct cleavage of the target (Azlan et al., 2016; Meister, 2013). Small RNAs that are loaded into Ago2 are – in contrast to those that are sorted into Ago1 – methylated at their 2'OH of the ribose at their 5' end. This reaction is catalyzed by the enzyme Hen1 and makes the siRNAs resistant to oxidation with periodate (Ji and Chen, 2012). Ago1 loaded small RNAs do not have this modification and react with the reagent (Figure 7—10A and B). The ribose ring is opened, the base removed and no ligation of the linker can occur during deep sequencing library generation. In these beta-eliminated libraries, Ago1 loaded small RNAs and small RNAs that are not loaded in any Argonaute protein are thus depleted. Only Ago2 sorted siRNAs are resistant.

Figure 7—10C shows clearly that in the Blanks SD situation the Ago1 loaded miRNAs are decreased after beta-elimination as expected. Also U6 snRNAs degradation products that are neither loaded onto Ago1 nor onto Ago2 are depleted. However, TE-mapping siRNAs as well as Dcr-2 dependent hp-RNAs (CG4068, probe B) are still abundant in the beta-eliminated sample and the levels are unchanged compared to the untreated condition. This proves that upon Blanks depletion siRNAs are efficiently loaded onto Ago2, arguing that Blanks is not involved in the sorting or loading of the small RNAs in their effector proteins.

Altogether, Blanks is not mainly involved in the catalytic step of processing of dsRNA coming from endogenous TEs into siRNAs by Dcr-2 or in the loading of them onto Ago2. This is consistent with the observation that Dcr-2 and Blanks do not co-immunoprecipitate. Moreover, the effect of Blanks on the silencing of the reporter genes is much stronger than on natural, endogenous TEs, when measuring the transcript levels. This suggests that the main function of Blanks is beyond the canonical RNAi.

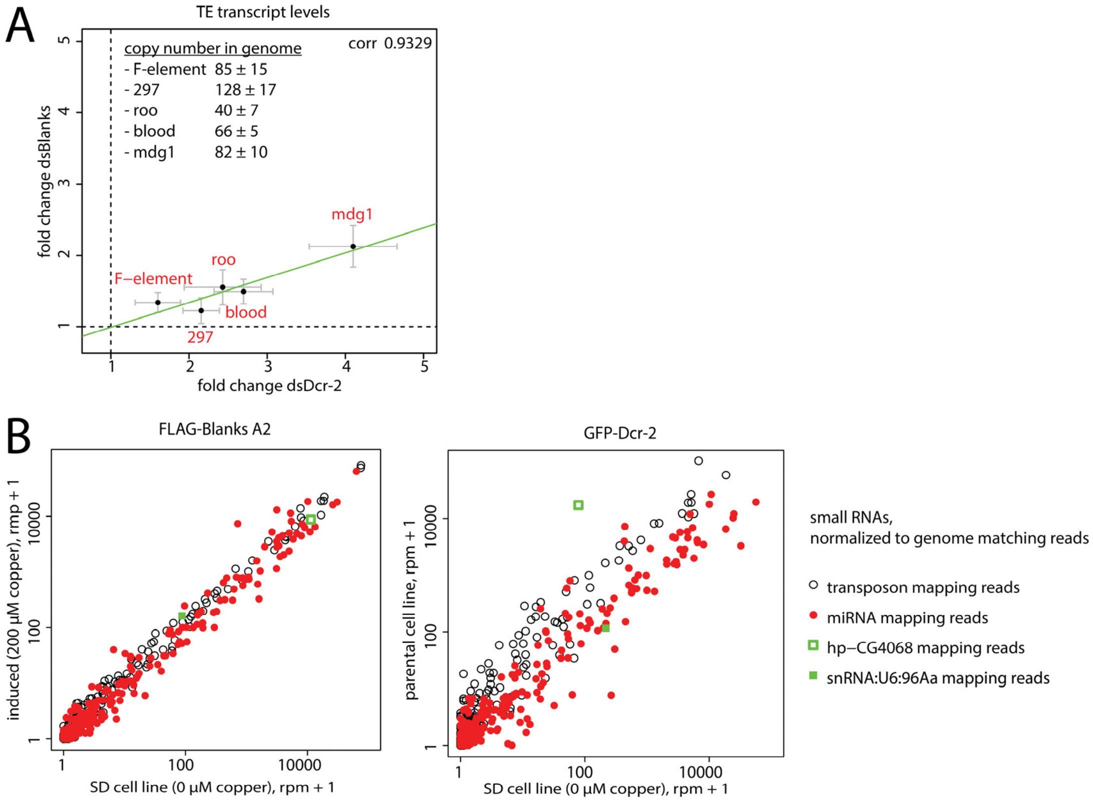

Figure 7—8: Blanks depletion results in a slight de-repression of endogenous TEs, while the siRNA levels are only minorly affected. (A)RT-qPCR of transcripts of selected endogenous TEs after Blanks or Dcr-2 knockdown. Fold change was calculated relative to *rp49* levels. Copy number of the corresponding TEs in the genome was determined by qPCR using gDNA extracted from S2 cells. n≥ 3 replicates, error bars represent standard error. (B) Quantification of miRNAs and TE-mapping siRNAs in Blanks and Dcr-2 SD cell lines in comparison to induced or parental cells. Reads were normalized to genome matching reads.

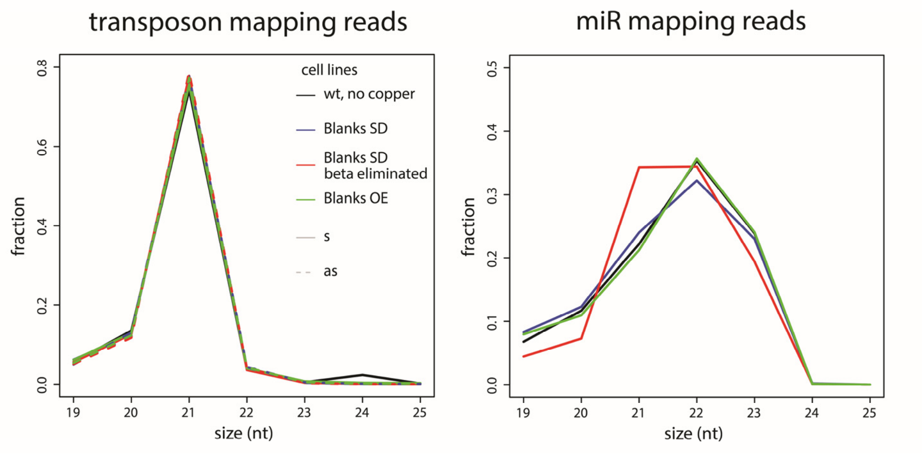

Figure 7—9: Length distribution of 19-25nt long small RNAs mapping to either TEs or miRNAs. The reads of Blanks-SD, Blanks-SD$^{\text{beta-eliminated}}$, induced Blanks-SD and wt cells were analyzed and mapped for sense (solid line) and antisense (dashed line) direction.

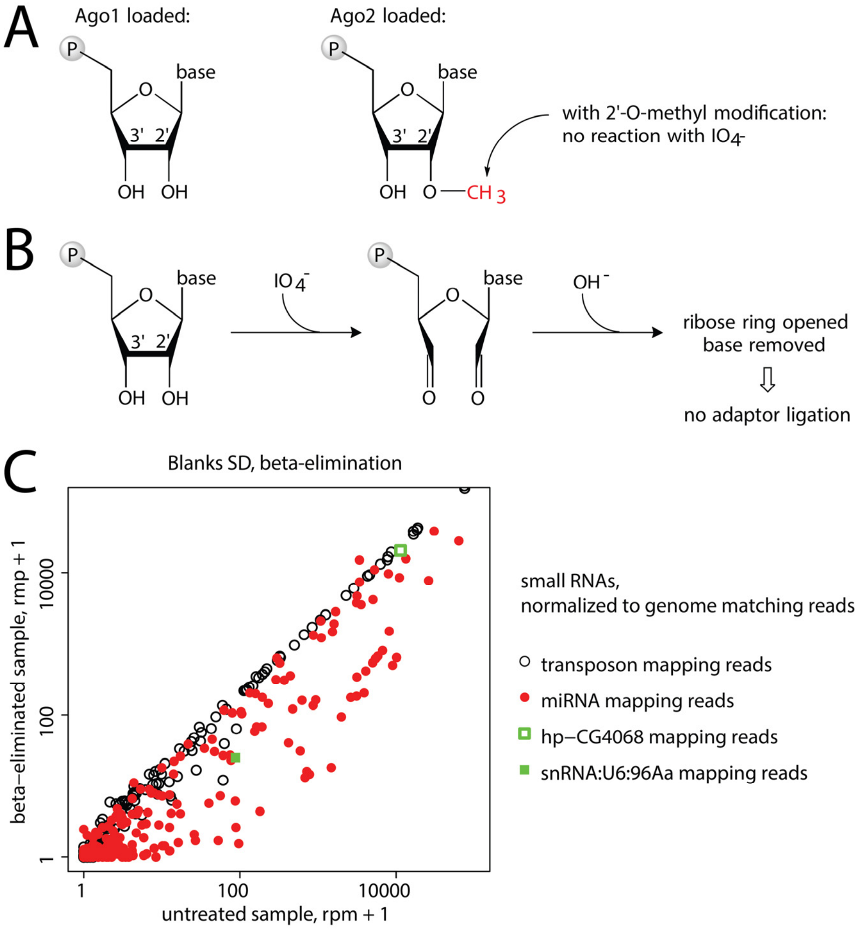

Figure 7—10: TE-mapping siRNAs are loaded Blanks-independent into Ago2. (A) Chemical nature of the 3'end of small RNAs that are either loaded into Ago1 or Ago2. Ago2-laoded siRNAs are methylated at the 2' OH of the ribose which makes them resistant to the oxidation with periodate. Un-methylated 2'OH of the ribose reacts with periodate what results in the opening of the ribose ring (B). A subsequent shift of the pH value mediates the removal of the base called beta-elimination. (C) Abundance of miRNAs and TE-mapping siRNAs after beta-elimination in comparison to the untreated sample in Blanks shutdown cells. While miRNAs which are loaded into Ago1 are depleted as expected, the amount of TE-mapping siRNAs remains unchanged.

7.2.4 Blanks is not involved in the translational repression of TEs

Usually siRNAs are loaded onto Ago2, the endonuclease which cleaves the target mRNA thereby leads to repression of the gene expression. As already introduced, miRNAs function differently. They are loaded onto Ago1 which together with GW proteins and factors mediates the repressive function on the mRNA by destabilizing the mRNA or interfere with translation initiation (Carthew et al., 2016).

Since many ribosomal proteins co-purified with Blanks in immunoprecipitation experiments (for details see Figure 7—16), we reasoned that Blanks might repress the translation of TE proteins – maybe in an Ago1 and GW protein comparable manner. Moreover, depletion of the ribosomal release factor Pelo results in derepression of mRNAs and proteins from selective TEs (Yang et al., 2015a). This suggests the possibility that so far unknown proteins are involved in TEs silencing beyond the classical RNAi pathways or the establishment of heterochromatin at TE loci.

In order to check if Blanks functions in such a novel fashion beyond the canonical RNAi pathway, I measured if the levels of TE proteins are elevated in a Blanks depleted background. To this end, I used the Blanks and Dcr-2 shutdown cell lines and performed shot-gun proteomics (Figure 7—11A). Induced and non-induced cells were harvested, lysed with stringent conditions (SDS, urea and DTT buffer, sonication) and mass spec samples were prepared following the FASP protocol (Wisniewski et al., 2009). Protein levels were quantified using the label-free-quantification approach by MaxQuant (Cox and Mann, 2008). In total, 3901 proteins were identified for all cell lines and conditions, for details see Table 3.

When comparing the protein levels of TEs in Dcr-2 shutdown cells with the induced situation, a clear derepression of the protein abundance can be detected (Figure 7—11B). However, this seems not to be the case for the Blanks shutdown. Upon Blanks depletion the abundance of TE proteins does not increase, suggesting that Blanks has no influence on the translational regulation of TEs. Moreover, fewer proteins (90 proteins) differ in their expression levels after manipulating the Blanks expression compared to the Dcr-2 setting, where approx. 186 proteins are upregulated upon Dcr-2 shutdown. Only two proteins are upregulated in both data sets. Among the upregulated proteins no specific GOterms were significantly enriched.

All in all, we found no evidence that Blanks mediates the repression of TE proteins.

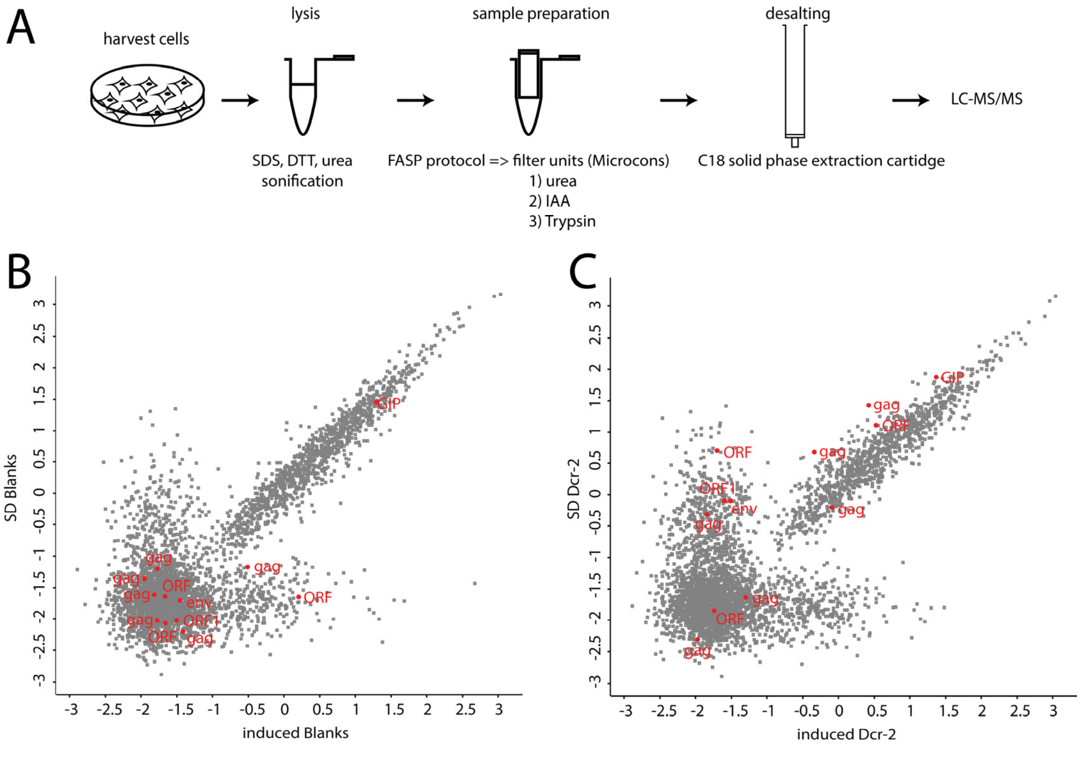

Figure 7—11: Protein abundance changes after depletion of Blanks and Dcr-2. (A) Overview of the shotgun proteomics approach in order to quantify the abundance of proteins following the depletion of Blanks and Dcr-2 expression. Sample preparation was performed following the FASP protocol (Wisniewski et al., 2009), subsequently, the peptides were desalted and analyzed by LC-MS/MS. Label-freely quantified and normalized protein levels are plotted in a scatterplot (log-scale) in order to compare the abundance for Blanks (B) and Dcr-2 (C) depletion. Induced expression of Blanks or Dcr-2 is plotted against shutdown of the protein. Proteins that derive from TEs are marked and annotated in red.

