Centromere-associated RNAs in *Drosophila melanogaster*

Iris Valent

2022

Inaugural dissertation

for

obtaining the doctoral degree

of the

combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht – Karls – University

Heidelberg

Presented by

M.Sc. Iris Anne Valent

born in Amstelveen, The Netherlands

Oral examination: October 4th, 2022

Centromere-associated RNAs in *Drosophila melanogaster*

Referees:

Prof. Dr. Georg Stoecklin

Prof. Dr. Sylvia Erhardt

Summary

The centromere is an essential locus for chromosome segregation. It serves as a platform for kinetochore assembly and subsequent spindle microtubule attachment during cell division. Histone H3 variant CENP-A, or CID in *Drosophila*, is incorporated into centromeric chromatin to epigenetically mark its location. To maintain the centromere, new CID-containing nucleosomes are loaded by the chaperone Cal1, and targeted to the centromere by the inner kinetochore protein CENP-C. Even though centromeres are epigenetically defined and thus independent of the underlying DNA sequences, most are localized on repetitive DNA. In *Drosophila,* the centromere of each chromosome has a unique combination of transposable elements (TEs), flanked by both simple and complex satellite repeats. In all eukaryotes studied so far, centromeric transcription and the derived centromeric RNAs (cenRNAs) seem to be a structural or regulatory component of the centromere. In *Drosophila*, the (peri)centromeric RNA SATIII is a part of centromeric chromatin and important for CID loading.

Due to the wide variety of repeats and TEs present at *Drosophila* centromeric DNA, it is likely that we do not have the full picture of the RNAs that are associated with centromeres. Therefore, I set out to identify all centromere-associated RNAs using RNA chromatin immunoprecipitation followed by deep sequencing (RNA-ChIP-seq) in embryos. RNAs from several TEs were enriched in this centromeric pull down, most significantly gypsy element Blastopia. I validated the centromeric localization of Blastopia with single molecule RNA fluorescent in situ hybridization (smRNA-FISH), but was unable to observe pronounced mitotic defects after knockdown.

Additionally, I set out to identify which of the factors involved in the CID loading machinery are RNA-binding proteins. Using an RNA-protein complex isolation method called XRNAX, I showed that Cal1 and CID, but not CENP-C, directly bind RNA in S2 cells. By sequencing the RNAs linked to Cal1, I identified the centromere-associated TE Copia as the interaction RNA of Cal1. smRNA-FISH confirmed the presence of Copia at centromeres.

Taken together, I identified novel centromere-associated RNAs as a structural or regulatory component of centromeric chromatin, some of which directly interact with the key centromeric protein Cal1.

Zusammenfassung

Das Zentromer ist ein wesentlicher Ort für die Chromosomentrennung. Es dient als Plattform für den Aufbau des Kinetochors und die anschließende Befestigung der Spindelmikrotubuli während der Zellteilung. Die Histon H3-Variante CENP-A, in *Drosophila* CID genannt, wird in das zentromerische Chromatin eingebaut, um dessen Position epigenetisch zu markieren. Um das Zentromer zu erhalten, werden neue CID-haltige Nukleosomen durch das Chaperon Cal1 auf die DNA geladen und durch das innere Kinetochorprotein CENP-C zum Zentromer gelenkt. Obwohl Zentromere epigenetisch definiert und somit unabhängig von den zugrunde liegenden DNA-Sequenzen sind, befinden sich die meisten auf repetitiver DNA. Bei *Drosophila* weist das Zentromer jedes Chromosoms eine einzigartige Kombination von Transponiblen Elementen (TEs) auf, die sowohl von einfachen als auch komplexen Satelliten-DNA-Sequenzen flankiert werden. In allen bisher untersuchten Eukaryonten scheinen die zentromerische Transkription und die daraus entstehenden zentromerischen RNAs (cenRNAs) eine strukturelle oder regulatorische Komponente des Zentromers zu sein. In *Drosophila* ist die (peri)zentromerische RNA SATIII ein Teil des zentromerischen Chromatins und wichtig für die CID-Ladung.

Aufgrund der großen Vielfalt an repetitiven Sequenzen und TEs, die in der zentromerischen DNA von *Drosophila* vorkommen, ist es wahrscheinlich, dass wir kein vollständiges Bild von den RNAs haben, die mit Zentromeren verbunden sind. Daher hab ich in dieser Arbeit das Ziel verfolgt, alle Zentromer-assoziierten RNAs mit Hilfe einer RNA-Chromatin-Immunpräzipitation gefolgt von tiefer Next-Generation Sequenzierung (RNA-ChIP-seq) in *Drosophila* Embryonen zu identifizieren. RNAs von mehreren TEs wurden in dieser zentromerischen Fällung angereichert, vor allem vom Gypsy-Element Blastopia. Ich validierte die zentromerische Lokalisierung von Blastopia mit Einzelmolekül-RNA-Fluoreszenz-in-situ-Hybridisierung (smRNA-FISH), konnte aber nach Verringerung der RNA mittels Knockdown keine ausgeprägten mitotischen Defekte beobachten.

Des Weiteren wollte ich herausfinden, welche der Faktoren, die an der CID-Lademaschinerie beteiligt sind, RNA-bindende Proteine sind. Mithilfe einer RNA-Protein-Komplex-Isolierungsmethode namens XRNAX konnte ich zeigen, dass Cal1 und CID, nicht aber CENP-C, in S2-Zellen direkt RNA binden. Durch Sequenzierung der mit Cal1 verknüpften RNAs identifizierte ich das Zentromer-assoziierte TE Copia als Interaktions-RNA von Cal1. smRNA-FISH bestätigte das Vorhandensein von Copia an Zentromeren.

Zusammenfassend habe ich neue Zentromer-assoziierte RNAs als strukturelle oder regulatorische Komponente des zentromerischen Chromatins identifiziert, von denen einige direkt mit dem zentralen zentromerischen Protein Cal1 interagieren.

Table of Contents

1. Introduction

1.1 The centromere

Every living organism is made out of cells and cell division is one of the hallmarks of life, needed to reproduce and develop into complex multicellular organisms (WILSON, 1925). In eukaryotes, during each division chromosomes have to be properly replicated and distributed evenly between the newly formed daughter cells (NICKLAS, 1997). Aberrant chromosome segregation or faulty regulation can result in aneuploidy or inheritance of DNA damage, which can promote tumorigenesis or cause genetic disabilities (HASSOLD & HUNT, 2001; WEAVER & CLEVELAND, 2007). The locus on each chromosome that is essential for chromosome separation is called the centromere (ALLSHIRE & KARPEN, 2008). Centromeres are located at the primary constriction of mitotic chromosomes and are the platform onto which kinetochores can form, which in turn connect to spindle microtubules (ALLSHIRE & KARPEN, 2008).

1.1.1 Centromeric chromatin – CENP-A

Centromere identity is epigenetically maintained by the incorporation of the histone H3 variant CENP-A, or CID in *Drosophila* (BLACK & BASSETT, 2008). CENP-A is conserved in the fast majority of eukaryotes, independent on whether the species has point, regional or holocentric centromeres (ALLSHIRE & KARPEN, 2008). CENP-A levels have to be tightly regulated, both reduction and overexpression are deleterious, resulting in centromere loss or formation of ectopic centromeres, respectively (BLOWER ET AL., 2006; HEUN ET AL., 2006; HOWMAN ET AL., 2000; JÄGER ET AL., 2005). At regional centromeres, CENP-A-containing nucleosomes are interspersed with H3-containing nucleosomes (BLOWER ET AL., 2002; BODOR ET AL., 2014). The centromere core is embedded in large regions of pericentromeric heterochromatin (ALLSHIRE & KARPEN, 2008). However, centromeric H3-containing nucleosomes lack the post-translational histone modifications and DNA methylation associated with this surrounding heterochromatin, such as methylation of lysine 9 of histone H3 (H3K9me) (ALTEMOSE ET AL., 2022; SULLIVAN & KARPEN, 2004). There are also clear differences from euchromatin. For example, while the typical euchromatic H3K4me2 modification is observed at centromeres, acetylation is missing (SULLIVAN & KARPEN, 2004).

Unlike canonical histones, new CENP-A is not loaded onto DNA in S-phase. During DNA replication, CENP-A-containing nucleosomes are retained randomly at one of the two sister chromatids, with the resulting gaps on the other chromatid filled by histone H3 variants H3.3 and H3.1 (DUNLEAVY ET AL., 2011). CENP-A levels are replenished during or just after mitosis depending on the species or cell type (JANSEN ET AL., 2007; MELLONE ET AL., 2011; SCHUH ET AL., 2007). For example, in human cells, new CENP-A

is targeted to centromeres in telophase and properly incorporated in early G1 (JANSEN ET AL., 2007), whereas loading of new CID in *Drosophila* S2 cells occurs during metaphase (MELLONE ET AL., 2011). By contrast, CID replenishment occurs during anaphase in early *Drosophila* embryogenesis (SCHUH ET AL., 2007).

CENP-A/CID loading is mediated by its dedicated chaperone, called HJURP in human (DUNLEAVY ET AL., 2009; FOLTZ ET AL., 2009) and Cal1 in flies (CHEN ET AL., 2014; ERHARDT ET AL., 2008). HJURP binds both free CENP-A and the Mis18 complex (FUJITA ET AL., 2007). In anaphase, the Mis18 complex is recruited to centromeres by interacting with the constitutive inner-kinetochore protein CENP-C (MOREE ET AL., 2011). Untimely localization of the Mis18 complex and HJURP to centromeres is inhibited by phosphorylation by CDK1/2 (SILVA ET AL., 2012; STANKOVIC ET AL., 2017).

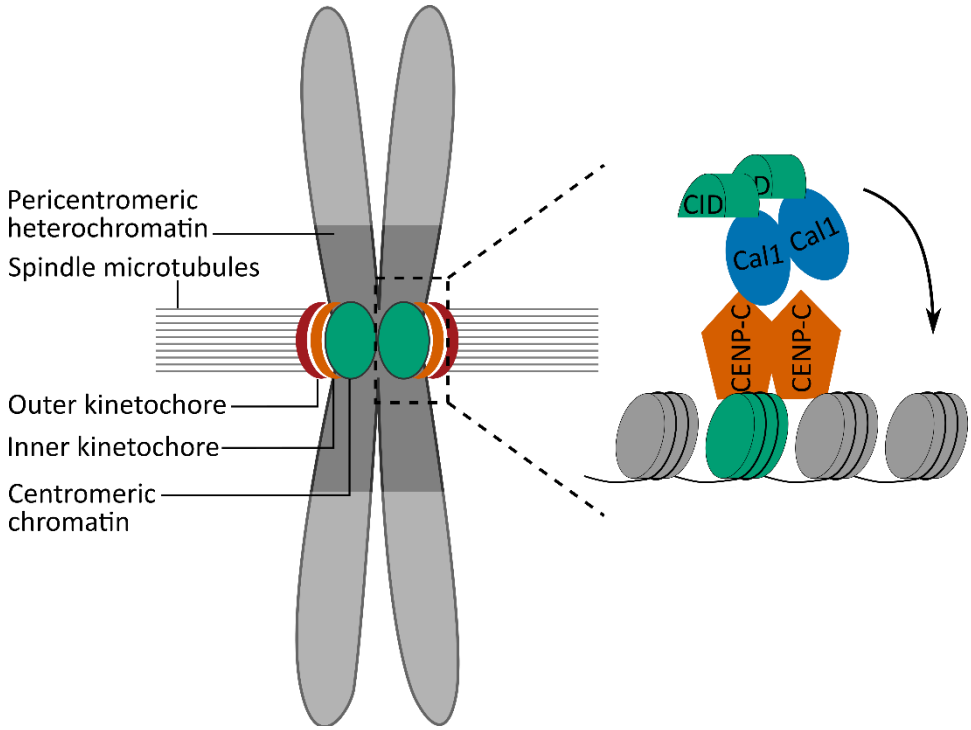


Figure 1. Centromere loading in *Drosophila*.
Centromeres are located at the primary constriction of mitotic chromosomes, embedded in pericentromeric heterochromatin. Inner and subsequently outer kinetochore proteins assemble onto the centromeres, allowing spindle microtubules to attach. Histone H3 variant CID epigenetically defines centromeres and is loaded during mitosis by its chaperone Cal1 (arrow). A centromere-associated CENP-C dimer recruits a Cal1 dimer, both already bound with CID. Inspired by Allshire & Karpen (2008) and Roure at al. (2019).

The *Drosophila* CID chaperone Cal1 performs the tasks of both HJURP and the Mis18 complex (Figure 1) (PHANSALKAR ET AL., 2012). The N-terminal part of Cal1 binds soluble CID/H4 dimers, while the C-terminal part of Cal1 interacts with CENP-C, targeting CID to the centromere (PHANSALKAR ET AL., 2012;

SCHITTENHELM ET AL., 2010). CENP-C dimerization through its cupin domain stabilizes the interaction of CENP-C with one Cal1 molecule (MEDINA-PRITCHARD ET AL., 2020; ROURE ET AL., 2019). In turn, Cal1 forms a homodimer to ensure two CID/H4 dimers are available for the newly formed CID-containing histone octamer (ROURE ET AL., 2019). CID, Cal1 and CENP-C are mutually dependent for their centromeric localization and are essential for CID inheritance (ERHARDT ET AL., 2008; ROURE ET AL., 2019). Although Cal1 is present at centromeres throughout the cell cycle, new Cal1 is recruited to centromeres during prophase, likely regulating the timing of CID loading (MELLONE ET AL., 2011; SCHITTENHELM ET AL., 2010).

1.1.2 The kinetochore

A large structure called the kinetochore assembles at centromeres to facilitate microtubule attachment (ALLSHIRE & KARPEN, 2008). This structure is subdivided in the inner and outer kinetochore (Figure 1) (MUSACCHIO & DESAI, 2017). As the name suggests, the inner kinetochore is located closest to the centromere and its proteins remain associated with centromeric chromatin throughout the cell cycle (MUSACCHIO & DESAI, 2017). In human, this Constitutive Centromere Associated Network (CCAN) is composed of 16 proteins, several of which can either bind DNA directly or interact with CENP-A (GASCOIGNE & CHEESEMAN, 2011). While most of the inner kinetochore proteins are conserved in yeast, all but one of the CCAN proteins, CENP-C, are lost in *Drosophila* (DRINNENBERG ET AL., 2016; PRZEWLOKA ET AL., 2011). The inner kinetochore serves as a stable platform linking the centromere to the dynamically recruited outer kinetochore (MUSACCHIO & DESAI, 2017; PRZEWLOKA ET AL., 2011).

The outer kinetochore assembles during mitosis and is responsible for binding the spindle microtubules (GASCOIGNE & CHEESEMAN, 2011). The main structural components of the outer kinetochore are composed of the Knl1/Spc105 complex, the Ndc80 complex and the Mis12 complex, also known as the KNM network (MUSACCHIO & DESAI, 2017). Unlike the inner kinetochore, most of the KMN network proteins are conserved in *Drosophila* (DRINNENBERG ET AL., 2016).

Besides these structural components, several protein complexes localize to the kinetochore to monitor proper microtubule attachments and regulate mitotic progression (CHEESEMAN, 2014). The former is regulated by the chromosomal passenger complex (CPC), of which Aurora B kinase targets outer kinetochore proteins in absence of the tension caused by proper microtubule attachments (CHEESEMAN, 2014). The three other CPC components, INCENP, Borealin and Survivin, function as scaffolds and activators of the complex (CARMENA ET AL., 2012).

Without proper kinetochore assembly and microtubule attachments, the spindle assembly checkpoint (SAC) prevents progression into anaphase (MUSACCHIO & SALMON, 2007). The mitotic checkpoint complex (MCC) is activated at unattached kinetochores to sequester Cdc20 (SUDAKIN ET AL., 2001). Once the SAC is satisfied, Cdc20 is released to activate the ubiquitin ligase anaphase-promoting

complex/cyclosome (APC/C) (SUDAKIN ET AL., 2001). The APC/C ubiquitinates cyclin B and securin, targeting these proteins for degradation to initiate anaphase (PETERS, 2006). With the degradation of securin, seperase becomes active and cleaves centromeric cohesion, which up to that point had been protected by Sugosin1 and the CPC, to release the cohesion between the sister chromatids (HENGEVELD ET AL., 2017). In *Drosophila* S2 cells, CID chaperone Cal1 interacts with SAC component Zw10, suggesting a link between CID loading and the mitotic checkpoint (PAULEAU ET AL., 2019). Indeed, there is a link between the amount of CID loaded at centromeres and duration of mitosis, the more CID is loaded the faster anaphase is initiated (PAULEAU ET AL., 2019).

1.1.3 (Peri)centromeric DNA

As previously stated, centromeres are epigenetically defined by the incorporation of CENP-A and are thus independent of the underlying DNA sequence. The lack of DNA sequence specificity or homology between species is surprising, considering that centromeres are essential and conserved genomic features (DEBOSE-SCARLETT & SULLIVAN, 2021). Even within the same species, centromeres from different chromosomes or individuals have different DNA sequence compositions (ALTEMOSE ET AL., 2022; CHANG ET AL., 2019; HARTLEY & O'NEILL, 2019). Still, centromeres of the vast majority of eukaryotes are localized on repetitive DNA, composed of satellite repeats and transposable elements (TEs) (FUKAGAWA & EARNSHAW, 2014; HARTLEY & O'NEILL, 2019). Neocentromere formation, a process of *de novo* centromere establishment in absence of the original CENP-A locus which can occur naturally or be induced experimentally, has been observed mostly at repetitive-rich pericentromeres and telomeres (DEBOSE-SCARLETT & SULLIVAN, 2021). However, functional neocentromere formation and maintenance has also been reported on non-repetitive DNA (ALONSO ET AL., 2010; LOGSDON ET AL., 2019; MURILLO-PINEDA ET AL., 2021; SHANG ET AL., 2013). These studies on neocentromeres suggest centromeres are preferentially, but not exclusively, formed at repetitive sequences. It is still unknown how the location for *de novo* centromere formation is selected.

An alternative explanation for the co-occurrence between centromeres and repeats involves the specific evolutionary forces that are at play at centromeric repeats. Since centromere function is independent from its DNA sequence, rearrangements can introduce new satellite repeats and TEs can insert into the core centromere and mutate without a loss of fitness (ALTEMOSE ET AL., 2022; HARTLEY & O'NEILL, 2019). These newly introduced centromeric repeats can homogenize or expand, for example through replication slipping or non-homologous or unequal crossing over, pushing older repeats toward the pericentromere (ALTEMOSE ET AL., 2022; HARTLEY & O'NEILL, 2019). Repeat expansion can further be enhanced by a process called centromere or meiotic drive, during which centromeric DNA and CENP-A (loading) co-evolve in a molecular arms race (ROSIN & MELLONE, 2017). During female meiosis I, only one of the homologous chromosomes is retained in the egg and inherited, while the

other ends up in the polar body or disintegrates (ROSIN & MELLONE, 2017). Selfishly expanded centromeric DNA regions are able to attract more kinetochore proteins and thus more microtubules, leading to preferential pulling of their homolog towards the egg compared to chromosome with the "weaker" centromere (IWATA-OTSUBO ET AL., 2017; ROSIN & MELLONE, 2017). As this asymmetry can be detrimental for proper alignment and segregation during male meiosis, CENP-A or its loading is also rapidly evolving to restore the balance, although the underlying mechanism remains unclear (ROSIN & MELLONE, 2017). Evidence for rapid co-evolution of the CENP-A loading machinery to counter the meiotic drive has been observed in *Drosophila*, where CID and Cal1 of related species, like *D. melanogaster* and *D. bipectinata,* which diverged 12 million years ago, are incompatible (ROSIN & MELLONE, 2017).

The repetitive nature of (peri)centromeres made the genomic analysis of these regions difficult. Only in the recent years, with the rise of long-read sequencing techniques, are the (peri)centromeric sequences starting to get resolved (ALTEMOSE ET AL., 2022; CHANG ET AL., 2019; JAIN ET AL., 2018). The first complete and gapless reference genome assembly for human was only published this year by the Telomere-to-Telomere consortium (NURK ET AL., 2022). However, it has been known for decades that the human centromeric DNA is composed of large arrays of AT-rich alpha satellite repeats (ALTEMOSE ET AL., 2022; WILLARD, 1985). An alpha satellite monomer is approximately 171 bp and individual monomers share between 50-80 % sequence identity with one another, which can be grouped in different alpha satellite subtypes (HARTLEY & O'NEILL, 2019). At centromeres, specific combinations of alpha satellite subtypes are arranged in head-to-tail orientation in chromosome-specific higher-order repeats (HORs) (MCNULTY ET AL., 2017). These HORs are repeated many times with nearly 99 % sequence identity in HOR arrays spanning megabases (ALTEMOSE ET AL., 2022).

A 17 bp motif within roughly half of the centromeric alpha satellites, called the CENP-B box, can be directly bound by inner-kinetochore protein CENP-B (GAMBA & FACHINETTI, 2020). CENP-B is the only sequence-specific DNA binding protein of the CCAN, and has an important role in *de novo* CENP-A loading (GAMBA & FACHINETTI, 2020; OHZEKI ET AL., 2019). However, the absence of CENP-B boxes, and therefore CENP-B, at the Y chromosome suggests that CENP-B is expendable for centromere function (GAMBA & FACHINETTI, 2020).

At human pericentromeres, the HOR structure is lost and alpha satellites are organized in smaller arrays of more diverse monomer subtypes (ALTEMOSE ET AL., 2022). There are also other satellite repeat families, such as HSat1, HSat2, HSat3 and βSat, and ribosomal DNA repeats present at pericentromeres (ALTEMOSE ET AL., 2022). At both centromeres and pericentromeres, TEs are found inserted in the repeat arrays (HARTLEY & O'NEILL, 2019). Surprisingly, while the few TEs present at centromere cores are

predominantly active, TEs present at pericentromeric regions, which contain many more TEs, are often inactivated TEs or fragments of ancient TEs (ALTEMOSE ET AL., 2022; HARTLEY & O'NEILL, 2019). TE accumulation at the (peri)centromere can be explained by it being a safe insertion site for TEs, since they are less likely to disrupt gene expression in this relative gene desert (KLEIN & O'NEILL, 2018). Furthermore, homologous recombination is suppressed at centromeres, protecting TEs from mutations that will render them immobile (KLEIN & O'NEILL, 2018).

1.1.4 The (peri)centromeric DNA of *Drosophila melanogaster*

Although a complete and gapless genome reference for *Drosophila melanogaster* is still unavailable, a joint effort of the Mellone and Larracuente labs recently provided the sequence contigs of all core centromeres (CHANG ET AL., 2019). In *Drosophila* embryos, they combined long read sequencing with chromatin immunoprecipitation (ChIP) of CID-containing chromatin to assemble a centromeric contig for each chromosome. Unlike most mammalian centromeres, which mainly consist of satellite repeats (HARTLEY & O'NEILL, 2019), *Drosophila* centromeres are located on islands of TE-rich DNA (Figure 2) (CHANG ET AL., 2019). Each chromosome has a unique combination of TEs and a variable CID-containing chromatin domain size. The only TE present in all centromeric contigs is a non-LTR long interspersed nuclear element (LINE) called G2/Jockey-3, making it the most enriched element in the centromeric chromatin (CHANG ET AL., 2019). The number of copies of G2/Jockey-3 varies between the different centromeres, with only 2 truncated fragments at the centromere of chromosome 2 and many dozens full-length and truncated copies at the centromere of chromosome Y. Other highly enriched TEs are Doc and Doc2, also LINE elements, present on centromeres of chromosome X and 4. Surprisingly, the centromere of chromosome 3 contains 240 copies of the intergenic spacer (IGS) of rDNA repeats, the non-transcribed region containing promotor sequences usually located between the ribosomal genes of different rDNA repeats (CHANG ET AL., 2019; GRIMALDI & DI NOCERA, 1988). It is noteworthy that none of these repetitive elements are unique to centromeres, they are all also found at other genomic loci. Simple satellite repeats are found at the periphery of the CID-containing chromatin domains. Centromeres of chromosomes X and 2 contain the AAGAG simple repeat, the centromere of chromosome 4 contains the AAGAT simple repeat and the centromere of chromosome 3 contains prodsat (AATAACATAG) and dodeca ((C)GGTCCCGTACT) (CHANG ET AL., 2019; SHATSKIKH ET AL., 2020). Surprisingly, in S2 cells, centromeric ChIP identified additional repetitive elements and complex satellite repeats which are pericentromeric in embryos, suggesting centromere expansions or rearrangements in this cell line (CHANG ET AL., 2019).

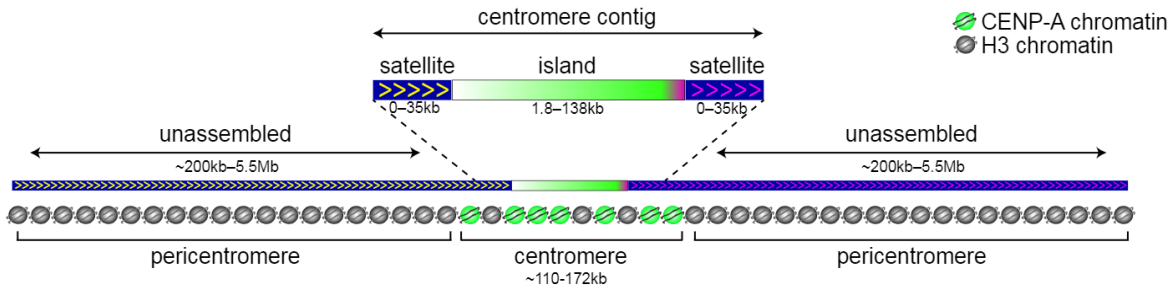


Figure 2. *Drosophila melanogaster* (peri)centromere organization.
The core centromeres of *Drosophila melanogaster* are located on islands enriched for transposable elements (TEs), with each chromosome having a unique combination of repetitive elements. The chromatin domain with CID (CENP-A) containing nucleosomes ranges in size from 110 kb to 172 kb. The surrounding pericentromeres are composed of megabases of both simple and complex satellite repeats, interspersed with TEs, but remain unassembled. From Chang et al. (2019).

The pericentromeric regions of the *Drosophila* genome remain unassembled (CHANG ET AL., 2019). Nevertheless, a rough sequence composition was established using techniques like *in situ* hybridization (LOHE ET AL., 1993; SHATSKIKH ET AL., 2020). Pericentromeres are composed of large regions of simple and complex satellite repeats, interspersed with TEs, ribosomal DNA and histone genes (SHATSKIKH ET AL., 2020). Roughly 20 % of the *Drosophila* genome is composed of satellite repeats, long tandem arrays stretching up to megabases in the pericentromeric heterochromatin (LOHE & BRUTLAG, 1986). There are 17 different satellite types or families, most of whom are relatively simple in sequence (SHATSKIKH ET AL., 2020). The most abundant complex satellite family is the 1.688 repeat family, covering 4 % of the *Drosophila* genome (HSIEH & BRUTLAG, 1979). This family includes the 260 bp satellite repeat, located mainly on the 2L pericentromere, the 353 bp and 356 bp satellite repeats, located mainly on the 3L pericentromere, and the 359 bp satellite repeat, also known as SATIII, located mainly on the X pericentromere (ABAD ET AL., 2000; KHOST ET AL., 2017; LOSADA & VILLASANTE, 1996). Furthermore, the latest genome reference showed the different 1.688 satellite repeat members are intermingled with one another in different combinations across different unassigned contigs (CHANG ET AL., 2019). Another complex repeat family is called Responder, located mainly on the 2R pericentromere (KHOST ET AL., 2017; LARRACUENTE, 2014). Responder is organized as a dimer of two 120 bp repeats (Left and Right), which are AT-rich and are 84 % identical in sequence (LARRACUENTE, 2014). Shorter clusters of all these complex satellite repeats are also present in the euchromatin of chromosomes X, 2 and 3 (KUHN ET AL., 2012; LARRACUENTE, 2014).

1.2 Transposable elements

1.2.1 Classes of transposable elements

It is clear that TEs are an integral part of most, if not all, eukaryotic centromeres (CHANG ET AL., 2019; KLEIN & O'NEILL, 2018). They are also highly enriched in pericentromeric heterochromatin (KAMINKER ET AL., 2002; KLEIN & O'NEILL, 2018). Transposable elements were first described in the 1940s as mobile elements in maize by Barbara McClintock (MCCLINTOCK, 1950). There are different types of TEs, classified by their transposition intermediate and mode of transposition (Figure 3) (WELLS & FESCHOTTE, 2020). The main subdivision is between Class I elements, also called retrotransposons, and Class II elements, also known as DNA transposons (WELLS & FESCHOTTE, 2020).

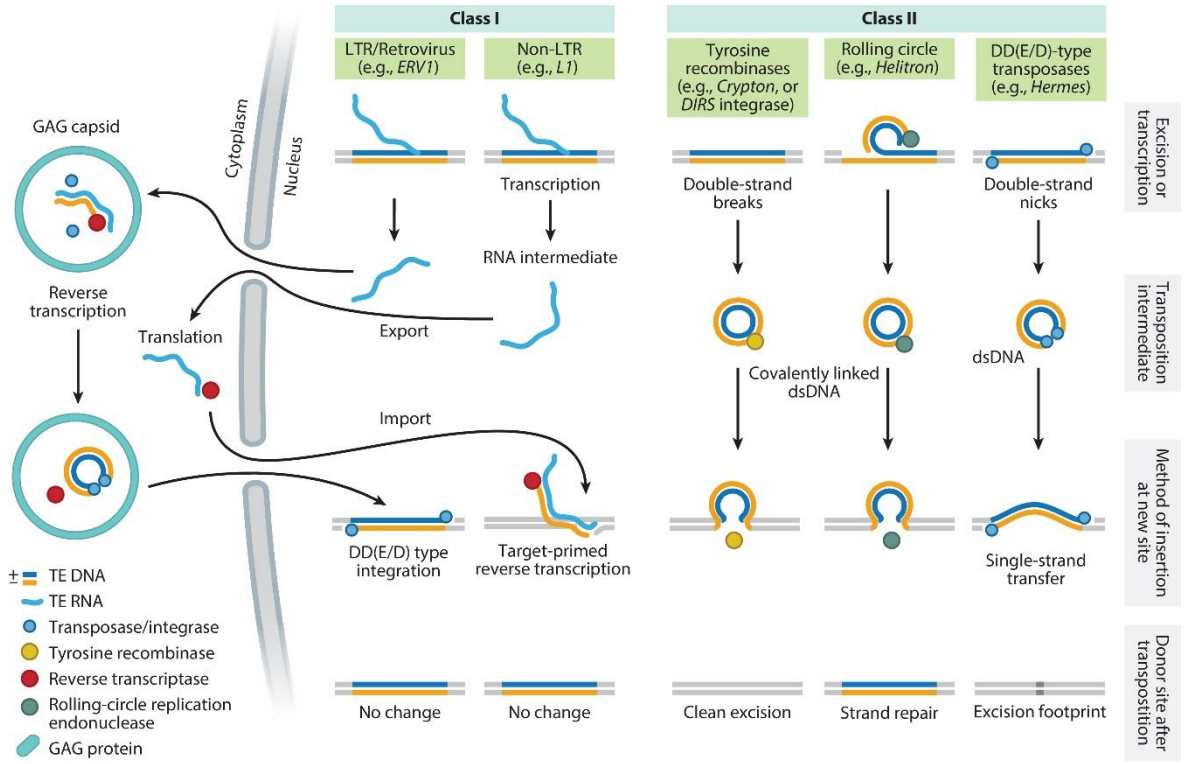


Figure 3. Transposable element classification and mode of transposition.
Transposition mechanisms of the different classes and subclasses of TEs. Transposition intermediates and key replication steps are indicated, see text for details. From Wells & Feschotte (2020). dsDNA = double-stranded DNA, LTR = long terminal repeat, TE = transposable element.

Class I TEs have an RNA intermediate which is reverse-transcribed into a cDNA copy that is integrated into the genome (WELLS & FESCHOTTE, 2020). This process is also referred to as a copy-and-paste mechanism, as the original TE copy remains intact (WELLS & FESCHOTTE, 2020). Retrotransposons are further subdivided into LTR and non-LTR subclasses, which differ in their integration method (BOURQUE ET AL., 2018; WELLS & FESCHOTTE, 2020). The *Drosophila* genome consist of roughly 12 % LTR and 5 %

non-LTR TEs (MÉREL ET AL., 2020). LTR elements are closely related to retroviruses, however, are usually missing the envelop proteins which are required for intercellular movement (WELLS & FESCHOTTE, 2020). They are called LTR elements after the two flanking long terminal repeats (LTR), which contain an RNA polymerase II promoter and a polyadenylation signal in the 5' LTR and 3' LTR, respectively (COFFIN JM ET AL., 1997; MÉREL ET AL., 2020; WELLS & FESCHOTTE, 2020). From this promoter a single polycistronic RNA is transcribed, containing the *gag* and *pol* genes (WELLS & FESCHOTTE, 2020). *Gag* encodes a structural protein which polymerizes to form a cytosolic capsid in which further processing steps occur (MÉREL ET AL., 2020; WELLS & FESCHOTTE, 2020). *Pol* encodes a polyprotein containing a protease, a reverse transcriptase, an RNaseH and an integrase, which is cleaved by the mentioned protease (WELLS & FESCHOTTE, 2020). Inside the capsid, reverse transcriptase uses a tRNA or tRNA fragment from the host as a primer for cDNA synthesis, after which the TE RNA is digested away by RNaseH (WELLS & FESCHOTTE, 2020). The reverse transcriptase also has DNA-dependent DNA polymerase activity to complete the generation of double-stranded DNA (dsDNA) (MCCULLERS & STEINIGER, 2017). Once the dsDNA is ready, its two ends are bound by integrase, which mediates the transport to the nucleus and the integration into a new location of the host genome (WELLS & FESCHOTTE, 2020). There are three superfamilies of LTR retrotransposons, namely Ty1/copia, Ty3/gypsy and endogenous retroviruses (ERV) (BOURQUE ET AL., 2018). ERV elements contain an additional gene encoding envelope proteins and is therefore able to mobilize between cells (MCCULLERS & STEINIGER, 2017). Single LTRs, lacking the internal part of the TE, are abundant in the genome as a result of recombination events (MCCULLERS & STEINIGER, 2017).

The non-LTR retrotransposons have internal RNA polymerase II promoters at their 5' end, compensating for the lack of LTRs, and at their 3' end there is an adenine-rich tail followed by a polyadenylation signal (KAZAZIAN, 2004; WELLS & FESCHOTTE, 2020). In general, non-LTR TEs contain two open reading frames (ORF) (WELLS & FESCHOTTE, 2020). While the exact function of the protein encoded by ORF1 remains unclear, it does have RNA binding and nucleic acid chaperone properties (MÉREL ET AL., 2020). ORF2 encodes the reverse transcriptase, which also has endonuclease activities (MÉREL ET AL., 2020; WELLS & FESCHOTTE, 2020). After translation, the TE RNA is bound by the ORF2 protein and transported back to the nucleus (MÉREL ET AL., 2020). At the insertion site, the endonuclease creates a single-stranded nick after which the exposed 5' end of the host DNA is hybridized with the 3' end of the TE RNA and used to initiate reverse transcription (MÉREL ET AL., 2020; WELLS & FESCHOTTE, 2020). This process is called target-primed reverse transcription, creating the cDNA at the site of insertion (WELLS & FESCHOTTE, 2020). 5'-truncated copies are frequent in non-LTR TEs due to early termination of target-primed reverse transcription (MÉREL ET AL., 2020). Such truncations usually inactivate the element due to the loss of its RNA polymerase II promoter (WELLS & FESCHOTTE, 2020). Non-LTR superfamilies include long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs) and Penelope-

like elements (BOURQUE ET AL., 2018). SINEs are non-autonomous TEs, as they require proteins from other TEs for their transposition (WELLS & FESCHOTTE, 2020).