Table 3: Overview of proteins that are most abundant in S2 cells or upregulated after Dcr-2 or Blanks depletion.

50 most abundant proteins in 5-3 cells (wt situation), untreated	Top 50 proteins upregulated in Dcr-2 depleted background	Top 50 proteins upregulated in Blanks depleted background
Act42A	CG16817	Nlp
Ef1α48D	RpS15Aa	Tapdelta
Hsc70-4	Cyt-c1	Nat1
His4	dUTPase	CG3662
Pdi	CG9953	CG42748
EF2	RpL13	Mpcp
ERp60	CG1354	Sgt1
Hsp83	RhoL	Vap-33B
β-Tub56D	RpL35	CG9977
Crc	Snx1	ND-B15
14-3-3ε	smt3	eIF-2beta
RpL4	Caf1	CG5174
α-Tub84D	PCNA	CG12304
Eno	Sec61beta	Galphai
Act57B	nclb	Abp1
Zip	eff	dco
ATPsyn-β	Src42A	Rpd3
PyK	shot	anon-i1
Gale	bai	Syx1A
Gapdh2	CG6907	ND-ASHI
Inos	Chro	Stat92E
Gp93	CG11377	su(f)
Hsp60	RpL27A	yin
α-Spec	ric8a	CD98hc
Chc	hrg	Ero1L
Hsc70-3	Nlp	CG4365-RC
RpS7	twr	AnxB11
eIF-4a	CG7048	Trx-2
Ald-RH	CG7456	elm
awd	Alph	CG7324
RpS5a	PlexA	Pvf2
RpS3	B52	CG11148
Akap200	CG2051	mRpL24
capt	Stat92E	Atg16
Act5C	RfC3	CG5745
Vha68-2	nito	Ns3
Past1	AdenoK	Nnp-1
RpS12	Nmt	CG8771-RA
cher	CG9318	Psi
Uba1	mit(1)15	anon-37B-2
Clic	sds22	sip2
TER94	CG7791	Tom20
CG17259	CG2021	Vps28
Mtpα	Oscillin	pgant6
Gpo-1	ND-MLRQ	CG6523
bic	CG7519	qkr58E-1
CG31664	Rrp6	CG9248
RpLP0	CG7787	ZnT86D
	Cyp12a5	Su(dx)
	Usp5	beg

Upregulated proteins were upregulated for at least 2-fold after Blanks or Dcr-2 depletion and do not show any response to copper addition (tolerance: 0.8 – 1.2 fold change).

7.2.5 Some genomic loci generate *bona fide* siRNAs in a Blanks-dependent fashion

As already mentioned above, Blanks might be involved in the silencing of transcripts that derive from specific genomic loci as it is necessary for efficient repression of GFP or luciferase based reporters that are present in the genome in low and high copies (Zhou et al., 2008).

In order to test if there are genomic loci whose siRNA production specifically depends on Blanks, I reanalyzed the sRNA-sequencing data of the Blanks shutdown cell lines. Since *bona fide* siRNAs are predominantly 21nt in length and sense and antisense reads are comparably abundant, I filtered for 19-25nt long reads and mapped them to the *Drosophila* genome. After binning the reads into windows of 100nt's, I isolated those loci that have approximately comparable amounts of reads in sense and antisense direction. An excess of reads in one direction means that these small RNAs are rather degradation products of transcripts than siRNAs. After normalization of the reads to genome matching reads, I compared their amount between the Blanks induced cell line and the Blanks shutdown situation and was thereby able to identify twelve loci that produce notably more siRNAs upon Blanks induction (see Table 4).

Table 4: Strength of siRNA generation and genomic features of the Blanks-dependent siRNA loci. The genomic features were identified using the data provided by Flybase GBrowse.

Genes in neighborhood	Genomic coordinates (flybase release 5.23)	siRNA ratio Blanks^{ind} / Blanks^{SD}	siRNA ratio wt^{ind}/ wt	Convergent transcription *ov = overlapping (cis-NAT)* *ad = adjacent*	Annotated heterochromatin *g = siRNA locus* *n = in neighbor-hood*	TE insertions in neighborhood
ap	2R:1614667-1616606	2.21	1.22	-	g	multiple INE-1
Clc-b	2R:8462726-8463520	2.39	0.68	ov	-	-
Med15 cbt	2L:474996-478346	3.85	0.79	ad	-	-
Mitf Dyrk3	4:1225213-1231332	11.28	0.77	ad	g	multiple INE-1
ppk13	2L:21087843-21089267	2.62	1.04	-	-	-
Psa / cue	3L:1502715-1506033	7.15	1.10	ov	-	17.6
ref(2)P	2L:19543467-19545881	2.35	0.56	ad	-	-
RpS8	3R:25689274-25689685	2.71	1.71	ov	-	-
Sam-S	2L:108208-114500	3.53	0.89	-	n	-
Snoo	2L:7969040-7977349	1.27	0.93	-	-	297, pogo
unc-13	4:889576-904347	6.39	0.77	-	g	INE-1, Cr1a, Bari,
zfh-1	3R:26599518-26606594	3.15	0.98	ad	-	-

The Blanks shutdown situation at these loci mimics very well the depletion of Dcr-2, as depicted for two example loci in Figure 7—12A and B. Since the induction of Blanks or Dcr-2 in the shutdown cell lines is mediated via addition of copper to the medium, the observed phenomenon could be due to an indirect effect. Accumulation of heavy metals such as copper challenges the cellular integrity and may result in altered gene expression profiles (Yepiskoposyan et al., 2006). However, when comparing the siRNA levels at these loci with the parental cell line that was treated with copper with untreated cells, no difference in siRNA abundance could be detected (Figure 7—12C).

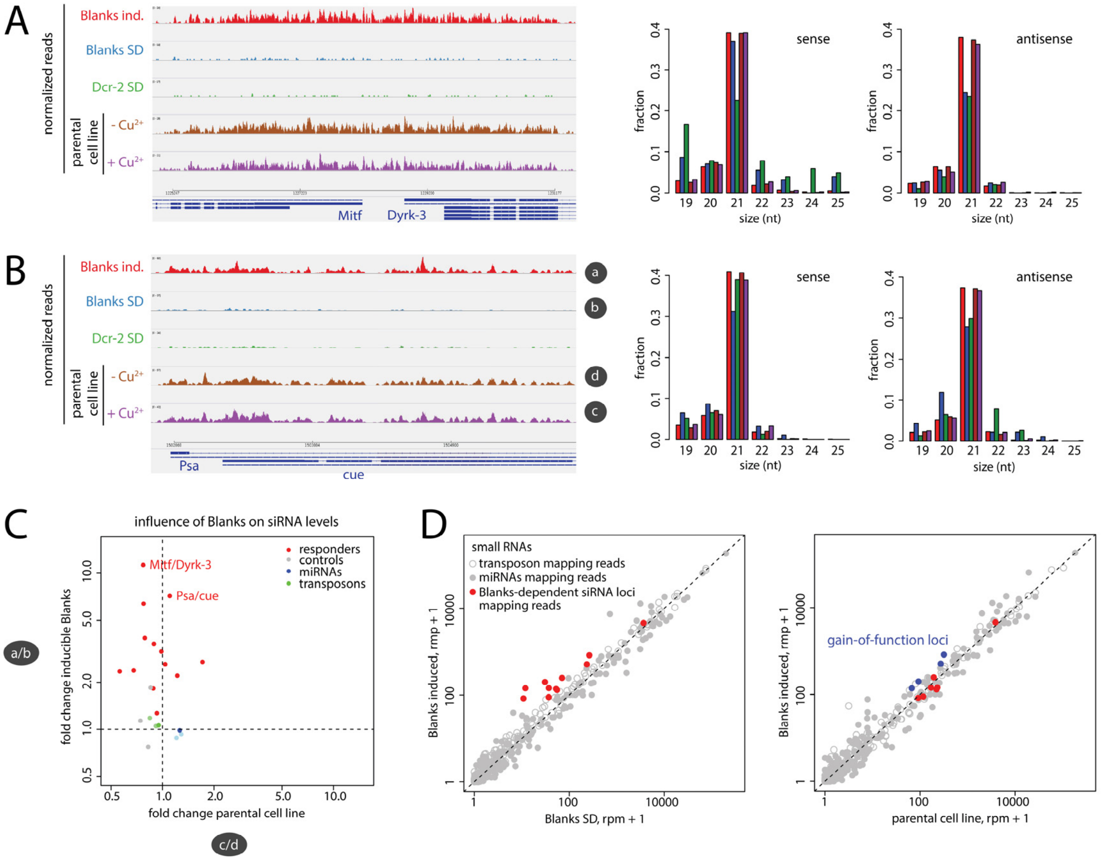

Figure 7—12: Blanks-dependent siRNA loci produce *bona fide* siRNAs. (A,B) Sequencing traces of 19-25nt long siRNAs mapping to Blanks dependent loci. Reads were normalized to genome matching reads. Length distributions of the reads per cell line are depicted in the right panel in sense and antisense orientation, color coding as in the sequencing traces. (C) The abundance of siRNAs mapping to identified Blanks-dependent siRNA loci is plotted as fold change that compares the amount of reads in the Blanks induced vs. the Blanks shutdown state (y-axis). The fold change of the parental cell line is depicted on the x-axis and shows the effect of the copper addition that is necessary for the induction of Blanks expression in the shutdown cell line on the amount of siRNAs. (D) Scatter plots of normalized TE-mapping siRNAs, miRNAs and Blanks-dependent siRNAs. Small RNAs derived from the Blanks-dependent siRNA loci are more abundant after Blanks induction. When comparing the amount of reads of induced Blanks cells to the parental cell line representing the wt situation, four gain-of-function loci could be identified, that give rise to more siRNAs when Blanks is slightly overexpressed.

The small RNAs which derive from these loci exhibit a clear peak at 21nt length and are comparably abundant in sense and antisense orientation (Figure 7—12A and B). Due to the fact that they are also Dcr-2 dependent, they are *bona fide* siRNAs. Eleven loci are more than 2-fold inducible upon Blanks induction (Figure 7—12C). Only one locus (*snoo*) shows a weaker response to the Blanks induction, but this can be due to the fact that the overall amount of siRNA is low so that the induction of siRNA generation is weaker. The identified loci are named Blanks-dependent siRNA loci. When comparing the change in siRNA abundance of these loci with the variation in TE-mapping siRNAs and miRNAs, it is very clear that Blanks very specifically affects the generation of siRNAs that derive from these genomic regions.

Based on the sequencing and annotation data that is provided by Flybase, many of the loci are between genes whose transcription converges, so that read-through transcription events can give rise to dsRNA. Three Blanks-dependent siRNA loci (*Psa, Clc-b* and *RpS8*) were previously known cis-NAT loci with annotated overlapping transcripts. Furthermore, the identified loci are often close to regions or overlapping with regions of annotated heterochromatin in S2 cells or adjacent to TE insertion sites. Predominantly, multiple insertions of the INE-1 element, a SINE-1-like non-LTR retrotransposon, could be detected. Thus, it seems that either convergent transcription or heterochromatic regions with TE insertions are characteristic features that license a locus as Blanks-dependent siRNA locus. Taken together, the loci can be grouped into four classes:

1.) Convergent transcription, no heterochromatic region:
Clc-b, Med15, Psa, ref(2)P, RpS8, zfh-1
2.) Heterochromatic region and/or TE insertions (predominantly non-LTR-retrotransposons):
app, unc-13, Snoo
3.) Convergent transcription, heterochromatic region and TE insertions:
Mitf / Dyrk3
4.) No obvious feature / data is missing:
ppk13, Sam-S

While the *Mitf/Dyrk3*locus that is characterized by both features (convergent transcription, heterochromatin and TE insertions) shows the strongest Blanks-dependence, no quantitative correlation between the Blanks-dependence and the different classes could be observed.

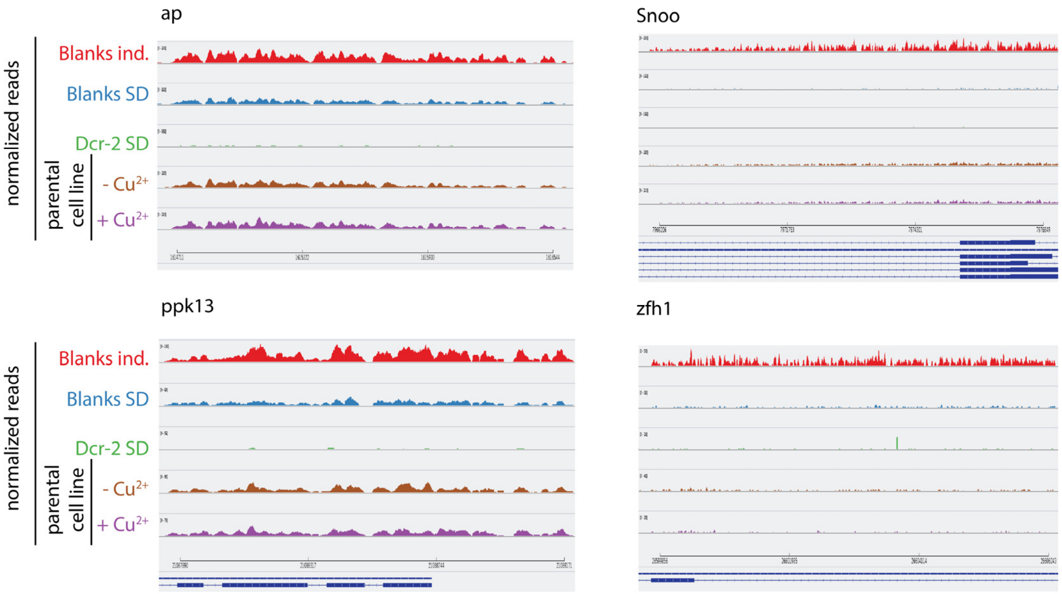

Figure 7—13: Sequencing traces of the four gain-of-function loci as described in Figure 7—12 (D).