Class II TEs mobilize with a DNA intermediate (WELLS & FESCHOTTE, 2020). DNA transposons are grouped in 4 subclasses, namely the DDE transposases, the Tyrosine recombinases (also known as Cryptons), the helitrons (also known as rolling-circle elements), and the Mavericks/Polintons (BOURQUE ET AL., 2018; WELLS & FESCHOTTE, 2020). DDE transposases and Tyrosine recombinases make use of the cut-and-paste mechanism of transposition, where the TE is removed from its original location and inserted into a new location (WELLS & FESCHOTTE, 2020). Both these elements have a single ORF encoding a recombinase, a DDE transposase or a Tyrosine recombinase, respectively, flanked by short terminal inverted repeats (TIRs) (WELLS & FESCHOTTE, 2020). These transposases or recombinases have to be transcribed and translated, before they can return to the TE in the DNA and bind the two TIRs (MCCULLERS & STEINIGER, 2017; WELLS & FESCHOTTE, 2020). The two bound enzymes dimerize to bring the two ends of the TE together triggering its excision from the genome (MCCULLERS & STEINIGER, 2017; WELLS & FESCHOTTE, 2020). Subsequently the TE is inserted into a new genomic location (WELLS & FESCHOTTE, 2020). While most of these cut-and-paste elements have lost their ability to mobilize autonomously in mammals, *Drosophila* still has several active and autonomous DDE TEs (1.5 % of the genome), of which the P elements are the most well-known (MCCULLERS & STEINIGER, 2017; MÉREL ET AL., 2020). Helitrons use the so-called peel-and-paste mechanism of transposition, one of the DNA strands is peeled away into a circular DNA and used as a template to create a new complementary DNA strand (BOURQUE ET AL., 2018; WELLS & FESCHOTTE, 2020). This circular dsDNA is moved and inserted into a new genomic location, while the original complementary strand stays to be repaired (WELLS & FESCHOTTE, 2020). Helitrons do not have flanking TIRs and encode for a protein with a DNA helicase and a replicator domain (MCCULLERS & STEINIGER, 2017; MÉREL ET AL., 2020). In *Drosophila*, only truncated non-autonomous copies of helitrons have been identified, surprisingly still occupying 0.8 % of the genome (MÉREL ET AL., 2020). The last subclass of DNA TEs, Mavericks/Polintons, are still poorly understood (WELLS & FESCHOTTE, 2020). They are remarkably large, 15-20 kb, encode for up to twenty proteins and are flanked by long TIRs (WELLS & FESCHOTTE, 2020). Most eukaryotes only have a few copies of Mavericks/Polintons TEs incorporated into their genomes (WELLS & FESCHOTTE, 2020).

1.2.2 Transposable elements impact the genome

Only a small percentage of TEs inserted in the genome retains their ability to mobilize (BOURQUE ET AL., 2018). Either due to truncations that occur during transposition, recombination events or natural accumulation of mutations, essential parts of the TE are lost or lose their functionality (BOURQUE ET AL., 2018; KLEIN & O'NEILL, 2018). However, TE activity from this small percentage can have a large impact, mutating genes or altering gene expression depending on where the insertion occurs (BOURQUE ET AL.,

2018; HANCKS & KAZAZIAN, 2016). Gene disruptions due to insertions into gene exons is the most obvious cause of mutations, called insertional mutagenesis (HANCKS & KAZAZIAN, 2016; KLEIN & O'NEILL, 2018). However, also intronic insertions can disrupt genes by inducing alternative splicing or alternative polyadenylation (BOURQUE ET AL., 2018; KLEIN & O'NEILL, 2018). Even when inserted upstream of genes, TEs can act as enhancers or alternative promoters, altering gene expression levels (BOURQUE ET AL., 2018; KLEIN & O'NEILL, 2018). Alternatively, epigenetic repression of the TE can result in reduced expression of the surrounding genes (KLEIN & O'NEILL, 2018). Last, due to the presence of TE copies on different chromosomes, TEs can induce non-allelic homologous recombination, resulting in deletions, inversions and segmental duplications (KLEIN & O'NEILL, 2018).

Several diseases are known to be caused or aggravated by TE insertions. In human, LINE-1 insertions alone are linked to at least 124 heritable genetic diseases (HANCKS & KAZAZIAN, 2016). Recent advances in sequencing of patient samples uncovered links of TE activity with cancer, autoimmune diseases and neuropsychiatric disorders (AHMADI ET AL., 2020; HANCKS & KAZAZIAN, 2016). For example, a *de novo* TE insertion in tumor suppressor retinoblastoma 1 was reported to cause retinoblastoma (RODRÍGUEZ-MARTÍN ET AL., 2016). A general derepression of TEs that is observed in several cancers is linked to genomic instability and aggravate tumor progression (ANWAR ET AL., 2017).

However, not all TE insertions are detrimental. During evolution, TEs have also been domesticated into new non-coding and coding genes (BOURQUE ET AL., 2018; KAESSMANN ET AL., 2009). A relevant example of a TE derived proteins is the mammalian inner kinetochore protein CENP-B, which has evolved from a DNA transposases of the pogo-like family (GAMBA & FACHINETTI, 2020). In *Drosophila*, three Jockey TEs of the LINE superfamily, HeT-A, TART and TAHRE, are responsible for telomere maintenance by specifically inserting at chromosome ends (MCCULLERS & STEINIGER, 2017). Furthermore it is suggested that TEs can be incorporated into the cis-regulatory network (BOURQUE ET AL., 2018). For example, TE sequences can contribute new enhancers for neighboring genes, as they are able to recruit transcription factors for their own expression (BOURQUE ET AL., 2018).

1.2.3 Transposable element repression

To prevent the potentially mutagenic effects of active transposition, all eukaryotic host genomes have evolved several different defense mechanisms (SLOTKIN & MARTIENSSEN, 2007). In *Drosophila*, two pathways involving different types of TE-derived small RNAs are responsible for silencing TEs and inducing transcriptionally repressive heterochromatin formation (MTEIREK ET AL., 2014). First, the RNAi pathway, which is active in somatic cells, makes use of endogenous small interfering RNAs (endo-siRNAs) (MTEIREK ET AL., 2014). Low levels of convergent antisense TE transcripts are produced by RNA polymerase II, which interact with regular sense TE transcripts to form double-stranded RNA (dsRNA)

(RUSSO ET AL., 2016). These dsRNAs are cleaved by Dicer-2 into endo-siRNAs, which together with Argonaute-2 assemble into RNA-induced silencing complex (RISC) (MTEIREK ET AL., 2014; SLOTKIN & MARTIENSSEN, 2007). RISC binds already transcribed TEs that are complementary to its loaded endo-siRNA and targets them for degradation (SLOTKIN & MARTIENSSEN, 2007).

The second pathway, the piRNA pathway, involves Piwi-interacting RNAs (piRNAs) and silences TEs in the germline, where suppressing the mutagenic effects of TEs is essential to protect the offspring (OZATA ET AL., 2019). Most piRNAs are derived from long RNAs transcribed from piRNA clusters, containing fragments of a wide variety of TEs (GUZZARDO ET AL., 2013; OZATA ET AL., 2019). Three members of the Argonaute family, PIWI, Aubergine and Argonaute-3, are able to process piRNA precursor transcripts into mature piRNAs (GUZZARDO ET AL., 2013; OZATA ET AL., 2019). Subsequently, piRNA-loaded Aubergine and Argonaute-3 are able to covalently target already transcribed TEs in the cytoplasm and neutralize these TEs by processing them into new piRNAs, which is called the ping-pong cycle (GUZZARDO ET AL., 2013). PIWI binds mature piRNAs and is transported to the nucleus where it can target complementary nascent TE RNAs (BATKI ET AL., 2019; OZATA ET AL., 2019). There, the piRNA pathway can induce transcriptional silencing and heterochromatin formation by recruiting H3K9 methyltransferase Eggless and heterochromatin protein 1 (HP1) (BATKI ET AL., 2019; LE THOMAS ET AL., 2013; OZATA ET AL., 2019; SIENSKI ET AL., 2015; Y. YU ET AL., 2015).

Components of the piRNA pathway are also maternally deposited into *Drosophila* embryos to inhibit TE transposition during early development (FABRY ET AL., 2021). During the first hours of embryogenesis, rapid syncytial nuclear divisions occur without zygotic transcription, thus completely relying on maternally deposited transcripts (SELLER ET AL., 2019). Heterochromatin marks are also absent in these nuclei (FABRY ET AL., 2021; SELLER ET AL., 2019). In stage 5 embryos, zygotic transcription is activated, which necessitates heterochromatin formation to suppress TE expression (FABRY ET AL., 2021; SELLER ET AL., 2019). The maternally deposited piRNA pathway and Eggless are required for this heterochromatin establishment in the somatic cells of the embryo (FABRY ET AL., 2021; SELLER ET AL., 2019). Zygotic transcripts subsequently induce cellularization and gastrulation, and somatic heterochromatin is maintained through a positive feedback loop between H3K9 methyltransferase SU(VAR)3-9 and HP1 (MTEIREK ET AL., 2014; SELLER ET AL., 2019).

1.3 Centromeric RNA

Despite the transcriptionally repressive chromatin environment, the pericentromeres are transcriptionally active, with the transcripts playing a crucial role in heterochromatin formation

(SLOTKIN & MARTIENSSEN, 2007). Next to the pericentromeric TEs, also pericentromeric satellites are transcribed and important for heterochromatin formation in many species (ARUNKUMAR & MELTERS, 2020; CORLESS ET AL., 2020). For example, satellites of the 1.688 family and Responder are expressed in the *Drosophila* germline and subsequently processed by the piRNA pathway to promote heterochromatin formation at pericentromeric DNA containing these repeats (USAKIN ET AL., 2007; WEI ET AL., 2021). Likewise, centromeric transcription and the derived transcripts, cenRNAs, are a conserved component of centromeres in eukaryotes (ARUNKUMAR & MELTERS, 2020; CORLESS ET AL., 2020). Centromeric transcription by RNA polymerase II is found in all eukaryotic species examined, among others yeast, maize, fruit fly, xenopus, mouse and human (ARUNKUMAR & MELTERS, 2020). In the last two decades, centromere biologists have begun exploring the role of centromeric transcription and cenRNAs in different species.

1.3.1 Yeast

Budding yeast, which has a point centromere of just one CENP-A containing nucleosome, specific cenRNA levels are required for centromere function, as both upregulation and suppression results in mitotic defects and chromosomal instability, even at chromosomes lacking that specific repeat (LING & YUEN, 2019; OHKUNI & KITAGAWA, 2011). Higher cenRNA levels resulted in lower cellular levels of CENP-A, CENP-C, chaperone HJURP and chromosomal passenger complex (CPC) components Survivin and INCENP (LING & YUEN, 2019). Consequently, CENP-A, CENP-C and CPC component Aurora B are also reduced at centromeres (LING & YUEN, 2019). Transcription factors Ste12, Cbf1 and Dig1 localize to centromeres and promote (Ste12) and suppress (Cbf1 and Dig1) RNA polymerase II dependent transcription at centromeres (OHKUNI & KITAGAWA, 2011). Cbf1 ensures centromeric transcription is limited to S phase, it stabilizes the kinetochore and herewith blocks polymerase read-through from the surrounding pericentromeres (HEDOUIN ET AL., 2022).

Although in fission yeast CENP-A maintenance at the centromere is dependent on RNA polymerase II transcription, cenRNAs themselves seem less important (CATANIA ET AL., 2015; SHUKLA ET AL., 2018; SINGH ET AL., 2020). In G2 phase, elongating RNA polymerase II gets stalled at the centromere core, which creates a chromatin environment needed for the exchange of placeholder H3 with new CENP-A (CATANIA ET AL., 2015). Transcription elongating factor TFIIS promotes continued transcription of stalled polymerases, preventing excessive CENP-A loading. Accordingly, fission yeast cenRNAs are hardly detectable, as TFIIS cleaves nascent transcripts to resolve the stalling. Only after exosome inhibition polyadenylated cenRNAs of different lengths from both strands can be observed, suggesting transcripts are degraded shortly after transcription (CATANIA ET AL., 2015; CHOI ET AL., 2011). RNA polymerase II is still present at centromeres during mitosis, indicating a potential unexplored function of transcription besides CENP-A loading in G2 phase (SHUKLA ET AL., 2018).

1.3.2 Maize

The discovery of cenRNAs transcribed from centromeric TEs (CRMs) and satellite repeats (CentC) in maize sparked (TOPP ET AL., 2004). It was proposed that these polyadenylated transcripts are transcribed by RNA polymerase II from bidirectional promoters encoded by CRMs (TOPP ET AL., 2004). Maize cenRNAs promote the interaction between inner kinetochore protein CENP-C with centromeric DNA (DU ET AL., 2010). Centromeric CENP-A levels were reduced after knockdown of CRMs, specifically circular CRM RNAs which promote chromatin looping through the formation of R-loops, three-stranded RNA-DNA hybrids (LIU ET AL., 2020).

1.3.3 *Drosophila*

In *Drosophila*, chaperone Cal1 recruits RNA polymerase II and FACT, a factor mediating nucleosome disassembly and reassembly during transcription, to centromeres (CHEN ET AL., 2015). FACT directly interacts with Cal1 and is involved in centromeric transcription and in the replacement of the placeholder H3.3 or H3.1 with new CID during mitosis (CHEN ET AL., 2015). Cal1 still recruits new CID to centromeres independently, however, it will not be incorporated into the centromeric chromatin in absence of transcription (BOBKOV ET AL., 2018). Nascent RNA, labeled with 5-ethynyluridine (EU), is visible at (peri)centromeres during mitosis and G1 (BOBKOV ET AL., 2018). TE G2/Jockey-3, present in the centromeric DNA of all chromosomes, is expressed at low levels (CHANG ET AL., 2019). Furthermore, RNA derived from pericentromeric SATIII, the 359 bp satellite repeat from the 1.688 family, is also important for centromere functioning (ROŠIĆ ET AL., 2014). SATIII RNA is a part of centromeric chromatin and interacts with CENP-C. CENP-C knockdown reduces SATIII RNA levels, indicating it is involved in SATIII transcription or SATIII RNA stabilization. Furthermore, SATIII knockdown results in mitotic defects, such as lagging strands, in both S2 cells and embryos. Interestingly even though the major locus of SATIII DNA is located on the X chromosome, SATIII knockdown leads to lagging of all major chromosomes during mitosis. These lagging strands have reduced levels of CID, CENPC and outer kinetochore protein Spc105, and overall recruitment of new CID and CENP-C was impaired (ROŠIĆ ET AL., 2014).

1.3.4 Xenopus

Similarly, in xenopus, frog centromeric repeat 1 (Fcr1) RNA seems to interact with kinetochore proteins. CENP-C and CPC components Aurora B and Borealin pull down Fcr1 RNA (BLOWER, 2016; GRENFELL ET AL., 2016). Additionally, Aurora B was shown to directly interact with Fcr1 RNA in an electrophoretic mobility shift assay (BLOWER, 2016). Fcr1 is transcribed by RNA polymerase II, processed by the spliceosome and locates to all centromeres, even of chromosomes lacking Fcr1 DNA (BLOWER, 2016; GRENFELL ET AL., 2016). Centromeric localization and activation of Aurora B is impaired

by perturbation of the Fcr1 RNA through the inhibition of its transcription initiation and splicing, or direct knockdown (BLOWER, 2016; GRENFELL ET AL., 2016). Furthermore, centromeric levels of CENP-A, CENP-C and outer kinetochore protein NDC80 are slightly reduced after inhibition of transcription initiation and splicing (GRENFELL ET AL., 2016).

1.3.5 Mouse

Aurora B promotes centromeric transcription in mouse by phosphorylating H3K9me3 at serine 10, which creates an open chromatin state (FERRI ET AL., 2009; MALLM & RIPPE, 2015). There is a variety of centromeric minor satellite transcript sizes, ranging from 120 nt, corresponding to a single repeat length, to 4 kb, in both sense and antisense direction (BOUZINBA-SEGARD ET AL., 2006). Since Aurora B influences centromeric expression, it is not surprising that minor satellite RNA follow a similar expression pattern to CPC proteins throughout the cells cycle, with expression starting in S-phase, the highest expression levels in G2/M phase and almost no expression in G1 phase (CARMENA ET AL., 2012; FERRI ET AL., 2009). Minor satellite RNA is associated with centromeric chromatin and interacts with CPC components Aurora B and Survivin (FERRI ET AL., 2009). Minor satellite RNA is needed for Aurora B activity and its interaction with Survivin and CENP-A (FERRI ET AL., 2009; MALLM & RIPPE, 2015). Surprisingly, even Aurora B activity at telomeres requires activation by minor satellite RNA at centromeres (MALLM & RIPPE, 2015). However, minor satellite RNA levels have to be tightly regulated, as ectopic overexpression of minor satellite results in mitotic defects, impaired sister chromatid cohesion and Aurora B mislocalization (BOUZINBA-SEGARD ET AL., 2006; CHAN ET AL., 2017). Maternally deposited minor satellite RNAs are also important for proper spindle attachment during meiosis in mouse oocytes (WU ET AL., 2021).

1.3.6 Human

Also in human, cenRNA derived from alpha satellites is a part of centromeric chromatin, as it is pulled down by both CENP-A and CENP-C (MCNULTY ET AL., 2017). CENP-C can even directly interact with alpha satellite RNA *in vitro,* using the same domain that binds DNA (TALBERT & HENIKOFF, 2018; WONG ET AL., 2007). RNA polymerase II inhibition or alpha satellite knockdown results in reduced centromeric CENP-A, CENP-C and HJURP levels, impaired loading of new CENP-A and mitotic defects (BERGMANN ET AL., 2011; CHAN ET AL., 2012; CHEN ET AL., 2021; MCNULTY ET AL., 2017; MOLINA ET AL., 2016; QUÉNET & DALAL, 2014; WONG ET AL., 2007). Furthermore, even prior to loading into chromatin, still soluble HJURP/CENP-A complex binds alpha satellite RNA (QUÉNET & DALAL, 2014). Interestingly, when the knockdown targets alpha satellite RNAs from HORs of specific chromosomes, CENP-A and CENP-C levels were only reduced at centromeres of those chromosomes (MCNULTY ET AL., 2017). Accordingly, both neocentromeres and centromeres of human artificial chromosomes were reported to require transcription for centromere maintenance and function (CHUEH ET AL., 2009; MOLINA ET AL., 2016; NAKANO ET AL., 2008; NAUGHTON ET

AL., 2021; OKAMOTO ET AL., 2007). However, Chen et al. (2021) did not observe a change in centromeric CENP-A levels after transcription inhibition.

Similarly to other species, CPC proteins are found connected to cenRNAs in human (MORAN ET AL., 2021; WONG ET AL., 2007). Centromeric localization of INCENP is affected by alpha satellite RNA and Aurora B localizes to centromeric alpha satellite R-loops (MORAN ET AL., 2021; WONG ET AL., 2007). Subsequently, Aurora B aids resolving these R-loops by interacting with R-loop regulator RBMX, suggesting a feedback loop between Aurora B and cenRNA R-loops (MORAN ET AL., 2021). The CPC promotes inner centromere localization of centromere cohesion protector Sugosin1 and this localization is therefore also dependent on transcription (CHEN ET AL., 2021; LIU ET AL., 2015; MORAN ET AL., 2021). Accordingly, centromeric cohesion defects were also observed after elongating RNA polymerase II inhibition (CHEN ET AL., 2021; LIU ET AL., 2015).

Taken together, these studies suggest alpha satellite RNA interacts with proteins involved in CENP-A loading and centromeric levels of these proteins are reduced upon knockdown and transcription inhibition. Furthermore, localization to the inner centromere of CPC and Sgo1 is dependent on centromeric transcription and the resulting chromatin structures (R-loops), which could affect proper mitotic spindle attachment and centromeric cohesion.

Recently, a first transcription factor was identified to be involved in human and mouse centromeric transcription, ZFAT (ISHIKURA ET AL., 2020). Both alpha satellites and minor satellites contain an 8 bp ZFAT binding motif and altered levels of ZFAT result in an equivalent change in satellite transcription. ZFAT recruits histone acetyltransferase KAT2B to centromeres, which acetylates histone 4 at lysine 8 in order to induce BRD4-mediated RNA polymerase II transcription (ISHIKURA ET AL., 2020). Another active histone mark H3K4me2, was also shown to promote centromeric transcription in human (BERGMANN ET AL., 2011; MOLINA ET AL., 2016).

A much debated question is when centromeric transcription occurs. Actively transcribing RNA polymerase II has been observed at human centromeres during mitosis and early G1 (CHAN ET AL., 2012; QUÉNET & DALAL, 2014). Moreover, alpha satellite nascent RNA, labelled with fluorescent UTPs or EU, can be clearly detected at centromeres of mitotic chromosomes (CHAN ET AL., 2012; CHEN ET AL., 2021; PALOZOLA ET AL., 2017). In contrast, several recent papers reported a lack of centromeric RNAs in human cells during mitosis (HOYT ET AL., 2022; NAUGHTON ET AL., 2021). The Telomere-to-Telomere consortium used PRO-seq, Precision Run-On sequencing, a method for high resolution nascent RNA sequencing, to assess centromere transcription throughout the cell cycle (HOYT ET AL., 2022). In general, they found active transcription of repetitive elements is dramatically repressed during mitosis, with alpha satellite transcription at almost non-detectable levels. However, during other cell cycle phases, the highest alpha satellite nascent RNA signal was detected at active, CENP-A-containing, HOR. Using their

complete human reference, they could determine active centromeric transcription mainly occurs at the embedded TEs (HOYT ET AL., 2022). Furthermore, at neocentromeres which were positioned on unique sequences, although RNA polymerase II is present, no RNAs from these unique sequences were detected (NAUGHTON ET AL., 2021). The high resolution of these latest studies would suggest that previous observations might have originated from closely surrounding pericentromeric alpha satellite sequences. However, further investigation is needed to clarify this question.

1.3.7 cenRNAs in cancer and stress

Given that cenRNAs seem to have an important function at the centromere, it almost inevitable to see them misregulated in cancer and after cellular stress. Indeed, several tumors have elevated levels of (peri)centromeric repeats and TEs. In mouse breast, ovarian and epithelial cancers pericentromeric major and centromeric minor satellites and TE LINE-1 can be seen upregulated (TING ET AL., 2011; Q. ZHU ET AL., 2011, 2018). In human, not only alpha satellites but also the pericentromeric satellites HSat2 and HSat3 have been observed to be overexpressed in a multitude of cancers (HALL ET AL., 2017; ICHIDA ET AL., 2018; KANNE ET AL., 2021; TING ET AL., 2011; Q. ZHU ET AL., 2011, 2018). Induced overexpression of alpha satellite RNA results in mitotic defects and chromosomal instability, which can induce tumor formation or aggravate tumor progression (CHAN ET AL., 2017; ICHIDA ET AL., 2018; LANDERS ET AL., 2021; ZHU ET AL., 2011, 2018). Tumor suppressor BRCA1 localizes to centromeres by interacting with alpha satellite R-loops and is involved in resolving these R-loops together with SETX (RACCA ET AL., 2021; Q. ZHU ET AL., 2018). BRCA1 knockdown has detrimental effects on centromeres, resulting in R-loop accumulation, alpha satellite overexpression, reduced centromeric levels of CENP-A and Aurora B, and mitotic defects (RACCA ET AL., 2021; Q. ZHU ET AL., 2011). HSat2 and HSat3 are known to accumulate in nuclear stress bodies in cancer and after a wide variety of stress treatments, from heat shock to a variety of DNA damaging treatments (HALL ET AL., 2017; JOLLY ET AL., 2004; KANNE ET AL., 2021; RIZZI ET AL., 2004; VALGARDSDOTTIR ET AL., 2005, 2008). These nuclear stress bodies remain associated with the site of transcription and their size correlates to the severity of HSat overexpression, which varies depending on the type of stress treatment (JOLLY ET AL., 2004; VALGARDSDOTTIR ET AL., 2008). RNA processing factor, epigenetic factors and even cancer drugs can accumulate in HSat-containing nuclear stress bodies, sequestering them from elsewhere in the nucleus and rendering the cancer drugs ineffective (HALL ET AL., 2017; KANNE ET AL., 2021; VALGARDSDOTTIR ET AL., 2005). Also in mouse it has been observed that cellular stress, such as differentiation, DNA damage or aging, results in accumulation of minor and major satellite RNAs (BOUZINBA-SEGARD ET AL., 2006; DE CECCO ET AL., 2013; HÉDOUIN ET AL., 2017).

In summary, in all eukaryotes cenRNAs and/or RNA polymerase II transcription are involved in centromeric processes that potentially varies depending on the species examined. First, they can be required for CENP-A loading at centromeres, as shown in maize, budding and fission yeast, xenopus fruit fly and human, which is essential for maintaining centromere identity. Second, centromeric recruitment of inner and outer kinetochore proteins can be affected either by altered CENP-A levels or directly by interacting with cenRNAs, as seen in maize, budding yeast, xenopus, fruit fly and human. This affects kinetochore structure and proper spindle attachment. Third, components of the chromosomal passenger complex are found to interact with cenRNAs, affecting the proper localization and activation of CPC components, as seen in budding yeast, xenopus, mouse and human. Surprisingly, this not only affects proper microtubule attachments and inner centromere cohesion, but can also act as a feedback loop to regulate centromeric transcription. Last, stress, DNA damage and cancer have been linked to elevated cenRNA expression in mouse and human, and can potentially influence cancer severity and treatment effectiveness. It is also clear that disruption of centromeric transcription or knockdown of cenRNAs can cause mitotic defects. However, there is still uncertainty about the mechanism underlying these defects and about whether some of these observations might be secondary due to the destabilization of the region. Furthermore, there is still an active debate on whether the act of transcription alone is important for centromere biology or whether the derived transcripts also play a role in centromere functioning and maintenance. The repetitive nature of centromeric DNA and thus cenRNAs complicate analysis and perturbations to elucidate functions further, especially since these repeats are not unique to centromeres.

1.4 Aim

So far in *Drosophila*, the study of (peri)centromeric RNAs on centromere functioning has mainly focused on SATIII, while it has become clear that the *Drosophila* centromere consists of a wide variety of different repeats and TEs (CHANG ET AL., 2019). It is thus important to get a full picture of RNAs at centromeric chromatin through an unbiased study of all Drosophila cenRNAs. The identification of RNAs that localize at the centromere was the first aim of my thesis. Since stress was shown to affect cenRNA levels in mouse and human, I also examined whether the composition and levels of *Drosophila* centromere-associated RNAs change after DNA damage.

Given the clear effect of transcription and/or cenRNAs on CID maintenance in *Drosophila*, it is also important to further elucidate the role of cenRNAs in CID loading. The three main factors involved in this process are CID, Cal1 and CENP-C. In human the homologs of these proteins were show to directly or indirectly interact with cenRNAs (MCNULTY ET AL., 2017; QUÉNET & DALAL, 2014; WONG ET AL., 2007). Likewise, in *Drosophila*, CENP-C pulls down SATIII RNA in a native immunoprecipitation (ROŠIĆ ET AL.,

2014). Both Cal1 and CENP-C have large unstructured domains, called intrinsically disordered regions, which often facilitate interactions between RNA-binding proteins and RNA (CALABRETTA & RICHARD, 2015; MEDINA-PRITCHARD ET AL., 2020). Therefore, my second aim was to reexamine the direct interaction of the CID loading factors and RNA, and to identify the specific RNA components that are bound. Together, my thesis aims to further our understanding of the functional relevance of cenRNAs in *Drosophila* centromere function.

2. Results

2.1 Identification of centromere-associated RNAs

Despite being embedded in large regions of heterochromatin, centromeres are actively transcribed. In *Drosophila melanogaster*, SATIII RNA has been shown to localize to the centromere and have a role in centromere maintenance (ROŠIĆ ET AL., 2014). Furthermore, low amounts of centromeric TE transcripts were detected in RNA isolated from whole embryos (CHANG ET AL., 2019). However, the full range of RNAs localized at the centromeric region is unknown. There are several other repetitive elements in the (peri)centromeric DNA that could be transcribed and/or localize to the centromere. RNAs may be recruited to centromeres from more distant regions, in fact from any region of the genome. To determine which RNAs are localized at centromeric chromatin, I carried out an RNA Chromatin Immunoprecipitation (RNA-ChIP) followed by deep sequencing. In order to assess whether there are any DNA damage-specific centromere-associated RNAs, I performed RNA-ChIP with chromatin from both untreated embryos and embryos that have been exposed to γ-irradiation to induce DNA damage in a controlled manner.

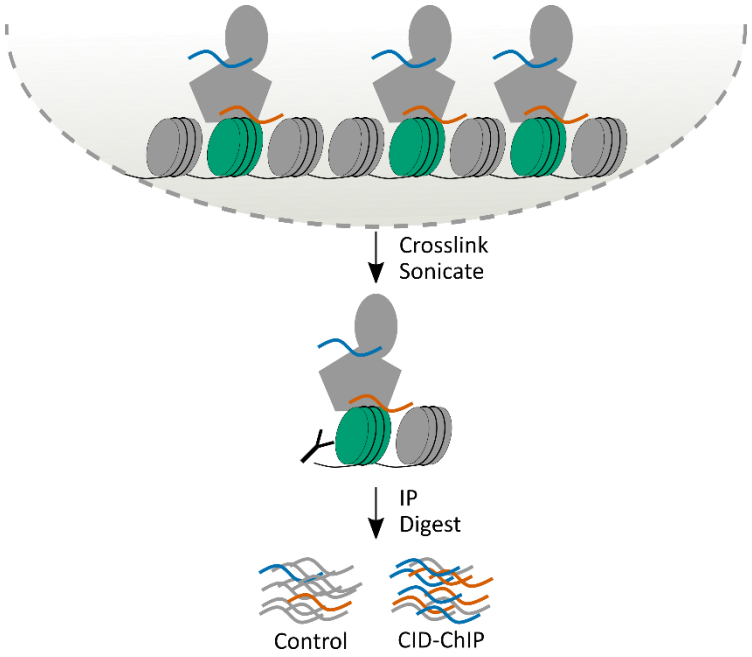


Figure 4. Schematic overview of the RNA-ChIP experiment.
Nuclei from formaldehyde crosslinked embryos were isolated and sonicated to fragment the chromatin. CID-containing nucleosomes (green) and the crosslinked environment were immunoprecipitated (IP), after which DNA and protein was digested away. Isolated RNA from this IP should be enriched for centromere-associated RNAs (blue and orange).

2.1.1 Embryo CID RNA-ChIP method optimization

To develop the embryo RNA-ChIP protocol (Figure 4), I combined and adapted published protocols for drosophila embryo DNA-ChIP (DOMSCH ET AL., 2019; SANDMANN ET AL., 2007) and human cell RNA-ChIP (SUN ET AL., 2006). In order to identify centromeric RNA in Drosophila embryos, I collected OregonR embryos 2-8 hours after egg deposition (AED) and fixed them with formaldehyde, which creates reversible cross-links between macromolecules that are within a 2 Å distance of each other (protein-protein, protein-DNA and protein-RNA)(HOFFMAN ET AL., 2015). Nuclei were then isolated and the chromatin was fragmented to a size of about 300 bp by sonication. Next, I enriched for centromeric chromatin by immunoprecipitation (IP) before digesting away DNA and proteins and reversing the remaining cross-links with heat. The subsequent isolated RNA is presumably specifically enriched for RNAs associated with centromeres.

Several steps had to be optimized in order to obtain a specific pull down while maintaining RNA integrity, for instance reducing the duration of the reverse cross-linking heating step. I also had to ensure that the CID antibody used for the ChIP is able to recognize formaldehyde cross-linked epitopes. For this purpose, I fixed OregonR embryos collected 2-3 hour AED and performed an immuno-staining with this CID antibody and a CENP-C antibody to independently label inner-kinetochores (Figure 5A). The strong overlap of the signals of both antibodies suggests that the CID antibody recognizes the desired epitopes, even though it also produced some weak off-target signals both inside and outside the nuclei of formaldehyde fixed embryos.

To isolate the nuclei, the fixed embryos had to be dissociated and fractioned using a dounce homogenizer. To control this step, I performed western blotting with antibodies against the nuclear proteins H3 and Lamin, which were, indeed, present in the nuclear and absent in the cytosolic fraction (Figure 5B). As expected, Actin is present in both the nuclear and cytosolic fraction. Subsequently, the chromatin of these nuclei was sonicated, which had to be optimized as well to obtain chromatin fragments of approximately 300 bp (Figure 5C). Exposing lysed nuclei to 15 cycles of 30 seconds on and 30 seconds off high intensity sonication resulted in the desired fragment size.

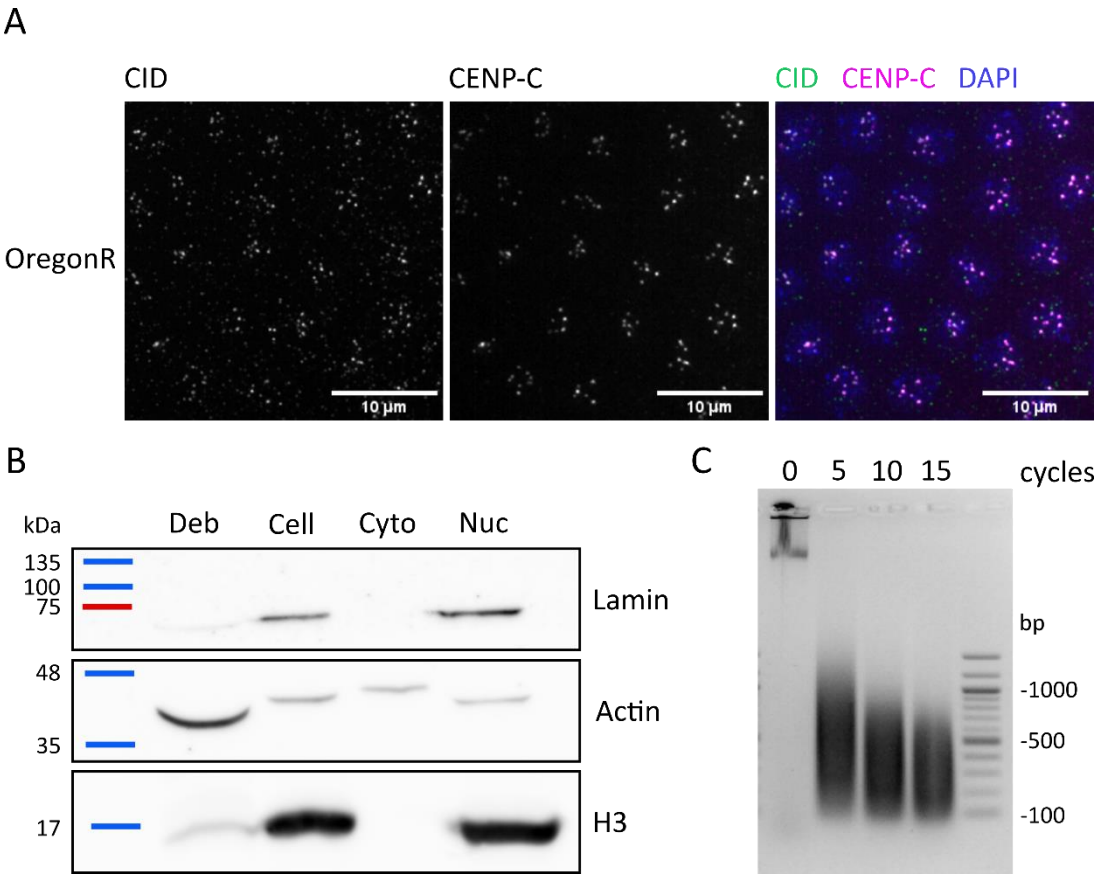


Figure 5. Centromeric RNA-ChIP-seq in drosophila embryos method optimization.

A) Specificity of the CID antibody used for RNA-ChIP in embryos confirmed with IF in fixed stage 5 OregonR embryos together with CENP-C antibody. Images are of a single z-slice. Composite image shows α-CID in green, α-CENP-C in magenta and DAPI in blue. Scale bar = 10 µm.

B) Successful retention of Lamin and H3 in the nuclear fraction was evaluated with WB after fractionation of 2-8 h OregonR embryos with the douncer. Deb = Embryo debris, Cell = Whole dissociated cells, Cyto = Cytoplasmic fraction, Nuc = nuclear fraction.

C) Chromatin fragment sizes after 0, 5, 10 and 15 cycles of shearing 30 seconds on and 30 seconds off on high intensity with the Bioruptor were assessed on a 1 % agarose gel.

Next, I had to establish the appropriate γ-irradiation dose which induces DNA damage in embryos, without being lethal. This will activate the cellular DNA damage response and potentially alter the composition of centromere-associated RNAs. As can be seen in Figure 6A, there is a dose-dependent increase of the DNA damage chromatin marker γH2Av in γ-irradiated embryos (ISMAIL & HENDZEL, 2008). Even without DNA damage induction there is always a low level of this marker present in the cells, but it increases 3.6-fold with 2 Gray, 4.5-fold with 5 Gray and up to 11.5-fold with 10 Gray γ-irradiation. The majority of embryos were able to recover from milder DNA damage, with 85 % of embryos hatching after exposure to 2 Gray γ-irradiation and 57 % hatching after 5 Gray γ-irradiation (Figure 6B). However, only 8 % were able to survive exposure to 10 Gray γ-irradiation, implying that the inflicted

damage was too severe to be repaired. Therefore, I chose to use 5 Gray γ-irradiation on 1-7 hour-old embryos. Embryos were aged for another hour after DNA damage induction, to give cells time to respond to the damage and potential DNA damage-specific centromere-associated RNAs time to accumulate.

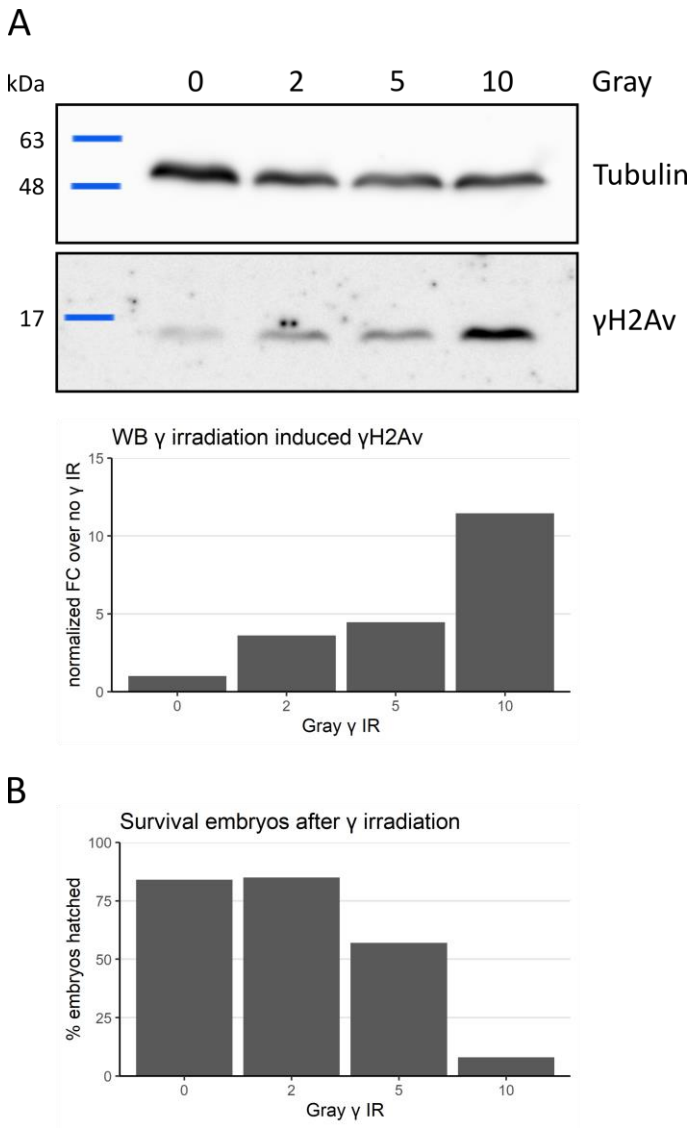


Figure 6. DNA damage titration in OregonR embryos.
1-7 h OregonR embryos were exposed to increasing intensities of γ irradiation and aged for another hour.

A) The DNA damage response was evaluated by WB for γH2Av in whole embryo lysate. The increase of γH2Av expression was quantified and normalized against Tubulin.

B) Percentage of embryos that survive and hatch into larvae after increasing intensities of γ irradiation. IR = irradiation.

2.1.2 Embryo CID RNA-ChIP sequencing

I performed the RNA-ChIP for CID together with two control pull downs. To be able to exclude general chromatin-associated RNAs, an H3 RNA-ChIP was used as the first control. The second control was a mock RNA-ChIP without any added antibodies, which allows the identification of non-specifically bound RNAs. Furthermore, I included the input samples. The fragment size of the chromatin used for all four RNA-ChIP replicas was checked on an agarose gel (Supplementary Figure 1D).

For all four replicas, half of each sample was used to assess the efficiency of the pull down with western blot (Figure 7 and Supplementary Figure 1A-C). It is apparent that CID is strongly enriched in the CID

RNA-ChIP of both untreated control and irradiated samples. Unsurprisingly, H3 is also pulled down with the CID RNA-ChIP, since neighboring nucleosomes of the CID containing nucleosomes can contain H3. Some western blots show that the CID chaperone Cal1 was pulled down in the CID RNA-ChIP. However, its large size and low abundance makes it generally difficult to detect. Lamin is pulled down as well, suggesting that some of the nuclear scaffold is cross-linked to the chromatin. Clearly, both H3 and Lamin are enriched in the H3 RNA-ChIP, while CID is not, suggesting a successful pull down of general non-centromeric chromatin. Last, the mock RNA-ChIP did not pull down any of the indicated proteins. Taken together, these results show a specific CID and H3 RNA-ChIP.

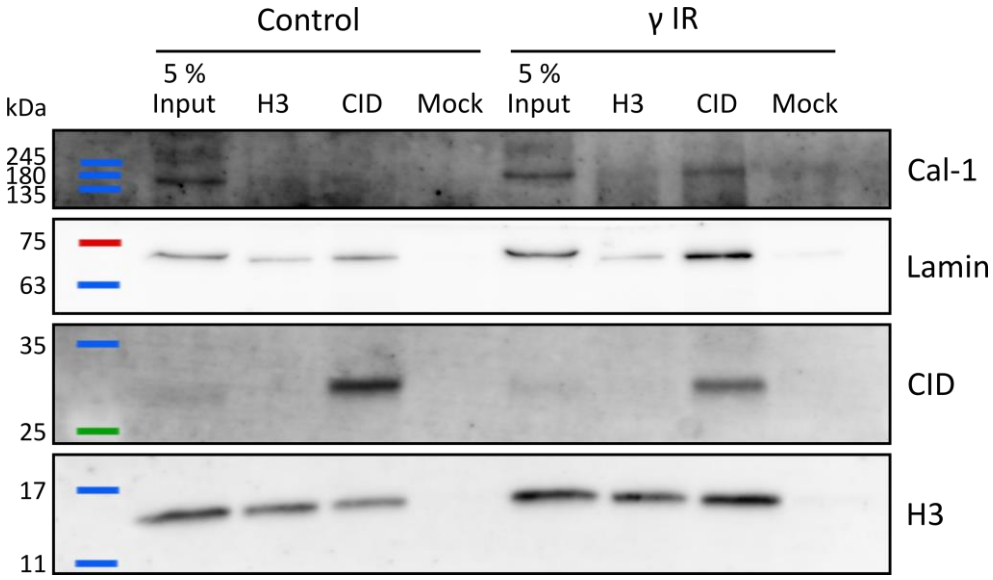


Figure 7. Embryo RNA-ChIP control western blot show successful CID and H3 pulldown.
Input and indicated ChIP samples of replicate 1 from both control and γ irradiated embryos were checked with WB using the indicated antibodies. Formaldehyde crosslinked and sonicated chromatin (Input) was immunoprecipitated (IP) to pull down centromeric chromatin (CID) and general chromatin (H3). No antibody IP (Mock) was used as a negative control. IR = irradiated.

The other half of each RNA-ChIP sample and input was treated with DNaseI and Proteinase K, and subsequently heated up to 65°C to reverse the formaldehyde cross-link. The released RNA was isolated and used for library preparation. Four replicas of each condition, untreated control and irradiated, and input, mock, H3 and CID RNA-ChIP, were sequenced.

Supplementary Figure 2 provides an overview of the technical sequencing metrics, the sequencing depth (A) and the mapping rates (B). On average I obtained close to 10 million raw reads per sample, which is sufficient to detect differentially enriched RNAs, especially with 4 replicas and the relatively small *Drosophila* genome (Y. LIU ET AL., 2014). The reads were mapped with STAR (DOBIN ET AL., 2013) to a reference genome published by the Mellone and Larracuente labs, which is the first *Drosophila melanogaster* reference to include annotated centromeric sequences (CHANG ET AL., 2019). This

reference genome was assembled from PacBio long reads and CID DNA-ChIP derived centromeric contigs. It also includes an extensive repeat and TE annotation. It is the most complete *Drosophila melanogaster* reference available, but unfortunately the pericentromeric regions remain unassembled. Because centromere-associated RNAs potentially also originate from the highly repetitive (peri)centromere, it was essential to include reads that map to multiple loci in the reference, also known as multimappers. However, most reads were still assigned to a unique location. On average, 83.8 % of the reads mapped uniquely and 5.4 % were multimappers. Subsequently, TEcount (JIN ET AL., 2015) was used to count the reads for both genes and repetitive elements.

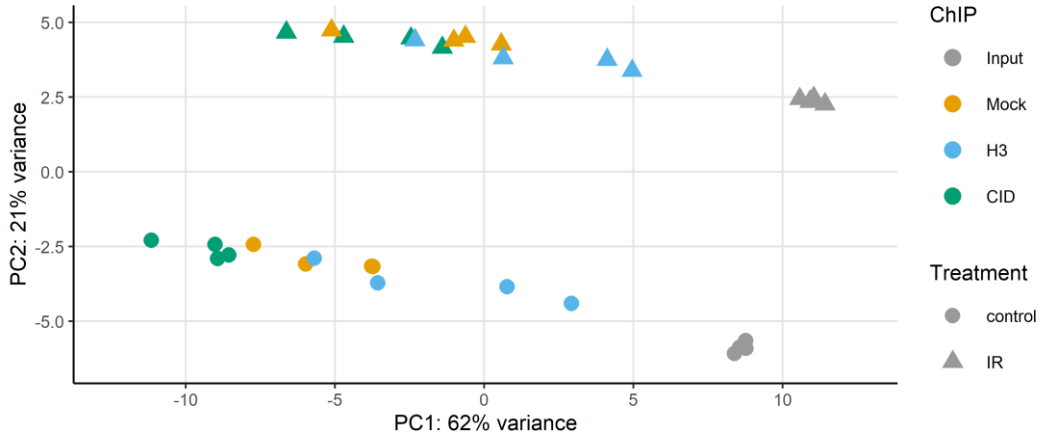


Figure 8. PCA plot indicating main variation in the RNA-ChIP-seq dataset is derived from ChIP sample type and stress treatment.
Principal Component Analysis (PCA) was performed to assess the main causes of variation between the RNA-ChIP-seq samples. The first two principal components were plotted which account for respectively 62 % and 21 % of the variation between the samples. Samples are separated along the first principle component according to the ChIP sample type and along the second principle component according to stress treatment. Samples cluster based on their similarity. IR = irradiated, PC = principle component.

A principal component analysis was performed as an initial exploration of the count data. In Figure 8 all samples are plotted according to the first two principal components, which account for 62 % and 21 % of the variation in the data, respectively. As can be seen in Figure 8, the first principal component separates the samples according to sample type, namely input, mock, H3 or CID RNA-ChIP. The second principal component separates the samples according to treatment. As expected, the H3 RNA-ChIP samples, representing general chromatin-associated RNAs, are closest to the input samples, which are the sonicated nuclei. The CID RNA-ChIP samples are more distinct from the input samples, potentially due to a specific enrichment of centromere-associated RNAs. However, since there is proximity with

the Mock RNA-ChIP samples, it is possible that there are many non-specifically bound RNAs pulled down as well. Interestingly, the proximity of the RNA-ChIP samples is more pronounced after γ-irradiation, which could indicate that stress-specific reads are masking locus specific reads.

2.1.3 Identification of stress-specific centromere-associated RNAs

Having identified the stress treatment as a major source of variation in the dataset, I wanted to ascertain that the γ-irradiation indeed induced a general DNA damage response on the transcript level. To identify significantly differentially expressed RNAs, I analyzed the count data with DESeq2 (LOVE ET AL., 2014). DESeq2 is a tool that can test differential expression or enrichment across different RNA-sequencing samples and estimate fold changes. Significantly differentially expressed genes or repetitive elements after exposure to γ-irradiation are listed in Supplementary Table 1. There are many genes with elevated expression levels, 61 of which have at least a 1.5-fold increase in expression level and an average normalized count number of at least 100 reads across the different samples. I used these particularly elevated genes for a gene ontology (GO) term enrichment analysis, of which the top results are summarized in Figure 9A. More than half of the top 20 enriched GO terms are related to cellular stress responses, in particular response to radiation. Furthermore, surprisingly many muscle development GO terms are enriched.

The top three DNA damage-specific genes that were upregulated after γ-irradiation are shown in Figure 9B. Xrp1 has a 4.4-fold increase in expression in γ-irradiated embryos compared to untreated control embryos and is a basic leucine Zipper (bZip) transcription factor (FRANCIS ET AL., 2016). Expression of Xrp1 is induced by p53 after exposure to X-irradiation (BRODSKY ET AL., 2004). It has an important role in maintaining genome stability and can suppress cell proliferation (AKDEMIR ET AL., 2007). Next, the expression of Sickle is increased 3.7-fold after γ-irradiation. Sickle is an antagonist of Inhibitor of Apoptosis Proteins and is one of the proteins responsible for triggering apoptosis after DNA damage (CHRISTICH ET AL., 2002). Last, Companion of Reaper has a 3.2-fold increase after exposure to γ-irradiation. Also this gene is known to be transcriptionally induced by p53 after DNA damage (CHAKRABORTY ET AL., 2015). It promotes the survival of somatic cells by suppressing the pro-apoptotic effects of p53 in a negative feedback loop (CHAKRABORTY ET AL., 2015). It is important to note that in all cases, these three examples and others, the increased count number after stress occurs in all sample types and is, therefore, most likely not specific to a particular sample or chromatin type but a general effect of irradiation. This result is consistent with the above mentioned separation according to treatment by the second principal component.

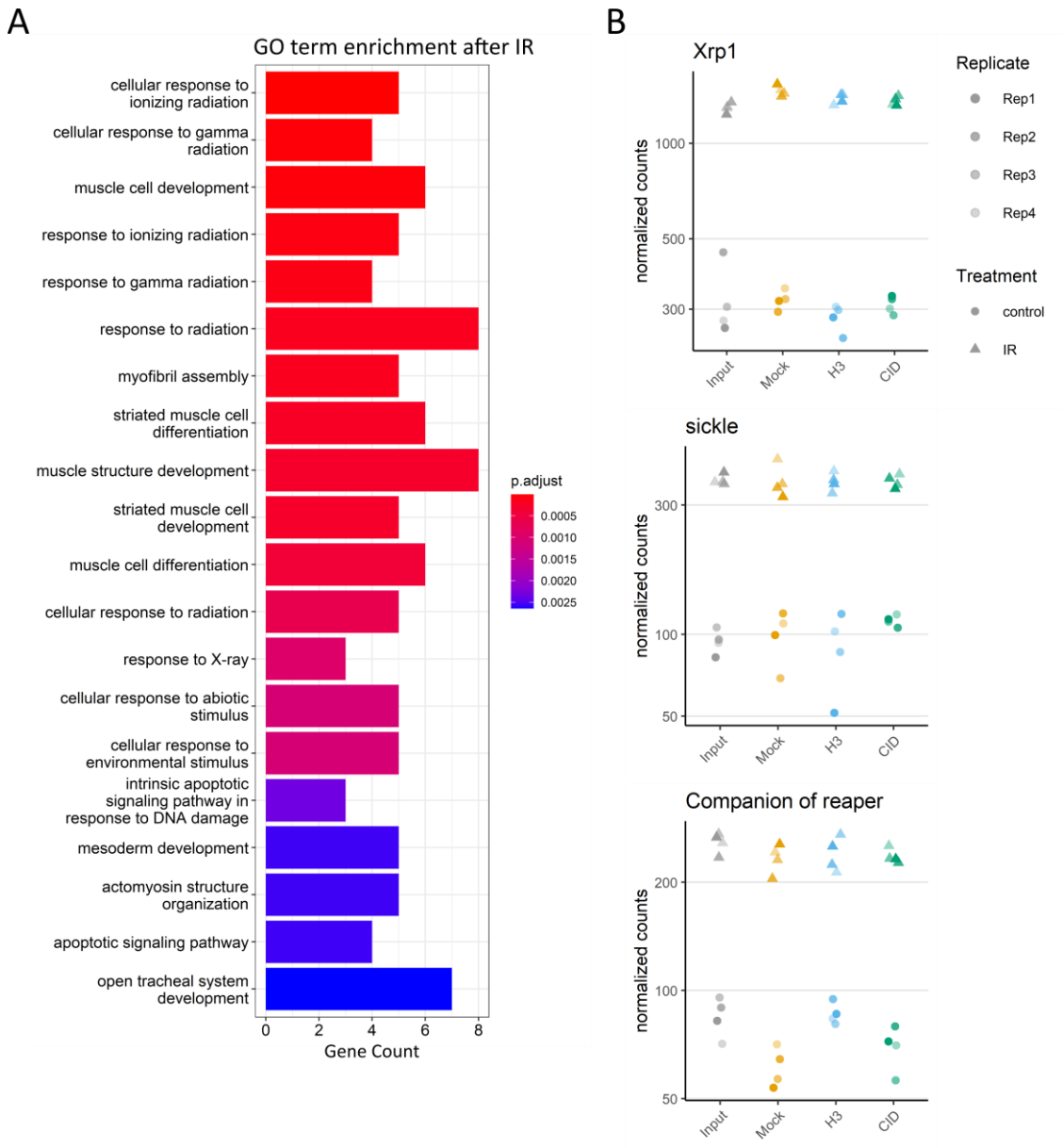


Figure 9. Several stress response pathways were activated in the γ-irradiated embryos.

A) Gene ontology (GO) term enrichment analysis for biological processes of the top differentially expressed genes. The top differentially expressed genes used for this analysis had a significant enrichment with at least a 1.5 Fold Change and an average normalized count above 100. p.adjust = p-values adjusted with the Benjamini and Hochberg method for multiple testing.

B) Examples of DNA damage related genes upregulated after γ-irradiation. Normalized counts were plotted for each sample for the genes Rab3-GEF, Xrp1 and sickle. Rep = replicate, IR = irradiated, Rep = replicate.

Now that I have established there is a transcriptional response to the DNA damage induced by γ-irradiation, I turn to identifying stress-specific centromere-associated RNAs. For this I performed a differentially enrichment analysis with DESeq2 comparing the irradiated CID RNA-ChIP samples with both the untreated control CID RNA-ChIP and the irradiated H3 RNA-ChIP samples (Supplementary

Table 2). It is important to make both comparisons to identify RNAs that are specifically enriched at centromeres after stress induction and are not also enriched at general chromatin. The log2 fold changes of both comparisons are plotted against each other in Figure 10. To reduce the effect of noisy fold changes caused by genes with low count numbers without using an arbitrary cut-off, I shrunk the log2 fold changes with the apeglm shrinkage estimator (A. ZHU ET AL., 2019). There are a total of 17 genes and repetitive elements significantly enriched in the irradiated CID RNA-ChIP samples compared to both the untreated control CID RNA-ChIP and the irradiated H3 RNA-ChIP samples, which are depicted as red dots in the top right quadrant. These genes are potentially recruited to or additionally maintained at centromeres upon genotoxic stress. However, the fold changes are relatively small, especially compared to the irradiated H3 RNA-ChIP.

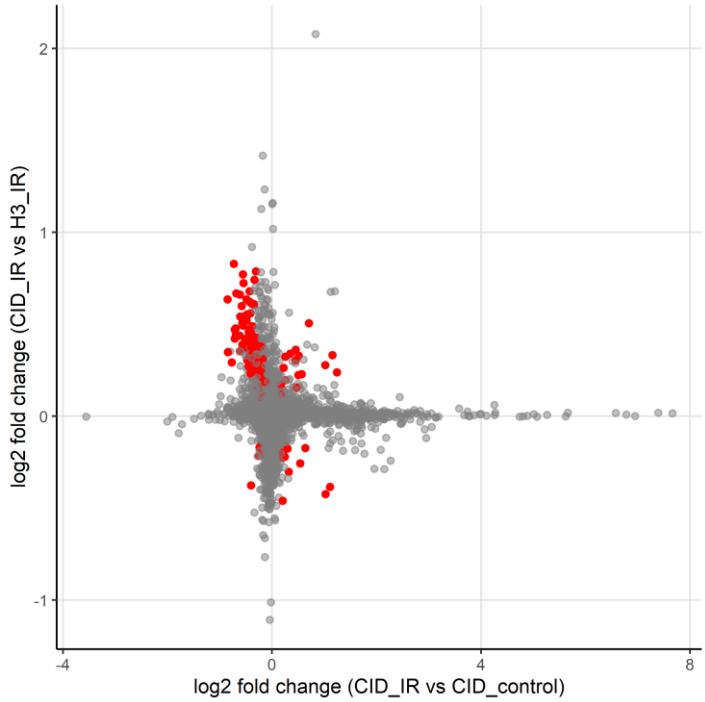


Figure 10. Analysis of DNA damage-specific centromere-associated RNAs.
For each gene the log2 fold change of the irradiated CID RNA-ChIP over the untreated control CID RNA-ChIP was plotted against the log2 fold change of the irradiated CID RNA-ChIP over the irradiated H3 RNA-ChIP. Genes that are significantly differentially enriched and/or depleted in both comparisons are indicated red. In the top right quadrant are the potential centromere-associated RNAs recruited during DNA damage. Log2 fold changes were shrunken with the apeglm shrinkage estimator in order to remove noisy fold changes from RNAs with low count numbers. IR = irradiated.

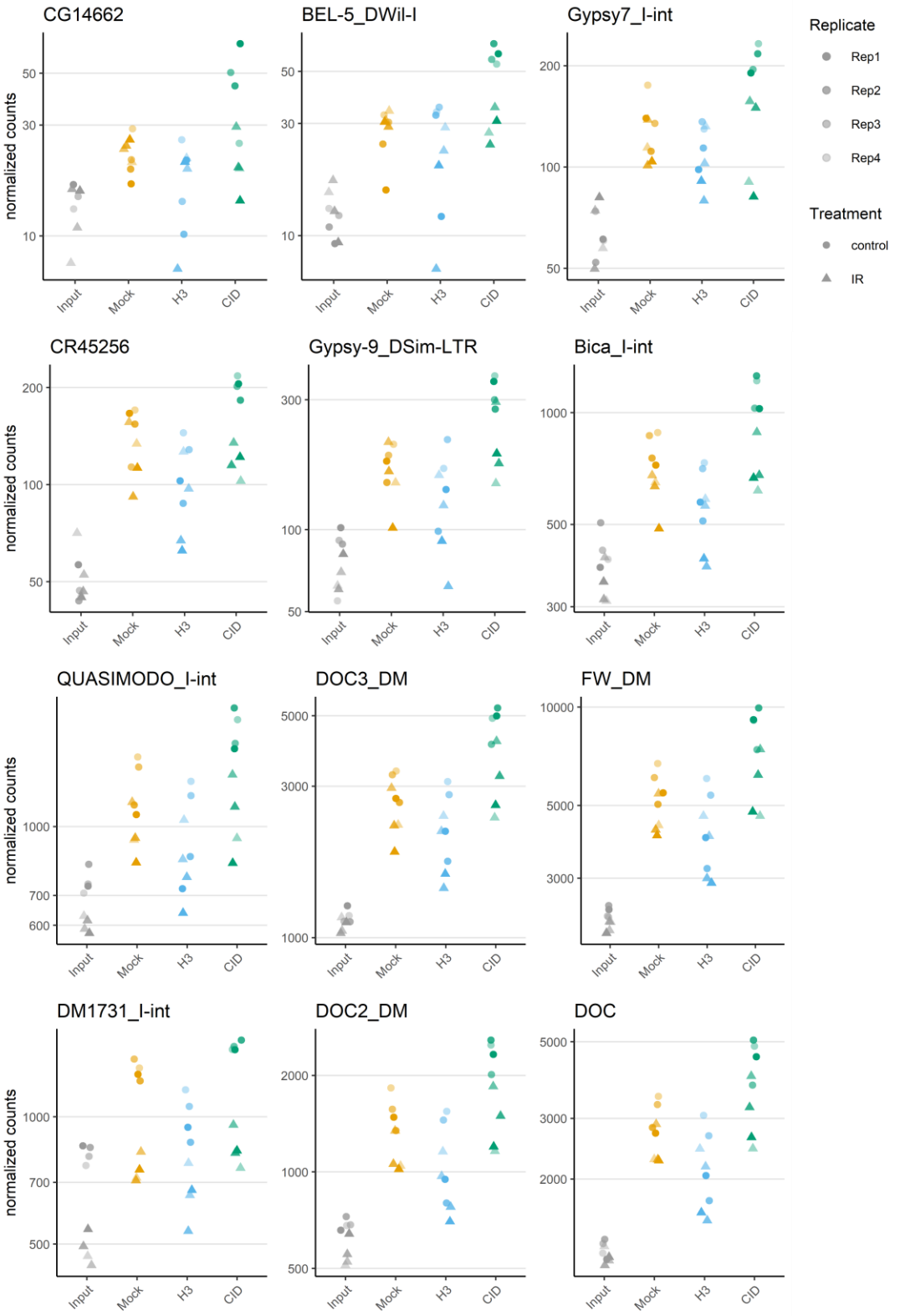
To assess the validity of these putative candidates, it is important to compare not only to the H3 RNA-ChIP, the general chromatin control, but also to the other control samples, namely Input and Mock RNA-ChIP. Therefore, I plotted the individual normalized counts for each sample for the top candidates. The top 12 candidates that are enriched at centromeres in γ-irradiated embryos are shown in Supplementary Figure 3. These genes have an increase of at least 1.2-fold in comparison to both the irradiated H3 RNA-ChIP and untreated control CID RNA-ChIP. Most of these genes did not show a difference in normalized count number when comparing the irradiated CID RNA-ChIP (green triangles) to the irradiated Mock RNA-ChIP (yellow triangles), suggesting that these candidates are rather a result

of unspecific binding during the pull down. Only for Mid1 RNA there is a clear separation of the irradiated CID RNA-ChIP (green triangles) from all other samples across all replicas. However, it is still a questionable candidate due to the low fold changes, 1.5-fold compared to the irradiated H3 RNA-ChIP and 1.7-fold compared to the control CID RNA-ChIP, and the low count numbers. Overall, these results indicate that a general DNA damage response is induced in γ-irradiated embryos, but this did not result in a meaningful accumulation of stress-specific centromere-associated RNAs.

2.1.4 Stress-induced depletion of centromere-associated RNAs

While analyzing the data for stress-specific centromere-associated RNAs, I noticed a strikingly large group of genes and repetitive elements showing a decrease at centromeres in irradiated samples (Figure 10, top left quadrant). To identify centromere-associated RNAs which are depleted after γ-irradiation, I performed a similar differential enrichment analysis with DESeq2 as described above, now comparing the untreated control CID RNA-ChIP samples with both the irradiated CID RNA-ChIP and the untreated control H3 RNA-ChIP samples (Supplementary Table 3). There is a group of just over 270 genes and repetitive elements that are significantly enriched in untreated control CID RNA-ChIP samples compared to H3 RNA-ChIP samples, but depleted in the irradiated CID RNA-ChIP samples. These genes are potential centromere-associated RNAs that are displaced or degraded after exposure to γ-irradiation.

As described above, normalized counts of all the different samples were plotted for each candidate. It became clear that most of these genes are also reduced in the controls. Therefore, I compared the reduction after γ-irradiation in the CID RNA-ChIP samples with the reduction after γ-irradiation in all samples, which are reported in Supplementary Table 1. A group of 31 RNAs have a larger DNA damage induced reduction in the CID RNA-ChIP than in all samples. There is a clear overrepresentation of transposable elements in this group, 21 are TEs, 6 are protein-coding RNAs and 4 are non-coding RNAs (ncRNAs). The top 9 candidates are plotted in Figure 11. I also plotted 3 additional TEs which were reported by Chang et al., 2019 as part of the centromeric DNA (CHANG ET AL., 2019), which just fell short of the cut-offs used to select the top candidates. Although the fold-changes are not very high and some of the selected RNAs have very low average count numbers, the high number of genes following this trend is interesting. These results suggest there might be a broad depletion of RNAs from centromeres after DNA damage induction. Despite this being an interesting result, I did not follow up with additional experiments due to a more interesting candidate arising from the following analysis.



Legend on next page

Figure 11. Potential centromere-associated RNAs depleted by DNA damage.
Normalized counts plotted per sample for the top candidates indicated as both significantly enriched in untreated control CID RNA-ChIP over untreated control H3 RNA-ChIP and significantly depleted in irradiated CID RNA-ChIP over untreated control CID RNA-ChIP. The top 9 candidates were selected for having at least a 1.5-fold respective increase or decrease in the previous mentioned analysis and only an up to 1.4-fold decrease after irradiation when analyzing all samples together. The bottom 3 transposable elements showed a similar trend and were previously reported as part of the centromeric DNA. IR = irradiated, Rep = replicate.

2.1.5 RNAs from transposable elements are enriched at the centromere

The next step was to identify novel centromere-associated RNAs in the absence of genotoxic stress. I analyzed the untreated control samples to identify RNAs enriched in the CID RNA-ChIP compared to H3 RNA-ChIP with DESeq2 (Supplementary Table 4). To get an overview of the differentially enriched RNAs, I plotted the apeglm shrunken log2 fold changes reported by DESeq2 against the mean of normalized counts (Figure 12A). The RNAs were color coded according to RNA type. The RNAs that have a positive log2 fold change are enriched at centromeric chromatin compared to general chromatin. As can be seen in Figure 12A, many of these enriched RNAs are TEs or repetitive elements (shown as red dots). There are also quite some protein-coding RNAs (blue dots) and non-coding RNA (ncRNAs, green dots) present in this cloud of enriched transcripts. In the group of significantly enriched RNAs with at least a 1.5-fold enrichment, there are 150 TEs/repeats, 75 protein-coding RNAs and 33 ncRNAs. Considering the difference in prevalence of each RNA type present in the dataset, this trend becomes even more pronounced, with 12.9 % of the annotated TEs present in the centromere-enriched group, compared to 0.6 % protein-coding RNAs and 1.4 % ncRNAs.

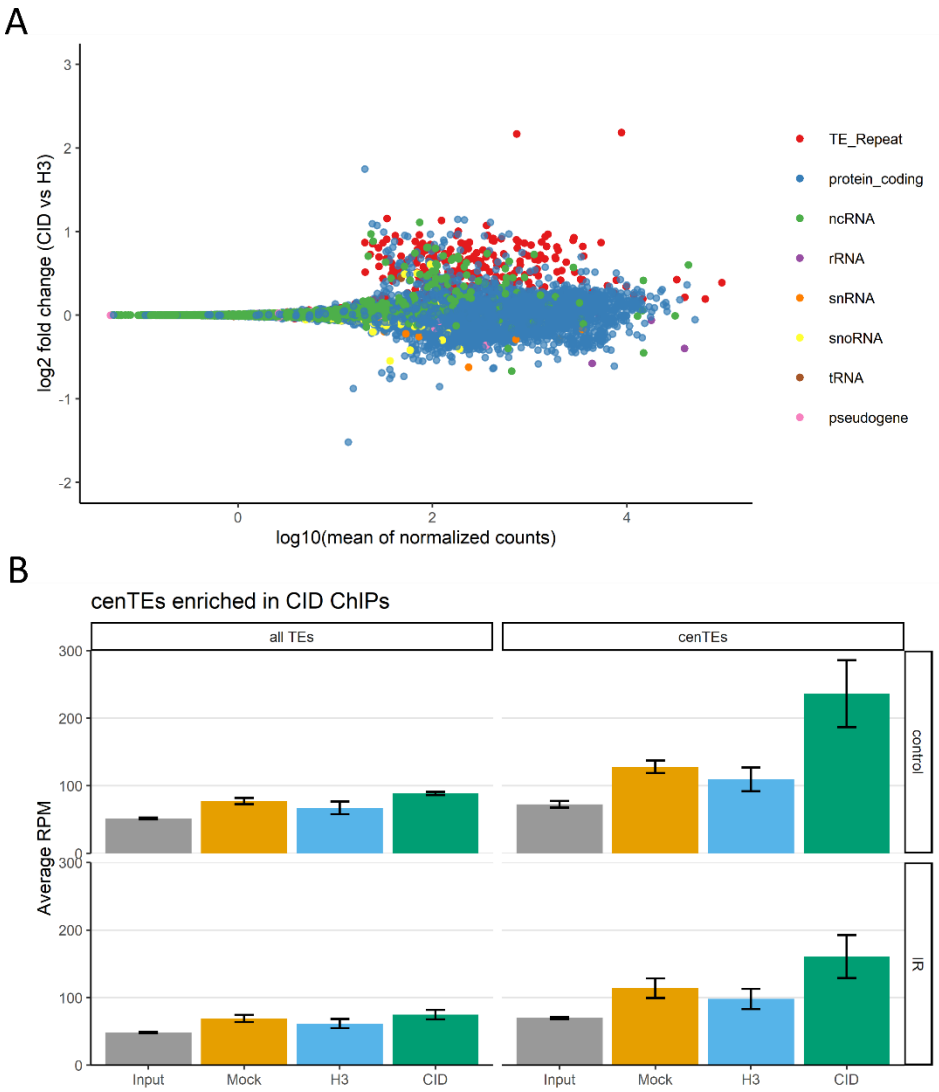


Figure 12. Transposable elements are enriched in the untreated control CID RNA-ChIP.

A) Plot of the log2 fold change and the log10 normalized counts from the differential enrichment analysis by DESeq2 comparing untreated control CID RNA-ChIP samples to untreated control H3 RNA-ChIP samples. Log2 fold changes were shrunken with the apeglm shrinkage estimator in order to remove noisy fold changes from RNAs with low count numbers. Genes are color coded by RNA type as indicated.

B) The average reads per million (RPM) of all TEs (left) and the TEs enriched in centromeric DNA (cenTEs, right) plotted per sample type. The bar indicates the mean over the 4 replicas per sample type. Error bars show the standard deviation across the 4 replicas. TE = transposable element, IR = irradiated, RPM = reads per million.

Because of the high number of enriched TEs, I assessed whether TEs which are reported to be present in centromeric DNA were specifically amongst those enriched TEs. As mentioned, Chang et al., 2019 reported that centromeres are located on islands of TE-rich DNA. Table 1 list the 33 TEs that are reportedly present in centromeric DNA. Although none of these TEs are unique to centromeres, they were enriched in the CID DNA-ChIP of Chang et al., 2019 or part of their assembled centromeric contigs.