Moreover, four identified Blanks-dependent RNA loci are also gain-of-function loci (*ap*, *snoo*, *ppk13*, *zfh1*), see Figure 7—12D and Figure 7—13. Comparing the siRNA abundance at the characterized loci in the induced, two-fold overexpressed situation with the parental cell line, which has wildtype expression levels, an increase of the small RNAs can be observed. In other words, an overexpression of Blanks seems even to facilitate the production of siRNAs from these loci. This argues that wildtype Blanks levels are limiting siRNA production from these regions and for an RNA chaperone effect of Blanks on its substrates.

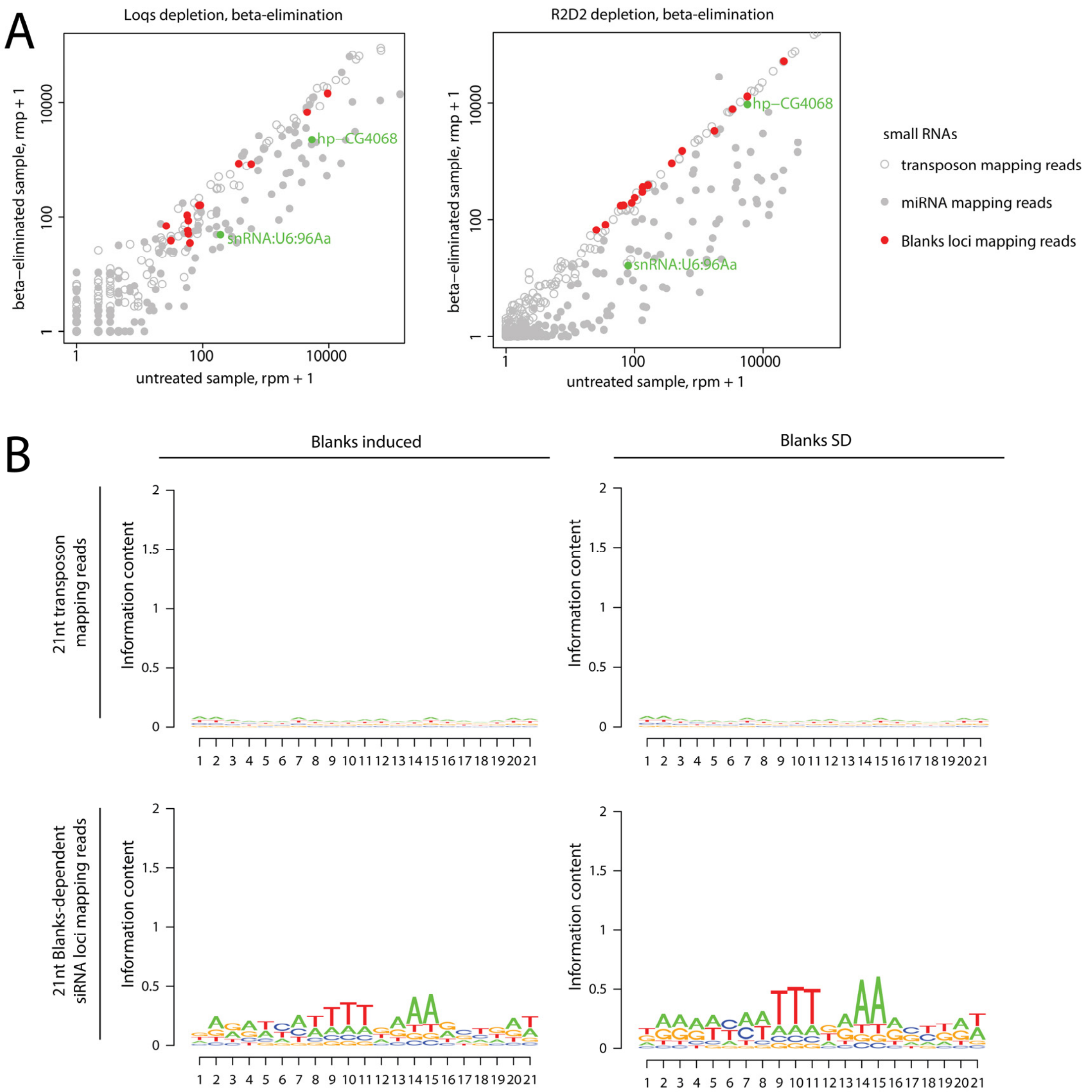

Figure 7—14: (A) Abundance of miRNAs, TE-mapping and Blanks-dependent siRNAs after beta-elimination in comparison to the untreated sample in Loqs and R2D2 depleted cells. Reads were normalized to genome matching reads and plotted on a logarithmic scale. Controls are annotated and colored in green. (B) Sequence logo of 21nt long reads mapping to either TEs or Blanks-dependent siRNA loci in Blanks shutdown cells with and without induction of Blanks expression. The information content correlates with the conservation of the specific nucleotide at each position.

Next, I wanted to check whether these Blanks-dependent siRNA loci give rise to small RNAs that are loaded onto Ago2. In *Drosophila*S2 cells, the loading of siRNA into Ago2 can be mediated either by Loqs or R2D2(Fesser, 2013). In contrast, in flies the RISC-loading complex consists of Dcr-2 and R2D2; Loqs cannot substitute R2D2 for this job (Liang et al., 2015; Mirkovic-Hosle and Forstemann, 2014).

In Loqs and R2D2 knockout cell lines (characterized in Tants et al., 2017, *manuscript in revision*), small RNA levels were quantified to check for proper loading of TE-derived siRNAs and small RNAs that were generated from Blanks-dependent siRNA loci. In addition, beta-elimination of the samples was performed to investigate the loading state. Reads mapping to these specific loci cluster with TE-derived siRNAs in both cell lines are still sufficiently loaded onto Ago2 after Loqs or R2D2 knockout (Figure 7—14A). Thus, Blanks-dependent siRNAs are comparable to canonical endo-siRNAs that target TEs with respect to their length distribution, their Dcr-2 dependency and their loading onto Ago2.

Furthermore, I analyzed the sequence of the small RNAs that derive from Blanks-dependent siRNA loci and compared the results with TE-mapping siRNAs. 21nt reads were filtered and the prevalence of specific nucleotides at each position was determined using sequence logos (Figure 7—14B). While the sequence of TE-mapping reads is highly diverse and shows no conservation of bases at specific positions, the Blanks-dependent siRNAs exhibit a higher overall A/T-content. Thymidine is more frequent than the other nucleotides at position 9, 10 and 11, and adenosine at position 14 and 15. However, the higher A/T-content can also be due to the fact that the Blanks-dependent siRNA loci are predominantly intergenic, within introns or 5'/3'-UTRs, which have *per se* a higher A/T-content than the CDS of protein coding genes.

Since Blanks seems to be specifically important for the generation of siRNAs from a small set of loci, the protein somehow has to recognize its target dsRNA. dsRNA-binding proteins have no sequence dependence since they recognize the shape of A-form dsRNA. Due to its less conserved sequence of the dsRBD2, Blanks may recognize specifically modified dsRNA. This dsRNA may have a slightly different structure that allows Blanks to distinguish between different substrates. A very frequently occurring modification in the nucleus is the deamination of adenosine to inosine by ADAR (Nishikura, 2010). Frequent targets of ADAR are within UTRs. Thus, potential regions of convergent transcription, which is a characteristic feature of Blanks-dependent loci, are hotspots of ADAR activity. There is also experimental evidence that dsRNA that is fed into the RNAi machinery is preferential substrate for ADAR activity (Hundley and Bass, 2010). Therefore, a reasonable hypothesis would be that the Blanks-dependent siRNA loci produce transcripts that are more often substrate for ADAR than RNAs from other genomic positions and thereby are specifically recognized and bound by Blanks.

When deep sequencing libraries of the modified RNAs are generated, the inosine templates the incorporation of a C, rather than a T during reverse transcription. Hence, A-to-G conversions are introduced, which can be detected when the reads are mapped back to the genome. If Blanks binds specifically to inosine containing RNAs, the amount of mapping reads should increase dramatically when allowing mismatches during the mapping step. Indeed, more reads mapped to the Blanks-dependent siRNA loci. However, the effect is not more pronounced than for TE-derived siRNAs or miRNAs (Figure 7—15A). Moreover, there is no difference if Blanks is depleted or slightly overexpressed when comparing the results to the parental cell line. The nature of the mismatches is highly diverse and A-to-G mismatches are not enriched which would be characteristic for the ADAR activity (Figure 7—15B).

In summary, there are specific genomic loci that produce *bona fide* siRNAs in a Blanks-dependent manner. These siRNAs seem to be generated and behave like endo-siRNAs which are produced in order to silence TEs.

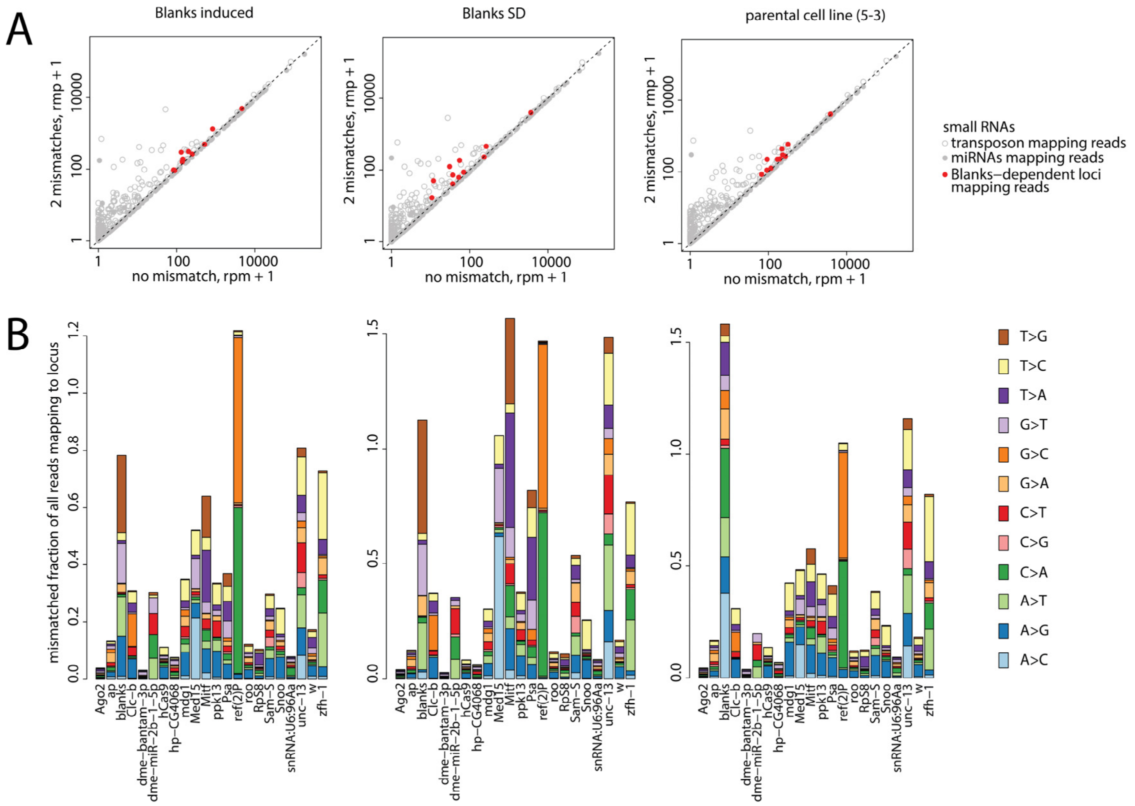

Figure 7—15: dsRNAs from Blanks-dependent loci seem not to be substrates for increased ADAR activity. (A) Amount of reads mapping to either TEs, miRNAs or Blanks-dependent siRNA loci when reads were mapped allowing no or two mismatches. Depicted are the scatter plots for Blanks shutdown cells and the parental cell line (5-3). Reads were normalized to genome matching reads. (B) Characterization of mismatches at different loci as fraction of all reads mapping to the locus.

7.2.6 Blanks interacts with proteins involved nuclear import and export and is a putative dsRNA-export factor

Since a functional role of Blanks in the generation of siRNAs was confirmed, we aimed at describing its mechanistic involvement in the RNAi process. I, therefore, performed immunoprecipitations of Blanks and identified interacting proteins by mass spectrometry.

Two cell lines with FLAG epitope tags introduced either at the C-terminus or at the N-terminus (shutdown cell line, mild overexpression) of the genomic locus were used to perform the pull downs in three biological replicates (Figure 7—16A). Prior to immunoprecipitation, weak and transient interactions were stabilized by mild *in vivo* cross-linking with 0.1 % formaldehyde. As controls and in order to identify unspecific binders, the immunoprecipitation was additionally performed either with a different antibody (isotype control) or with the parental cell line (5-3) that does not contain a FLAG-epitope tag.

Many proteins co-purified with Blanks, which can be already seen on the silver stain of the IP samples (Figure 7—16B and C). In order to identify the unknown interactors, the washed beads with the bound proteins were digested with trypsin into peptides, which were analyzed by LC-MS/MS (Thermo Scientific Orbitrap XL). For all conditions, about 500 proteins were identified. In order to check for potential interactors, the abundance of the identified proteins was compared in the IP samples with the isotype control. 125 proteins were significantly enriched in the FLAG-Blanks pulldown, 14 in the Blanks-FLAG pulldown and 6 in both (see Table 5).

Proteins that are associated with replication are enriched amongst the identified interactors, such as the ssDNA-binding proteins (RpA-70 and RPA2) or components of the MCM complex (Figure 7—16D and E). As an example, the interaction between Blanks and RpA-70 could be confirmed on a western blot by using a cell line that expressed FLAG-tagged Blanks and V5-tagged RpA-70 (Figure 7—16F).

Moreover, the *Drosophila* homologs of HP1a, Su(var)-205, and HP1b could be identified as Blanks interactors. HP1 binds to methylated histones and is crucial for the establishment and maintenance of heterochromatin. The physical interaction between Blanks and HP1 was already published by(Swenson et al., 2016), where they looked for binding partners of HP1 and described Blanks as a protein that is involved in heterochromatin function. Additionally, factors of the nuclear import and export were identified as Blanks binders (Figure 7—16G). Ran, Bj1 (the *Drosophila* RanGEF) and members of the importin family Kap-α3 and Karyβ3 co-purified with Blanks, as well as Mtor, a subunit of the nuclear basket of the nuclear pore complex.

The identified interaction between components of the nuclear export and import machinery and Blanks hints to a potential involvement of Blanks in the export of dsRNA from the nucleus to the cytoplasm. dsRNA is – apart from viral infections – exclusively generated in the nucleus where transcription takes place. However, the small RNAs are generated in the cytoplasm by Dcr-2, which is predominantly cytoplasmic. In the case of miRNAs, exportin-5 exports the miRNA precursor after Drosha-processing from the nucleus so that Dcr-1 can generate mature miRNAs. For siRNA precursors, long dsRNAs, the export factor is not yet known. For some specific substrates (e.g. dsRNA derived from Blanks-dependent siRNA loci) Blanks could be the responsible export factor of dsRNA.

To confirm this hypothesis, I first tested whether Blanks indeed shuttles between the nucleus and the cytoplasm. As illustrated in Figure 7—17A, the small molecule importazol blocks the import of factors into the nucleus (Bird et al., 2013; Song et al., 2014). If Blanks shuttles between cytoplasm and nucleus, Blanks protein should become detectable in the cytoplasm after application of importazol to the cells due to the blocked re-import. 16 hours after addition of the drug to the medium Blanks-GFP signal can be detected in the cytoplasm (statistically significant, see Figure 7—17C), while the GFP signal is still exclusively nuclear in control cells treated with the solvent DMSO. For the constitutively nuclear protein H2Av, no increase in cytoplasmic localization could be detected. Moreover, since the cells were split prior to the addition of the drug, partially synchronizing their cell cycle, and the readout was performed after 16h, the Blanks-GFP signal in the cytoplasm is unlikely due to the nuclear envelope breakdown during mitosis.

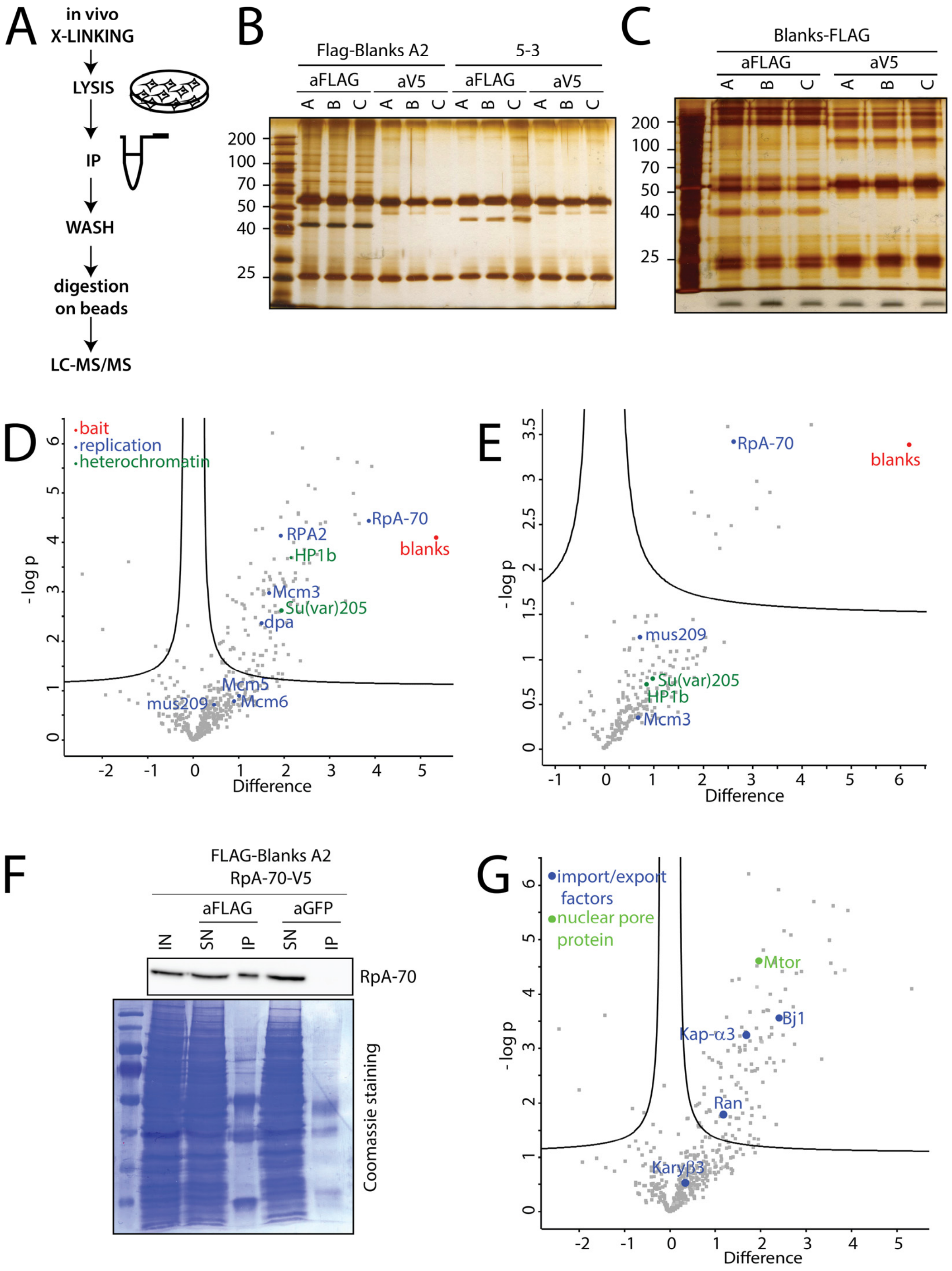

Figure 7—16: Analysis of interaction partners of Blanks using immunoprecipitation and mass spectrometry. (A) Cells were cross-linked *in vivo* with 0.1 % formaldehyde, lysed and immunoprecipitation was performed. The results of the co-IP were analyzed by mass spectrometry. IP samples were run on a 10 % SDS-PAGE and stained to visualize binding partners of FLAG-Blanks (B) and Blanks-FLAG (C). Pulldowns were either conducted with anti-V5 antibody or parental cell line extracts (5-3) as negative controls. Vulcano plots of FLAG-Blanks (D, G) and Blanks-FLAG (E) interactors. T-test was conducted to check for statistical significance, FDR = 0.05, s0 = 0.1. IP with anti-V5 was used as negative control. (F) Validation of the interaction between Blanks and RpA-70 by IP and Western Blotting. V5-tagged RpA70 was probed with anti-V5 antibody. The membrane was stained with Commassie blue as a loading control.

Table 5: Summary of significantly enriched proteins in Blanks-FLAG and FLAG-Blanks IPs compared to control condition (isotype control). Common factors are highlighted in bold.

Significantly enriched proteins in Blanks-FLAG IP	Significantly enriched proteins in FLAG-Blanks IP		
Hsc70-4	RPA2	smid	Uba2
14-3-3zeta	nop5	mbf1	bic
RpA-70	Mtor	RpL18	TER94
Blanks	Dek	La	mor
Chd64	CG30122-RB	hyd	CstF-64
Bacc	CkIIα	Trip1	lost
Hsp83	Cctγ	Kap-α3	dre4
βTub56D	T-cp1	Ef1β	RpS8
Droj2	Hrb98DE	Hsp68	blanks
Rm62	ran	His1	l(2)09851
αTub84D	CkIIβ	Hsp27	Uba1
Ef1α48D	CG4038	RpL32	lark
RpS3A	kay	Su(var)205	HP1b
Ef2b	RpL30	RpL4	eIF-5A
	Atpα	mod	Ref1
	RpS18	RpS17	Cbp20
	Dsp1	Ote	Jafrac1
	Cdc37	ncd	EndoGI
	RpS9	Pros28.1	Spt5
	Hsc70-4	Map205	Hcf
	eIF-4a	Bj1	RpL21
	hoip	B52	RpL27
	Eb1	**14-3-3zeta**	CG3353
	Hrb27C	RpL19	**Droj2**
	Mi-2	tsr	Art1
	RpS15Aa	RpL7A	rump
	CG10417	RpL23	RpL24
	RpS21	RpL9	FKBP59
	Fib	RpL22	Ssb-c31a
	RpS3	dod	cl
	FK506-bp2	**RpS3A**	pzg
	U2af50	RpS13	CG7564
	Lam	nonA	CG1316
	chic	Nap1	CG17737
	Mlc-c	RpS5a	CG12082
	Pep	Cp190	CG1240
	RpS2	**RpA-70**	RpS24
	Hsp83	dpa	CG15784
	baf	Hel25E	Mcm3
	Top2	Nlp	pic
	nocte	Geminin	BcDNA.LD23876
	CG10103	CG4747	

t-test, FDR = 0.05, s0 = 0.1

The results of this experiment allow us to conclude that Blanks shuttles between both compartments and provides thereby the possibility to export dsRNA from the nucleus to the cytoplasm. Taken together, these findings suggest multiple roles of Blanks in cellular functions and regulation mechanisms. While the link to heterochromatin was already known, the interaction with replication proteins and nuclear shuttling factors is an exciting new discovery. It suggests mechanistic role of Blanks as potential dsRNA export factor. This model is consistent with the identification of siRNA loci whose function is fully Blanks-dependent.

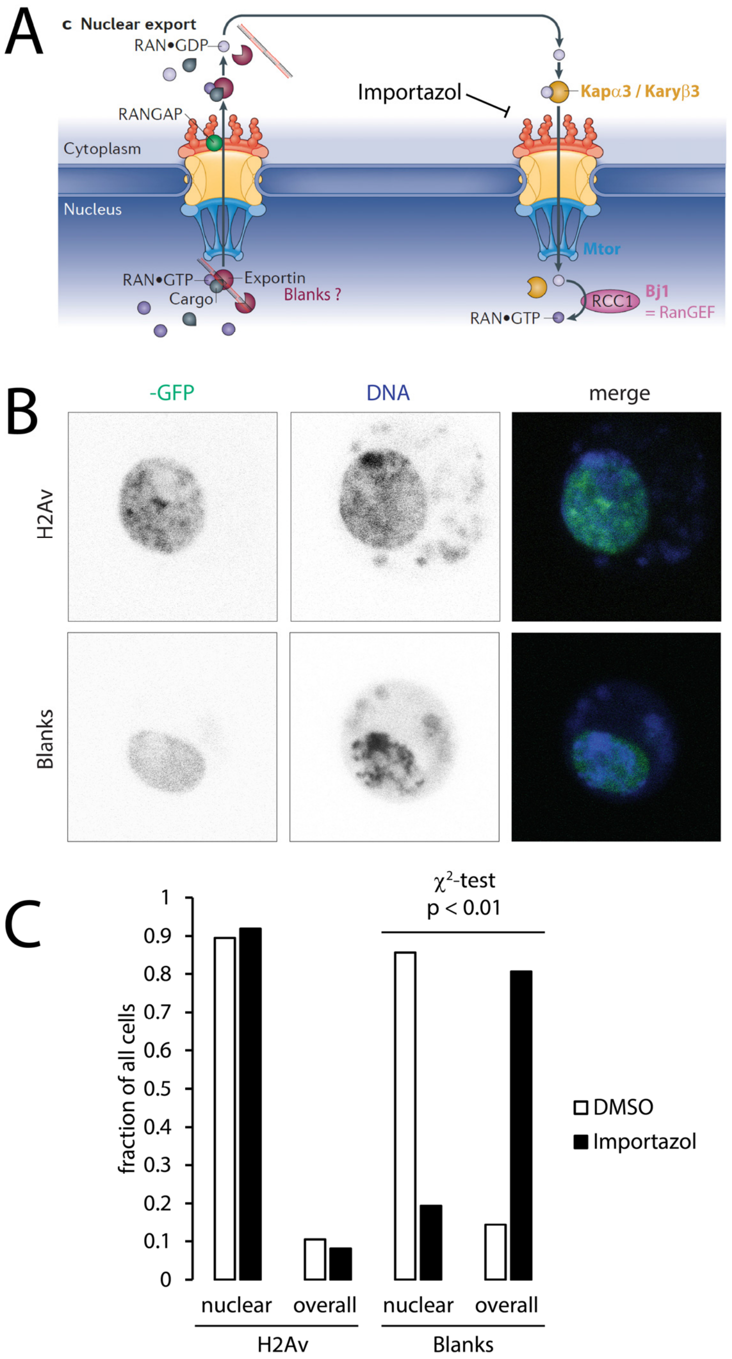

Figure 7—17: Blanks shuttles between the nucleus and the cytoplasm. (A) Schematic of the nuclear export and import of proteins in Drosophila and the potential involvement of Blanks in the process. Blanks could bind to dsRNA and export the molecule Ran-dependent to the cytoplasm. The importins Kapα-3 or Karyβ3 associate with Blanks and import the protein back to the nucleus. This import process is inhibited by the addition of importazol. (Picture adapted from (Raices and D'Angelo, 2012)) (B) H2Av-GFP and Blanks-GFP cell lines were used to monitor the shuttling of the Blanks protein. (C) Importazol treatment results in increase of Blanks-GFP in the cytoplasm while H2Av-GFP remains nucleoplasmic. Approx. 100 cells in total were analyzed in three different experiments, χ^2-test was calculated to validate the statistically significant change in the Blanks-GFP localization after addition of the drug.

7.3 Conclusions

Summing up the data that comes from publications and from my own experiments, one can conclude that Blanks is involved in the efficient RNAi when the dsRNA derives from nuclear sources (Figure 7—18). Blanks is not involved in the biogenesis of siRNAs by interacting with Dcr-2 nor in the loading of siRNAs onto Ago2. Nevertheless, a small effect on the abundance of the siRNA targeting TEs could be observed, although this seems not to be the main function of Blanks.

Blanks is rather important for the generation of siRNAs derived from specific loci which I was able to identify. The 12 loci are either characterized by surrounding heterochromatic regions and TE insertions nearby or convergent transcription. In eukaryotic genomes, overlapping transcription units are rather frequent events. In *Drosophila*, approximately 900 loci with annotated overlaps were identified. Around 800 loci have overlaps at their 3′ ends (convergent transcription) and 100 loci overlap at their 5′ ends (divergent transcription) (Zhang et al., 2006). Most of these loci give rise to *bona fide* siRNAs, however, the amount of generated small RNAs in *Drosophila* is less than in mammalian cells. In addition to these rather short siRNA loci, two genes with extensive transcription overlap are the source of endogenous siRNAs: *klarsicht* and *thickveins*(Okamura et al., 2008a). Although some of the Blanks-dependent siRNA loci are characterized by convergent transcription (annotated overlap or read through events), only three so far published cis-NAT loci show the Blanks-dependency. The other cis-NAT regions do not or only marginally depend on Blanks. This suggests that Blanks is not *per se* necessary for the generation of siRNA from cis-NAT loci but rather needed for the generation of siRNAs from novel, so far uncharacterized siRNA generating regions. Overexpression of Blanks can even boost the generation of siRNAs at some of these loci.

Proteins of the nuclear import and export such as Ran or proteins of the importin-β family were identified as interactors of Blanks. This supports the model that Blanks might shuttle between the nucleus and the cytoplasm in order to export dsRNA that comes from specific loci to the cytoplasm where it can be further processed by the RNAi machinery (Figure 7—18). By inhibiting the importin-dependent re-import of nuclear proteins, I was able to show that Blanks accumulates indeed in the cytoplasm, thus confirming the model that Blanks is a potential dsRNA export factor.

Next, it would be interesting to learn more about the substrate specificity of Blanks and if Blanks can discriminate between dsRNAs from different loci. An initial insight would be gained by performing *in vitro* assays with recombinant Blanks and distinctly modified RNAs to measure differences in the K_D value. The results could provide hints towards the physiological function. Alternatively, RNA immunoprecipitation with Blanks followed by deep sequencing could be performed in order to learn more about the RNA substrate that is bound by Blanks *in vivo*.