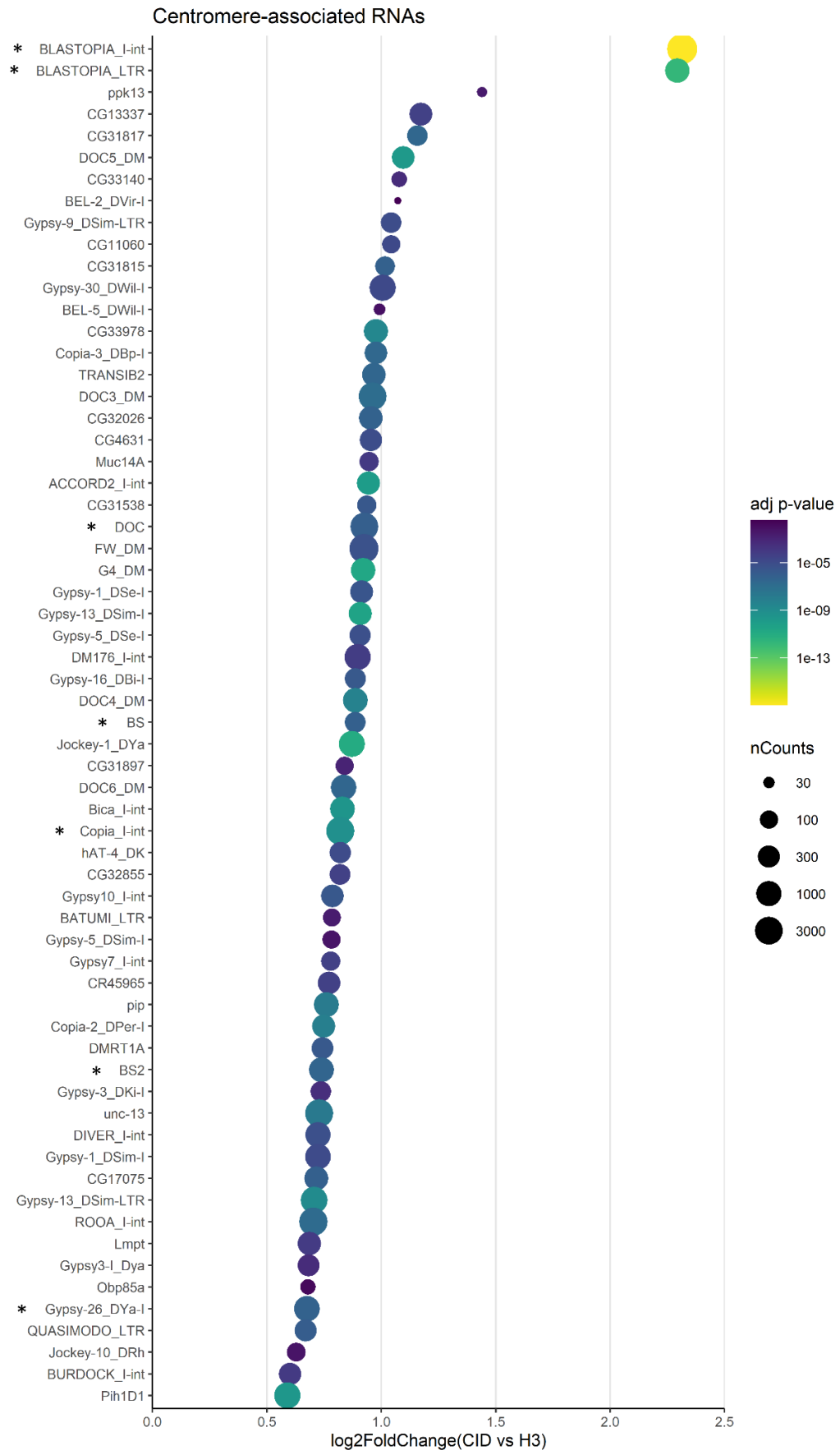
All but three of these centromeric TE (cenTEs) are enriched in the CID RNA-ChIP compared to the H3 RNA-ChIP, although not all with large fold changes or with significant enrichment. I normalized the counts of the cenTEs to the total number of counted reads in each sample, resulting in Reads per Million (RPM) normalized counts. Next, I took the average RPM for all cenTEs and averaged this average across the 4 replicates. I did the same for all TEs with at least one counted read across all samples, in order to compare the average levels of the cenTEs to the overall TE levels. Figure 12B shows that the average RPM of all TEs is at a similar level across the different RNA-ChIP samples. However, cenTEs are clearly at much higher levels in the CID RNA-ChIP samples, most apparent in the untreated control but also in the irradiated samples. Taken together, RNAs from TEs, and especially TEs present at centromeric DNA, are enriched in the CID RNA-ChIP samples.

Table 1. TEs in centromeric DNA reported by Chang et al., 2019

BEL-22_DTa-I	Chimpo_I-int	DOC2_DM	Gypsy2_I	Jockey-3_DSim
BEL-6_DYa-I	Copia_I-int	DODECA_SAT	Gypsy-24_DYa-I	NOMAD_I-int
Bica_LTR	DM1731_I	G_DM	Gypsy-26_DYa-I	NTS_DM
BLASTOPIA_I-int	DM412B_LTR	G2_DM	Gypsy-27_DYa-I	R2_DM
BLASTOPIA_LTR	DMLTR5	G5_DM	Gypsy-7_DSe-I	TART-A
BS	DMRT1B	G6_DM	Gypsy8_LTR	
BS2	DOC	Gypsy-2_DSim-I	Jockey-1_DSi	

2.1.6 Overview of putative centromere-associated RNAs

The differential enrichment analysis described above showed that a large group of genes is enriched in the untreated control CID RNA-ChIP samples compared to untreated control H3 RNA-ChIP samples. Previous analysis showed that there are many RNAs that have similar count numbers in the CID RNA-ChIP and Mock RNA-ChIP samples. In order to separate centromere-associated RNAs from non-specifically pulled down RNAs, I performed a differential enrichment analysis comparing untreated control CID RNA-ChIP samples with untreated control Mock RNA-ChIP samples. The significantly enriched RNAs from this analysis were cross referenced with the significantly enriched RNAs from the comparison with untreated control H3 RNA-ChIP samples reported in Supplementary Table 4. This provided me with a list of genes and repetitive elements enriched in CID RNA-ChIP samples compared to both H3 and Mock (Supplementary Table 5). The top candidates were selected to have at least a 1.5-fold enrichment over the H3 RNA-ChIP. Among them are 1 ncRNA, 19 protein-coding mRNAs and 43 TEs, of which 7 cenTEs. The log2 fold changes compared to the H3 RNA-ChIP were plotted for these candidates in Figure 13, with dot size and color depicting the mean of normalized counts and the adjusted p-value, respectively. DESeq2 adjusts the p-value with the Benjamini and Hochberg method to reduce false discovery rates associated with multiple testing (BENJAMINI & HOCHBERG, 1995).



Legend on next page

Figure 13. Top candidates identified as centromere-associated RNAs.
The log2 fold change of potential centromere-associated RNAs were plotted, with their size indicating the mean of normalized counts for that gene and the color indicating the p-value adjusted for multiple testing with the Benjamini and Hochberg method. Top hits were selected as having at least a 1.5-fold enrichment in untreated control CID RNA-ChIP compared to untreated control H3 RNA-ChIP and significant enrichment compared to untreated control Mock RNA-ChIP. TEs found in centromeric DNA are indicated with an asterisk in from of their name. nCounts = mean of normalized counts , adj p-value = Benjamini and Hochberg adjusted p-value.

From Figure 13, it can be seen that by far the most enriched RNA is Blastopia. The internal part (I-int) has a 5-fold and the long terminal repeat (LTR) has a 4.9-fold enrichment in CID RNA-ChIP compared to H3 RNA-ChIP and they have an average normalized count number of respectively 8783 and 734. Both also have a strong enrichment compared to Mock RNA-ChIP, with respectively a 4.2-fold and a 3.7-fold enrichment. Gypsy TE Blastopia is a cenTE, since it was reported to be enriched in the CID DNA-ChIP by Chang et al., 2019. Blastopia is a very promising putative novel centromere-associated RNA and will be discussed in more detail in the section 2.2.

The other top candidates do not have as strong of an enrichment in CID RNA-ChIP compared to H3 RNA-ChIP. In Supplementary Figures 4 to 6, I plotted the normalized counts per sample for the candidates that have at least a 1.5-fold enrichment compared to both H3 and Mock RNA-ChIP. mRNA ppk13 has the largest fold change after Blastopia, with a 2.7-fold enrichment compared to H3 and 3.4-fold enrichment compared to Mock. However, the small count numbers make it a less convincing candidate. Protein-coding RNA CG13337 and TEs DOC5_DM and Gypsy-30_DWil-I still have fold change above 2 and have substantially higher count numbers than ppk13. CG13337 has a 2.3-fold enrichment compared to H3 and 2-fold enrichment compared to Mock. Jockey TE DOC5 has a 2.1-fold enrichment compared to H3, but only a 1.6-fold enrichment compared to Mock. Gypsy-30 has the highest normalized counts of the three, with a 2-fold enrichment compared to H3 and 1.7-fold enrichment compared to Mock.

Normalized counts per sample of significantly enriched cenTEs are plotted in Supplementary Figure 7. Here, the differences with Mock are lower than 1.5-fold, but the count numbers of especially DOC, Copia_I-int and DOC2_DM are large. I have also plotted G2/Jockey-3, the most prominent cenTE, and several satellite repeats in Supplementary Figure 7. The RNA from G2/Jockey-3 and the different satellite repeats were not significantly enriched in the CID RNA-ChIP samples. This is especially surprising for SATIII (359bp_SAT), which was previously identified as a centromere-associated RNA (ROŠIĆ ET AL., 2014). One explanation could be that SATIII RNA localizes strongly at the pericentromere next to the core centromere and is therefore equally present in the H3 RNA-ChIP.

It is remarkable that all of the six annotated DOC transposable elements are significantly enriched in the CID RNA-ChIP samples compared to H3 RNA-ChIP samples, with DOC6 having the lowest enrichment with 1.8-fold. All of them have high count numbers and three of them, DOC3, DOC4 and DOC5, are also enriched compared to Mock. Both DOC and DOC2 are cenTEs. It is interesting that this group of TE RNAs from the Jockey superfamily is collectively enriched at the centromere.

2.2 Blastopia

2.2.1 Blastopia counts and coverage

As indicated above, Blastopia is evidently the most enriched putative centromere-associated RNA identified by the embryo RNA-ChIP-seq analysis. When looking more closely at the counts of Blastopia in Figure 14A, it becomes clear that Blastopia is a highly expressed TE with specifically higher count numbers in the CID RNA-ChIP samples, both from untreated control and irradiated embryos. Interestingly, Blastopia seems to be slightly reduced at centromeres after exposure to γ-irradiation, although this reduction was not significant according to the analysis performed in section 2.1.4.

Blastopia, also known as Flea, is a 5034 bp retrotransposon of the gypsy superfamily (FROMMER ET AL., 1994). Full length copies of Blastopia, including its two flanking LTRs, with a transcriptional start site in the 5' LTR and a poly-A site in the 3' LTR, are annotated 19 times in the Mellone and Larracuente labs genome reference (CHANG ET AL., 2019). However, it is important to remember that the highly repetitive pericentromeric regions are still not fully assembled in this reference, which could potentially harbor additional copies. When looking at read coverage at a representative Blastopia locus in chromosome arm 2R with the Integrative Genomics Viewer, it can be seen that sequencing reads from untreated control Input and RNA-ChIP samples of replicate 4 cover the entire length of the TE (Figure 14B).

Like most retrotransposons, Blastopia encodes a polyprotein that contains the proteins required for retrotransposition. In the case of Blastopia this is a 1333 amino acid polyprotein, encoding the *gag* and *pol* genes, of which *pol* can be cleaved further into a peptidase, reverse transcriptase and integrase (FROMMER ET AL., 1994). The locations of these enzymes are indicated in Figure 14C (NCBI_PROTEIN:CAA81643.1, UNIPROT: Q24262). I have also indicated the locations of the qPCR amplicon and the short hairpin RNA (shRNA) and double-stranded RNAs (dsRNAs) used later in this study.

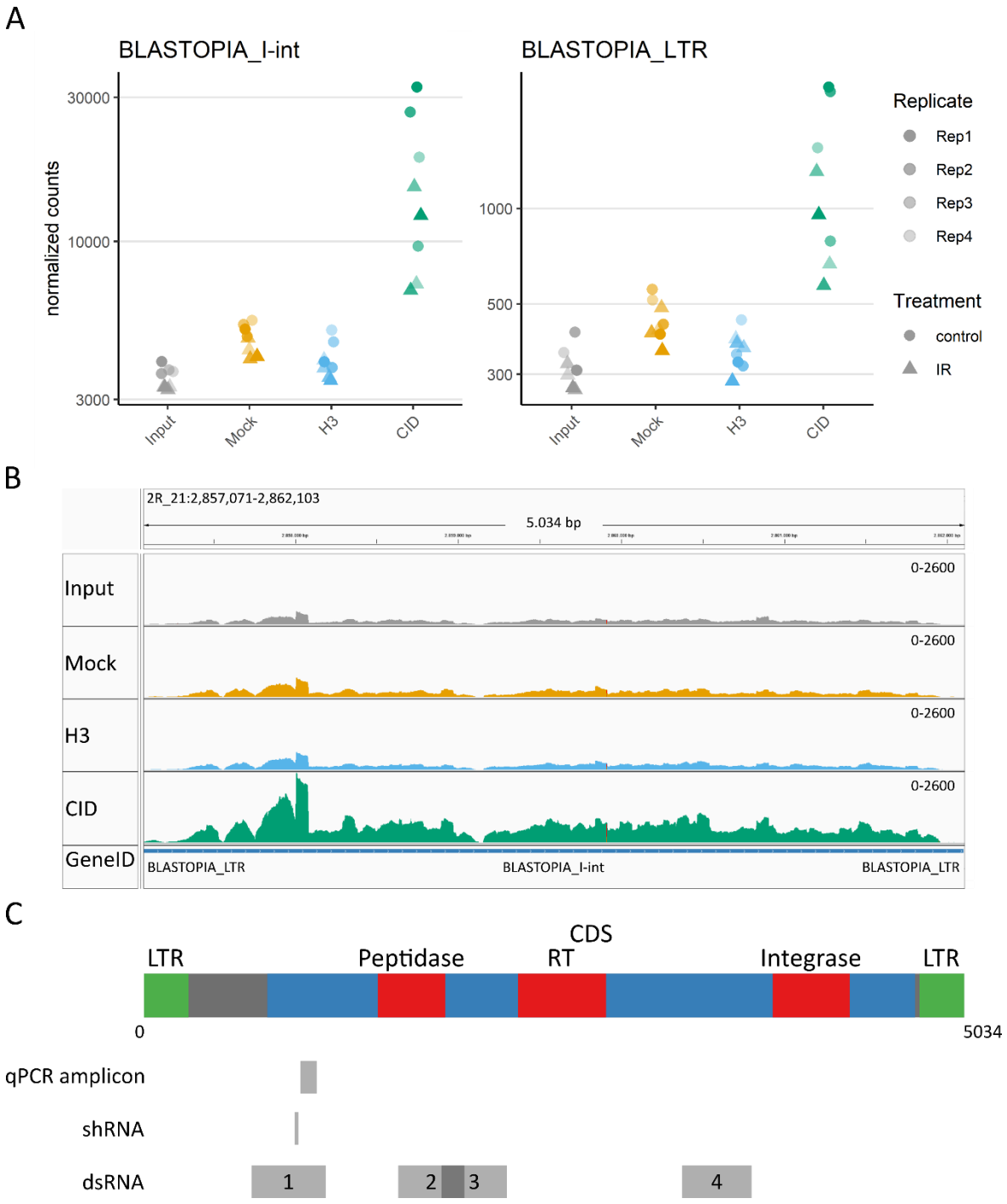


Figure 14. Blastopia significantly enriched in CID RNA-ChIP over all controls.

A) Normalized counts plotted per sample for Blastopia_I and Blastopia_LTR. IR = irradiated, Rep = replicate.

B) A representative Blastopia locus in the genome shown in Integrated Genomics Viewer with the read coverage tracks of Rep4_control.

C) Graphical overview of the Blastopia gene, with the coding region (CDS) of the polyprotein in blue and the Long Terminal Repeats (LTR) in green. The locations of the enzymes are indicated in red, the Peptidase, the Reverse Transcriptase (RT) and the Integrase. Underneath are the locations of the qPCR primers, shRNA and dsRNAs used in this study indicated.

2.2.2 Blastopia in the *Drosophila* genome

Blastopia was first identified as an active transposable element of *Drosophila melanogaster* by Frommer et al., 1994. They detected the incorporation of Blastopia into DNA of both euchromatic arms and the heterochromatin chromocenter by in situ hybridization to polytene salivary gland chromosomes (FROMMER ET AL., 1994). Blastopia copies were found in all chromosome arms, except chromosome 4 (KAMINKER ET AL., 2002). As mentioned, Blastopia was pulled down in the CID DNA-ChIP reported by Chang et al., 2019 and there was an exceptionally strong signal near centromeres in the above mentioned in situ hybridization assay by Frommer et al., 1994, suggesting that Blastopia is present in centromeric DNA in *Drosophila melanogaster*.

Figure 15. Blastopia insertions into intronic regions of CG3777, jing and tipE homolog 1.

A) Integrated Genomics Viewer of the Blastopia insertion into the intronic region of CG3777 as reported by Deloger et al., 2009. No or very few reads map outside the Blastopia area of this intron.

B) Normalized counts plotted per sample for CG3777, jing and tipE homolog 1 (Teh1). All have Blastopia inserted in sense orientation into one of their introns. IR = irradiated, Rep = replicate.

One of the insertion sites is in the intronic region of protein coding gene CG3777 (DELOGER ET AL., 2009). I checked all 19 full length insertion sites to assess whether Blastopia copies were inserted into other intronic regions. Besides CG3777, Blastopia is incorporated in sense direction into introns of protein-coding genes jing and tipE homolog 1. In antisense direction, Blastopia is incorporated into introns of protein-coding genes CG43861, CG10947, Diacyl glycerol kinase, Adenylyl cyclase 78C, twin and Odorant receptor 1a. With Integrative Genomics Viewer I examined the read coverage of CG3777, specifically surrounding the Blastopia insertion (Figure 15A). Throughout this copy of Blastopia the regular coverage of reads is detected, which is expected for repetitive elements when allowing for multimappers. Neighboring intronic regions have very few reads aligned to them, and the normalized counts assigned to CG3777 are very low (Figure 15B). Similarly to CG3777, tipE homolog 1 (Teh1) is expressed at very low levels (Figure 15B). Although jing is highly expressed, the intronic region surrounding the Blastopia insertion site show very few aligned reads. Together this implies that the reads were correctly assigned to Blastopia, and did not originate from highly expressing intronic regions.

Copies of Blastopia have also been identified in the genomes of *Drosophila simulans* and *Drosophila sechellia*, but not in the genome of *Drosophila yakuba* (LERAT ET AL., 2011). All these species are part of the melanogaster subgroup together with *Drosophila melanogaster*. They have a last common ancestor roughly 11 million years ago after which *Drosophila yakuba* separates into a different branch. *Drosophila melanogaster* diverged from *Drosophila simulans* and *Drosophila sechellia* approximately 5 million years ago (timetree.org). This suggests that Blastopia appeared as a novel TE only after *Drosophila yakuba* diverged or that Blastopia was lost in the *Drosophila yakuba* branch. To make this distinction, one would have to examine the genomes of a species in the same branch as *Drosophila yakuba*, such as *Drosophila erecta*, and a species just outside the melanogaster subgroup, such as *Drosophila ananassae* or *Drosophila pseudoobscura*.

2.2.3 Blastopia RNA localization in embryos

Frommer et al., 1994 described the expression pattern of Blastopia across embryonic development using in situ hybridization. From as early as the blastoderm stage, when zygotic transcription has just started, they describe a band of transcripts visible near the anterior pole of the embryo (FROMMER ET AL., 1994). Upon gastrulation, this band of expression is directly anterior of the newly formed cephalic furrow and starts to become stronger towards the ventral side, an area which will develop into the head. In later developmental stages, they see Blastopia expression in several different areas of the embryo, however, it is still clearly present in the central nervous system. Frommer et al., 1994 showed that the blastoderm stage anterior expression band does not appear in Bicoid mutants, suggesting Blastopia is under direct or indirect control of Bicoid.

Having identified Blastopia as a putative centromere-associated RNA in embryos, I set out to explore Blastopia's expression pattern in embryonic developmental stages used for the RNA-ChIP-seq experiment. By using single molecule RNA fluorescent *in situ* hybridization (smRNA-FISH), I corroborated the Frommer al al., 1994 expression pattern, but at a higher resolution than previously described (Figure 16): In blastoderm/stage 5 embryos, the smRNA-FISH signal of Blastopia is mainly nuclear. The strongest signal is indeed visible in an anterior band where the cephalic furrow will form. However, cells throughout stage 5 embryos show Blastopia foci as well, although weaker. The cephalic furrow is a transient ring-like cleft of epithelial cells of which the function is still unknown (GHEISARI ET AL., 2020; HARTENSTEIN, 1993). In stages 6 and 8, the signal becomes stronger in a ventral area which is anterior of the cephalic furrow and posterior of the anterior midgut invagination. Still a lot of the signal is nuclear, but there is also signal in the cytoplasm of these highly expressing cells. The cells with high levels of expression are located both on the outside of the embryos and in the invaginations. Stage 8 cells expressing Blastopia in these invaginations have yet to lose the characteristic columnar shape of epithelial cells (Figure 17A). However, as can be seen in Figure 17B, in a slightly later developmental stage, a solid cluster of rounded, apolar cells are expressing Blastopia in the ventral anterior region, which could be the anterior mesoderm or the endoderm of the anterior midgut rudiment (GHEISARI ET AL., 2020; HARTENSTEIN, 1993). Later in development, stages 11 and 13, the strong anterior signal is lost, but there are still distinct cells with higher expression levels throughout the embryo (Figure 16). From these images, it is not possible to identify the specific cell types expressing Blastopia in these later stages. The expression pattern of Blastopia up to stage 8 is strikingly similar to the expression pattern of Hox gene Deformed (FISHER ET AL., 2012), suggesting Blastopia expression in these early stages might be regulated by Deformed or together with Deformed.

For smRNA-FISH of repetitive elements like Blastopia, it is particularly important to verify the signal is derived from hybridization to RNA. Hybridization to the multiple DNA copies could result in a strong inaccurate signal. For this protocol, I used a hybridization temperature of 30°C, making hybridization with DNA already unlikely. Furthermore, the differences in smRNA-FISH signal between different cells supports that the signal is derived from RNA molecules, as each cells should have the same genome and thus the same Blastopia copy number. In order to confirm the Blastopia smRNA-FISH probe specifically recognizes RNA, I treated the fixed embryos with RNaseA. Indeed, Blastopia RNA-FISH signal in stage 5 and stage 9 OregonR embryos was no longer detected after RNaseA treatment, confirming the specificity of the signal (Figure 18).

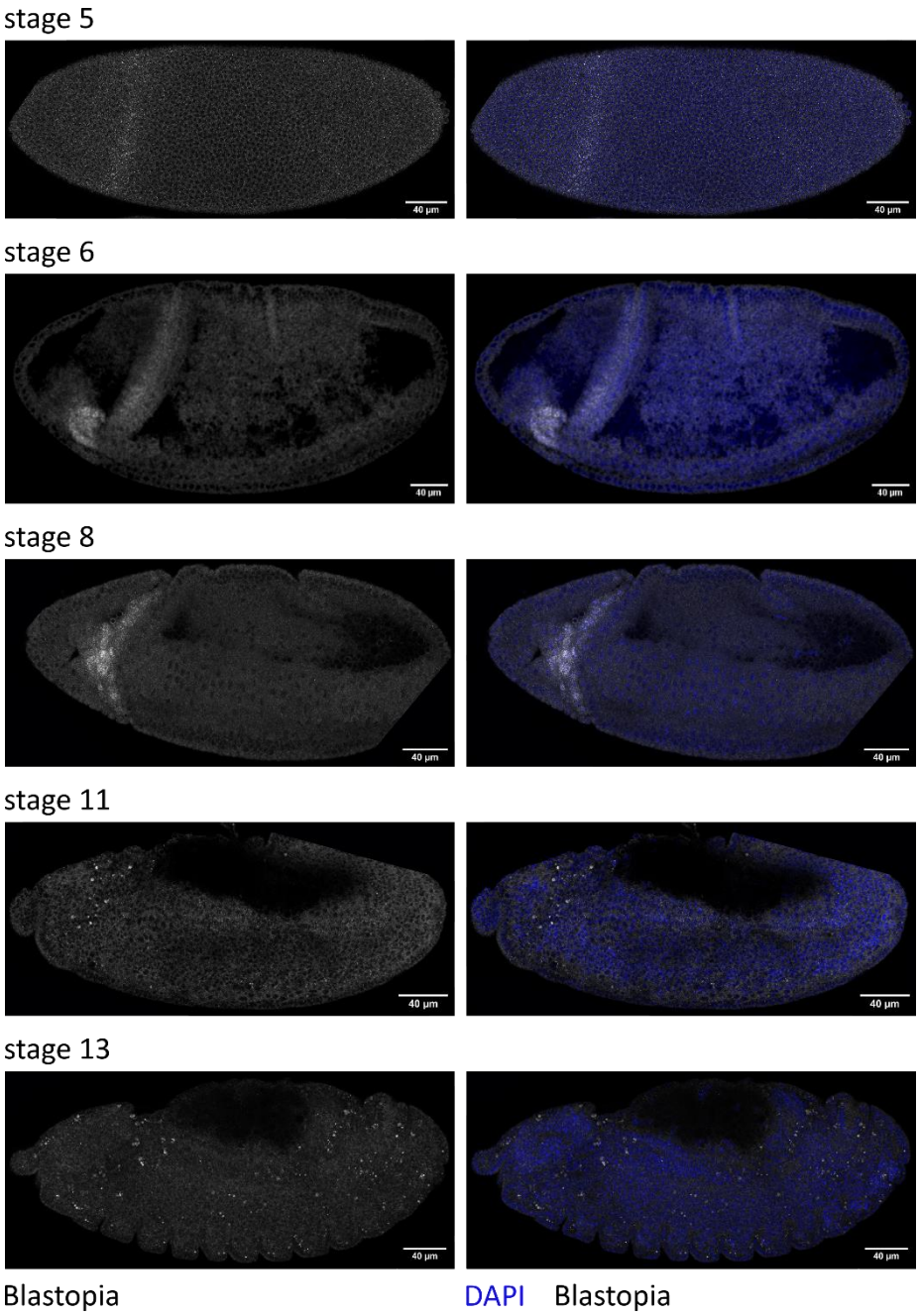


Figure 16. Blastopia smRNA-FISH in different embryo stages.
Single z slices of Blastopia smRNA-FISH staining (gray) in wild-type embryos. Embryos are positioned with their anterior pole to the left and dorsal side up, with exception of the stage 8 embryo which is slightly rotated with its ventral side towards the microscope. Embryos were counterstained with DAPI (blue) to visualize the nuclei. These images are from several different experiments, making it difficult to compare signal intensity between the different stages. Blastopia expression pattern observed is similar to the pattern described by Frommer at al., 1994. Scale bar = 40 µm.

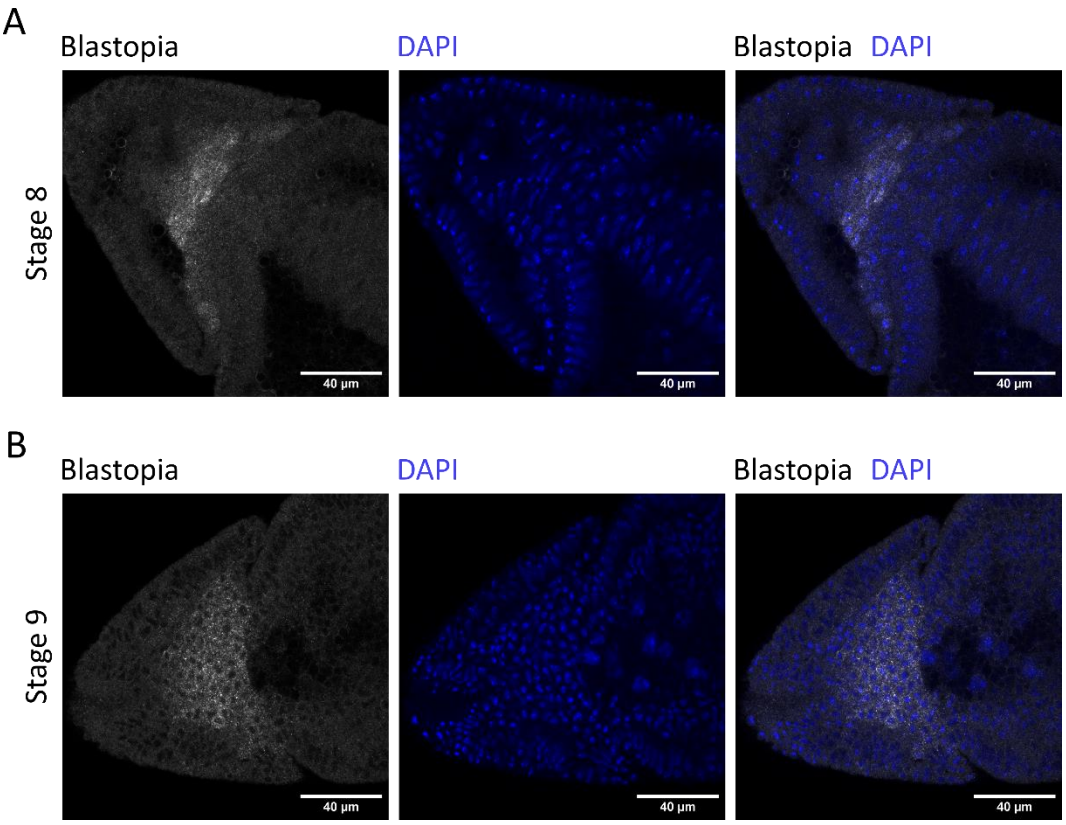


Figure 17. Blastopia smRNA-FISH signal in both epithelial cells and endodermal or mesodermal cells.

A) Strong ventral anterior Blastopia RNA-FISH signal found at the intersection of the cephalic furrow, the ventral furrow and the anterior midgut invagination late gastrulation embryos. At stage 8, cells still have their epithelial morphology, meaning they still have their columnar shape.

B) Shortly after gastrulation cells that express Blastopia has adopted a round shape, indicating that these cells could be part of the anterior mesoderm or the endoderm of the anterior midgut rudiment.

Shown are single z-slices of wildtype embryos, with their anterior pole to the (top) left and the ventral side facing the camera. Blastopia smRNA-FISH in gray, DAPI in blue. Scale bar = 40 µm.

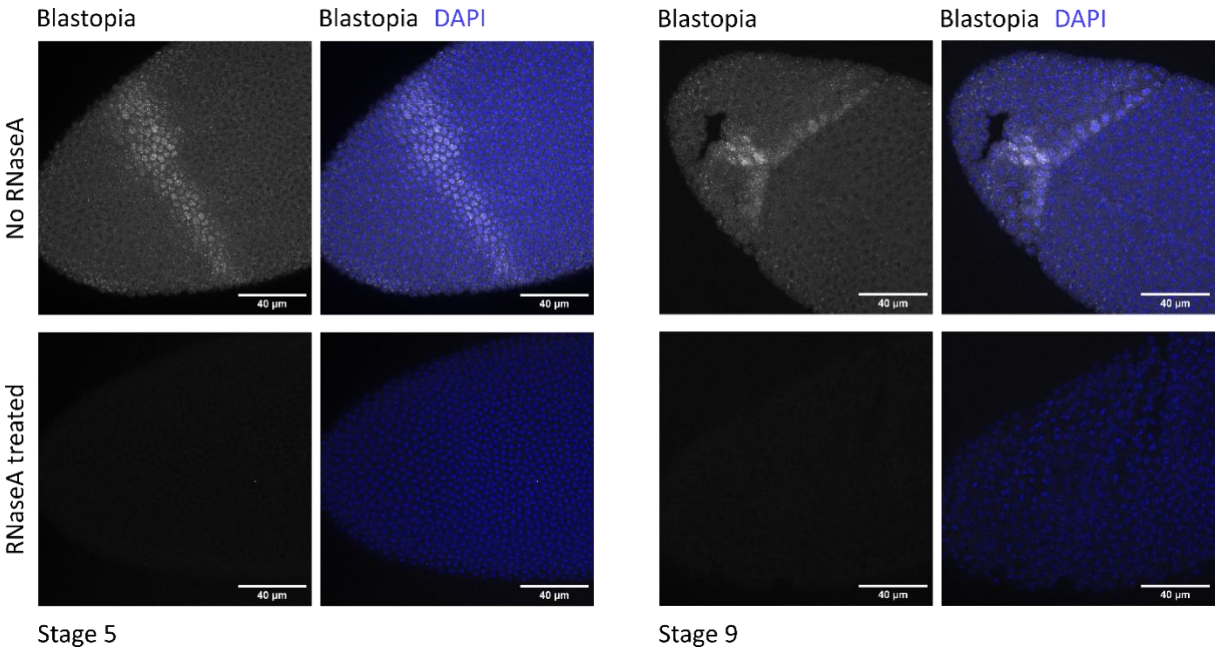
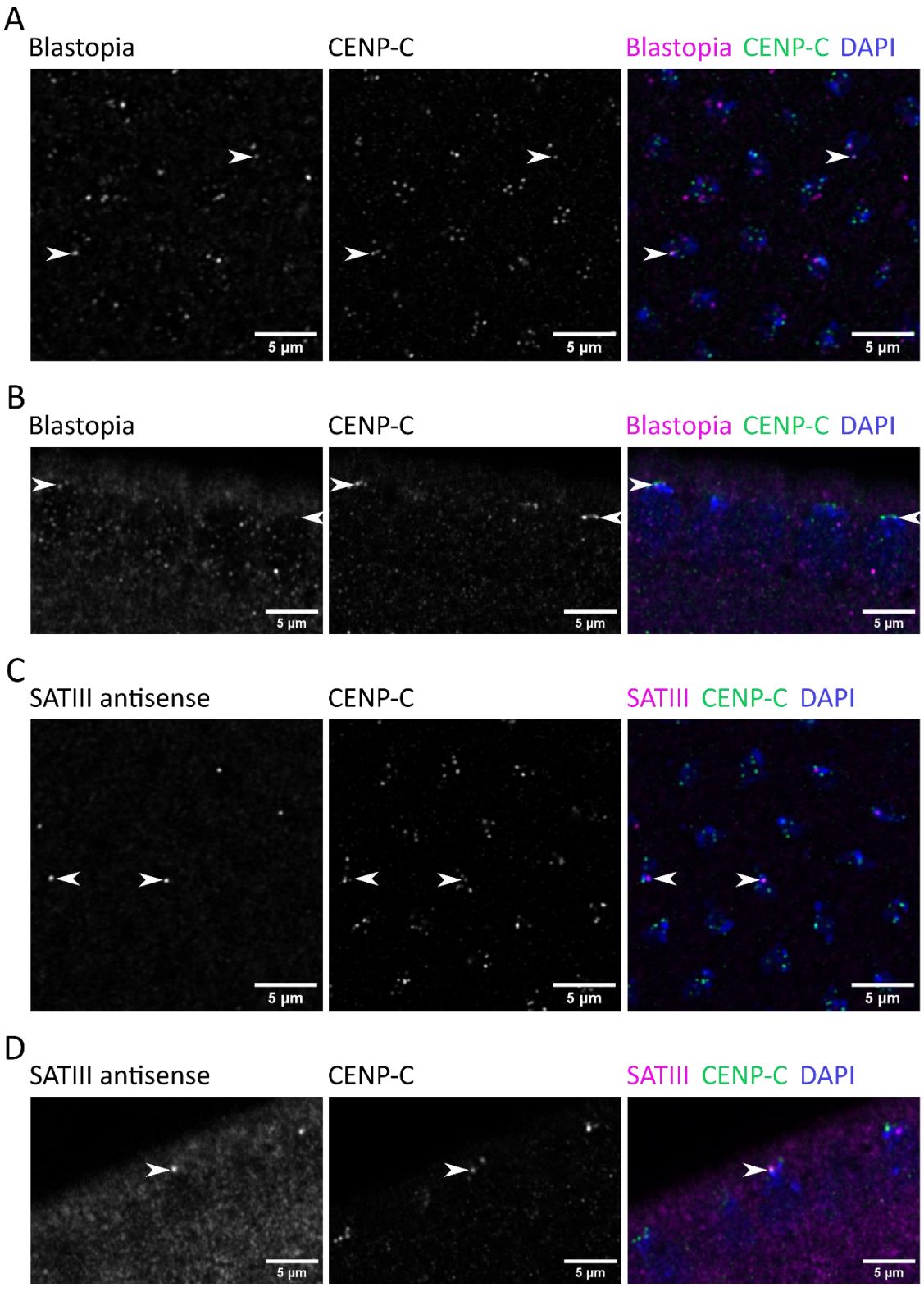


Figure 18. RNaseA treatment eliminates Blastopia smRNA-FISH signal.
Maximum intensity z-projections of the anterior side of stage 5 (left) and stage 9 (right) OregonR embryos. After fixation and permeabilization some embryos received an RNaseA treatment while others were left untreated. The subsequent Blastopia smRNA-FISH staining resulted in no signal after RNaseA treatment. Embryos were counterstained with DAPI to visualize the nuclei. The staining was performed in parallel and imaging was done on the same day. Images were concatenated while adjusting the brightness and contrast. Blastopia smRNA-FISH in gray, DAPI in blue. Scale bar = 40 µm.