If Blanks works as a dsRNA export factor, its overexpression in tissues that normally have no Blanks expression may enhance RNAi phenomena. A handy system would be to study transgenic flies whose *white* gene is silenced in the eye by an inverted repeat which is constitutively expressed (*w*IR). These flies have orange eye color. Ectopic expression of Blanks should then even enhance the silencing effect, so that even less *white* is expressed. This might result in brighter eyes. If so, different mutants of Blanks proteins can be used to study the relevance of the distinct domains and the NLS signal. As a starting point, the introduction of point mutations in dsRBD2 can disrupt dsRNA binding and should therefore work as a negative control. Similar, the introduction of a NES should impair the enhancing of the RNAi phenomena. However, this has to be tested, if the *w*IR system is the right approach to address this question. Alternatively, the experiments could be also conducted as rescue experiments in Blanks shutdown cell lines.

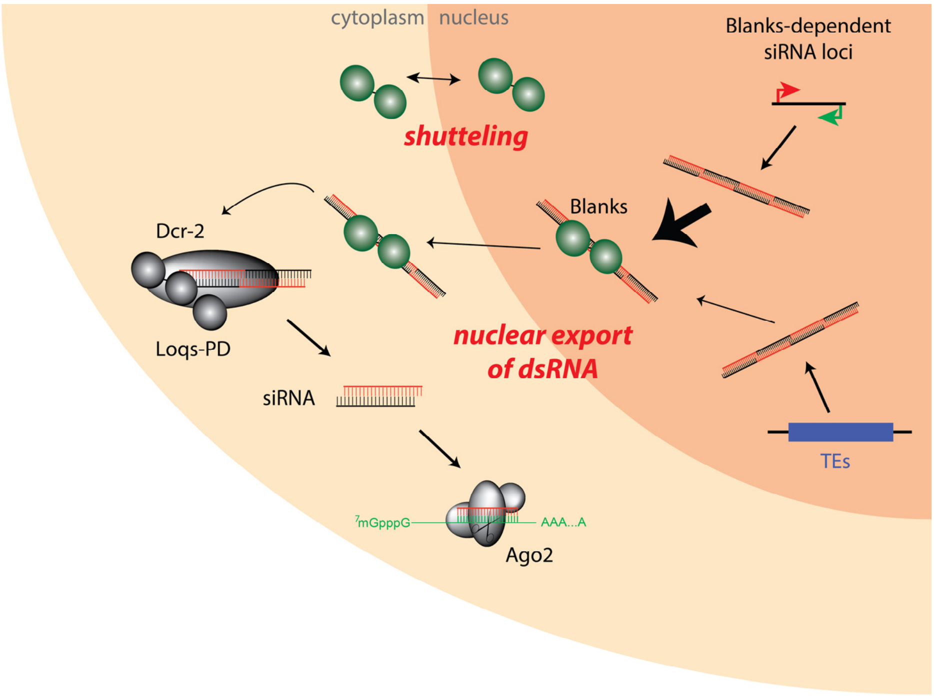

Figure 7—18: Model describing the function of Blanks. Blanks shuttles between the cytoplasm and the nucleus and exports dsRNA which mainly derives from specific loci but also to some smaller extent from TEs. In the cytoplasm the dsRNA is handed over to Dcr-2 which generates *bona fide* 21 nt long siRNAs that are loaded in Ago2.

Nevertheless, it remains questionable if S2 cells are the right model system to study the physiological role of Blanks. Beside S2 cells, the protein is solely expressed in testes. Genome integrity of germ line cells is ensured by the piRNA pathway. The piRNA cluster contains the information of all TEs that have to be silenced like a database. But do testes have a higher TE load, especially of recently integrated TEs or TEs that are not yet silenced efficiently by the piRNAs? Maybe, Blanks evolved as part of a new line of defense against these TEs that pose a severe problem for genome integrity of the germ line. This hypothesis is supported by the fact, that several annotated TE integrations are in close proximity to some Blanks-dependent siRNA loci. Moreover, the repressive effect of Blanks is much stronger on reporter genes which may resemble a "novel" TE insertion than on endogenous TEs. Additionally, the piRNA pathway is less active in the male germline than in the female resulting in the need of additional silencing mechanisms (Quenerch'du et al., 2016).It would be interesting to study the siRNA abundance and the characteristic of the Blanks-dependent siRNA loci in testes as well in order to gain more insight into the physiological role of Blanks.

In addition, it remains unclear if the interaction between Blanks and HP1 has a functional and physiological impact. Does Blanks represent the adaptor that connects RNAi with heterochromatin in flies? In fission yeast, a transcriptional silencing mechanism exists that links siRNAs with heterochromatic proteins. Moreover, the interaction of Blanks with ssDNA binding proteins and factors involved in replication remains elusive. Maybe Blanks bridges DNA – RNA interactions in order to epigenetically regulate the expression of distinct genomic regions. ChIP-seq could be a conclusive method to investigate these hypotheses.

The deregulation of the expression of several proteins upon Dcr-2 and Blanks depletion provides evidence that their shutdown has either indirect effects on the gene regulation or that both proteins are involved in additional processes that have to be characterized further. Since none of the proteins that are encoded by the genes adjacent to the Blanks-dependent siRNAs are upregulated upon Blanks depletion, this may argue for an siRNA-independent function of Blanks. The interaction of Blanks with proteins involved in replication and chromatin condensation would support this hypothesis. Very recently it was discovered that RNA-binding proteins can bind to the 3'UTR of mRNAs and function as scaffolds to mediate protein-protein interactions. These recruited proteins can interact with the nascent polypeptide chain and have an influence on stability, localization or translation of the protein (Mayr, 2016). Maybe Blanks binds normally to structured regions of the 3'UTR of the upregulated proteins and mediates thereby the translational repression of them.

Although no effect of Blanks on the exo-siRNA pathway was detectable, it would be worth to infect cells with viruses to study the effect of Blanks on viral defense. It is known that RNA viruses that are able to infect *Drosophila* can give rise to cDNA fragments which are generated from the viral RNA genome (Goic et al., 2013). The reverse transcription is probably mediated by reverse transcriptases of endogenous TEs. The viral cDNA might be the source of dsRNA generation and can serve as a "long-term" immunity of the flies against a re-infection by the viruses. This phenomenon might be Blanks dependent since the source of the dsRNA was nuclear.

Finally, a region within Blanks has homology to the mammalian NF90 protein, an important factor involved in virus defense and expression regulation. This might point towards an (at least) functional conservation of Blanks in higher eukaryotes and mammals. One first interesting experiment would be to ectopically express *Drosophila* Blanks in human cells and check if RNAi triggered by the expression of hairpin RNAs is more efficient than in the wt situation.

All in all, the described findings emphasize that the RNAi pathway and the regulation of gene expression is a highly complex system. It becomes successively clear that many so far unknown factors are involved in the mediation and fine-tuning of these processes. The results of this study are likely only a small piece within the whole picture of TE and virus defense.

8 Concluding remarks and relevance

In this study, I was able to contribute on the one hand to the development of state-of-the-art methods in molecular biology and on the other hand to the characterization of the RNAi pathway.

CRISPR/*cas*-mediated genome engineering and mass spectrometry-based analysis of interactomes are very powerful and versatile methods that have transformed biology from the focused investigation of a single process to the description of networks and global mechanisms. Although both techniques are very well suited for the comprehensive analysis of a single protein or pathway, their real power is within the "-omics" field. Proteins that are epitope tagged at their chromosomal locus can be used amongst others for chromatin-immunoprecipitation, CLIP-experiments and interactomics. Most importantly, there is no risk of artifacts due to overexpression and large cell-to-cell variation. With the transient expression of the bait proteins, this was a considerable problem. All in all, this may offer the possibility that many experiments which so far were performed *in vitro* can much easier be addressed in more physiological and systemic contexts.

The question how *Drosophila* cells deal with foreign genetic elements was the central theme of my thesis. By studying either the features on the level of gene structure (*cis*) or the involvement of Blanks in the RNAi pathway (*trans*), I observed that the genomic defense mechanism against invading sequences is highly complex, adaptive and efficient. Before cells trigger an siRNA response, mechanisms seem to evaluate the threat. For elements that are integrated into the genome, the copy number appears to be assessed. Furthermore, episomal DNA has a high priority and thus a strong RNAi response is triggered. Similarly, genome editing by CRISPR/Cas9 and subsequent integration of an HR donor results in a strong RNAi response. Although this situation resembles a low copy phenomenon, the RNAi response is rather strong in the beginning. The difference in the mechanisms how the genetic information is integrated into the genome may provide an explanation for the varying response strategies: While the reporter cell lines for the investigation of the *cis*-elements were generated using an integrase, the introduced epitope tag at the chromosomal locus occurs after the introduction of a DSB by the Cas9 nuclease. The integrase-mediated integration process of the plasmid, however, has no DSB intermediate. The fact that upon a DSB a strong siRNA response is triggered in *Drosophila* (Michalik et al., 2012; Schmidts et al., 2016) and also in other organisms supports the hypothesis that the cell perceives these as distinct situations. The siRNAs possibly mark the integrated sequence as a genomic scar.

At first sight, an interesting contradiction becomes apparent when one compares the following findings: As described in section 4.2.4, the amount of siRNAs depends on the transcription rate of the FLAG-Blanks locus under the control of the inducible *mtnDE* promoter. However, for the transient transfection of the *cis*-element reporter plasmids in section 5.2.2, no correlation between promoter strength of the reporter gene and the amount of siRNAs could be detected; rather, the whole expression cassette is covered with siRNAs. While genomic siRNA loci show a copy number threshold and transcription of the locus is required, episomal DNA triggers an RNAi response in an apparently less transcription-dependent fashion.

The finding that genome engineering can cause genome defense mechanisms is on the one hand an alarming discovery because many researchers may not consider this effect upon interpretation of their data; on the other hand, the process can be used to study the underlying mechanisms further. The requirements to trigger the siRNA generation *in cis* can be easily studied using this system. This proves the high flexibility of the genome defense strategies that have evolved.

All in all, the investigation of genome defense mechanisms and TE silencing in particular is highly important also for medical reasons. TE activity and new insertions are a severe risk for genome integrity and can result in altered gene expression, chromosomal rearrangements or impaired splicing, which all are hallmarks of human cancer (Chenais, 2015). A recent study characterized the tumor-specific TE insertions in humans and provided evidence for their causal role in tumorigenesis. For example, the authors found an *Alu* insertion in proximity of the enhancer of the *CBL* gene, which functions as a tumor suppressor, in the genome of a breast-cancer patient (Clayton et al., 2016). Consequently, the understanding of TE recognition is essential to understand why transposition activity increases and confers a higher oncogenic risk. Although the presented work was performed in cultured *Drosophila* cells, the underlying principles may well be conserved in humans despite considerable differences in genome defense strategies between the two organisms. Therefore, the results can serve as a starting point to develop hypotheses for further research.

9 Experimental Procedures

9.1 Molecular biological methods

9.1.1 Used plasmids in this study

	Insert / features	resistance	source
pRB10	*ubi*-GFP-polyA, Blasti-R, *attB*	Amp	AG Förstemann
pSK5	*sucb*-GFP-polyA, Blasti-R, *attB*	Amp	AG Förstemann
pSK7	*H2A*-GFP-polyA, Blasti-R, *attB*	Amp	AG Förstemann
pSK9	*sucb*-GFP-HSL, Blasti-R, *attB*	Amp	AG Förstemann
pSK11	*H2A*-GFP-HSL, Blasti-R, *attB*	Amp	AG Förstemann
pSK13	*ubi*-GFP-HSL, Blasti-R, *attB*	Amp	AG Förstemann
pSK20	*tctp*-mini-white-intron-GFP, -AGgt…agGT-	Amp	this study
pSK21	*tctp*-mini-white-intron-GFP, -AGgt…ccGT-	Amp	this study
pSL2	*tctp*-GFP	Amp	AG Förstemann
pSL3	*tctp*-mini-white-intron-GFP, -gt…ag-	Amp	AG Förstemann
pSL4	*tctp*-mini-white-intron-GFP, -gt…cc-	Amp	AG Förstemann
pSK28	as pSL2, Blasti-R, *attB*	Amp	this study
pSK34	as pSK20, Blasti-R, *attB*	Amp	this study
pSK35	as pSK21, Blasti-R, *attB*	Amp	this study
pSK33	as pSL3, Blasti-R, *attB*	Amp	this study
pSK36	as pSL4, Blasti-R, *attB*	Amp	this study

Plasmids for genomic tagging are described in section 9.3.2

9.1.2 Molecular cloning of the reporter plasmids

The cloning of the plasmids pSK5, pSK7, pSK9, pSK11 and pSK13 and the features of the other plasmids that were used in this study were previously described(Kunzelmann, 2013).

The used splice reporter plasmids were cloned according to established laboratory practice (Sambrook and Russel, 2000). They are listed in section 9.1.1.To this end, the mini-white intron was amplified by PCR using primers that contain restriction enzyme recognition sites at their ends and the corresponding splice signals / mutations. Prior to ligation, the PCR products and the Plasmid pSL2 were digested with the *Sph*I-HF (NEB) and *Nhe*I-HF (NEB). The digested plasmid backbone was treated with FastAP (Thermo) to remove the 5′ phosphate at the ends. Linearized and dephosphorylated plasmid backbones and PCR products were purified by using the Wizard SV Gel and PCR Clean-Up System (Promega) according to the manufacturer's instructions.

Insert and backbone DNA was ligated using T4 DNA ligase (NEB) according to the manufacturer's instructions. *E. coli* cells (XL2) were transformed with the ligation product. After selection on LB-Amp-plates, positive clones were validated by colony PCR, subsequent restriction digest and sequencing (Eurofins MWG Operon, Anzinger Str. 7a, 85560 Ebersberg). Sequencing results were analyzed using the ApE plasmid editor (v2.0.45).

In order to generate plasmids that can be integrated into the genome of cells via the ΦC31 integrase system, the *attB* site and the blasticidin resistance gene were amplified by PCR and blunt end cloned into the pJET vector (Thermo). The cassette was excised from the resulting plasmid with *Eco*RI (NEB) and ligated into the equally linearized plasmids pSK20, pSK21, pSL2, pSL3 and pSL4. Cloning success was confirmed by restriction digest and sequencing.

9.1.3 gDNA isolation, RNA isolation and reverse transcription, qPCR

Genomic DNA was isolated from *Drosophila* S2 cells either using the Wizard Genomic DNA Purification Kit (Promega) as described in the manufacturer's protocol or by using the Wizard SV Gel and PCR Clean-Up System (Promega). To this extent, 100 µl resuspended cells were incubated with 200 µl membrane binding buffer for 5 min. Subsequently, the DNA was purified following the manufacturer's instructions for the purification of PCR products. The 1 µL of concentrated gDNA can be used for PCR reactions.

RNA was isolated according to manufacturer's protocol using Trizol Reagent (Thermo), reverse transcription was performed using Superscript III (Invitrogen) as described by the company's protocol. The reaction was primed with random hexamers (Eurofins-MWG).

Quantitative PCR was carried out with SYBR green mixes (DyNAmo-Flash, Thermo) using approximately 10 ng of genomic DNA or cDNA as template. The qPCR was performed in a Biometra TOptical Thermocycler (Analytik Jena, Germany). Primers for GFP and rp49 were used as previously described (Hartig et al., 2009). Obtained data was analyzed using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001).

qPCR primer Blanks:

blanks_102+	tgctgtaattccgctcgcaga
blanks_297-	acggccattggttgcgtcat

9.1.4 In vitro transcription in order to generate dsRNA

In vitro transcription of DNA templates was performed as described in(Elmer, 2013). DNA templates for IVT were generated by PCR with the listed primers on cDNA of S2 cells. dsRNA was generated using home-made T7 RNA polymerase.

The following primers were used in the study:

dsRluc	taatacgactcactatagggatggcttccaaggtgtacgacc
	taatacgactcactatagggcattttctcgccctcttcgctc
dsDcr-2	taatacgactcactatagggATTGTTGACCAAAGCGGAAC
	taatacgactcactatagggATTCCCAAAACGCTCAACAC

dsAgo2	taatacgactcactatagggGCTGCAATACTTCCAGCACA
	taatacgactcactatagggCTCGGCCTTCTGCTTAATTG
dsBlanks	taatacgactcactatagggtgtggatagtcggttccaaa
	taatacgactcactatagggatggcctcttatgccattca

9.1.5 Deep sequencing library generation and analysis

sRNA libraries were generated as previously described (Elmer et al., 2014). However, the ZR small RNA PAGE Recovery Kit (Zymo Research, USA) was used after PAGE-purification of the small RNAs. Deep sequencing was performed on an Illumina HiSeq instrument by LAFUGA (Gene Center, LMU Munich, Germany) and the sequences were analyzed using custom Galaxy, bowtie, perl and R scripts: Preprocessing and quality control of sequencing data was performed on a Galaxy Server (Giardine et al., 2005) hosted by LAFUGA (Gene Center, Munich). The reads were mapped using Bowtie (version 1.0.0) (Langmead et al., 2009). Bowtie-built index files of the *D. melanogaster* genome (release 6.08), mature miRNAs (miRBase download 2017-01-19), transposons (Berkley Drosophila Genome Project, download 2017-01-19) and of the transfected plasmids were used to determine the matching sequences allowing no mismatches (-v0). Mapped reads were visualized by using RStudio 0.97.551 / R version 3.0.2 and the R Bioconductor packages ShortRead (version 1.20.0) and Gviz (version 1.6.0).

9.2 Protein biochemistry

9.2.1 SDS-PAGE and western blotting

SDS-polyacrylamide gel electrophoresis (SDS-PAGE) and western blotting was performed as previously described in (Aumiller et al., 2012). The proteins were separated by a 10 % PAG (National Diagnostics, USA) at 150 V in a BioRad electrophoresis tank. Gels were stained using colloidal Commassie blue solutions or a silver staining kit by Thermo Scientific.

Blotting to a PVDF membrane (Millipore) was performed by tank blotting (100 V, 90 min). The membrane was blocked in 5 % milk-TBS/T for at least 20 minutes at room temperature. The incubation with the primary antibodies was conducted over night (4°C) in 1 x TBS-solution with 0,02 % Tween (TBS/T) + 5% milk. Subsequently, the membrane was washed 3 times in TBS-T at room temperature for approximately 10 min each and incubated with the appropriate secondary antibody for 2 h at room temperature. Again, the membrane was washed three times and afterwards the Enhanced Chemiluminescence (ECL) substrate (Thermo Scientific) was applied. The resulting signal was detected on a GE Amersham Imager 600 or a Fuji LAS 3000 mini system.

Following antibodies were used:

Primary antibodies:

anti-FLAG M2 (Sigma)	mouse	1:10,000 in milk - TBS + 0.02 % Tween
anti-V5 (Biorad)	mouse	1:10,000 in milk - TBS + 0.1 % Tween
anti-Dcr-2 (from Siomi Lab)	mouse	1:1,000 in milk - TBS + 0.02 % Tween

Secondary antibody:

anti-mouse-HRP (Jackson Immuno Research)	1:10,000 in TBS + 0.02% Tween

9.2.2 Immunoprecipitation and mass spectrometry

S2 cells were harvested, washed twice with 1x PBS and resuspended in an appropriate volume of 1xPBS to gain 10^7 cells / ml. 37 % formaldehyde stock solution was added to the medium to reach a final concentration of 0.1 % formaldehyde. The cell suspension was incubated for 5 min at room temperature on a rotating wheel. Finally, glycine (stock 1.25 M) was added to reach a concentration of 125 mM to quench the cross-linking reaction, the suspension was centrifuged and the supernatant discarded.

For lysis, the cells were resuspended in an appropriate volume of lysis buffer (150 mM KAc pH7.4, 30 mM Hepes pH 7.4, 5 mM MgAc, 1 mM DTT, 15% glycerol, 1% tergitol, per 10 ml: 1 tablet Protease Inhibitor (complete EDTA-free, Roche)). Cells were mechanically lysed using the Bioruptor (Diagenode, 20 cycles a 20'' ON, 40'' OFF). Insoluble debris was pelleted by centrifugation, the supernatant contains the protein extract.

For immunoprecipitation, 20 µL Dynabeads Protein G (Invitrogen) were washed with lysis buffer, resuspended in 200 µl in lysis buffer and incubated with 2 µl anti-FLAG M2 (Sigma) or anti-V5 (Biorad) antibody for 30 min at 4°C on a rotating wheel. Cell extract containing 5 mg protein was added to the beads and incubated for 1 h at 4°C. The beads were washed twice with buffer 1 (150 mM KAc pH 7.4, 30 mM Hepes pH 7.4, 5 mM MgAc, 0.1 % tergitol) and twice with buffer 2 (150 mM KAc pH 7.4, 30 mM Hepes pH 7.4, 5 mM MgAc).

The beads were either analyzed via SDS-PAGE and western blotting or via mass spectrometry. For the latter method, the beads were additionally washed 3 times with 50 mM ABC (ammonium bicarbonate) and resuspended in 100 µL of 5 ng/µl trypsin (NEB) in 1 M urea, 50 mM ABC. The digestion reaction was incubated for 30 min at 25°C, shaking at 800 rpm. The supernatant was saved, beads were washed twice with 50 µL of 50 mM ABC and finally, the supernatant was fused with the wash fractions. After the addition of DTT to a final concentration of 1 mM, the reaction was incubated at 25°C over night, shaking at 800 rpm. On the next day, 10 µL of 5 mg/ml iodoacetamide was added and the solution was incubated for 3 min in the dark at 25°C. 1 µL of 1 M DTT was added to the reaction, incubated for 10 min and in the end 2.5 µL trifluoroacetic acid was added to stop the reaction. The samples were lyophilized using a vacuum centrifuge.

LC-MS/MS was performed on an EASY-nLC 1000 chromatography system (Thermo Scientific, Waltham, MA, USA) connected to an Orbitrap XL instrument (Thermo Scientific). Peptides were diluted in 10 µl 0.1% formic acid (FA), transferred on a trap column (PepMap100 C18, 75 µm × 2 cm, 3 µm particles, Thermo Scientific) at a flow rate of 5µl/min and separated at a flow rate of 200 nL/min (Column: PepMap RSLC C18, 75 µm × 50 cm, 2 µm particles, Thermo Scientific) using a linear gradient from 2% to 35% solvent B (0.1% formic acid, 100% ACN) in 60 min. For data acquisition, a top five data dependent CID method was used. MS spectra were acquired from 300-2000 m/z at a resolution of 60.000. Collision-induced dissociation was performed using normalized collision energy of 35%.

Mass spec data was analyzed using MaxQuant v1.5.3.8 and Perseus v1.5.2.4: The data that was recorded by the mass spectrometer was analyzed using MaxQuant and its label-free quantification algorithm. After removing identified known contaminants and by the algorithm speculatively identified proteins, the resulting protein levels were transferred to logarithmic values and normalized to the z-score of the sample. Proteins that were not identified or only to such a small amount that no quantification was possible were assigned imputed values. These values were randomly calculated from a normal distribution of values that are close to the normalized values of the sample. The downstream analysis was performed with data prepared in that way.

9.2.3 Recombinant expression of GST-Blanks, protein purification and RNA binding assay

Buffers and media:

Bacterial culture medium	LB medium
	0.5 % glucose
	25 µg/ml chloramphenicol
	100 µg/ml ampicillin
Lysis buffer	50 mM TRIS pH 8.0
	150 mM NaCl
	10 mM beta-mercaptoethanol
	10 µg/ml lysozym
	0.1 U/ml DNase I
	per 10 ml: 1 tablet protease-inhibitor cocktail
	(Roche complete Mini)
High salt washing buffer	50 mM TRIS pH 8.0
	500 mM NaCl
	10 mM beta-mercaptoethanol
Elution buffer	50 mM TRIS pH 8.0
	150 mM NaCl
	1 mM beta-mercaptoethanol
	20 mM glutathione (reduced)

E.coli BL21 (DE, pLysS) were transformed with the plasmid pHZ1 that expresses the GST-Blanks fusion protein under the control of the *lac* operon. A single colony was used to inoculate 200 ml medium which was grown at 25°C and 120 rpm overnight. The culture was used to inoculate 2l of medium with an optical density of OD_{600} = 0.1, incubation was done at 25°C and 120 rpm. When the culture reached OD_{600} = 0.6, the expression was induced by the addition of 1 mM IPTG to the culture; cells were grown over night.

Cells were harvested by centrifugation (10 min, 4000 rpm), washed twice with 500 ml 1x PBS. The pellet was resuspended in 10 ml lysis buffer, 3 times sonicated for 1 min each and incubated for 1 h at 4°C. After centrifugation (max. speed, 8 min), the supernatant was saved (= protein extract).

The protein extract was added to 500 µl glutathione sepharose beads (Glutathione Sepharose 4 Fast Flow, Amersham Biosciences) in a reaction tube and incubated on a rotating wheel for 2 h at 4°C. Subsequently, the beads were transferred to a column and washed twice with 2 ml of lysis buffer each, with 2 ml of high salt washing buffer and again 3 times with 2 ml of lysis buffer. GST-Blanks was eluted from the beads by incubation in 1 ml elution buffer for 10 min at 4°C.

RNA binding assays were performed as previously described by (Fesser, 2013).

9.2.4 Shot-gun proteomics

Approximately 10^7-10^8 cells were harvested, washed twice in 1x PBS and the pellet was resuspended in 300 µL lysis buffer (4 % SDS, 100 mM TRIS pH 7.6, 0.1 M DTT) and boiled for 4 min at 95°C. The samples were sonicated using the Bioruptor (Diagenode, 15 cycles a 30′′ ON, 60′′ OFF) and afterwards centrifuged at maximal speed. 200 µg of the cell extract was used for sample preparation following the FASP protocol (Wisniewski et al., 2009). Finally, the solution was desalted using C18 stage tips (Empore Solid Phase Extraction cartridge, Millipore) and the peptides were lyophilized. The measurement was performed by "Zentrallabor für Proteinanalytik ZfP", LMU Munich (extended LC MS/MS run 210 minutes top 6). Mass spec data was analyzed using MaxQuant v1.5.3.8 and Perseus v1.5.2.4 as described above.

9.3 Cell culture

9.3.1 Culture conditions, transfection and cloning

D. melanogaster S2 cells (laboratory stocks) were cultured in Schneider's Medium (Bio&Sell; Nürnberg, Germany) containing 10% fetal bovine serum (FBS, Thermo Fisher; Waltham, USA) and Penicillin/Streptomycin in cell culture dishes at 25°C. The cells were split 1-2 times per week in a 1:10 ratio into new dishes and medium up to 20 passages.

Transfection of cells was performed as described in (Shah and Forstemann, 2008) using 4 µL FuGENE HD Transfection Reagents (Promega) and 0.5 µg plasmid DNA per 500 µL of culture medium.

Generation of ΦC31 integrase mediated cell lines was performed as previously described (Kunzelmann, 2013): F09 f.c. cells were transfected with the *attB* site containing reporter plasmids. F09 f.c. cells are S2 derivatives and contain several *attP* landing sites in their genome and stably express the ΦC31 integrase. ΦC31 integrase mediates the interaction of *attB* and *attP* sites and catalyzes the targeted integration of the plasmid into the genome. Positive clones were selected by adding 10 µg/ml blasticidin to the culturing medium 1.5 weeks after transfection. For 3 weeks, the cells were split once a week 1:5 into fresh medium. As GFP was used as the reporter gene, transfection and selection success was determined by either fluorescence microscopy or flow cytometry.

For single cell cloning, cells were seeded at 8,000 / ml in 50 % conditioned medium (1 day old) and plated in serial dilutions (1:2). Single cell colonies were picked and cultured for further analyses.

9.3.2 Genomic tagging

The genomic tagging was performed as described in the protocols on the webpage of the Förstemann lab (http://www.foerstemann.genzentrum.lmu.de/protocols/, version 2016/04/06) and in previous publications (Bottcher et al., 2014; Kunzelmann et al., 2016).