2.2.4 Validation of centromeric localization of Blastopia RNA in embryos

To validate that Blastopia is a centromere-associated RNA, I performed smRNA-FISH in combination with IF for inner kinetochore protein CENP-C in stage 5 embryos (Figure 19). I also did this experiment for SATIII antisense RNA, to be able to compare Blastopia localization to the localization of a known centromere-associated RNA. In stage 5 embryos, chromosomes are organized in a Rab1 orientation, with the (peri)centromeric heterochromatin localized towards the top of cells and the euchromatin and telomeres localized basally (WALTHER ET AL., 2020). In the anterior band where Blastopia is highly expressed, all cells have Blastopia RNA foci (Figure 19A). There are many foci located at the top of the cells, in the (peri)centromeric region, but also several foci are observed in euchromatic regions (Figure 19A, B). Also outside the anterior band where Blastopia is highly expressed I can observe Blastopia RNA foci, although with lower intensity. In total, 90 % of the cells express Blastopia (Figure 20A). SATIII antisense RNA is expressed throughout the embryo, but only in 21 % of the cells (Figure 20A). Most cells have only 1 SATIII antisense RNA focus, which is always located towards the top of the nuclei (Figure 19C, D). For both RNAs I can observe foci overlapping with CENP-C foci, indicated by arrows in

Figure 19. However, the fast majority of centromeres have neither RNA localized to them and both RNAs are not exclusively found at centromeres, especially Blastopia.



Legend on next page

Figure 19. Blastopia and SATIII antisense smRNA-FISH foci co-localize with CENP-C IF foci in some cells in stage 5 OregonR embryos.
Single z-slice of stage 5 embryos stained with smRNA-FISH for Blastopia (A and B) or SATIII antisense (C and D) together with IF for inner-kinetochore protein CENP-C imaged at the top of the embryo (A, at the highly expressing anterior band, and C) or at the side of the embryo (B, at the highly expressing band, and D). To visualize the nuclei, DNA was counterstained with DAPI. Colocalizing Blastopia/SATIII-antisense and CENP-C foci are indicated by arrows. Composite images show α-CENP-C in green, smRNA-FISH in magenta and DAPI in blue. Scale bar = 5 µm.

In order to get a better understanding of the centromeric localization of both RNAs, I quantified the number of cells in which there was at least one overlapping FISH and CENP-C IF signal. As can be seen in Figure 20B, of the cells which have expression of Blastopia or SATIII, respectively 31 % and 42 % have at least one overlap in signal. However, since there are fewer cells expressing SATIII, the absolute number of cells which have an overlapping FISH and IF signal is higher for Blastopia.

Next, I determined the Manders' coefficients with JaCOP. JaCOP is a plugin for Fiji which allows for automated colocalization evaluation in three-dimensional images (BOLTE & CORDELIERES, 2006). The Manders' overlap coefficient is one of the intensity correlation coefficient-based analysis methods which uses the Pearson's correlation coefficient, but allows for variations in fluorescence intensities between the two channels that are being compared. The analysis will compute two coefficients, M1 and M2. M1 is the ratio of the 'summed intensities of pixels from channel A for which the intensity of channel B is above zero' to the 'total intensity of channel A', for M2 the channels are reversed. In other words, it will provide the proportion of smRNA-FISH signal that overlaps with CENP-C IF signal over the total smRNA-FISH intensity and vice versa. A manual threshold was applied to set background signal to zero. Both Manders' coefficients are plotted as percentages in Figure 20C. From this analysis it becomes clear that inner kinetochore protein CENP-C overlaps equally to or even slightly more with Blastopia RNA compared to known centromere-associated RNA SATIII, with 0.6 % and 0.45 % of the CENP-C IF signal overlapping respectively Blastopia and SATIII smRNA-FISH signal. However, the smRNA-FISH signal of Blastopia is far less likely to overlap with the CENP-C IF signal than SATIII antisense, with 0.67 % and 5.25 % respectively. Which was expected, since Blastopia is present as multiple foci in the cells and these foci are also located throughout the nucleus.

Together these results indicate that at least some Blastopia RNA localizes to centromeres. The number of CENP-C foci overlap with Blastopia to a similar extent as to known centromere-associated RNA SATIII. Besides this centromeric localization, Blastopia has a broader distribution throughout the nucleus, which could be derived from the euchromatic copies of Blastopia. The distinct higher expression of Blastopia in certain cell types is also surprising and could indicate a cell type specific function beside its centromeric function.

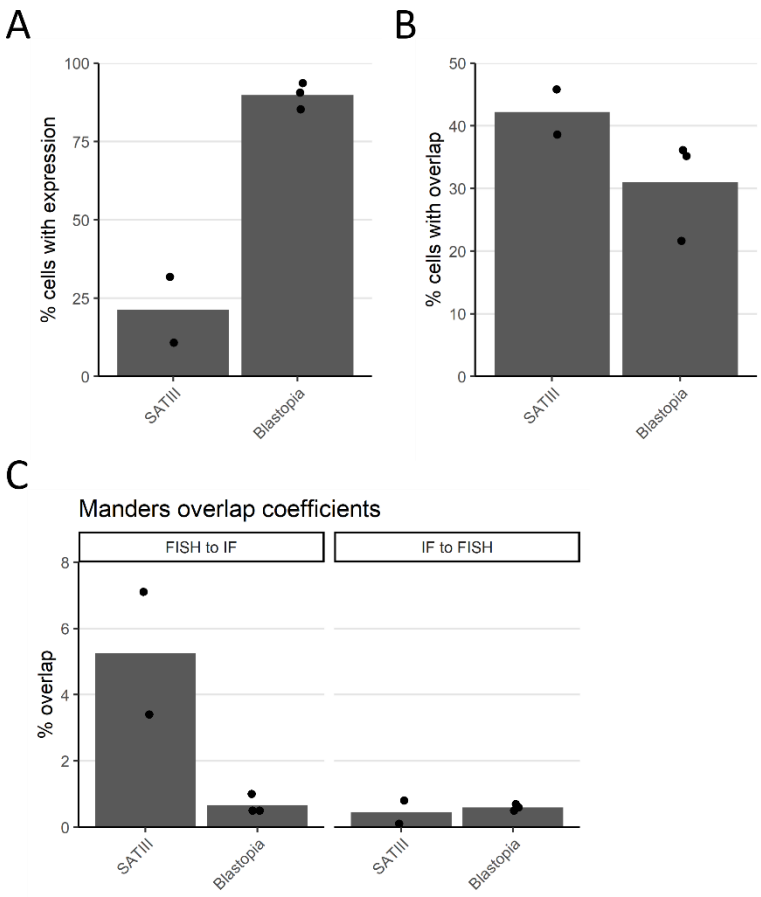


Figure 20. Quantification of Blastopia and SATIII antisense smRNA-FISH localization to centromeres.

A) Percentage of cells with SATIII antisense or Blastopia RNA expression quantified from images at the top of the embryo.

B) Percentage of cells which have at least one overlapping FISH and IF foci of the cells that have SATIII antisense or Blastopia expression. Per embryo, around 250 cells were analyzed for expression and overlap.

C) Manders' overlap coefficients that were measured with JaCOP to provide the proportion of RNA-FISH signal that overlaps with CENP-C IF signal over the total RNA-FISH intensity and vice versa. The overlap coefficients were plotted as percentages for both SATIII antisense and Blastopia.

For Blastopia 3 embryos were quantified and for SATIII 2 embryos.

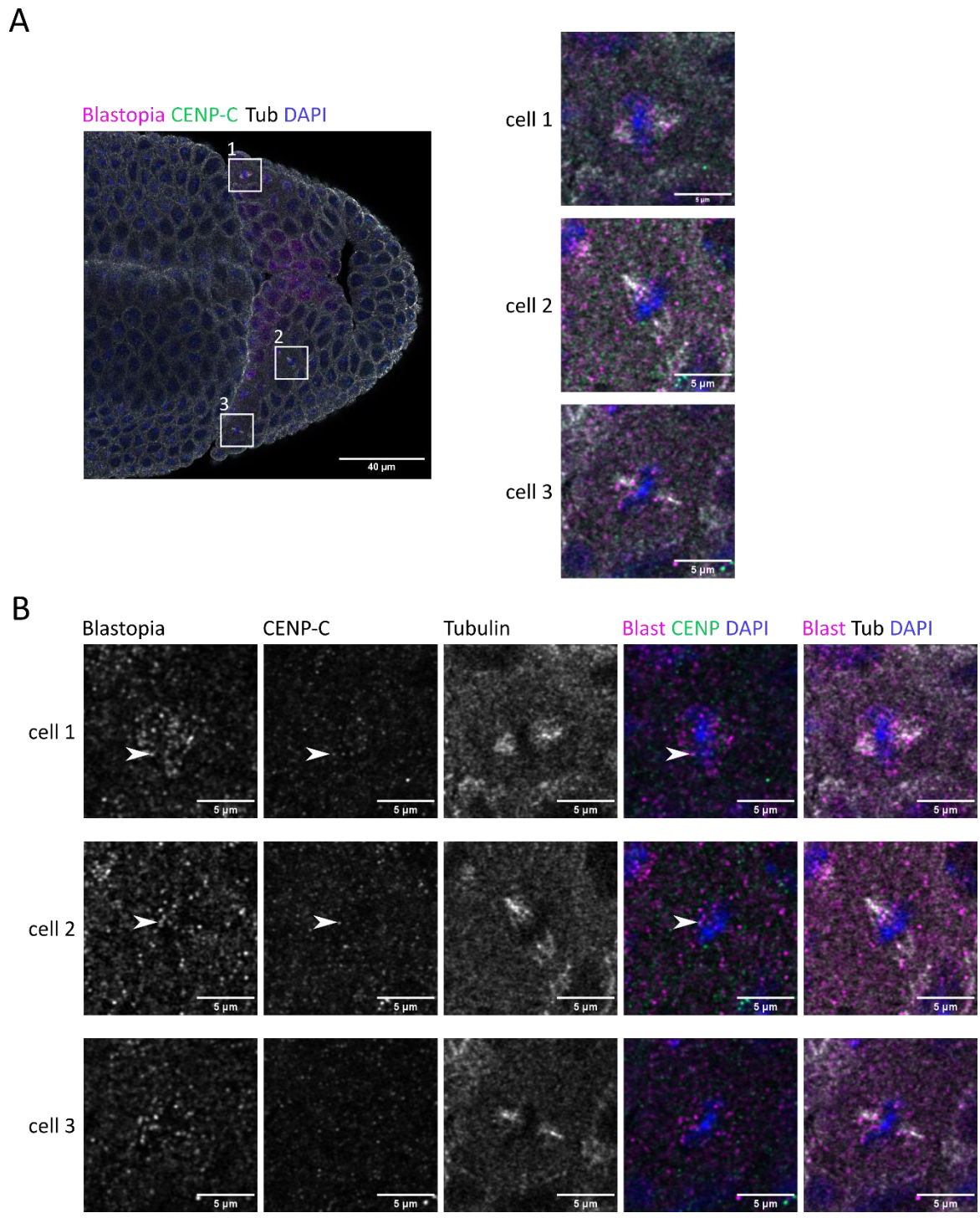


Figure 21. Blastopia smRNA-FISH localization in dividing cells of stage 8 embryos.

A) Single z-slice of a stage 8 embryo stained with smRNA-FISH for Blastopia together with IF for inner-kinetochore protein CENP-C and microtubule protein Tubulin. Boxes indicate dividing cells which are enlarged on the right.

B) Individual channels and composite images of the dividing cells from A, again depicting single z-slices. Arrows indicate the location of overlapping Blastopia FISH and CENP-C IF signal.

Composite images show Blastopia smRNA-FISH in magenta, α-CENP-C in green, α-Tubulin in gray and DAPI in blue. Scale bar = 40 µm and 5 µm for the zoomed images.

2.2.5 Localization of Blastopia RNA in dividing cells

Since at least some centromeres have Blastopia RNA located to them in interphase, it is now interesting to explore Blastopia RNA localization in dividing cells. *Drosophila* embryos undergo frequent and synchronous nuclear divisions in stage 1-4, during which mitosis is easily observed. However, during this period zygotic transcription is not yet activated and Blastopia RNA is not observed with smRNA-FISH, making this period unusable for exploring centromere-associated RNAs during mitosis. Therefore, I had to capture the less frequent and asynchronous divisions that occur during later developmental stages. To identify dividing cells, I performed Blastopia smRNA-FISH in combination with IF for inner kinetochore protein CENP-C and microtubule protein Tubulin. As can be seen in Figure 21, I was able to image several dividing cells in a stage 8 embryo in the region of high Blastopia expression. Blastopia foci were found both at the chromatin and on the microtubules, even several near microtubule ends. Unfortunately, the CENP-C staining in these dividing cells was weak, but I was still able to identify overlapping Blastopia FISH and CENP-C IF foci (arrows). These preliminary results suggest that Blastopia RNA could still be present at centromeres during mitosis. More data has to be collected in order to confirm these findings.

2.2.6 Knockdown of Blastopia and G2/Jockey-3 in embryos inefficient

The Transgenic RNA interference (RNAi) Project (TRiP) has developed vectors for tissue specific *in vivo* knockdown through short hairpin RNA (shRNA) expression (NI ET AL., 2008). The TRiP constructs contain several upstream activating sites (UAS), which can be targeted by tissue specific Gal4 driver lines to induce expression of the shRNA, and thus the knockdown, in that tissue (BRAND & PERRIMON, 1993). Transgenic TRiP fly lines are created by integrating the constructs into the genome through phiC31-mediated attB/attP recombination (GROTH ET AL., 2004). For many target genes, TRiP fly lines are commercially available, this is unfortunately not the case for TEs like Blastopia. Therefore, I created my own TRiP fly lines for Blastopia and for G2/Jockey-3, the most abundant TE in the centromeric DNA (CHANG ET AL., 2019).

I designed 21nt shRNAs to the consensus sequences of Blastopia and G2/Jockey-3 using the DSIR tool with its default settings (VERT ET AL., 2006). I performed a BLAST search of the by the tool suggested sequences against the drosophila reference and a custom BLAST database containing the TE consensus sequences, in order to exclude shRNAs with a perfect off-target match of 16 nt or more (MOORE ET AL., 2010). For each TE I selected one shRNA sequence which I cloned into a UASz-TRiP vector, a vector designed to function in both germline and somatic cells (DELUCA & SPRADLING, 2018). Both vectors were injected into *nos-int; ; attP2* embryos for phiC31-mediated integration on the 3rd chromosome (68A4) by the Cambridge University Drosophila Microinjection Service, which provided us with balanced fly lines. Additionally, the same Blastopia TRiP vector was also injected into VK33 (*vas-int; ; attP-3B)*

embryos for phiC31-mediated integration on the 3rd chromosome by Sandra Müller of the Teleman lab (DKFZ). From this injection I received the surviving larvae, with which I made single crosses to create balanced fly lines. Resulting flies were genotyped by PCR and crossed until homozygous stocks were obtained with the UASz-TRiP constructs integrated on the 3rd chromosome.

To induce a knockdown in embryos, I used a driver line with Gal4 under the control of an Armadillo promoter (Arm-Gal4). Arm-Gal4 is ubiquitously expressed in all tissues of embryos of all developmental stages from early gastrulation on (VINCENT ET AL., 1994). Arm-Gal4 virgins were crossed with males of the Blastopia TRiP lines that have the construct integrated at two different loci on the 3rd chromosomes, the G2/Jockey-3 TRiP line and several controls. As negative controls I used a TRiP-mCherry line, a GFP dsRNA line with the construct on the 3rd chromosome and W1118 wildtype line. A TRiP line for a protein, CG32344, was used as a positive control. The crosses are kept at 29°C, a temperature at which the yeast UAS/Gal4 system works most effectively. It takes time for the Gal4 transcription factor to get expressed after zygotic transcription activation, which then has to induce the expression of the shRNA to subsequently induce the knockdown. Therefore, I used embryos from an overnight collection to ensure the vast majority of the embryos have had time to induce the knockdown. With the RNA isolated from these embryos, I performed an RT-qPCR to assess the knockdown efficiency (Figure 22). I used primers against the Blastopia consensus sequence to assess Blastopia expression levels in the different crosses (males of the crosses indicated on the x-axis). Even though Blastopia RNA levels in embryos from crosses with the two Blastopia TRiP lines are less than half compared to the W1118 cross, the levels are similar in the other TRiP crosses. In general, Blastopia expression levels are extremely variable in the different control crosses. For G2/Jockey-3 I also had designed a primer pair against its consensus sequence, however the Ct values with this primer pair are too low and thus unreliable. Therefore, I used two published primer pairs which recognize centromere specific copies of G2/Jockey-3, Cen4-G2 is specific for the centromere of chromosome 4 and CenX-G2 is specific for the centromere of chromosome X (CHANG ET AL., 2019). Embryos from the Arm-Gal4 x TRiP-G2/Jockey3 cross did not show a reduction in either Cen4-G2 or CenX-G2 RNA levels compared to the other TRiP crosses and the W1118 and GFP control crosses. Again, embryos from the TRiP-mCherry cross have lower levels of both Cen4-G2 and CenX-G2. The positive control CG32344 expression levels are halved in the TRiP-CG32344 knockdown cross compared to the W1118 cross. Although not to the same extend, the other crosses do also have lower levels of CG32344 RNA compared to the W1118 cross. I also checked the RNA levels of two not targeted repeats, SATIII and DOC2, the latter with published primer pair against its centromeric copy on chromosome X. SATIII showed a reduction in all the TRiP crosses compared to the W1118 cross. CenX-DOC2 levels did not

change between the different crosses. Taken together these results suggest that I was not able to induce a specific or effective knockdown of Blastopia and G2/Jockey-3 in embryos.

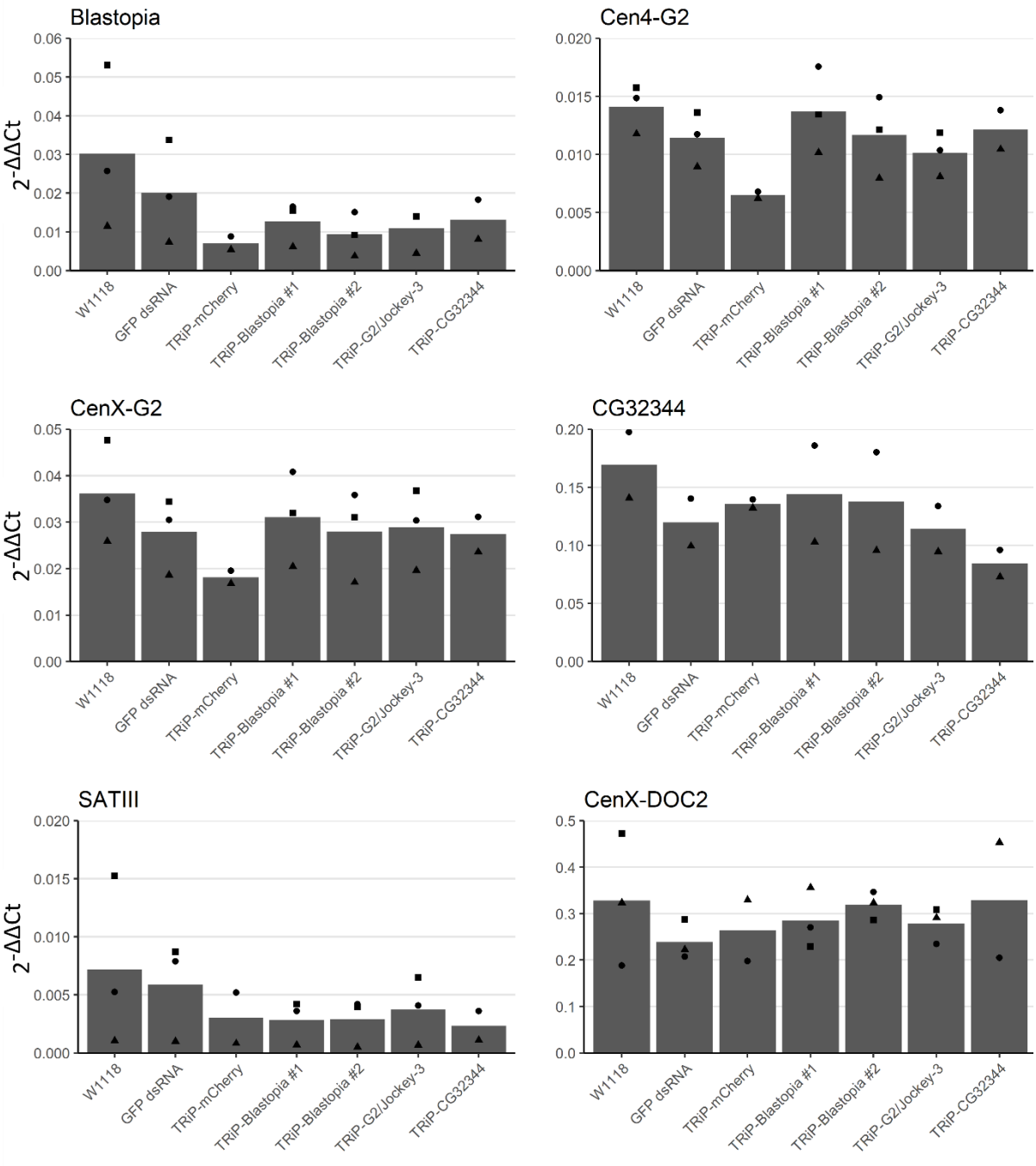


Figure 22. Blastopia and G2/Jockey-3 knockdown not efficient in embryos.
Average relative RNA levels in embryos, collected overnight at 29°C, normalized to housekeeping genes Actin, GAPDH and rp49/RpL32. Males from the on the x-axis indicated fly lines were crossed with Arm-Gal4 virgins to induce a knockdown in all embryonic tissues. Dots indicate RNA levels of the individual replicates, N = 3.

Despite these inefficient knockdowns, I checked survival rate of the embryos laid by these crosses. Again Arm-Gal4 virgins were crossed with males of wildtype W1118, GFP dsRNA control line, the two

TRiP-Blastopia line and the TRiP-G2/Jockey-3 line at 29°C. The percentage of hatched embryos from these crosses are shown in Figure 23A. No differences in embryo survival is observed. However, due to the inefficient knockdown I am not able to draw any conclusions from these results.

I also used the Maternal Triple Driver Gal4 line (MTD-Gal4), which has three copies of Gal4, two controlled by a nos promoter and one by an otu promoter (MAZZALUPO & COOLEY, 2006). This driver line can induce a knockdown in the female germline of the F1 flies and as a consequence in stage 1-4 embryos of the F2 generation. I crossed MTD-Gal4 flies with W1118, GFP dsRNA, TRiP-Blastopia and TRiP-G2/Jockey-3 fly lines and subsequently crossed the resulting F1 virgins with W1118 males at 29°C. There were no differences in the number of embryos laid by this cross (not shown). The percentage of hatched embryos from these crosses are shown in Figure 23B. No differences in embryo survival is observed. I did not check the knockdown efficiency by RT-qPCR for these crosses. Blastopia RNA levels in stage 1-4 embryos were also not detectable in the above Blastopia smRNA-FISH experiments.

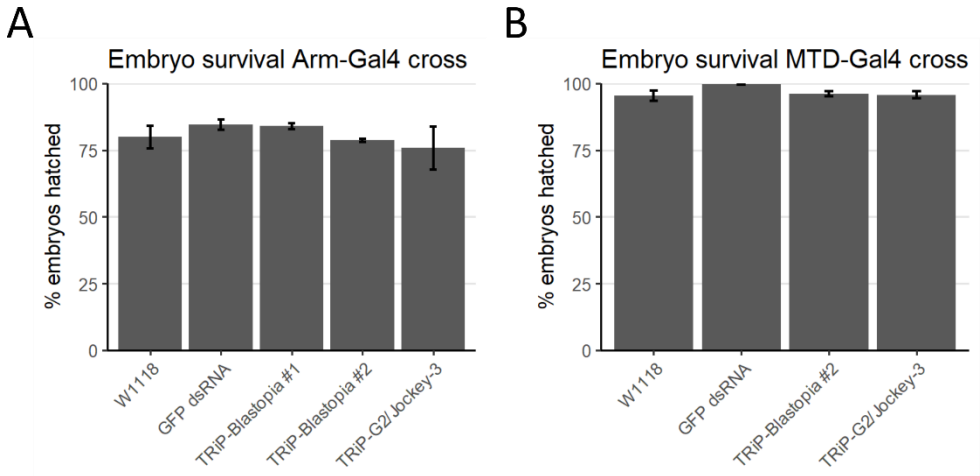


Figure 23. Embryos survival not affected by inefficient Blastopia or G2/Jockey-3 knockdown.

A) Hatch rate of embryos laid by Arm-Gal4 crossed with the indicated TRiP or control males at 29°C. N=4.

B) MTD-Gal4 flies were crossed with the indicated TRiP and control line to induce a knockdown in the female germline. The graph depicts the hatch rate of embryos laid by F1 females of those crosses which were all crossed with W1118 males at 29°C. N=3.

Last, I used the Gal4 driver line which was under the control of an Act5C promoter, which is a very strong and ubiquitous promoter (ITO ET AL., 1997). I crossed Act5C-Gal4 virgins with males of the two Blastopia TRiP line and all the mentioned negative controls. There was no observable difference in number of offspring, developmental timing of this offspring and their feeding or flying behavior (not shown). The offspring was allowed to mate with their siblings, which resulted in a normally developing F2 generation (not shown). Again, I did not check the knockdown efficiency by RT-qPCR for these crosses. All this suggest that either the knockdown is not working even with this strong driver line, or a Blastopia knockdown has no noticeable effect of development, survival and fertility.

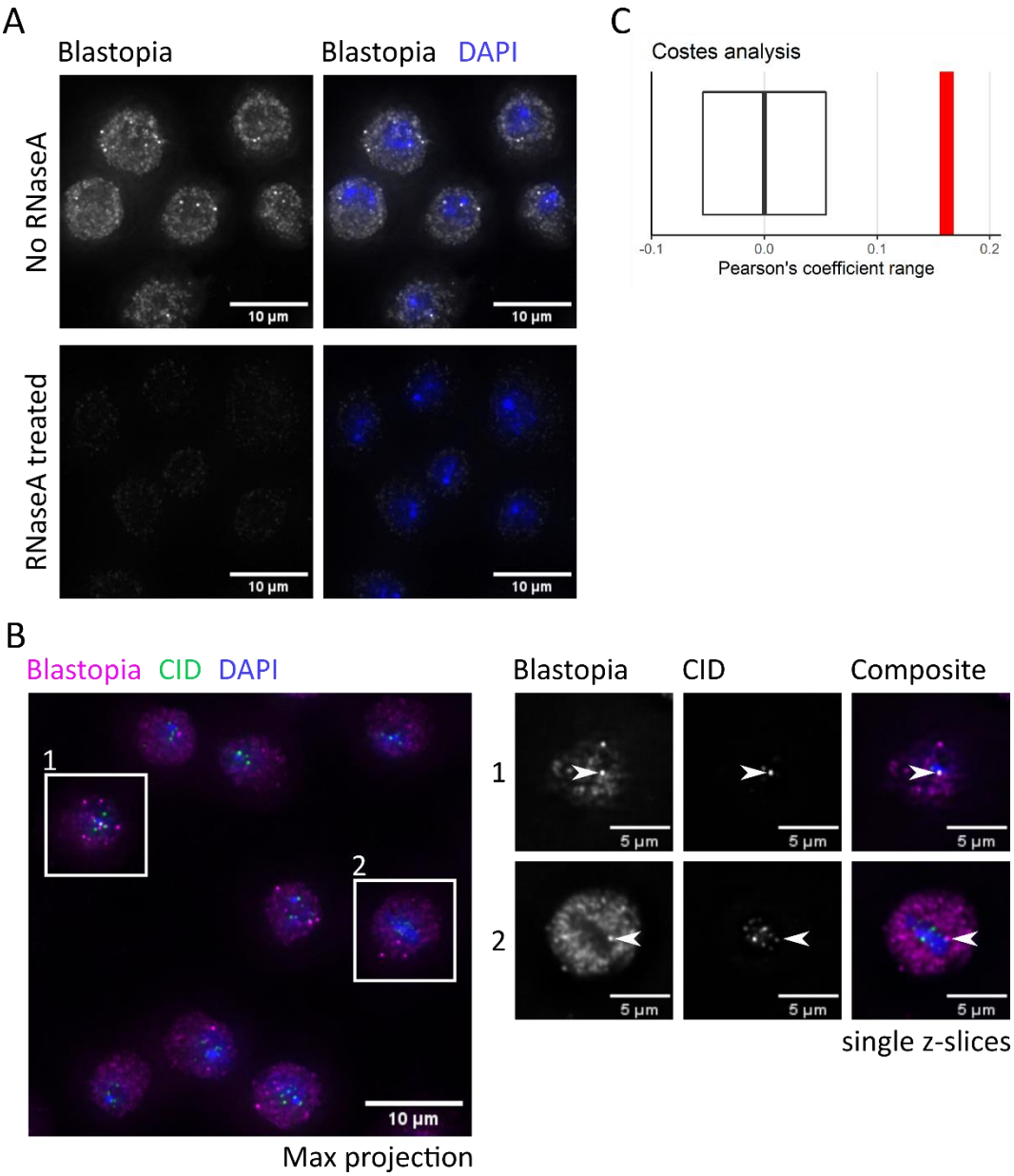


Figure 24. Blastopia smRNA-FISH and CID IF co-localize in some S2 cells.

A) Max intensity projection of Blastopia smRNA-FISH (gray) in S2 cells, counterstained with DAPI (blue) to visualize the nuclei. RNaseA treatment (bottom) eliminates the FISH signal that is visible without that treatment (top). Scale bar = 10 µm.

B) Left: Max intensity projection of Blastopia smRNA-FISH (magenta) and α-CID IF (green) in S2 cells, counterstained with DAPI (blue) to visualize the nuclei. Boxes highlight two cells which have Blastopia localized to the centromeres. Scale bar = 10 µm. Right: Single z-slices of cell marked by the boxes on the left. Arrows indicate overlapping Blastopia smRNA-FISH and CID foci. Of the 306 cell quantified, 19.4 % display a co-localization. Composite images show α-CID in green, Blastopia smRNA-FISH in magenta and DAPI in blue. Scale bar = 5 µm.

C) Costes' overlap analysis show a statistically significant overlap between Blastopia smRNA-FISH signal and CID IF singal in S2 cells. The red bar indicates the average pearson's correlation between the two channels from 18 images analyzed, being 0.162. This is significantly different, with an average p-value = 0.036, from the correlation range computed when one of the two channels is randomized 500 times (box, with middle black line the average randomized correlation being 0, range -0.055 to 0.055).

2.2.7 Blastopia RNA localizes at centromeres in S2 cells

Next, I explored Blastopia RNA localization in S2 cells. I confirmed the specificity of the Blastopia staining for RNA with an RNaseA treatment, as presented in Figure 24A. Multiple Blastopia RNA foci are present in both the nucleus and cytoplasm of S2 cells and these foci disappear after treatment with RNaseA. I performed Blastopia smRNA-FISH in combination with CID IF to assess the centromeric localization of Blastopia in S2 cells (Figure 24B). As with embryos, Blastopia RNA localizes to the centromere in only a subset of cells, 19.4 % of S2 cells show an overlap of the Blastopia smRNA-FISH and CID IF signal. This overlap can be seen in both interphase cells (example cell 1) and mitotic cells (example cell 2).

To quantify this overlap, I used the Costes' approach of automated colocalization evaluation provided by JaCOP, which is more appropriate when there is only a single condition available (BOLTE & CORDELIERES, 2006). Costes' overlap analysis calculates the Pearson's coefficient to describe the correlation between the pixel intensities of two channels, in this case the Blastopia smRNA-FISH signal and the CID IF signal. Next, JaCOP creates 500 randomized images of one of the two channels and computes the range of Pearson's coefficients to this set of randomized images, and whether the original Pearson's coefficient is significantly different. The average Pearson's coefficient between Blastopia and CID is 0.162, which corresponds to the occasional co-localization described above (Figure 24C). The average correlation range from the randomized images was -0.055 to 0.055, with a mean of 0. The original correlation was significantly different from this correlation range (p-value = 0.036). This would suggest that the observed colocalization is too high to be dismissed as chance. Overall, these preliminary results suggest that also in a subset of S2 cells, Blastopia is associated with centromeres.

2.2.8 Blastopia RNAi in S2 cells

In order to assess whether Blastopia RNA is important for centromere function or maintenance in S2 cells, I induced a knockdown of Blastopia. I designed four different dsRNAs targeting different regions of the Blastopia transcript with SnapDragon long dsRNA designing tool (Figure 14C), all around 400 bp long (FLOCKHART ET AL., 2012). As controls I used a dsRNA against G2/Jockey-3 and GFP. G2/Jockey-2 RNAs are undetectable in S2 cells by both smRNA-FISH and RT-qPCR. Cells were incubated with these dsRNAs to induce RNAi for 4 days before assessing knockdown efficiency with RT-qPCR and centromere function and maintenance with IF. In Figure 25A there is a clear reduction visible in Blastopia RNA levels with all four dsRNAs that target Blastopia. Especially dsRNA #1 reduces Blastopia RNA levels to 21.9 % of the GFP control knockdown levels. I also performed RT-qPCR for SATIII, which showed that Blastopia dsRNA #1 did not affect pericentromeric sequences in general.