The following primers were used to generate the tagged cell lines:

Blanks (C-terminal)	CRISPR	cctattttcaatttaacgtcgCGAAAACATCACATGATTCgtttaagagctatgctg
	HR_s	ATCCAAGAAGGCAGCCCGTTACAAACTATCCGCTTTAGTTTGTAACAAACTATTTGGAACCGACTATCCACAAAAAGgatcttccggatggctcgag

	HR_as	GAATGCCGCTAATACGTTTTAAAGTACTACCGGGTGCCCAGAgTtATGTGATGTTTTCGTGAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCCATATG
Blanks (N-terminal)	CRISPR	TAATACGACTCACTATAGCTAATTTGTGTTTTAAATAAGTTTTAGAGCTAG
	HR_s	GTTTTGTGGAAGCGGAGCTAATTTGTGTTTTAAAAGTTTTGTAAAGGCGGAGCTAATTTGTGTTTTAAATAGAAGTTCCTATACTTTCTAGAGAATAGGAACTTCCATATG
	HR_as	GGATCAATGCCTTCTTCATCTCTAAGATGTCCACCTTTGCACAACTTTCAGCCAACAATTGCTTTGCTTCACCGCCGCTTGGAGCAGC
pAbp (C-terminal)	CRISPR	cctattttcaatttaacgtcgTGAGCTGTTCGAGCTTAGTgtttaagagctatgct
	HR_s	GCCAAGGTGGAGGAGGCTGTGGCCGTGCTTCAGGTGCACCGCGTCACCGAGCCCGCCAACggatcttccggatggctcgag
	HR_as	acttacTTTTTGGGTGTAAAGTGTTCTTGTAAAATGTTCATACGCTTGAGCTGTTCGAGCgaagttcctattctctagaaagtataggaacttccatatg
Hrb27C (C-terminal)	CRISPR	tcaatttaacgtcgCAGGCTGTCTAAAGAGAGAgttttagagctag
	HR_s	CGAACTACGGAGCAGGGCCGCGATCAGCGTACGGCAACGACAGCTCCACGCAGCCACCCTATGCAACCTCGCAGGCTGTCggatcttccggatggctcgag
	HR_as	TCATTCGCCGAGGACGACATGCTACTCCGCTCCTCTCTGCTCCGCTACTCCTACTTCTCCTCCACACGATCCTTCTCTCTGAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCCATATG
RpA70 (C-terminal)	CRISPR	cctattttcaatttaacgtcgTGTTTTGGCGCACTGCGCAgtttaagagctatgctg
	HR_s	GAGTACAATAAGCACTTGCTCAAGGAGCTGCAGGAGCTAACCGGCATTGGCTCATCAAACggatcttccggatggctcgag
	HR_as	TAAATGTATAGCCTATTCCTAATTTATGGGAAACGTGTTTTGGCGtACTGaGCATGGAACgaagttcctattctctagaaagtataggaacttccatatg
SpnA (C-terminal)	CRISPR	cctattttcaatttaacgtcgGCTAATTGTGCTCACTTATgttttagagctag
	HR_s	CCGGAATCGGAGGCCATGTTCGCCATTCTGCCGGATGGAATAGGAGACGCCAGGGAGAGCggatcttccggatggctcgag
	HR_as	AATCATTAGAAAGTTAGGGAGCTCTATTCCTAGCATTAATTAACCTATAAGTGAGCACAAGAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCCATATG
Act5C (C-terminal)	CRISPR	cctattttcaatttaacgtcgACCGCAAGTGCTTCTAAGAgtttaagagctatgctg
	HR_s	TGGATCTCCAAGCAGGAGTACGACGAGTCCGGCCCCTCCATTGTGCACCGCAAGTGCTTCggatcttccggatggctcgag
	HR_as	CCTCCAGCAGAATCAAGACCATCCCGATCCTGATCCTCTTGCCCAGACAAGCGATCCTTCGAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCCATATG
	HR_as marker free	TCCTCCTCCTCCTCCAGCAGAATCAAGACCATCCCGATCCTGATCCTCTTGCCCAGACAAGCGATCCTTCAAAGTATAGGAACTTCACTAGT

Rtf1 (C-terminal)	CRISPR	cctattttcaatttaacgtcgGCGCGGTCTTATCTAGAAAgtttaagagctatgctg
	HR_s	GAAACAGTCTCCAAACGTTCGCTGAATCTAGAGGATTACAAAAAAAAGCGCGGTCTTATCggatcttccggatggctcgag
	HR_as	AAACTAGGACGTAAACATGCGACGCTTTGTCAGTATATGTATGACCAGCGGCATCCTTTTgaagttcctattctctagaaagtataggaacttccatatg

9.3.3 Knock down reporter assay and flow cytometry

dsRNA is used to knock down gene expression and so deplete different RNAi factors. *Drosophila* cells incorporate dsRNA by soaking and process the dsRNA precursor into effective siRNAs that knock down gene expression of complementary mRNAs (Caplen et al., 2000; Saleh et al., 2006). Cells were seeded at $0.5 \cdot 10^6$/ml and 3 µg/ml dsRNA was added to the culturing medium. After 1 week cells were split 1:10 to new dishes and 3 µg/ml dsRNA was again added. After 2 weeks, cells were ready for experiments and analysis.

The GFP fluorescence was read out at the Becton Dickinson FACSCalibur flow cytometer and the BD HTS plate reader. To this end, the cell suspension was diluted in an equal volume of FACS-Flow. GFP fluorescence was quantified using the following settings:

FSC detector:	E00 Voltage	1.00 AmpGain	Lin Mode
SSC detector:	340 Voltage	1.00 AmpGain	Lin Mode
FL1 detector:	325 or 400 Voltage	1.00 AmpGain	Log Mode
Threshold:	primary parameter:	FSC	
	value:	199	

CellQuest Pro (Version 6.0) and BD PlateManager (Version 2.0) were used to access the cytometer. Data was analyzed using Rstudio 0.97.551 / R version 3.0.2, the Bioconductor packages flowCore (version 1.28.09) and prada (version 1.38.0) and flowing software version 2.5.0 and 2.5.1 (http://www.flowingsoftware.com/).

9.3.4 Microscopy and importazol assay

Cells were split to an approximately concentration of 2x 10^6 /ml. Either DMSO or 200 µM importazol (Sigma, stock 6.28 mM) was added to the cells. 16 h later florescence microscopy was conducted. To this end, the cells were resuspended, 10 µl suspension was stained with 1 µl of 10 µg/ml Hoechst 33342 and investigated using a Leica CS SP2 confocal microscope.

10 Literature

Ahlquist, P. (2002). RNA-Dependent RNA Polymerases, Viruses, and RNA Silencing.

Anders, C., Niewoehner, O., Duerst, A., and Jinek, M. (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature *513*, 569-573.

Aumiller, V., Graebsch, A., Kremmer, E., Niessing, D., and Forstemann, K. (2012). Drosophila Pur-alpha binds to trinucleotide-repeat containing cellular RNAs and translocates to the early oocyte. RNA biology *9*.

Azlan, A., Dzaki, N., and Azzam, G. (2016). Argonaute: The executor of small RNA function. J Genet Genomics *43*, 481-494.

Barber, G.N. (2009). The NFAR's (nuclear factors associated with dsRNA): evolutionarily conserved members of the dsRNA binding protein family. RNA Biol *6*, 35-39.

Barber, G.N. (2011). Innate immune DNA sensing pathways: STING, AIMII and the regulation of interferon production and inflammatory responses. Curr Opin Immunol *23*, 10-20.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215-233.

Bassett, A.R., Tibbit, C., Ponting, C.P., and Liu, J.L. (2014). Mutagenesis and homologous recombination in Drosophila cell lines using CRISPR/Cas9. Biol Open *3*, 42-49.

Bhaya, D., Davison, M., and Barrangou, R. (2011). CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. Annu Rev Genet *45*, 273-297.

Bird, S.L., Heald, R., and Weis, K. (2013). RanGTP and CLASP1 cooperate to position the mitotic spindle. In Mol Biol Cell, pp. 2506-2514.

Byrne, S.M., Mali, P., and Church, G.M. (2014). Genome editing in human stem cells. Methods in enzymology *546*, 119-138.

Böttcher, R., Hollmann, M., Merk, K., Nitschko, V., Obermaier, C., Philippou-Massier, J., Wieland, I., Gaul, U., and Förstemann, K. (2014). Efficient chromosomal gene modification with CRISPR/cas9 and PCR-based homologous recombination donors in cultured Drosophila cells.

Caplen, N.J., Fleenor, J., Fire, A., and Morgan, R.A. (2000). dsRNA-mediated gene silencing in cultured Drosophila cells: a tissue culture model for the analysis of RNA interference. Gene *252*, 95-105.

Carthew, R.W., Agbu, P., and Giri, R. (2016). MicroRNA function in Drosophila melanogaster. Semin Cell Dev Biol.

Chawla, G., and Sokol, N.S. (2011). MicroRNAs in Drosophila development. International review of cell and molecular biology *286*, 1-65.

Chenais, B. (2015). Transposable elements in cancer and other human diseases. Curr Cancer Drug Targets *15*, 227-242.

Chiu, Y.H., Macmillan, J.B., and Chen, Z.J. (2009). RNA polymerase III detects cytosolic DNA and induces type I interferons through the RIG-I pathway. Cell *138*, 576-591.

Choi, E., Choi, E., and Hwang, K.C. (2013). MicroRNAs as novel regulators of stem cell fate. World journal of stem cells *5*, 172-187.

Clayton, E.A., Wang, L., Rishishwar, L., Wang, J., McDonald, J.F., and Jordan, I.K. (2016). Patterns of Transposable Element Expression and Insertion in Cancer. Front Mol Biosci *3*, 76.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature Biotechnology *26*, 1367-1372.

Czech, B., and Hannon, G.J. (2011). Small RNA sorting: matchmaking for Argonautes. Nature reviews Genetics *12*, 19-31.

Czech, B., Zhou, R., Erlich, Y., Brennecke, J., Binari, R., Villalta, C., Gordon, A., Perrimon, N., and Hannon, G.J. (2009). Hierarchical rules for Argonaute loading in Drosophila. Molecular cell *36*, 445-456.

Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F., and Hannon, G.J. (2004). Processing of primary microRNAs by the Microprocessor complex. Nature *432*, 231-235.

Doudna, J.A., and Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science *346*, 1258096.

Dumesic, P.A., and Madhani, H.D. (2013). The spliceosome as a transposon sensor. RNA biology *10*, 1653-1660.

Dumesic, P.A., Natarajan, P., Chen, C., Drinnenberg, I.A., Schiller, B.J., Thompson, J., Moresco, J.J., Yates, J.R., 3rd, Bartel, D.P., and Madhani, H.D. (2013). Stalled spliceosomes are a signal for RNAi-mediated genome defense. Cell *152*, 957-968.

Elmer, A.K. (2013). Modeling Transposon Recognition in Drosophila melanogaster.

Elmer, K., Helfer, S., Mirkovic-Hosle, M., and Forstemann, K. (2014). Analysis of endo-siRNAs in Drosophila. Methods Mol Biol *1173*, 33-49.

Eulalio, A., Tritschler, F., and Izaurralde, E. (2009). The GW182 protein family in animal cells: New insights into domains required for miRNA-mediated gene silencing. In RNA, pp. 1433-1442.

Fabrini, R., De Luca, A., Stella, L., Mei, G., Orioni, B., Ciccone, S., Federici, G., Lo Bello, M., and Ricci, G. (2009). Monomer-dimer equilibrium in glutathione transferases: a critical re-examination. Biochemistry *48*, 10473-10482.

Feinberg, E.H., and Hunter, C.P. (2003). Transport of dsRNA into cells by the transmembrane protein SID-1. Science *301*, 1545-1547.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics *9*, 397-405.

Fesser, S.M. (2013). Contribution of RNA binding proteins to substrate specificity in small RNA biogenesis. In Faculty of Chemistry and Pharmacy (Munich: LMU München).

Fetter, J., Samsonov, A., Zenser, N., Zhang, F., Zhang, H., and Malkov, D. (2015). Endogenous gene tagging with fluorescent proteins. Methods in molecular biology *1239*, 231-240.

Finocchiaro, G., Carro, M.S., Francois, S., Parise, P., DiNinni, V., and Muller, H. (2007). Localizing hotspots of antisense transcription. Nucleic acids research *35*, 1488-1500.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806-811.

Forstemann, K., Horwich, M.D., Wee, L., Tomari, Y., and Zamore, P.D. (2007). Drosophila microRNAs are sorted into functionally distinct argonaute complexes after production by dicer-1. Cell *130*, 287-297.

Forstemann, K., Tomari, Y., Du, T., Vagin, V.V., Denli, A.M., Bratu, D.P., Klattenhoff, C., Theurkauf, W.E., and Zamore, P.D. (2005). Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. PLoS biology *3*, e236.

Fu, Y., Reyon, D., and Joung, J.K. (2014). Targeted genome editing in human cells using CRISPR/Cas nucleases and truncated guide RNAs. Methods in enzymology *546*, 21-45.

Fukaya, T., and Tomari, Y. (2012). MicroRNAs mediate gene silencing via multiple different pathways in drosophila. Molecular cell *48*, 825-836.

Futatsumori-Sugai, M., Abe, R., Watanabe, M., Kudou, M., Yamamoto, T., Ejima, D., Arakawa, T., and Tsumoto, K. (2009). Utilization of Arg-elution method for FLAG-tag based chromatography. Protein Expr Purif *67*, 148-155.

Gao, M., McCluskey, P., Loganathan, S.N., and Arkov, A.L. (2014). An in vivo crosslinking approach to isolate protein complexes from Drosophila embryos. J Vis Exp.

Gerbasi, V.R., Preall, J.B., Golden, D.E., Powell, D.W., Cummins, T.D., and Sontheimer, E.J. (2011). Blanks, a nuclear siRNA/dsRNA-binding complex component, is required for Drosophila spermiogenesis. Proceedings of the National Academy of Sciences of the United States of America *108*, 3204-3209.

Ghildiyal, M., and Zamore, P.D. (2009). Small silencing RNAs: an expanding universe. Nat Rev Genet *10*, 94-108.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., *et al.* (2005). Galaxy: a platform for interactive large-scale genome analysis. Genome research *15*, 1451-1455.

Goic, B., Vodovar, N., Mondotte, J.A., Monot, C., Frangeul, L., Blanc, H., Gausson, V., Vera-Otarola, J., Cristofari, G., and Saleh, M.C. (2013). RNA-mediated interference and reverse transcription control the persistence of RNA viruses in the insect model Drosophila. Nat Immunol *14*, 396-403.

Halic, M., and Moazed, D. (2010). Dicer-independent primal RNAs trigger RNAi and heterochromatin formation. Cell *140*, 504-516.

Hartig, J.V., Esslinger, S., Bottcher, R., Saito, K., and Forstemann, K. (2009). Endo-siRNAs depend on a new isoform of loquacious and target artificially introduced, high-copy sequences. The EMBO journal *28*, 2932-2944.

Hartig, J.V., and Forstemann, K. (2011). Loqs-PD and R2D2 define independent pathways for RISC generation in Drosophila. Nucleic acids research *39*, 3836-3851.

Hartig, J.V., Tomari, Y., and Forstemann, K. (2007). piRNAs--the ancient hunters of genome invaders. Genes & development *21*, 1707-1713.

Huang, V., and Li, L.C. (2014). Demystifying the nuclear function of Argonaute proteins. In RNA Biol, pp. 18-24.

Hundley, H.A., and Bass, B.L. (2010). ADAR editing in double-stranded UTRs and other noncoding RNA sequences. Trends Biochem Sci *35*, 377-383.

Islam, M.S., Aryasomayajula, A., and Selvaganapathy, P.R. (2017). A Review on Macroscale and Microscale Cell Lysis Methods. Micromachines *8*, 1-25.

Jakob, L., Treiber, T., Treiber, N., Gust, A., Kramm, K., Hansen, K., Stotz, M., Wankerl, L., Herzog, F., Hannus, S., *et al.* (2016). Structural and functional insights into the fly microRNA biogenesis factor Loquacious. RNA *22*, 383-396.

Ji, L., and Chen, X. (2012). Regulation of small RNA stability: methylation and beyond. Cell Research *22*, 624-636.

Jiang, F., and Doudna, J.A. (2017). CRISPR-Cas9 Structures and Mechanisms. Annu Rev Biophys.

Kaboord, B., and Perr, M. (2008). Isolation of proteins and protein complexes by immunoprecipitation. Methods Mol Biol *424*, 349-364.

Kandasamy, S.K., and Fukunaga, R. (2016). Phosphate-binding pocket in Dicer-2 PAZ domain for high-fidelity siRNA production.

Kazazian, H.H., Jr. (2004). Mobile elements: drivers of genome evolution. Science *303*, 1626-1632.

Khurana, J.S., and Theurkauf, W. (2010). piRNAs, transposon silencing, and Drosophila germline development. The Journal of cell biology *191*, 905-913.

Kim, K., Lee, Y.S., Harris, D., Nakahara, K., and Carthew, R.W. (2006). The RNAi pathway initiated by Dicer-2 in Drosophila. Cold Spring Harbor symposia on quantitative biology *71*, 39-44.

Kim, V.N., Han, J., and Siomi, M.C. (2009). Biogenesis of small RNAs in animals. Nature reviews Molecular cell biology *10*, 126-139.

Kunzelmann, S., Bottcher, R., Schmidts, I., and Forstemann, K. (2016). A Comprehensive Toolbox for Genome Editing in Cultured Drosophila melanogaster Cells. G3 (Bethesda) *6*, 1777-1785.

Kunzelmann, S.B. (2013). The influence of functional genetic elements on triggering an RNAi response in Drosophila melanogaster (Master Thesis). In Faculty for Chemistry and Pharmacy, Department of Biochemistry (Munich: LMU München).

Lamontagne, B., Larose, S., Boulanger, J., and Elela, S.A. (2001). The RNase III family: a conserved structure and expanding functions in eukaryotic dsRNA metabolism. Current issues in molecular biology *3*, 71-78.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843-854.

Levin, H.L., and Moran, J.V. (2011). Dynamic interactions between transposable elements and their hosts. Nat Rev Genet *12*, 615-627.

Li, K., Wang, G., Andersen, T., Zhou, P., and Pu, W.T. (2014). Optimization of genome engineering approaches with the CRISPR/Cas9 system. PLoS ONE *9*, e105779.

Liang, C., Wang, Y., Murota, Y., Liu, X., Smith, D., Siomi, M.C., and Liu, Q. (2015). TAF11 assembles RISC loading complex to enhance RNAi efficiency. Mol Cell *59*, 807-818.

Lin, C.C., and Potter, C.J. (2016). Non-Mendelian Dominant Maternal Effects Caused by CRISPR/Cas9 Transgenic Components in Drosophila melanogaster. G3 (Bethesda).

Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods *25*, 402-408.

Lottspeich, F., and Engels, J.W. (2006). Bioanalytik, Vol 2 (Heidelberg: Spektrum Akademischer Verlag).

Luhur, A., Chawla, G., and Sokol, N.S. (2013). MicroRNAs as components of systemic signaling pathways in Drosophila melanogaster. Current topics in developmental biology *105*, 97-123.

Mann, M., Hendrickson, R.C., and Pandey, A. (2001). Analysis of proteins and proteomes by mass spectrometry. Annu Rev Biochem *70*, 437-473.

Mayr, C. (2016). Evolution and Biological Roles of Alternative 3'UTRs. Trends Cell Biol *26*, 227-237.

Meister, G. (2013). Argonaute proteins: functional insights and emerging roles. Nature reviews Genetics *14*, 447-459.

Michalik, K.M., Bottcher, R., and Forstemann, K. (2012). A small RNA response at DNA ends in Drosophila. Nucleic Acids Res *40*, 9596-9603.

Mirkovic-Hosle, M., and Forstemann, K. (2014). Transposon defense by endo-siRNAs, piRNAs and somatic pilRNAs in Drosophila: contributions of Loqs-PD and R2D2. PLoS One *9*, e84994.

Miyoshi, K., Miyoshi, T., Hartig, J.V., Siomi, H., and Siomi, M.C. (2010a). Molecular mechanisms that funnel RNA precursors into endogenous small-interfering RNA and microRNA biogenesis pathways in Drosophila. Rna.

Miyoshi, K., Miyoshi, T., and Siomi, H. (2010b). Many ways to generate microRNA-like small RNAs: non-canonical pathways for microRNA production. Molecular genetics and genomics : MGG *284*, 95-103.

Nishida, K.M., Miyoshi, K., Ogino, A., Miyoshi, T., Siomi, H., and Siomi, M.C. (2013). Roles of R2D2, a cytoplasmic D2 body component, in the endogenous siRNA pathway in Drosophila. Molecular cell *49*, 680-691.

Nishikura, K. (2010). Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem *79*, 321-349.

Okamura, K., Balla, S., Martin, R., Liu, N., and Lai, E.C. (2008a). Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in Drosophila melanogaster. Nature structural & molecular biology *15*, 581-590.

Okamura, K., Chung, W.J., Ruby, J.G., Guo, H., Bartel, D.P., and Lai, E.C. (2008b). The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. Nature *453*, 803-806.

Okamura, K., Hagen, J.W., Duan, H., Tyler, D.M., and Lai, E.C. (2007). The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. Cell *130*, 89-100.

Okamura, K., Ishizuka, A., Siomi, H., and Siomi, M.C. (2004). Distinct roles for Argonaute proteins in small RNA-directed RNA cleavage pathways. Genes Dev *18*, 1655-1666.

Okamura, K., and Lai, E.C. (2008). Endogenous small interfering RNAs in animals. Nature reviews Molecular cell biology *9*, 673-678.

Ooi, S.K.T., O'Donnell, A.H., and Bestor, T.H. (2009). Mammalian cytosine methylation at a glance.

Patel, D.J., Ma, J.B., Yuan, Y.R., Ye, K., Pei, Y., Kuryavyi, V., Malinina, L., Meister, G., and Tuschl, T. (2006). Structural biology of RNA silencing and its functional implications. Cold Spring Harbor symposia on quantitative biology *71*, 81-93.

Quenerch'du, E., Anand, A., and Kai, T. (2016). The piRNA pathway is developmentally regulated during spermatogenesis in Drosophila. Rna *22*, 1044-1054.

Raices, M., and D'Angelo, M.A. (2012). Nuclear pore complex composition: a new regulator of tissue-specific and developmental functions. Nature Reviews Molecular Cell Biology *13*, 687-699.

Rathinam, V.A., and Fitzgerald, K.A. (2011). Innate immune sensing of DNA viruses. Virology *411*, 153-162.

Rebollo, R., Romanish, M.T., and Mager, D.L. (2012). Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet *46*, 21-42.

Russo, J., Harrington, A.W., and Steiniger, M. (2016). Antisense Transcription of Retrotransposons in Drosophila: An Origin of Endogenous Small Interfering RNA Precursors. Genetics *202*, 107-121.

Sabin, L.R., Zheng, Q., Thekkat, P., Yang, J., Hannon, G.J., Gregory, B.D., Tudor, M., and Cherry, S. (2013). Dicer-2 processes diverse viral RNA species. PloS one *8*, e55458.

Saleh, M.C., van Rij, R.P., Hekele, A., Gillis, A., Foley, E., O'Farrell, P.H., and Andino, R. (2006). The endocytic pathway mediates cell entry of dsRNA to induce RNAi silencing. Nature cell biology *8*, 793-802.

Sambrook, J., and Russel, D.W. (2000). Molecular Cloning: A Laboratory Manual, Vol 1-3 (Cold Spring Harbor Laboratory).

Sanders, C., and Smith, D.P. (2011). LUMP is a putative double-stranded RNA binding protein required for male fertility in Drosophila melanogaster. PLoS One *6*, e24151.

Schmidts, I., Bottcher, R., Mirkovic-Hosle, M., and Forstemann, K. (2016). Homology directed repair is unaffected by the absence of siRNAs in Drosophila melanogaster. Nucleic Acids Res *44*, 8261-8271.

Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Soeding, J., and Cramer, P. (2013). Transcriptome surveillance by selective termination of noncoding RNA synthesis. Cell *155*, 1075-1087.

Shah, C., and Forstemann, K. (2008). Monitoring miRNA-mediated silencing in Drosophila melanogaster S2-cells. Biochimica et biophysica acta *1779*, 766-772.

Siomi, M.C., Miyoshi, T., and Siomi, H. (2010). piRNA-mediated silencing in Drosophila germlines. Seminars in cell & developmental biology *21*, 754-759.

Siomi, M.C., Sato, K., Pezic, D., and Aravin, A.A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. Nature reviews Molecular cell biology *12*, 246-258.

Smith, M.C.M., Brown, W.R.A., McEwan, A.R., and Rowley, P.A. (2010). Site-specific recombination by φC31 integrase and other large serine recombinases.

Song, L., Craney, A., and Rape, M. (2014). Microtubule-dependent regulation of mitotic protein degradation. Mol Cell *53*, 179-192.

Sternberg, S.H., LaFrance, B., Kaplan, M., and Doudna, J.A. (2015). Conformational control of DNA target cleavage by CRISPR-Cas9. Nature *527*, 110-113.

Stewart-Ornstein, J., and Lahav, G. (2016). Dynamics of CDKN1A in Single Cells Defined by an Endogenous Fluorescent Tagging Toolkit. Cell Rep.

Sun, M., Hurst, L.D., Carmichael, G.G., and Chen, J. (2006). Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense transcriptionand organismic complexity. Genome research *16*, 922-933.

Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., Snyder, M.P., and Wang, T. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res *24*, 1963-1976.

Sutherland, B.W., Toews, J., and Kast, J. (2008). Utility of formaldehyde cross-linking and mass spectrometry in the study of protein-protein interactions. J Mass Spectrom *43*, 699-715.

Swenson, J.M., Colmenares, S.U., Strom, A.R., Costes, S.V., and Karpen, G.H. (2016). The composition and organization of Drosophila heterochromatin are heterogeneous and dynamic. In eLife.

ten Have, S., Boulon, S., Ahmad, Y., and Lamond, A.I. (2011). Mass spectrometry-based immuno-precipitation proteomics - the user's guide. Proteomics *11*, 1153-1159.

Tomari, Y., Matranga, C., Haley, B., Martinez, N., and Zamore, P.D. (2004). A protein sensor for siRNA asymmetry. Science *306*, 1377-1380.

van Rij, R.P., and Berezikov, E. (2009). Small RNAs and the control of transposons and viruses in Drosophila. Trends in microbiology *17*, 163-171.

Vasilescu, J., Guo, X., and Kast, J. (2004). Identification of protein-protein interactions using in vivo cross-linking and mass spectrometry. Proteomics *4*, 3845-3854.

Verdel, A., and Moazed, D. (2005). RNAi-directed assembly of heterochromatin in fission yeast. FEBS Lett *579*, 5872-5878.

Wang, W., Han, B.W., Tipping, C., Ge, D.T., Zhang, Z., Weng, Z., and Zamore, P.D. (2015a). Slicing and Binding by Ago3 or Aub Trigger Piwi-Bound piRNA Production by Distinct Mechanisms. Mol Cell *59*, 819-830.

Wang, Z., Wu, D., Liu, Y., Xia, X., Gong, W., Qiu, Y., Yang, J., Zheng, Y., Li, J., Wang, Y.F., *et al.* (2015b). Drosophila Dicer-2 has an RNA interference-independent function that modulates Toll immune signaling. Sci Adv *1*, e1500228.

Werren, J.H. (2011). Selfish genetic elements, genetic conflict, and evolutionary innovation.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., *et al.* (2007). A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics *8*, 973-982.

Wilson, R.C., and Doudna, J.A. (2013). Molecular mechanisms of RNA interference. Annual review of biophysics *42*, 217-239.

Wisniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. Nat Methods *6*, 359-362.

Wyvekens, N., Tsai, S.Q., and Joung, J.K. (2015). Genome Editing in Human Cells Using CRISPR/Cas Nucleases. Curr Protoc Mol Biol *112*, 31 33 31-31 33 18.

Yamanaka, S., Siomi, M.C., and Siomi, H. (2014). piRNA clusters and open chromatin structure. Mobile DNA *5*, 22.

Yang, F., Zhao, R., Fang, X., Huang, H., Xuan, Y., Ma, Y., Chen, H., Cai, T., Qi, Y., and Xi, R. (2015a). The RNA surveillance complex Pelo-Hbs1 is required for transposon silencing in the Drosophila germline. EMBO Rep *16*, 965-974.

Yang, Q., Ye, Q.A., and Liu, Y. (2015b). Mechanism of siRNA production from repetitive DNA. Genes & development.

Yepiskoposyan, H., Egli, D., Fergestad, T., Selvaraj, A., Treiber, C., Multhaup, G., Georgiev, O., and Schaffner, W. (2006). Transcriptome response to heavy metal stress in Drosophila reveals a new zinc transporter that confers resistance to zinc. Nucleic Acids Res *34*, 4866-4877.

Yi, R., Qin, Y., Macara, I.G., and Cullen, B.R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. In Genes Dev, pp. 3011-3016.

Zhang, X.H., Tee, L.Y., Wang, X.G., Huang, Q.S., and Yang, S.H. (2015). Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. Mol Ther Nucleic Acids *4*, e264.

Zhang, Y., Liu, X.S., Liu, Q.R., and Wei, L. (2006). Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. Nucleic Acids Res *34*, 3465-3475.

Zhou, R., Hotta, I., Denli, A.M., Hong, P., Perrimon, N., and Hannon, G.J. (2008). Comparative analysis of argonaute-dependent small RNA pathways in Drosophila. Mol Cell *32*, 592-599.

Zhou, R., Mohr, S., Hannon, G.J., and Perrimon, N. (2014). Inducing RNAi in Drosophila cells by soaking with dsRNA. Cold Spring Harb Protoc *2014*.

11 Acknowledgements

Mein Dank gilt zuerst meinem Doktorvater Prof. Dr. Klaus Förstemann für die Betreuung und Begleitung während meiner Zeit am Genzentrum. Danke, dass ich so viel von dir lernen durfte, du immer für Gespräche und Diskussionen eine offene Tür hattest und mir motivierend zur Seite standest.

Außerdem möchte ich mich bei allen aktuellen und vergangenen Mitgliedern der AG Förstemann für die schönen gemeinsamen Stunden bedanken – allen voran Romy Böttcher für die Freundschaft und Unterstützung, aber auch Ines Schmidts, Karin Merk und Volker Nitschko für die fachlichen Diskussionen und die gemeinsamen lustigen Stunden. Die Zeit mit euch am Genzentrum werde ich so schnell nicht vergessen! Außerdem ein besonderes Danke an Volker für das Korrekturlesen und die konstruktiven Kommentare zu dieser Arbeit.

Danke auch an Dietmar Martin für die Übernahme des Zweitgutachtens und allen Mitgliedern meiner Prüfungskommission für ihre investierte Zeit, sowie den Mitgliedern meines TAC, Birgitta Beatrix und Stefan Krebs. Ein besonderer Dank geht noch an Thomas Fröhlich für die umkomplizierte Einführung in die Geheimnisse der Massenspektrometrie und die spontane Hilfe bei den Messungen.

Und nicht zuletzt möchte ich meiner Familie, meinen Eltern, meiner Schwester und meinen Großeltern, danke sagen, dass sie mich in meinen Studienzeiten unterstützt und mir damit all das erst ermöglicht haben.