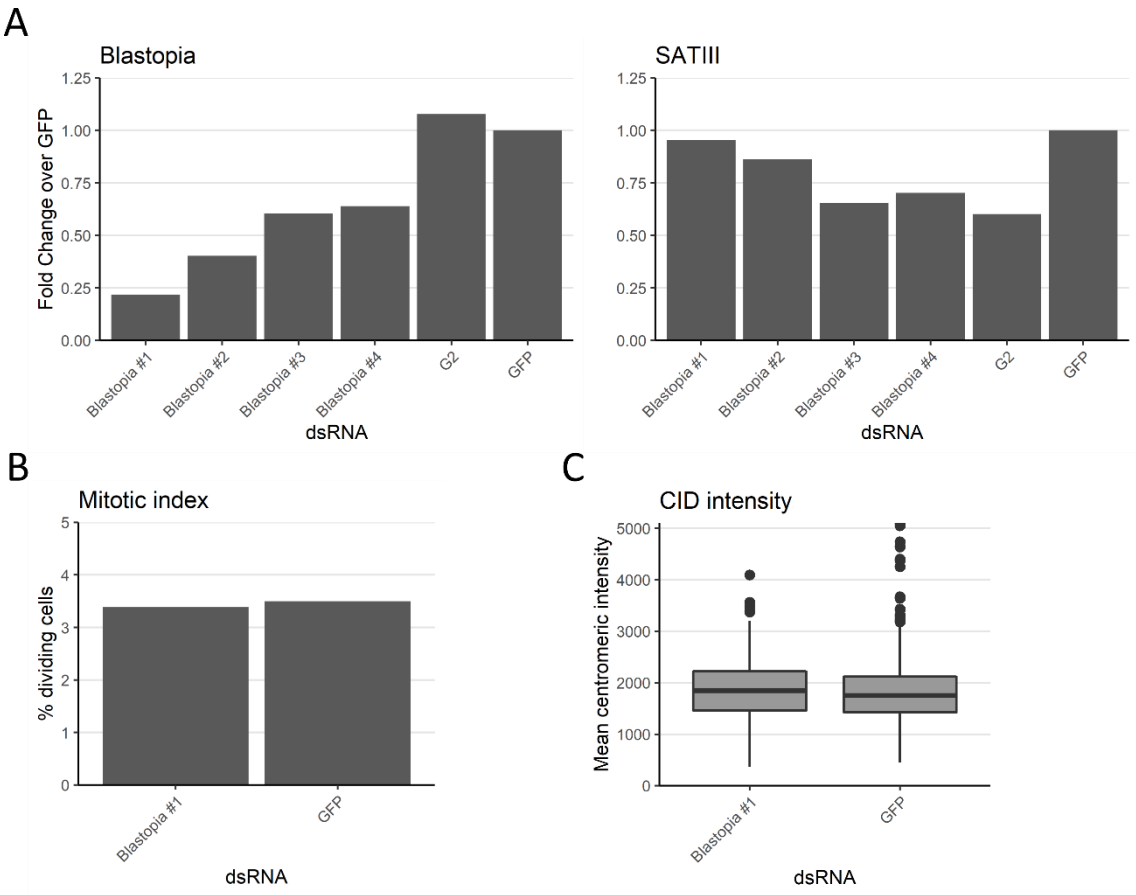


Figure 25. Blastopia RNAi efficient in S2 cells, but has no effect on mitosis.

A) RT-qPCR shows decreased Blastopia RNA levels in S2 cells treated for 4 days with four different Blastopia dsRNAs, compared to a GFP control dsRNA and an unrelated G2/Jockey-3 targeting dsRNA. Specifically Blastopia dsRNA #1 reduces the Blastopia RNA levels to 21.9 % of the cells treated with GFP control dsRNA. SATIII RNA levels are unaffected by the RNAi.

B) Mitotic index determined by the presence of mitotic marker H3 phosphor Ser10. Cells treated for 4 days with either Blastopia dsRNA #1 or GFP dsRNA were immuno-stained for H3 phosphor S10. Of the 709 and 914 cells counted, 3.39 % and 3.50 % were positive for this marker for the Blastopia RNAi and GFP RNAi, respectively.

C) Mean centromeric CID intensity per cell was determined in cells treated for 4 days with either Blastopia dsRNA #1 or GFP dsRNA. For Blastopia RNAi 321 cells were quantified, for GFP RNAi 262 cells.

The cells with RNAi Blastopia #1 and GFP control are immuno-stained with mitotic marker phosphorylated H3Ser10 in order to determine the mitotic index. As can be seen in Figure 25B, the number of cells undergoing mitosis is similar in cells treated with Blastopia dsRNA #1 or the control GFP dsRNA. Separately, I imaged 65 dividing cells for the Blastopia #1 RNAi and 50 dividing cells for the GFP control RNAi, which were identified by Tubulin immuno-staining. I was unable to observe an increase in mitotic defects in the Blastopia RNAi cells compared to the GFP RNAi cells (not shown). Unperturbed S2 cells have many mitotic defects of themselves, so potentially a larger sample size is

required to observe defects caused by the Blastopia knockdown. Especially considering that in only a fraction of cells Blastopia is centromere-associated. Next, I quantified the average CID intensity per cell (Figure 25C). Blastopia knockdown did not reduce overall centromeric CID intensities compared to the control knockdown. However, SATIII knockdown only affected CID maintenance in a subset of centromeres, this could also be the case for Blastopia (ROŠIĆ ET AL., 2014).

Even though these experiments have only been performed once, they suggest that longer dsRNAs are able to induce a successful knockdown of Blastopia in S2 cells. Initial observations did not indicate Blastopia knockdown results in severe disruptions of centromere function or maintenance.

2.3 Blastopia in other tissues

2.3.1 Re-analyzing modENCODE RNA-seq data

As described above, Blastopia is highly expressed in a specific subset of cells in embryos. Although this is probably independent from Blastopia's potential centromeric function, I was interested to explore this further. Very little is known about the expression patterns of repetitive elements like TEs, since repeats are usually removed while analyzing RNA-seq data. One of the most comprehensive RNA-seq datasets published for *Drosophila melanogaster* is from the modENCODE consortium, which sequenced RNA from thirty different developmental stage and nine different tissues from many of those developmental stages (DUFF ET AL., 2015). The expression patterns described by the modENCODE data is used as an important resource by the *Drosophila* research community.

I re-analyzed this dataset, mapping reads with STAR (DOBIN ET AL., 2013) to the Mellone and Larracuente labs reference genome (CHANG ET AL., 2019) while allowing for multimappers. I used TEcount (JIN ET AL., 2015) to count reads that mapped to repetitive elements. Counts were normalized to the sequencing depth of each sample (reads per million (RPM)). It is important to remember that the different repetitive elements are present in the genome with different copy numbers. Since the assembly of the heterochromatin, which harbors the majority of TEs and satellite repeats, is still missing, the exact copy numbers are unknown. Therefore, further normalization which takes into account "gene length", such as RPKM, is impossible. Comparing expression levels of different repetitive elements is for the same reasons problematic, high copy numbers can result in high read counts even with low expression levels. For this reason, I focused on the expression pattern of each repetitive element across the different samples, for which the copy numbers are stable. In order to visualize the expression patterns, without being overshadowed by highly expressing elements, I transformed the normalized counts to Z-scores for each repetitive element individually.

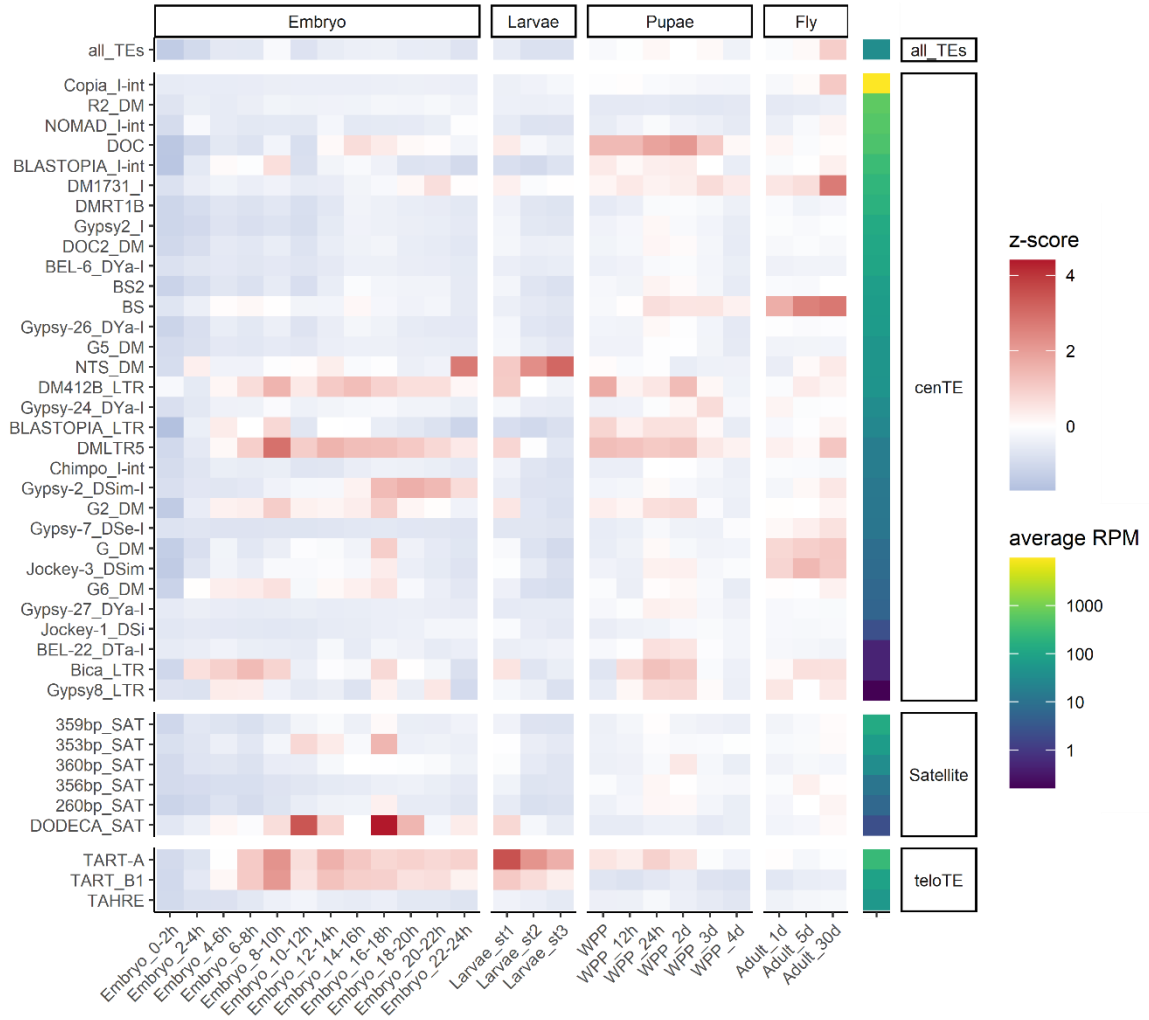


Figure 26. Expression pattern of TEs and satellite repeats in whole animals across development. Re-analysis of the modENCODE RNA-seq datasets of RNA isolated from entire animals at different developmental stages for TEs and satellite repeats. Counts were normalized to the sequencing depth of each sample, resulting in reads per million (RPM). For each gene the normalized counts are transformed to Z-scores so that the expression patterns are not overpowered by the expression levels. For each sample there were 2 replicates, which were averaged before plotting the z-score (blue to red fill) for all indicated repeats. The top row shows the results of all repetitive elements with counts combined in an average per sample, which can be used as a reference. The TEs and satellite repeats are grouped per centromeric TEs (cenTE), satellite repeats (Satellite) and telomeric TEs (teloTE), and are arranged by decreasing average RPM across all the samples. This average RPM per TE/repeat is also plotted in the final column in a log10 scale (blue to yellow fill).

I plotted the expression patterns across development of the cenTEs, satellite repeats and telomeric TEs in Figure 26. In the top row, I plotted the average of all annotated repetitive elements with at least one count as a reference. The repetitive elements are sorted within each group by descending average expression levels, with the average RPM depicted in the rightmost column (blue to yellow scale). The expression level of Copia is exceptionally high compared to the other repetitive elements, a 10-fold

higher than other cenTEs. The main component of the plot depicts the Z-scores of each repetitive element per sample (blue to red scale). It can be seen that most elements have a stable expression pattern throughout developments. Unsurprisingly, the lowest scores are often found in 0-2 h embryos, where zygotic transcription is not yet activated. Furthermore, expression levels of most repetitive elements increase in 30 day old adults. This corresponds with previous reports on activation of TEs with age (WOOD ET AL., 2016). Two telomeric transposable elements are higher expressed during earlier development, in embryos, larvae and pupae. And while other satellite repeats have a stable expression throughout development, Dodeca shows a striking deviation from this pattern with two waves of high expression in 10-12 h embryos and 16-18 h embryos.

Turning now to the expression patterns in different tissues, it can be seen that certain tissues have a general upregulation of TE expression (Figure 27). The data was processed together with the developmental data, so again Z-scores (blue to red scale) are used to visualize expression pattern across samples. In ovaries and testes, TEs and repeats are expressed at lower levels than average. This is to be expected, TE activity is tightly regulated by the gonad specific piRNA pathway to prevent inheritance of TE insertion defects through germ cells. Most TEs and repeats are highly expressed in the central nervous system, 3rd instar larvae imaginal discs and 2 day old pupae fat bodies. TEs have been shown to play a role in neuronal development, creating an extra layer of neural cell diversity (AHMADI ET AL., 2020; TREIBER & WADDELL, 2020). In several species, including *Drosophila,* TE derepression has been linked to neurodegenerative diseases (RAVEL-GODREUIL ET AL., 2021). Interestingly, Copia, the cenTE with by far the highest expression level, is the most active in the digestive system.

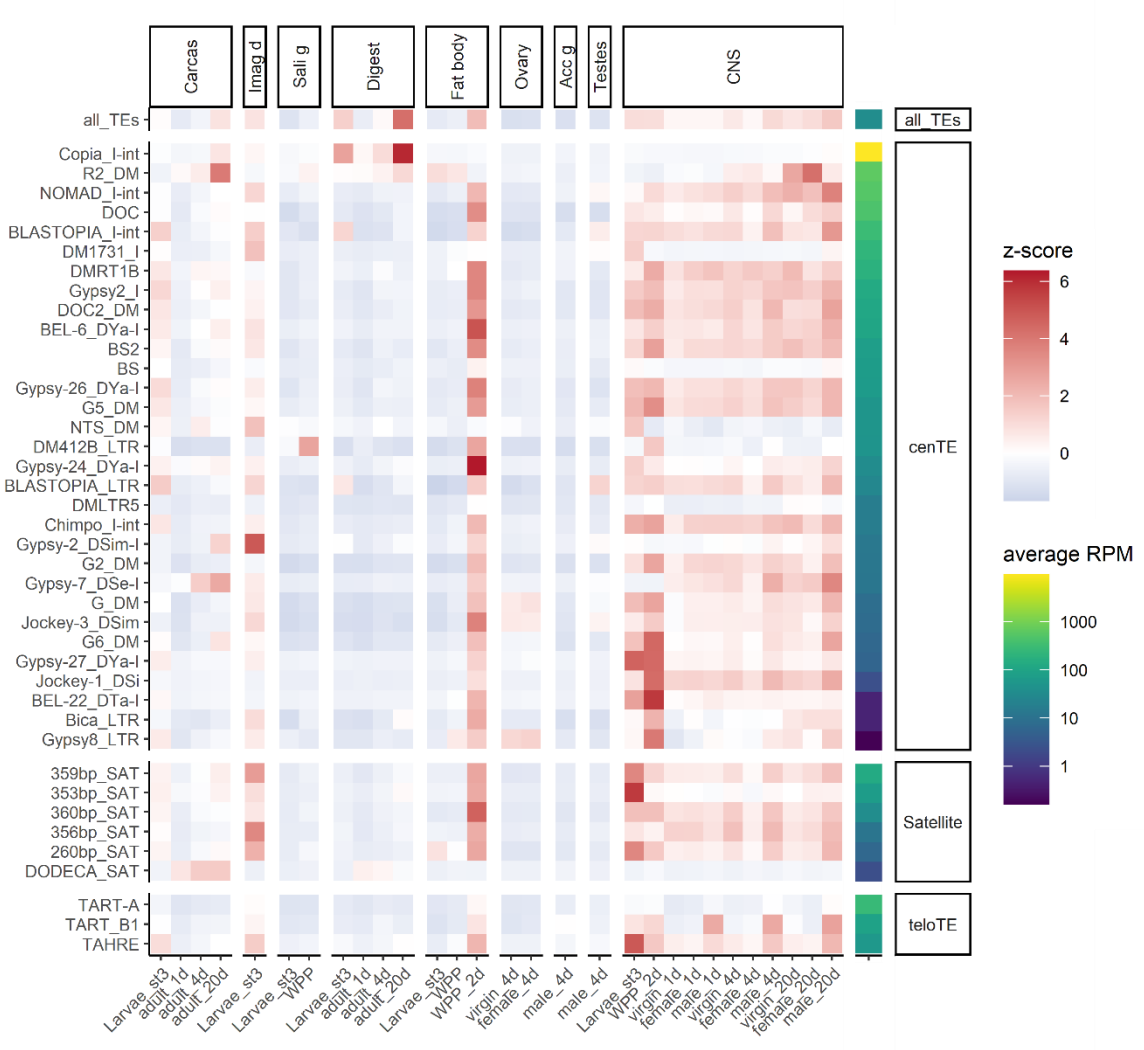


Figure 27. Expression pattern of TEs and satellite repeats in indicated tissues across development. Re-analysis of the modENCODE RNA-seq datasets of RNA isolated from the indicated tissues for TEs and satellite repeats. Counts were normalized to the sequencing depth of each sample, resulting in reads per million (RPM). For each gene the normalized counts are transformed to Z-scores so that the expression patterns are not overpowered by the expression levels. For each sample there were 2 replicates, which were averaged before plotting the z-score (blue to red fill) for all indicated repeats. The top row shows the results of all repetitive elements with counts combined in an average per sample, which can be used as a reference. The TEs and satellite repeats are grouped per centromeric TEs (cenTE), satellite repeats (Satellite) and telomeric TEs (teloTE), and are arranged by decreasing average RPM across all the samples. This average RPM per TE/repeat is also plotted in the final column in a log10 scale (blue to yellow fill). Imag d = Imaginal Discs, Sali g = Salivary Glands, Digest = Digestive System, Acc g = Accessory Glands, CNS = Central Nervous System.

Blastopia follows the expression patterns of most other cenTEs. In Figure 28, normalized counts of Blastopia are plotted across development and in indicated tissues. Blastopia has a higher expression in the central nervous system of all developmental stages. Which is in line with the smRNA-FISH observations in embryos described above. Blastopia expression has also been previously described in

adult heads(DELOGER ET AL., 2009). Also 3rd instar larvae imaginal discs and 2 day old pupae fat body have higher normalized counts, as was seen with the other cenTEs. To a lesser extent, 3rd instar larvae carcass and digestive system, and 4 day old testes are also slightly elevated.

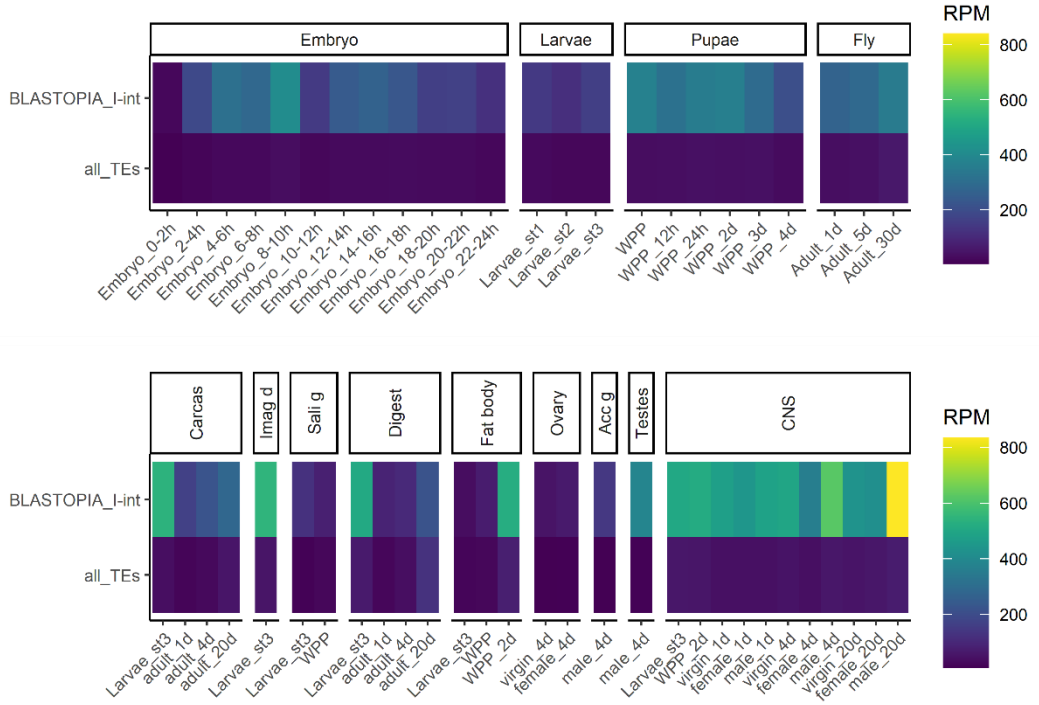


Figure 28. Expression pattern of Blastopia in whole animals and indicated tissues across development shows increased expression in the Central Nervous System (CNS).
Re-analysis of the modENCODE RNA-seq datasets of RNA isolated from the indicated tissues for Blastopia. Counts were normalized to the sequencing depth of each sample, resulting in reads per million (RPM), indicated with the blue to yellow fill. Imag d = Imaginal Discs, Sali g = Salivary Glands, Digest = Digestive System, Acc g = Accessory Glands, CNS = Central Nervous System.

2.3.2 Blastopia RNA localization in 3rd instar larvae brain

With the central nervous system expressing Blastopia throughout development, I set out to explore Blastopia localization in 3rd instar larvae brain. The 3rd instar larvae brain was chosen since it is an easily accessible tissue. I performed smRNA-FISH for Blastopia in combination with IF for neuronal marker Elav (ROBINOW & WHITE, 1991) and glial cell marker Repo (XIONG ET AL., 1994). I observed high Blastopia expression in some cells in the brain lobes, the ventral nerve cord and eye discs. Unfortunately, the antibody staining did not penetrate properly into the tissue, especially in the brain lobes. However, in the ventral nerve cord I was able to observe that Blastopia expressing cells were stained with neuronal marker Elav (Figure 29, top), and not with glial cell marker Repo (not shown). In these cells, signal is observed in both the nucleus and the cytoplasm. RNaseA treatment once more showed the smRNA-FISH probes are specifically hybridizing with RNA molecules (Figure 29, bottom).

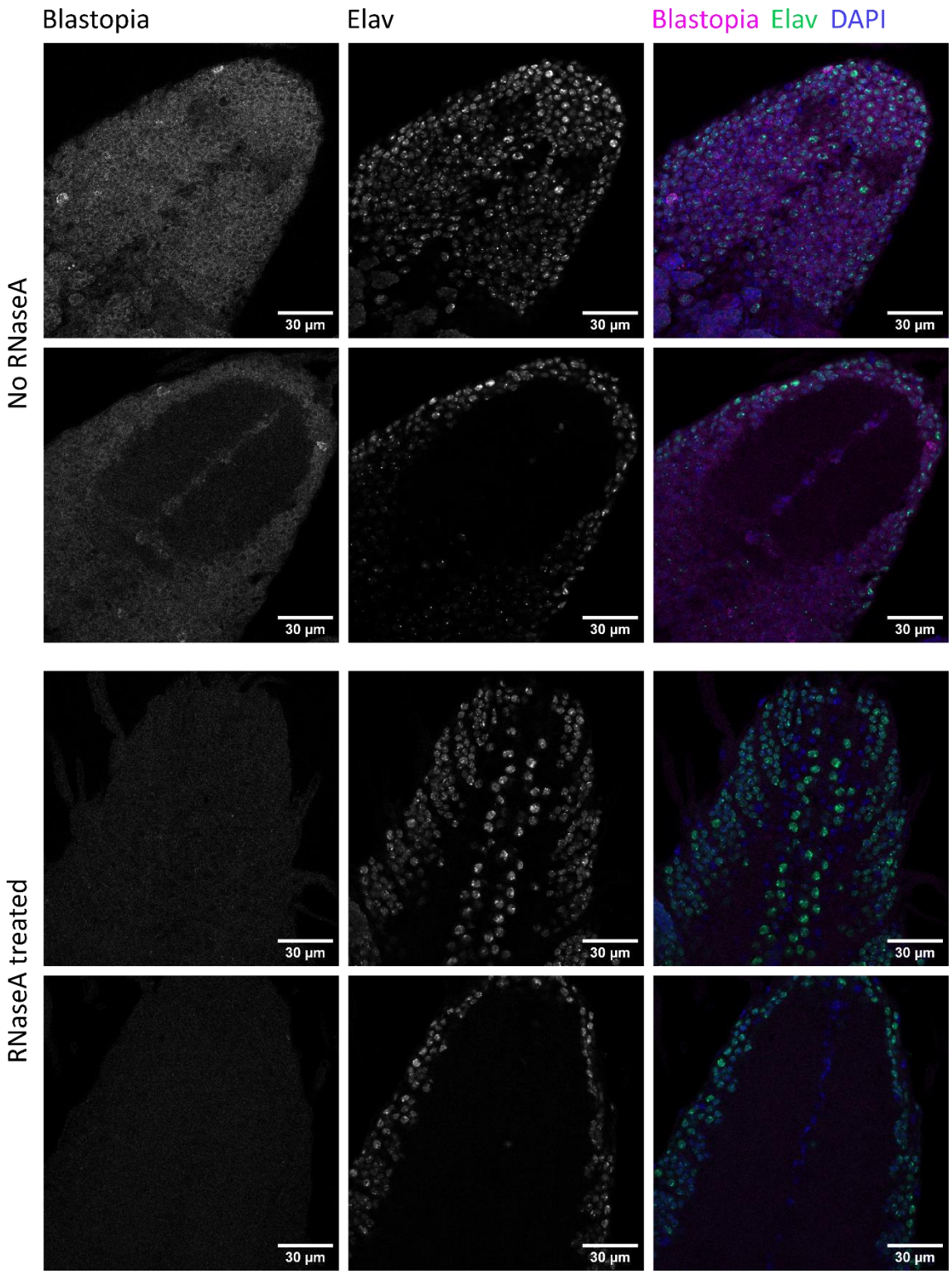


Figure 29. Blastopia smRNA-FISH expressing cells are labeled with neuronal marker Elav.
Single z-slices of 3rd instar larvae ventral nerve cord stained with Blastopia smRNA-FISH and α-Elav IF. Top: several cells express high levels of Blastopia, all these cells also show α-Elav IF staining. Bottom: after RNaseA treatment no smRNA-FISH signal is detected. Composite images show α-Elav in green, Blastopia smRNA-FISH in magenta and DAPI in blue. Scale bar = 30 µm.

In order to confirm that neurons are the main Blastopia expressing cells, also in the brain lobes, I made use of the UAS/Gal4 system to express GFP in neurons. Two Elav-Gal4 fly lines were used to drive the expression of UAS-NLS-GFP in post mitotic neurons, Bloomington number 458 and 8765 (OGIENKO ET AL., 2020). Line 458 is an enhancer-trap Gal4 line, with Gal4 inserted into the promoter region of the Elav gene, known as the elav^{C155} insertion. In line 8765, the Elav promoter region has been inserted on a separate chromosome together the Gal4 sequence. Both Elav-Gal4 drivers have a slightly different neuronal expression pattern, potentially due to some missing enhancer sequences, both not covering all neuronal cells. Figure 30 shows 3rd instar larval brains with neuronal cells expressing GFP, stained for Blastopia with smRNA-FISH. As with the co-staining described above, some neurons are expressing high levels of Blastopia in the ventral nerve cord. Blastopia smRNA-FISH signal is visible in Elav-Gal4 #458>GFP cells. Also in the brain lobe, there is Blastopia signal in some Elav-Gal4 #8765>GFP positive cells, particularly in the lamina plexus of the optical lobe. However, there are also some cells in the brain lobes with high Blastopia signal, which do not have GFP expressed by either driver.

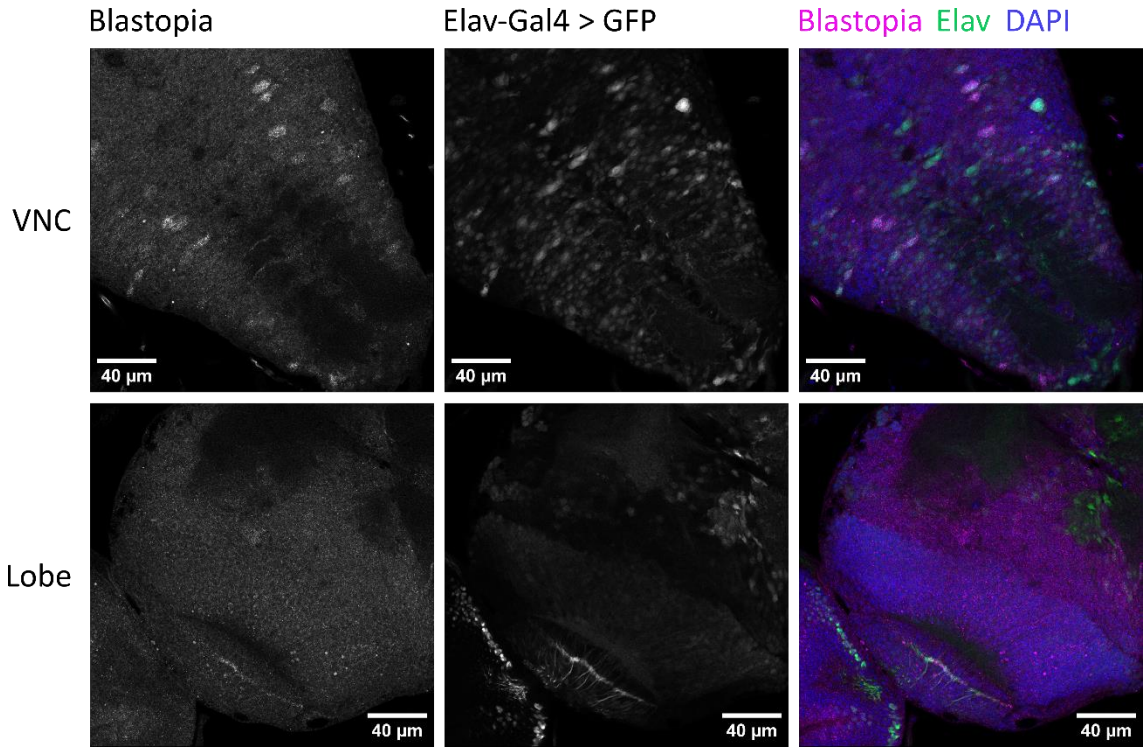


Figure 30. A subset of neurons express Blastopia in 3rd instar larvae brains.
Single z-slices of Blastopia smRNA-FISH in 3rd instar larvae brains expressing NLS-GFP in post mitotic neurons. UAS-NLS-GFP expression was driven by two Elav-Gal4 fly lines, top Bloomington Elav-Gal4 fly line #458, bottom Bloomington Elav-Gal4 fly line #8765. Composite images show GFP in green, Blastopia smRNA-FISH in magenta and DAPI in blue. Scale bar = 40 µm.

2.4 Identifying RNA-binding centromeric proteins

There are strong indications that both inner-kinetochore protein CENP-C and CID chaperone Cal1 are able to directly bind RNA. A direct interaction between either of these proteins and RNA could provide an explanation on how RNA functions at centromeres. I made use of a recently published method for isolating RNA-protein complexes from cells, called protein-Xlinked RNA eXtraction (XRNAX) (TRENDEL ET AL., 2019). The XRNAX extract was used to assess whether CENP-C and Cal1 are indeed direct RNA binders and as input for a pull down to identify which RNAs are bound.

2.4.1 XRNAX with S2 cells

XRNAX uses TRIzol to isolated RNA-protein complexes from UV-cross-linked cells (Figure 31). I adapted the method, which was initially developed for human cells, to work for *Drosophila* S2 cells. A monolayer of cells was exposed to UV light to create zero-distance covalent bonds between directly interacting proteins and RNAs (RAMANATHAN ET AL., 2019). Although UV is a less efficient crosslinker than formaldehyde, it is more specific in that it only cross-links directly interacting RNA-protein complexes at the light intensity used. So while the CID RNA-ChIP revealed the RNAs enriched at the centromeric region, with this method I can clarify the RNA-binding capacities of individual centromeric proteins. Cells are lysed in TRIzol and chloroform is added to separate the organic phase, containing denatured proteins, and aqueous phase, containing RNA. Cross-linked RNA-protein complexes get trapped in the interphase, together with the DNA. This interphase is dissolved and the DNA is digested away. After precipitation, the XRNAX extract contains highly concentrated RNA-protein complexes with some loose RNA and protein. The efficiency of the isolation was confirmed for each XRNAX extract with two quality control digestions, as shown in Figure 32. RNA-protein complexes are unable to migrate through an agarose gel and get trapped in the well (first lane). Treatment with Proteinase K releases the RNA from the complexes to migrate (second lane). The second control digestion is with RNaseA (third lane), which confirms that there is no DNA contamination left in the XRNAX extract.

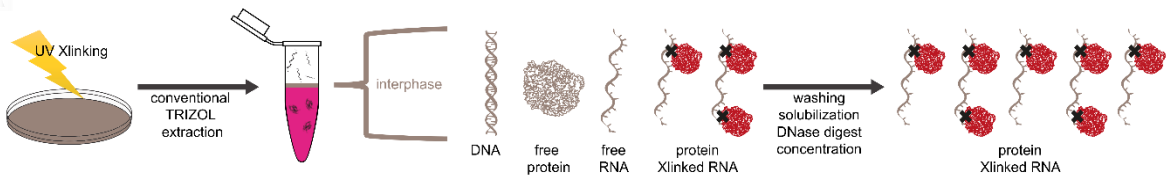


Figure 31. Schematic overview of Protein-Xlinked RNA eXtraction (XRNAX) of UV crosslinked RNA-protein complexes with TRIzol.
Cells are UV cross-linked and lysed with TRIzol. After phase separation, RNA-protein complexes are present in the interphase. The interphase is solubilized and DNA is digested away, resulting in a RNA-protein complexes containing extract. Figure from Trendel et al., 2019.

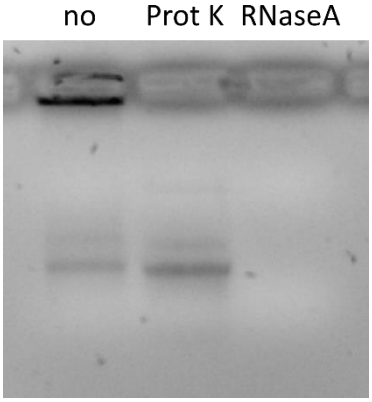


Figure 32. Quality control digestion of XRNAX extract.
Representative agarose gel of the quality control digestions performed on XRNAX extracts to show the presence of UV crosslinked RNA-protein complexes. XRNAX extract was seperated on an agarose gel either untreated (no), treated with Proteinase K (Prot K) or treated with RNaseA.

2.4.2 XRNAX identifies Cal1, not CENP-C, as an RNA-binding protein

I wanted to assess whether CENP-C and Cal1 are RNA-binding proteins by determining their enrichment in the XRNAX extract by western blot. Unfortunately, the antibodies for these two proteins are inefficient. Therefore, I created two new stable cell lines expressing C-terminally V5-His tagged CENP-C or Cal1 under the control of the by copper sulfate ($CuSO_4$) inducible metallothionein promoter (pMT) (BUNCH ET AL., 1988). For subsequent experiments, appropriate levels of $CuSO_4$ induction had to be determined. Whole cell lysate from cells exposed to increasing $CuSO_4$ concentrations overnight were analyzed with western blot and immunostaining. All tested concentration resulted in increased V5-His-tagged protein levels (Figure 33A and B). Importantly, even with the highest concentration, CENP-C-V5-His and Cal1-V5-His still localized specifically to centromeres marked by CID, not ectopically (Figure 33C). Therefore, overnight induction with 400 µM $CuSO_4$ was selected for both proteins, as it results in easily detectable, but still properly localizing proteins. Notable is that only 60.3 % of the cells show centromeric Cal1-V5-His expression, while for CENP-C-V5-His this is 96.7 % of the cells.

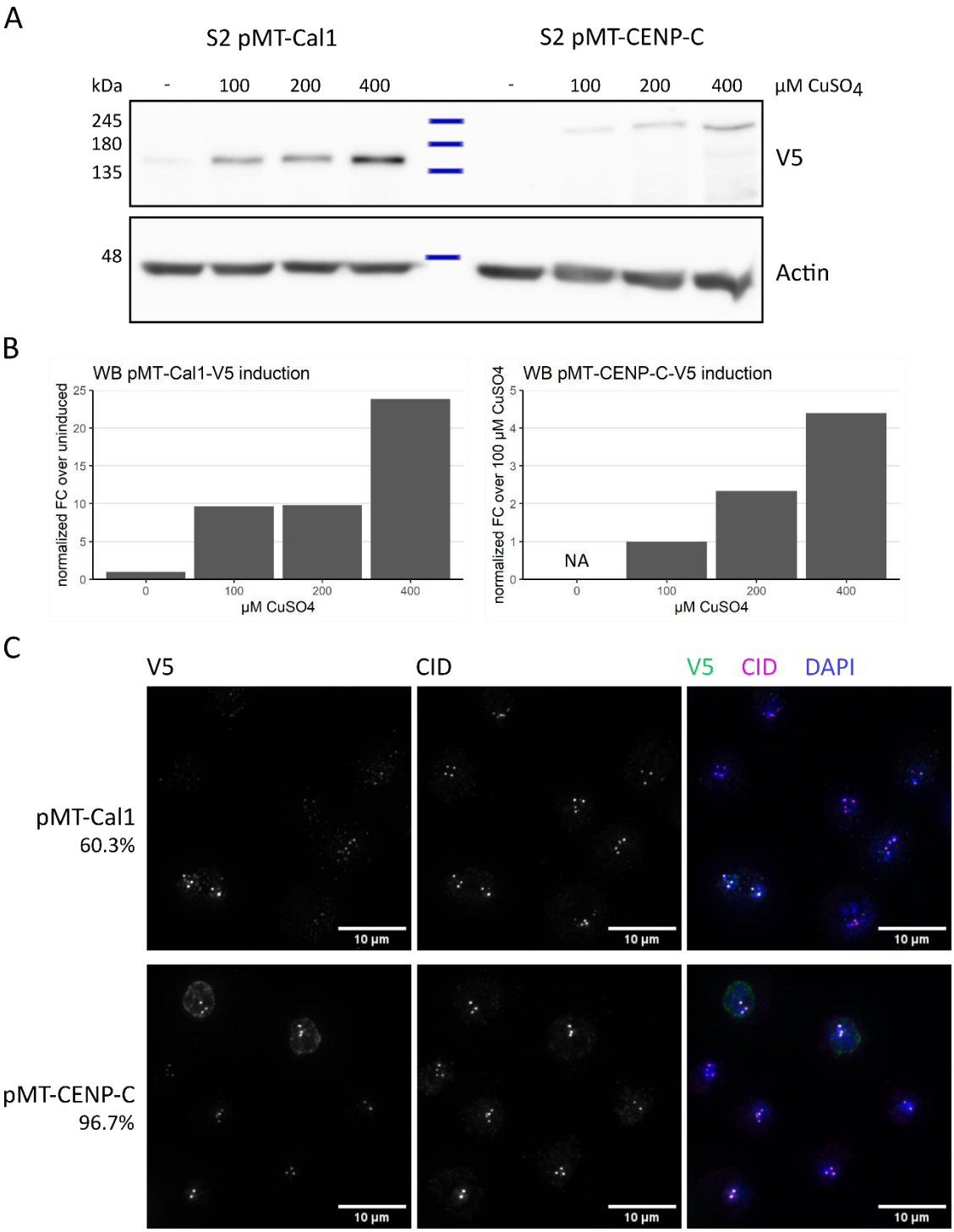


Figure 33. Cell lines expressing copper inducible V5-tagged Cal1 or CENP-C.

A) Stable S2 cell lines containing either the pMT-Cal1-V5 or pMT-CENP-C-V5 construct were exposed to the indicated increasing levels of $CuSO_4$ overnight. Whole cell lysate was analyzed with western blot to evaluate the induction levels of these exogenous proteins.

B) Quantification of western blots in A, exogenous V5 tagged protein levels normalized to Actin.

C) Transfection efficiency and centromere specificity was determined in normal cycling fixed cells induced with 400 µM $CuSO_4$ by immunostaining with the V5 and CID antibodies and counter staining with DAPI. The indicated percentages represent the proportions of cells with overlapping CID and V5 staining. Shown is a max intensity Z-projection. Composite images show α-V5 in green, α-CID in magenta and DAPI in blue. Scale bar = 10 µm.

Next, I isolated XRNAX extract from both cell lines after an overnight $CuSO_4$ induction. In order to obtain clear bands of RNA-binding proteins with western blotting, the cross-linked RNA was digested away with RNaseA before loading. Protein levels in the XRNAX extract were compared to protein levels in whole cell lysate from the same induced cells that were not exposed to UV light. I compared the enrichment of CENP-C and Cal1 in the XRNAX extract to a protein which is known RNA-binder, Rumpelstiltskin (KING ET AL., 2014), and a protein which, as far as we know, is not a RNA-binder, Actin. As can be seen in Figure 34, CENP-C is not enriched in the XRNAX extract. The band intensity ratio of XRNAX over the control whole cell lysate of CENP-C is similar to the ratio of Actin, and clearly lower than Rumpelstiltskin. This would suggest that CENP-C is not directly interacting with RNA in S2 cells.

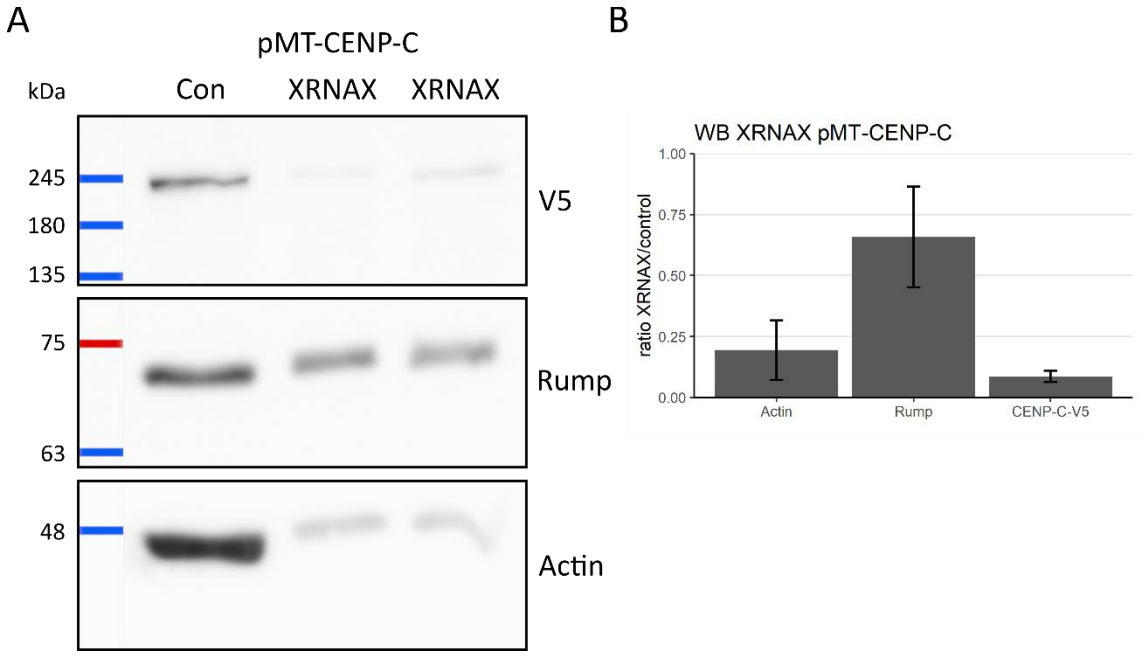


Figure 34. CENP-C is not enriched in XRNAX extract, which suggests it is not an RNA-binder.

A) Western blot showing exogenous CENP-C-V5, and endogenous Rumpelstiltskin and Actin in two XRNAX extracts from S2 pMT-CENP-C-V5 cells induced overnight with 400 µM $CuSO_4$ and a control whole cell lysate from the same induced cells.

B) Quantification of the band intensity ratios of XRNAX compared to the control whole cell lysate. The enrichment ratio of CENP-C-V5 is comparable to negative control Actin, not to known RNA-binding protein Rumpelstiltskin. N=3, error bars = SEM.

In contrast, Cal1 is significantly enriched in the XRNAX extract compared to Actin (Figure 35). The band intensity ratio of XRNAX over the control whole cell lysate of Cal1 is similar to the ratio of Rumpelstiltskin, indicating that Cal1 is, indeed, a strong RNA-binding protein. Last, I also used extracts from the pMT-Cal1-V5-His cell line to assess whether CID directly interacts with RNA. Figure 35 shows two bands of CID, one that is enriched in the XRNAX extract, one that is not. Knockdown experiments

performed by Bachelor students in the lab confirmed the upper band as the specific band for CID (arrowhead). This suggests that CID interacts directly with RNA also.

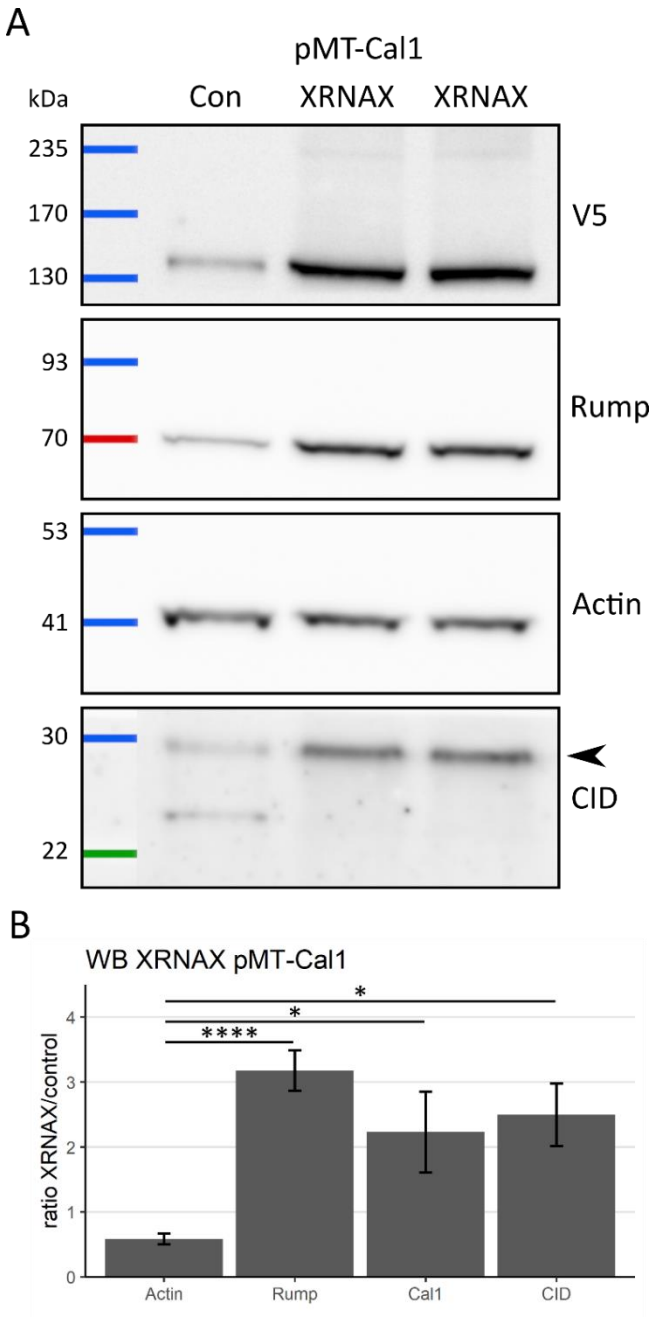


Figure 35. Cal1 and CID are enriched in XRNAX extract, and therefore RNA binding proteins.

A) Western blot showing exogenous Cal1-V5, and endogenous Rumpelstiltskin, Actin and CID in two XRNAX extracts from S2 pMT-Cal1-V5 cells induced overnight with 400 µM $CuSO_4$ and a control whole cell lysate from the same induced cells. Arrowhead indicates proper CID band.

B) Quantification of the band intensity ratios of XRNAX compared to the control whole cell lysate. The enrichment ratios of Cal1-V5 and CID are only slightly lower than known RNA-binding protein Rumpelstiltskin and much higher than negative control Actin. N=8, error bars = SEM, * $p<0.05$, ****<0.00005. Con = control whole cell lysate, Rump = Rumpelstiltskin.

2.4.3 Cal1 XRNAX-CLIP-seq

Having established that Cal1 is an RNA-binding protein, the obvious next question is to which RNAs Cal1 is binding. To answer this question, I performed an enhanced UV crosslinking and immunoprecipitation (CLIP) followed by deep sequencing (Figure 36) (VAN NOSTRAND ET AL., 2016). This method allows for transcriptome-wide identification of RNAs which are bound in vivo by RNA-binding proteins of interest. XRNAX extract was used as input, since it is already strongly enriched for RNA binding proteins and depleted for contaminating DNA. Especially the absence of DNA is important since Cal1 potentially interacts with repetitive RNAs, whose DNA copies could interfere with the data analysis. Furthermore, it allows for RNA fragmentation through sonication instead of the precarious limited RNase digestion. The optimal RNA fragment size for CLIP ranges from 30 to 200 nt (LEE & ULE, 2018). The fragmentation protocol I used resulted in larger fragments, average 180 nt with the largest fragments up to 500 nt. I used a V5 antibody to pull down Cal1 from sonicated XRNAX extract isolated from induced pMT-Cal-V5-His cells. IgG antibody was used for a control pulldown. In both cases, the IP and input samples are loaded on an SDS-PAGE gel to separate based on protein size. After transfer to a nitrocellulose membrane, a piece of the membrane was excised that corresponds to molecular weight of Cal1 and an area above which should contain Cal1-RNA complexes. This additional purification step removes RNA-protein complexes of different molecular weights. Input samples were processed similarly, creating size-matched input controls (SMI-control). The RNA is released from the membrane piece by Proteinase K digestion and isolated for library preparation.

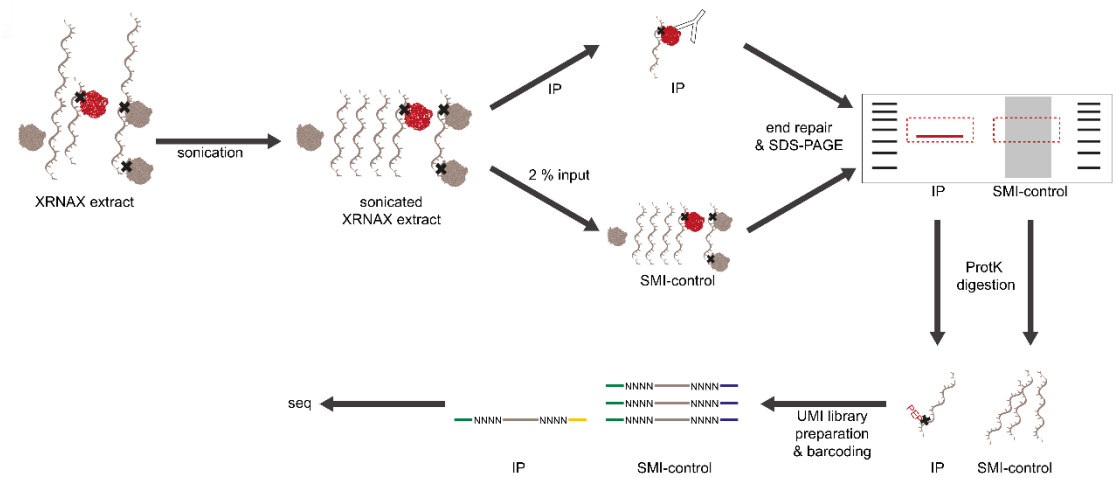


Figure 36. Schematic overview of the CLIP workflow using XRNAX extract as input.
The XRNAX extract was sonicated to fragment the RNA. V5 antibody was used to pull down V5-tagged Cal1, IgG antibody was used as a negative control. RNA in both the IP samples and sonicated XRNAX extract were repaired before running all samples on SDS-PAGE. Areas from the membrane corresponding to the molecular weight of Cal1, and an area above, were excised from which the RNA was isolated for library preparation and subsequent sequencing. SMI-control = size-matched input control. Figure from Trendel et al., 2019.

A fraction of each immunoprecipitation was used to assess the efficiency of the pulldown with western blot. As shown in Figure 37, V5-tagged Cal1 was successfully retained in the V5 IP compared to the control IgG IP. None of the other tested proteins were pulled down by the V5 antibody, suggesting that the V5 IP was specific. The western blots of the other two replicates are presented in Supplementary Figure 8.

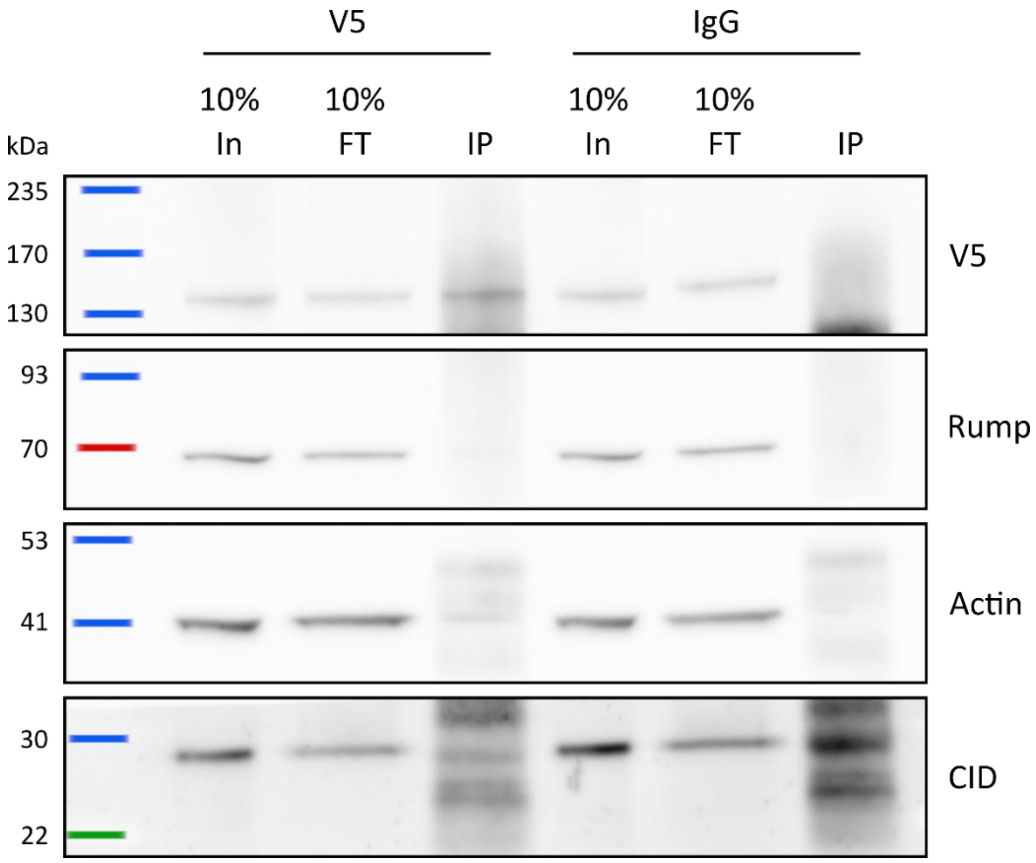


Figure 37. XRNAX-CLIP western blot shows specific enrichment of Cal1-V5 in the V5 CLIP.
XRNAX extract was used as input for a CLIP pulling down V5-tagged Cal1 with a V5 antibody. IgG antibody was used for a negative control pull down. Efficiency of the CLIP was checked with WB using the indicated antibodies. Representative western blot of replicate 2. In = Input, FT = flow through, IP = CLIP, Rump = Rumpelstiltskin.

After preparing the sequencing libraries, the concentration of two of the three IgG CLIP libraries were too low to be detected with both the Qubit and the Fragment analyzer. Even the third IgG CLIP library had a significantly lower concentration than the V5 CLIP libraries. Two IgG replicas were excluded, resulting in three V5 CLIP and SMI-control samples and one IgG CLIP and SMI-control sample being sequenced. Supplementary Figure 9 provides an overview of the technical sequencing metrics, the sequencing depth (A) and the mapping rates (B). The average sequencing depth of 29.1 million reads is excellent. Despite pooling all libraries to get an equal share of the reads, the IgG CLIP has only 16.1 million reads. This suggests that the IgG CLIP library was not quantified properly or was unable to amplify as efficiently on the sequencing chip. Before aligning the reads to the reference genome,

potential adapter sequences were clipped off the reads. Due to the extensive RNA fragmentation, it is more likely that some fragments were so short that the reads contain adapter sequences which could hinder the alignment. The clipped reads were mapped with STAR (DOBIN ET AL., 2013) to the reference genome published by the Mellone and Larracuente labs (CHANG ET AL., 2019) as described above, allowing for multimappers. Ligation based library preparation is always slightly less efficient, resulting in a lower mapping rates. Therefore, on average, only half of the reads mapped to the reference, with 28.2 % of the reads mapped uniquely and 20.6 % of the reads multimappers. Surprisingly, a larger percentage of the reads were able to map in the V5 CLIP samples. CLIP datasets usually have many PCR duplicates due to their low complexity (UHL ET AL., 2017), which have to be removed computationally. After all these processing steps, I end up with an average of 5.4 million counted reads per sample, again for both genes and repetitive elements.

I performed a principal component analysis to get an initial impression of the CLIP count data. In Supplementary Figure 10 all samples are plotted according to the first two principal components, which account for 42 % and 31 % of the variation in the data, respectively. The first principal component separates the samples according to replicates and the second principal component by sample type (SMI-control and CLIP). What stands out is that one of the replicates, replicate 1, is separated from the other replicates. This has to be kept in mind during subsequent data analysis, as this could indicate batch effects during the sample preparation. The control IgG CLIP sample clusters together with the two other Cal1-V5 CLIP replicates, indicating limited count differences between these samples. Importantly there is a clear distance between the SMI-control and CLIP samples, suggesting the CLIP did enrich for some specific RNAs.

2.4.4 Identifying Cal1-binding RNAs

In order to identify putative Cal1-binding RNAs from this dataset, I performed a differential enrichment analysis with DESeq2 (LOVE ET AL., 2014). Only those RNAs that are enriched in the Cal1-V5 CLIP over both SMI-control and the control IgG CLIP should be considered. Therefore, I plotted the apeglm shrunken log2 fold changes reported by DESeq2 from both comparisons in Figure 38A. In the top right quadrant are 16 genes and repetitive elements that are significantly enriched in the Cal1-V5 CLIP samples, indicated in red and listed in Supplementary Table 6. To get a better understanding of these putative Cal1-binding RNAs, I collapsed the DESeq2 results of both comparisons in order to plot the average log2 fold change for each of these significant genes together with the normalized counts and average Benjamini and Hochberg adjusted p-value. From Figure 38B, it can be seen that by far the most significantly enriched element is Copia, both the internal part and its long terminal repeats. Furthermore, the internal part of Copia is the only candidate that is also significantly enriched in the embryo CID RNA-ChIP. I will discuss Copia in more detail in the following section.

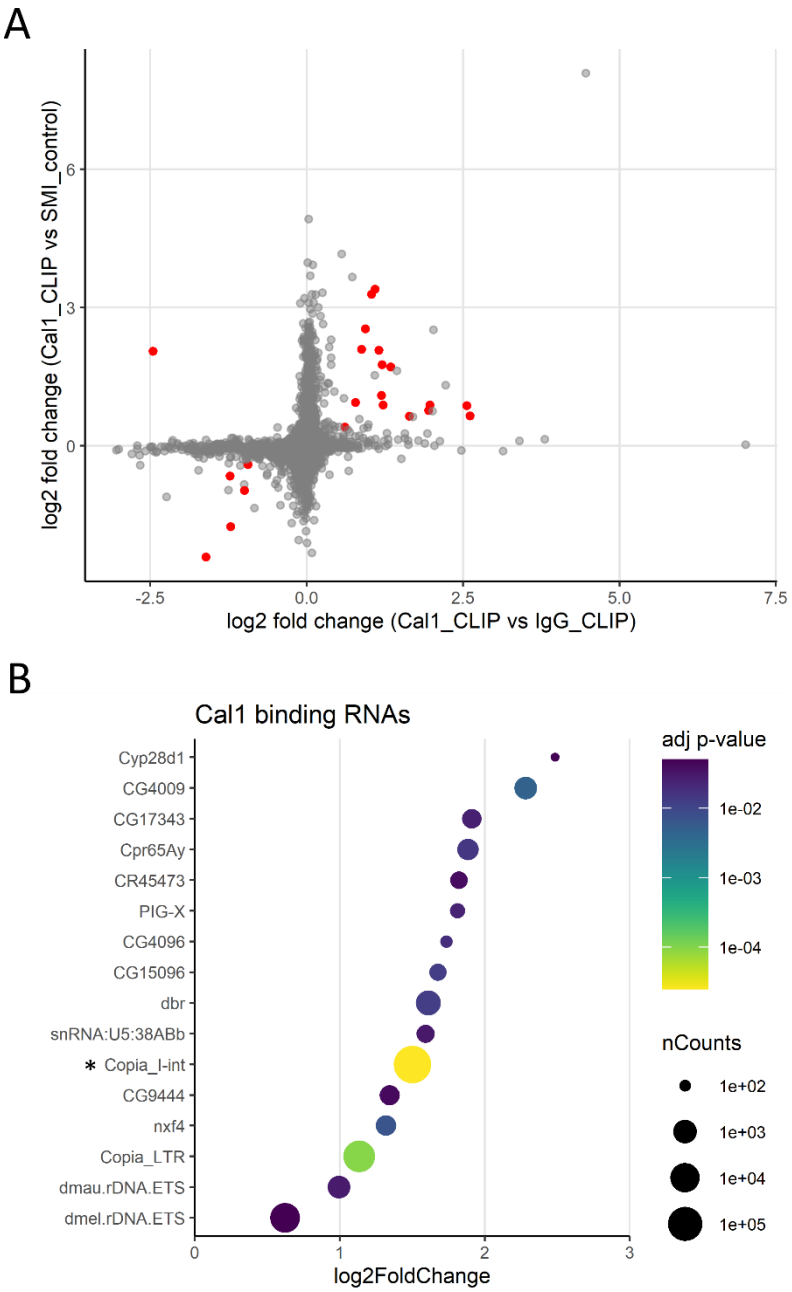


Figure 38. Putative Cal1-binding RNAs.

A) For each gene the log2 fold change of Cal1-V5 CLIP over SMI-control was plotted against the log2 fold change Cal1-V5 CLIP over IgG CLIP. Genes that are significantly differentially enriched and/or depleted in both comparisons are indicated red. In the top right quadrant, are the potential Cal1-binding RNAs. Log2 fold changes were shrunken with the apeglm shrinkage estimator in order to remove noisy fold changes from RNAs with low count numbers.

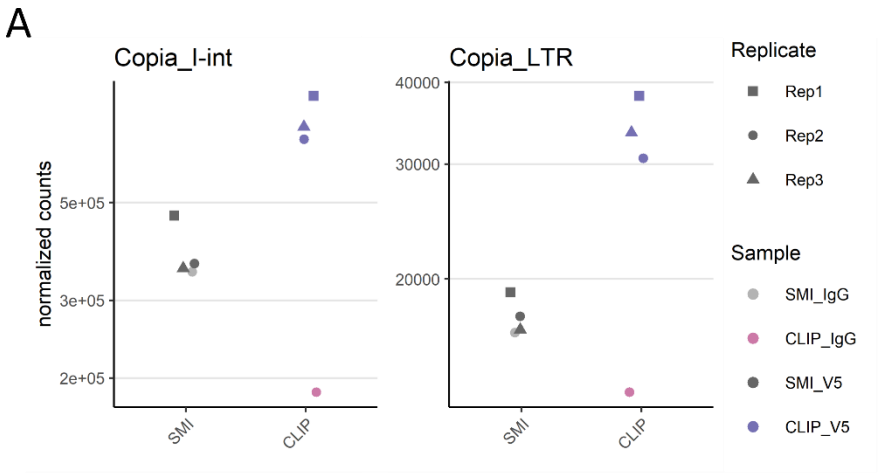
B) Putative Cal1-binding RNAs from the analysis in A. The fold changes and Benjamini and Hochberg adjusted p-values from both comparisons were averaged. This average log2 fold change was plotted for each of the significantly enriched genes or repetitive elements. Average adjusted p-value is indicated by color and normalized counts by dot size. TEs found in centromeric DNA are indicated with an asterisk in from of their name. nCounts = mean of normalized counts , adj p-value = Benjamini and Hochberg adjusted p-value.

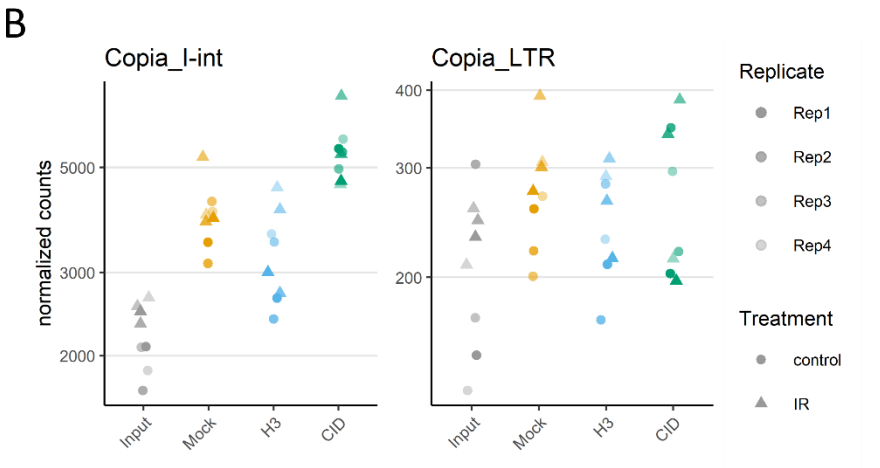
Individual normalized counts of the other putative Cal1-binding RNAs are plotted in Supplementary Figure 11. Most of these candidates have relatively low count numbers. Interestingly, two annotated external transcribed spacers (ETS) of the rDNA repeat unit are among the significantly enriched candidates. Centromeres are known to cluster around the nucleolus, the main locus of rDNA transcription and ribosome assembly (PADEKEN ET AL., 2013). Furthermore, Cal1 is known to locate to the nucleolus during interphase (DUNLEAVY ET AL., 2009; MELLONE ET AL., 2011; SCHITTENHELM ET AL., 2010). The ETS elements have higher count numbers, but only an enrichment of 1.38-fold over SMI-control and 1.72-fold over IgG CLIP for dmel.rDNA.ETS and an enrichment of 1.99-fold over SMI-control and 1.98-fold over IgG CLIP for dmau.rDNA.ETS.

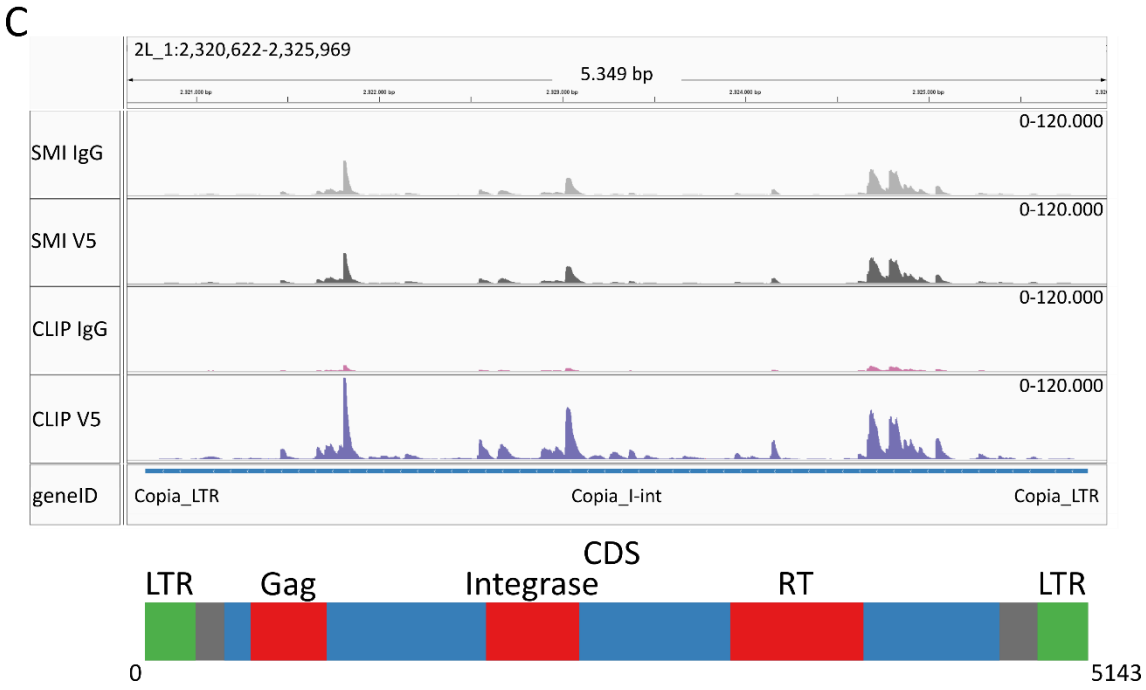
In Supplementary Figure 12, the individual normalized counts from the XRNAX-CLIP data of several interesting repeats were plotted. For Blastopia and G2/Jockey-3 there was no enrichment in the Cal1-V5 CLIP samples. Some of the members of the 1.688 satellite repeat family, like SATIII (359bp_SAT), 260bp_SAT and 353bp_SAT did show an enrichment in the Cal1-V5 CLIP samples. SATIII has an enrichment of 1.28-fold over SMI-control and 1.96-fold over IgG CLIP, 260bp_SAT has an enrichment of 1.36-fold over SMI-control and 2.90-fold over IgG CLIP, and 353bp_SAT has an enrichment of 1.35-fold over SMI-control and 1.45-fold over IgG CLIP. None of these were significantly enriched in both comparisons. Sequences of all 1.688 satellite repeat family members are found intermingled with one another in the genome, but SATIII, 260bp_SAT and 353bp_SAT are more frequently located together.

2.4.5 Copia

As mentioned above, Copia is the most significantly enriched RNA in the Cal1-V5 CLIP. The normalized individual counts are plotted in Figure 39A. The internal part of Copia has a 1.94-fold enrichment over SMI-control and 4.12-fold enrichment over IgG CLIP, while the LTR repeat has a 1.91-fold enrichment over SMI-control and 2.52-fold enrichment over IgG CLIP. Importantly, the internal part of Copia is also significantly enriched in the embryo CID RNA-ChIP, with a 1.77-fold and a 1.44-fold enrichment over H3 RNA-ChIP and Mock RNA-ChIP, respectively. The normalized counts of Copia from the embryo RNA-ChIP are shown in Figure 39B. As can be seen in the coverage tracks of Figure 39C, there are several peaks of reads which could represent the regions of Copia RNA that are bound by Cal1. Although at lower levels, the same regions also have reads clustering in the SMI-control samples. However, this is to be expected, while input samples have Cal1-RNA complexes and the SMI-control is processed on SDS-PAGE in the same manner as the CLIP samples.







Legend on next page

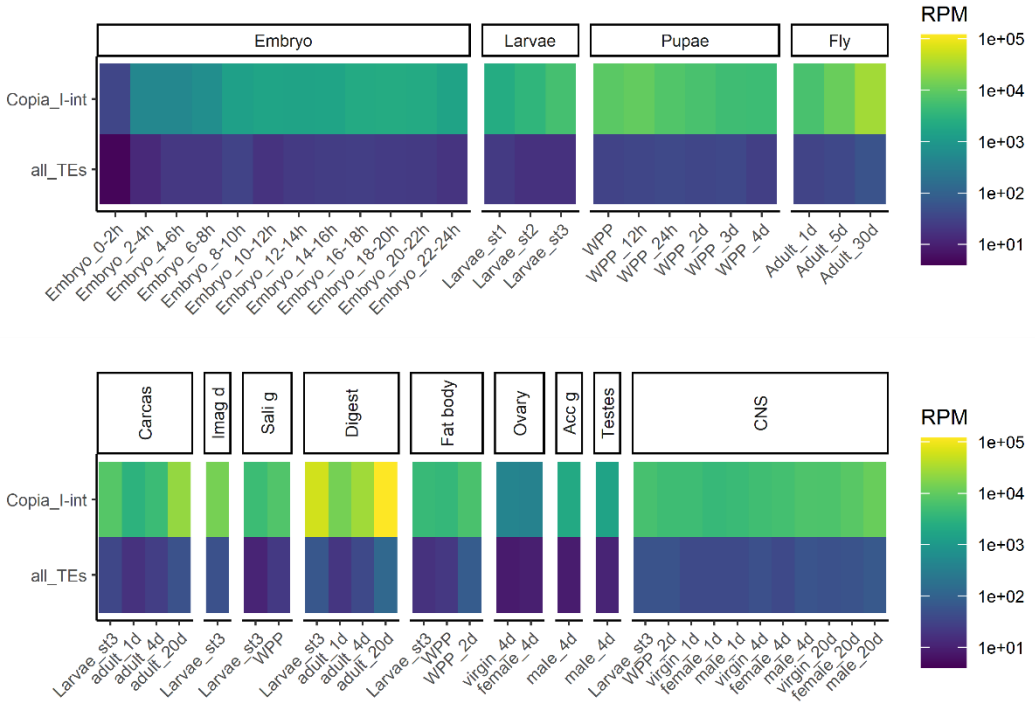
Figure 39. Copia RNA is the top candidate to interact with Cal1.

A) Normalized counts of the internal part and LTR repeat of Copia in Cal1 XRNAX-CLIP.
B) Normalized counts of those same candidates in the embryo RNA-ChIP-seq dataset.
C) Representative Copia locus in the genome shown in Integrated Genomics Viewer with the read coverage tracks of replicate 2 of XRNAX-CLIP (top) and a schematic overview of the Copia gene (bottom). The coding region (CDS) of the polyprotein is indicated in blue and the Long Terminal Repeats (LTR) in green. The locations of the individual proteins are indicated in red, the Gag protein (Gag), the Integrase and the Reverse Transcriptase (RT).

SMI = size-matched input control, Rep = replicate, IR = irradiated.

Copia is a 5143 bp retrotransposon of the Copia family of which the full sequence was first described in 1985 (MOUNT ET AL., 1985). The sequence composition is schematically depicted in the lower part of Figure 39C. It encodes a 1409 aa polyprotein, which encodes the usual enzymes required for retrotransposition. There are 70 full length copies of Copia annotated in the Mellone and Larracuente labs genome reference on all chromosomes, except chromosome 4. A truncated copy of Copia_I can be found in the centromeric DNA contig of the Y chromosome (CHANG ET AL., 2019). Copia copies are found in all studied species of the melanogaster subgroup (LERAT ET AL., 2011).

As discussed in section 2.3.1, I re-analyzed the modENCODE RNA-seq dataset, which showed an exceptionally high expression level of Copia. Figure 40 shows the normalized count of Copia across development and in the indicated tissues. While Copia RNA is present at high levels in all tissues throughout development, the digestive system has the highest expression levels.



Legend on next page

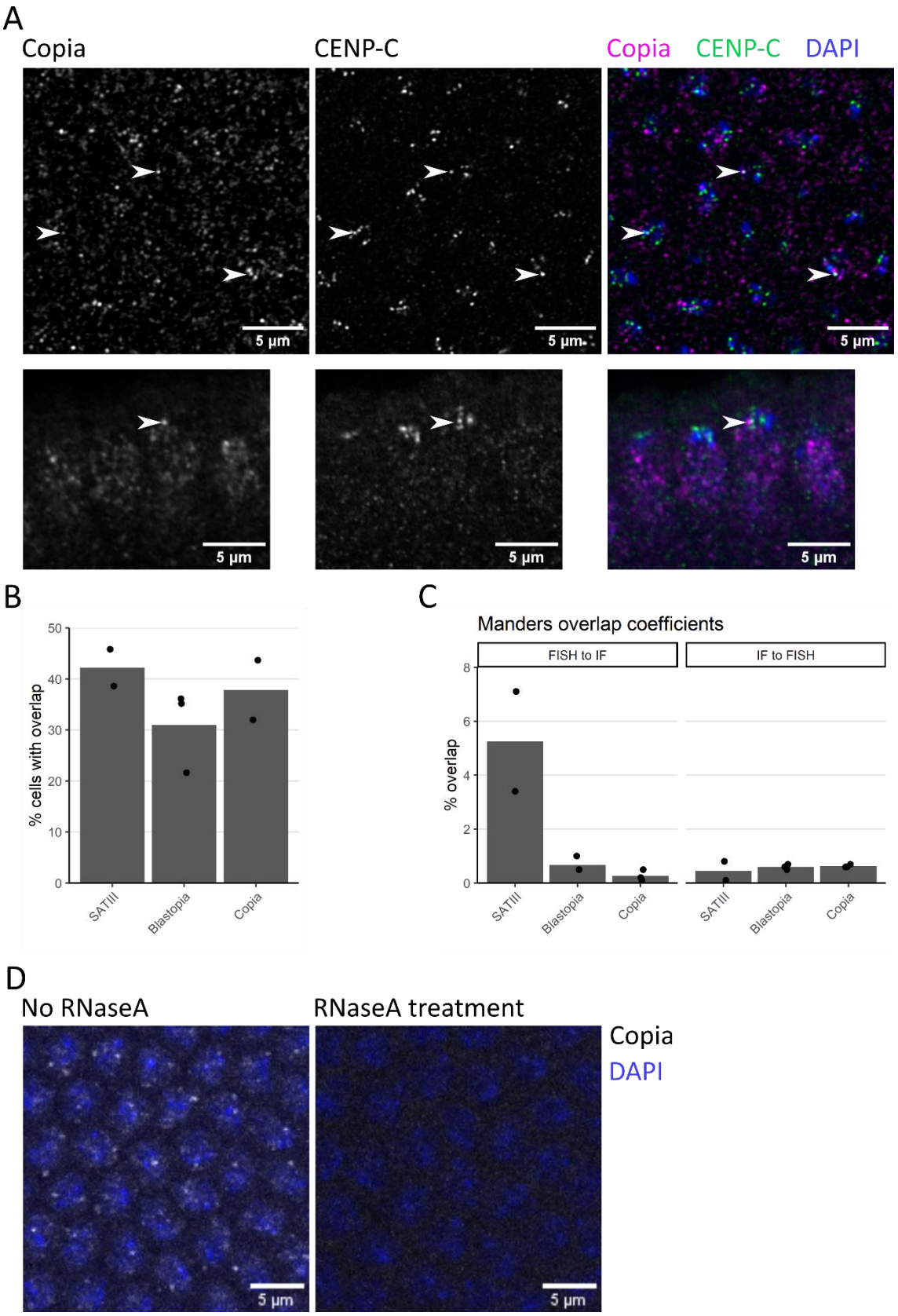
Figure 40. Expression pattern of Copia in whole animals and indicated tissues across development shows increased expression in the Digestive System.
Re-analysis of the modENCODE RNA-seq datasets of RNA isolated from the indicated tissues for Copia. Counts were normalized to the sequencing depth of each sample, resulting in reads per million (RPM), indicated with the blue to yellow fill. Imag d = Imaginal Discs, Sali g = Salivary Glands, Digest = Digestive System, Acc g = Accessory Glands, CNS = Central Nervous System.

2.4.6 Validation of centromeric localization of Copia

To confirm that Copia is localized to the centromere with Cal1, I performed smRNA-FISH in combination with IF for inner kinetochore protein CENP-C in stage 5 embryos (Figure 41). Copia RNA is visible as multiple strong foci in all nuclei of stage 5 embryos. Many of the foci are localized at the more basal euchromatin, but several foci are also present at the apical heterochromatin. As indicated by arrows in Figure 41A, in multiple cells Copia RNA foci are observed overlapping with CENP-C foci. In 38 % of the cells there was at least one overlapping Copia smRNA-FISH and CENP-C IF signal, which is comparable to SATIII and Blastopia RNA reported above, with 42 % and 31 %, respectively (Figure 41B).

Like for Blastopia and SATIII in section 2.2.4, I determined the Manders' coefficients for Copia smRNA-FISH and CENP-C IF foci overlap with JaCOP (BOLTE & CORDELIERES, 2006). As described in more detail above, the Manders' coefficients quantify the signal of RNA-FISH that overlaps with CENP-C IF signal as a proportion of the total RNA-FISH intensity and vice versa. Copia's coefficients are plotted as percentages along with the coefficients of SATIII and Blastopia in Figure 41C. Only a small portion of the Copia FISH signal, 0.27 %, overlaps with CENP-C IF signal (FISH to IF). This was to be expected with the large number of foci present throughout the nuclei. This is slightly lower than the FISH to IF coefficient of Blastopia and substantially lower than the FISH to IF coefficient of SATIII, which are 0.67 % and 5.25 % respectively. However, a similar proportion of inner kinetochore protein CENP-C IF signal overlaps with FISH signal of Copia, Blastopia and SATIII, namely 0.63 %, 0.6 % and 0.45 % respectively.

Overall, these results suggest a partial centromeric localization of Copia RNA. Copia is present in multiple foci throughout the nuclei. Importantly, CENP-C overlaps to a similar extent to Copia as to SATIII and Blastopia. Specificity of the Copia smRNA-FISH probes hybridization to RNA was confirmed by RNaseA treatment of the embryos prior to the staining (Figure 41D).



Legend on next page

Figure 41. Copia smRNA-FISH and CENP-C IF co-localize in some cells in stage 5 OregonR embryos.

A) Single z-slices of stage 5 embryos stained with smRNA-FISH for Copia together with IF for inner-kinetochore protein CENP-C imaged at the top of the embryo (top) and the side of the embryo (bottom). To visualize the nuclei, DNA was counterstained with DAPI. Colocalizing Copia and α-CENP-C foci are indicated by arrows. Composite images show α-CENP-C in green, Copia smRNA-FISH in magenta and DAPI in blue. Scale bar = 5 µm.

B) Percentage of cells which have at least one overlapping FISH and IF foci of the cells that have SATIII antisense, Blastopia or Copia expression. Per embryo, around 250 cells were analyzed for expression and overlap. For Blastopia and Copia 3 embryos were quantified and for SATIII 2 embryos. Blastopia and SATIII quantifications from Figure 20.

C) Manders' overlap coefficients that were measured with JaCOP to provide the proportion of RNA-FISH signal that overlaps with CENP-C IF signal over the total RNA-FISH intensity and vice versa. The overlap coefficients were plotted as percentages for SATIII antisense, Blastopia and Copia. For Blastopia and Copia 3 embryos were quantified and for SATIII 2 embryos. Blastopia and SATIII quantifications from Figure 20.

D) Max intensity projections of the top monolayer of cells in stage 5 OregonR embryos stained with Copia smRNA-FISH (gray) and counterstained with DAPI (blue). As indicated, embryos were treated with RNaseA or left untreated before the staining to confirm RNA specificity of the probes. Scale bar = 5 µm.

2.4.7 Human CENP-A chaperone HJURP is also an RNA-binding protein

The human orthologue of Cal1, HJURP, has also been reported to be a putative RNA-binding protein, since it contains a potential RNA binding motif and immunoprecipitates RNA (QUÉNET & DALAL, 2014). Since the XRNAX protocol was working so efficiently for *Drosophila* cells, it was repeated with the human cell line HEK293T (DUBRIDGE ET AL., 1987). A fellow PhD student in the lab, Vojtěch Dolejš, grew, UV crosslinked and lysed the cells in TRIzol. Next, I isolated the XRNAX extract and performed the western blot. Again, protein levels in the XRNAX extract were compared to protein levels in whole cell lysate cells that were not exposed to UV light. Heterogeneous nuclear ribonucleoprotein M (hnRNPm) was used as a positive control, as the name suggest it is a known RNA-binding protein (HO ET AL., 2021). As with *Drosophila* cell, Actin was used as a negative control. Given the similar molecular weights of HJURP and hnRNPm, samples were loaded twice. Similar loading was confirmed by blotting for Actin in both repeats. The human cell XRNAX extract verified HJURP as an RNA-binding protein (Figure 42). Despite having only 2 replicates, the enrichment ratio of HJURP in the XRNAX extract over the control whole cell lysate is significantly higher than both hnRNPm and Actin. As expected, hnRNPm has a significantly higher enrichment ratio than Actin.

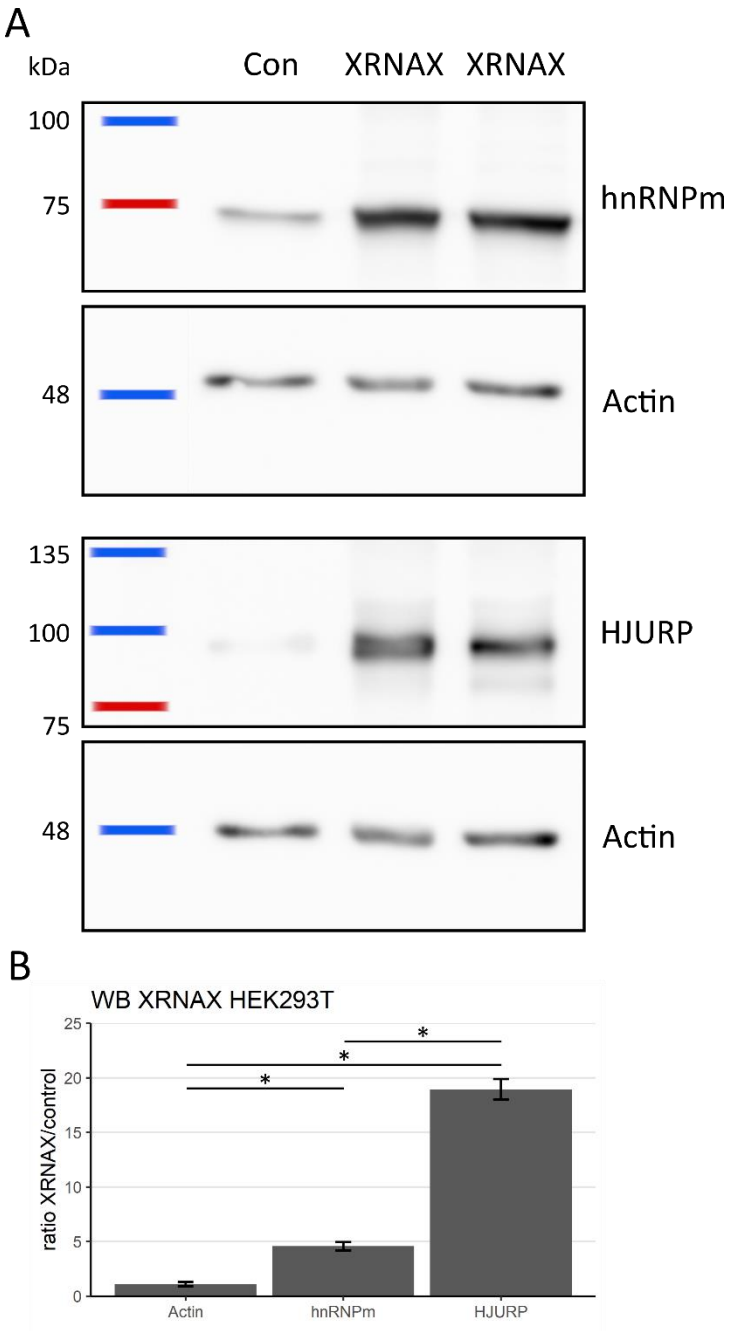


Figure 42. Human CENP-A chaperone HJURP is enriched in XRNAX extract from human HEK293T cells, and therefore an RNA binding protein.

A) Western blot showing HJURP, hnRNPm and Actin in two XRNAX extracts from HEK293T cells and a control whole cell lysate from the same cells. Samples were loaded twice, due to the similar molecular weights of HJURP and hnRNPm. Similar loading was confirmed by Actin.

B) Quantification of the band intensity ratios of XRNAX compared to the control whole cell lysate. The enrichment ratio of HJURP is significantly higher than both known RNA-binding protein hnRNPm and negative control Actin. N=2, error bars = SEM, * $p<0.05$.

2.5 Heterochromatin formation and SATIII

In a collaboration with the Iovino lab of the Max-Planck Institute of Immunobiology and Epigenetics in Freiburg, I evaluated the effect of improper heterochromatin establishment in *Drosophila* embryos on SATIII expression. A PhD student in the Iovino lab, Nazerke Atinbayeva, created fly lines which are mutant for all three H3K9 methyltransferases, Su(var)3-9, Eggless and G9a (SELLER ET AL., 2019). The establishment of heterochromatin during early embryogenesis is disrupted in these triple mutants. Given the large arrays of SATIII repeats in the heterochromatin of the X chromosome, we set out to assess SATIII expression in H3K9 methylation deficient stage 5 embryos.

2.5.1 RT-qPCR analysis of triple mutant embryos

Nazerke Atinbayeva provided me with purified RNA from stage 5 control and triple mutant embryos, which I used to determine relative expression levels of SATIII and G2/Jockey-3 from the centromere of chromosome 4 (Cen4-G2). As shown in Figure 43, the relative expression levels of both (peri-)centromeric repetitive elements was elevated in the triple mutant embryos. SATIII was increased by 2.9-fold while Cen4-G4 increase by only 1.8-fold, neither of them was significant. The lower increase of Cen4-G2 was expected, since this primer pair is specifically targets G2/Jockey-3 at one of the core centromeres. Centromeric chromatin, in contrast to pericentromeric heterochromatin, is naturally depleted from H3K9 methylation (SULLIVAN & KARPEN, 2004). Therefore, loss of this histone modification should have less impact.

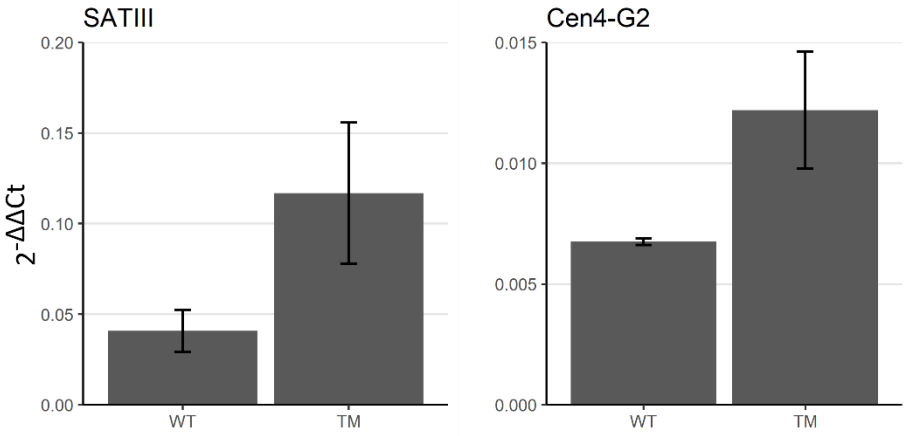


Figure 43. Relative expression levels of (peri-)centromeric repeats are elevated in stage 5 embryos that are mutant for all three H3K9 methyltransferases.
RT-qPCR analysis of SATIII and Cen4-G2/Jockey-3 RNA in triple mutant stage 5 embryos that lack H3K9 methylation (TM) compared to wild type (WT) embryos. The bar indicates the average expression level normalized to the housekeeping genes Actin, GAPDH and rp49/RpL32, the error bars are the SEM. N = 3 for WT and 4 for TM.

2.5.2 SATIII smRNA-FISH signal increased without heterochromatin

Next, SATIII antisense RNA levels were assessed with smRNA-FISH. For this I received fixed wildtype and triple mutant stage 5 embryos from Nazerke Atinbayeva. As shown in Figure 44A, SATIII antisense RNA was observed as single foci in a subset of cells as described above. These foci were not observed after RNaseA treatment. In stage 5 embryos that are triple mutant for all three H3K9 methyltransferases, more cells have SATIII antisense foci, sometimes also multiple foci per cell. Moreover, the average foci intensity was significantly increased (Figure 44A bottom). However, more strikingly was the significant increase of SATIII antisense smRNA-FISH signal at the posterior pole of the embryos, as presented in Figure 44B. Weak foci are observed in some cells in at the posterior pole of wildtype embryos. In contrast, large foci of high intensity are present in these cells of triple mutant embryos. The foci remain concentrated at the apical side of the cells, overlapping the more DAPI dense regions of pericentromeric regions. In only a fraction of embryos some pole cells, which are precursors of germ cells (UNDERWOOD ET AL., 1980), also have strong SATIII antisense smRNA-FISH foci.

Last, I used SATIII sense smRNA-FISH probes on fixed embryos received from Nazerke Atinbayeva. Observations in our lab suggest that this smRNA-FISH probe has some cross-reactivity with the other satellite repeats of the 1.688 family. However, since all members of this family are localized in the pericentromeric heterochromatin, this probe can still be used to assess the effect of H3K9 methylation absence on transcription at the pericentromere. SATIII sense smRNA-FISH signal increased tremendously throughout the entire stage 5 triple mutant embryo, not just at the posterior pole. As can be seen in Figure 45, in wildtype embryos single SATIII sense foci were observed in a subset of cells. In triple mutant embryos, on the other hand, the fast majority of cells display one or more foci of varying size. Average SATIII sense foci intensity is significantly higher in triple mutant embryos compared to wildtype embryos. Both triple mutant and wildtype embryos have a higher foci intensity compared to RNaseA treated triple mutant embryos.

Taken together, these results indicate an inadequate heterochromatin formation during early embryogenesis in the absence of all three H3K9 methyltransferases. Proper repression of Satellite repeat transcription is not established, resulting in a substantial overexpression of SATIII and potentially also other members of the 1.688 satellite family.

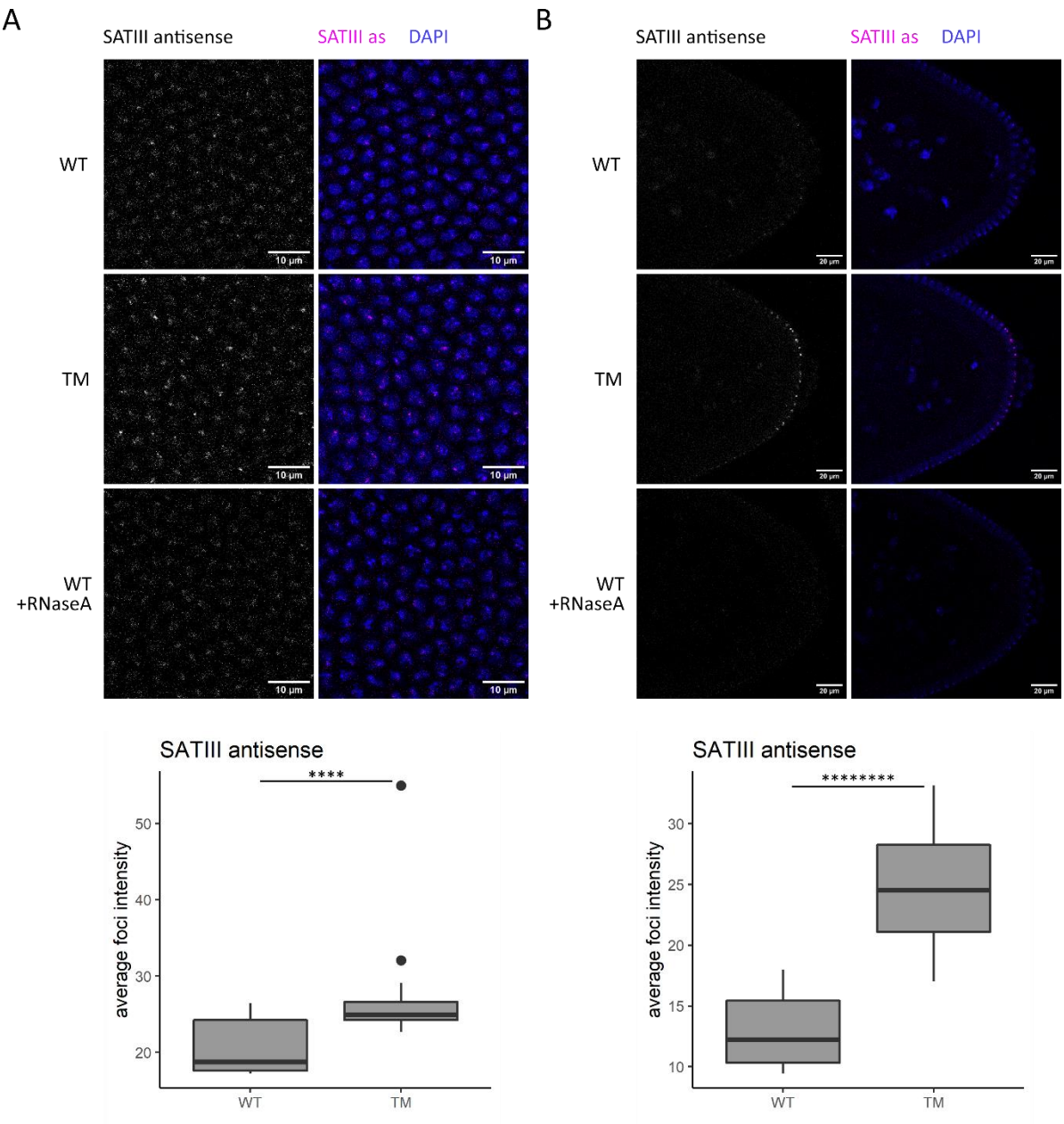


Figure 45. SATIII antisense smRNA-FISH signal increased in stage 5 embryos lacking heterochromatin.

A) Single z-slice of fixed stage 5 embryos stained with SATIII antisense smRNA-FISH and counterstained with DAPI to visualize the nuclei. Embryos from wildtype (WT) or H3K9 methyltransferase triple mutant (TM). Images are taken from the monolayer of cells in the middle of the embryo. Scale bar = 10 µm. Bottom: SATIII antisense smRNA-FISH foci intensity was quantified for WT and TM embryos.

B) Single z-slice of the same embryos with the same staining as in A. Images are taken from the median plane at the posterior pole of the embryo. Scale bar = 20 µm. Bottom: SATIII antisense smRNA-FISH foci intensity was quantified for WT and TM embryos.

Images are displayed with the same adjusted brightness and contrast. RNaseA treatment was performed to show the specificity of the smRNA-FISH foci. Composite images show SATIII antisense smRNA-FISH in magenta and DAPI in blue. Foci intensities from images of 28 WT and 28 TM embryos were quantified. **** p-value < 10^{4}, ******** p-value < 10^{13}.

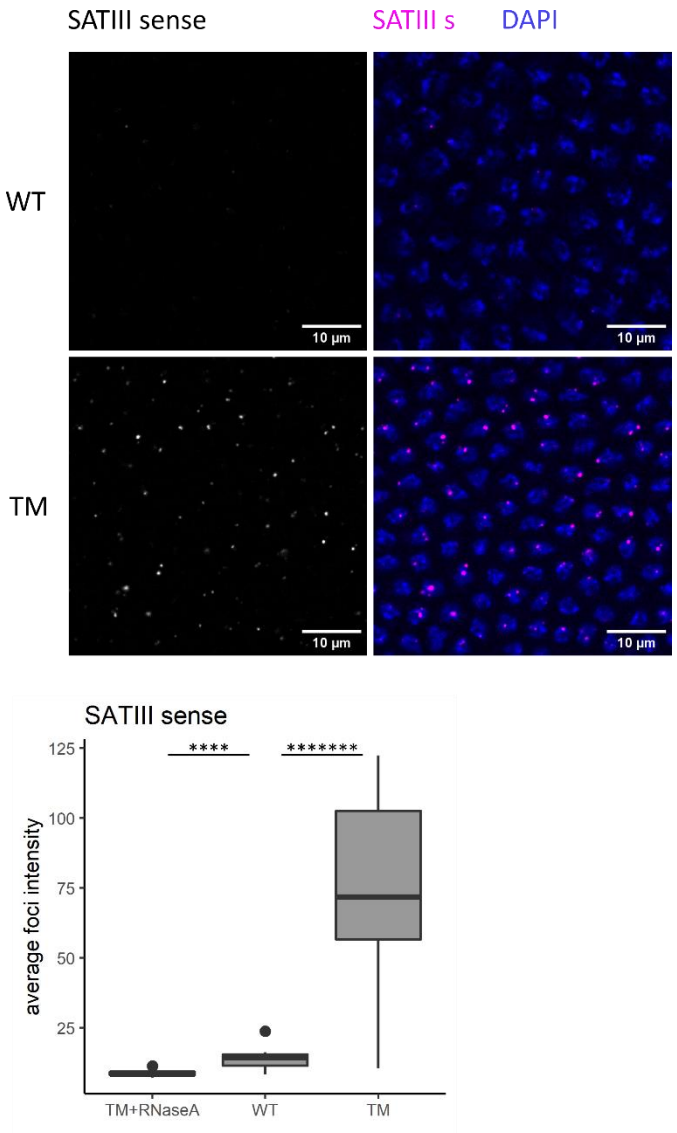


Figure 46. SATIII sense smRNA-FISH signal increased in stage 5 embryos lacking heterochromatin. Single z-slice of fixed stage 5 embryos stained with SATIII sense smRNA-FISH and counterstained with DAPI to visualize the nuclei. Embryos from wildtype (WT) or H3K9 methyltransferase triple mutant (TM). Images are taken from the monolayer of cells in the middle of the embryo. Images are displayed with the same adjusted brightness and contrast. Composite images show SATIII sense smRNA-FISH in magenta and DAPI in blue. Scale bar = 10 µm. Bottom: SATIII sense smRNA-FISH foci intensity was quantified for WT and TM, and TM embryos treated with RNaseA. Foci intensities from images of 9 TM+RNaseA, 19 WT and 20 TM embryos were quantified. **** p-value < 10^{4}, ******* p-value < 10^{7}.

3. Discussion

3.1 Centromere-associated RNAs

In all eukaryotes studied thus far, RNA is an integral component of the centromere. While in *Drosophila* the pericentromeric SATIII RNA is being studied and was shown important for centromere functioning (ROŠIĆ ET AL., 2014), other potentially centromere-associated RNAs have not been described or investigated. Besides SATIII at the X chromosome, there is a fast variety of other repetitive sequences, including transposable elements (TEs), present at centromeric DNA (CHANG ET AL., 2019). Given that actively transcribing RNA polymerase II is observed at all *Drosophila* centromeres (ROŠIĆ ET AL., 2014), other repetitive RNAs can potentially be present and play a fundamental structural or regulatory role at centromeres. Therefore, I set out to identify which RNAs are associated with centromeric chromatin by performing a centromeric RNA-ChIP.

3.1.1 SATIII RNA potentially not at centromeric chromatin in embryos

I successfully pulled down CID containing chromatin compared to H3 and mock ChIP controls, and sequenced the RNA associated with it. Surprisingly, SATIII is not among the top centromere-associated RNA candidates, there is no significant enrichment in the centromeric CID RNA-ChIP compared to the chromatin wide H3 RNA-ChIP and the mock RNA-ChIP. This finding contradicts a previous study from our lab, Rošić et al. (2014), which had identified SATIII as a centromere-associated RNA. They were able to specifically pull down SATIII RNA with a RNA-IP of inner-kinetochore protein CENP-C, which is located at centromeric chromatin throughout the cell cycle, and observed mitotic defects upon SATIII knockdown (ROŠIĆ ET AL., 2014). Therefore, I had anticipated to find SATIII enriched in the CID RNA-ChIP of this study also. However, a possible explanation for this discrepancy is that Rošić et al. (2014) performed the CENP-C RNA-IP with S2 cells, while I performed the CID RNA-ChIP in embryos. It has been shown that the core centromere in S2 cells have expanded into surrounding satellite repeat regions compared to embryos, where the core centromeres are restricted to transposable elements (CHANG ET AL., 2019). Additionally, the slightly different cellular localization of CID and CENP-C could explain the divergent RNAs. Furthermore, in metaphase chromosome spreads of S2 cells, SATIII RNA-FISH signal is largely pericentromeric, with only partial overlap with the CENP-C labelled core centromere (ROŠIĆ ET AL., 2014). Indicating the possibility that SATIII RNA might not directly interact with CID containing chromatin. The experimental setup of the CENP-C RNA-IP could explain this disparity, it was performed without crosslinking and by adding pre-isolated RNA to the pull down reaction (ROŠIĆ ET AL., 2014). This potentially could have introduced additional unspecific RNAs into the

centromeric complex, which would normally be unable to interact due to physical distance, such as the more pericentromeric SATIII. In the CID RNA-ChIP I performed, interactions were preserved by crosslinking. Therefore, it is possible that SATIII RNA does not localize at CID containing chromatin in embryos, but supports centromere function by creating the required pericentromeric chromatin (CORLESS ET AL., 2020).

3.1.2 Transposable element RNA enriched at centromeres

In this study, I identified that TEs are the majority of enriched RNAs at centromeric chromatin. Due to the abundance of TEs in the (peri)centromere, especially in *Drosophila*, this result was expected (CHANG ET AL., 2019; KLEIN & O'NEILL, 2018). Of the 63 top candidates I identified, 43 are TEs, of which seven are present in the centromeric DNA. Blastopia, a gypsy LTR retrotransposon which has a copy in the centromeric DNA, is the most enriched centromere-associated RNA. Both the internal part of Blastopia and its long terminal repeat are approximately 5-fold enriched in the CID RNA-ChIP compared to the H3 RNA-ChIP. Since this RNA is evidently the strongest candidate, I attempted to validate and characterize this candidate, which I will discuss in the following sections. The other five centromeric DNA-derived putative centromere-associated RNAs are DOC, BS, Copia, BS2 and Gypsy-26.

Unexpectedly, many of the putative centromere-associated RNAs are not derived from centromeric DNA. This would suggest that RNAs get recruited to centromeric chromatin from other genomic locations, for example together with centromeric proteins. The proper localization of several centromere or kinetochore proteins is dependent on RNA (BLOWER, 2016; MCNULTY ET AL., 2017; QUÉNET & DALAL, 2014; WONG ET AL., 2007). It is conceivable that these proteins interact with the by my CID RNA-ChIP TE RNAs, in order to get recruited to the TE-rich centromeres. Sequence similarities between different TEs could promote interactions between these recruited centromere-associated RNAs and centromeric DNA or newly transcribed RNA, aiding the proper localization centromere or kinetochore proteins.

3.1.1 Validation of Blastopia as a centromere-associated RNA

As Blastopia is clearly the most enriched, I continued to validate the centromeric localization of this RNA. As RNA sequencing is a PCR based technique, I used smRNA-FISH, an *in situ* hybridization and microscopy based technique, to validate in the localization of Blastopia in embryos, which we also used for the RNA-ChIP experiment. Blastopia smRNA-FISH signal is visible in approximately 90 % of the nuclei of stage 5 embryos. There is a ring of stronger expression visible near the anterior pole of the embryo, directly anterior of the newly formed cephalic furrow, consistent with results published by Frommer et al. (1994). In later embryonic stages, Blastopia RNA signal becomes more cytoplasmic and is the

strongest in the head area, specifically the anterior mesoderm or the endoderm of the anterior midgut rudiment.

Of the cells expressing Blastopia in stage 5 embryos, approximately 31 % show a colocalization of at least one Blastopia smRNA-FISH focus with a CENP-C IF focus marking the centromere. This confirms that there is at least some centromeric localization of Blastopia RNA. However, not in all cells and definitely not at all centromeres. Furthermore, there is Blastopia RNA located throughout the nuclei of stage 5 embryos, not only at or near centromeres. These other foci are potentially derived from and localized at the different genomic copies of Blastopia. However, it is surprising to observe strong Blastopia smRNA-FISH signals throughout the nucleus while only observing a centromeric colocalization in 1/3th of the cells, considering the strong enrichment I consistently observed in the centromeric RNA-ChIP data. Similarly, in only 20 % of S2 cells there is at least one colocalization between Blastopia smRNA-FISH and CID IF foci, with Blastopia foci also present in the cytoplasm.

I performed the same colocalization experiment in stage 5 embryos for SATIII RNA. Although SATIII RNA is not enriched in my centromeric RNA-ChIP, it was reported as centromeric by Rošić et al. (2014) and knockdown of SATIII RNA resulted in mitotic defects, also in embryos. With smRNA-FISH, I am able to observe a colocalization of SATIII RNA with CENP-C, with 42 % of the cells expressing SATIII RNA having at least one focus overlapping with CENP-C. This would suggest that SATIII RNA has a stronger centromeric localization than Blastopia. Therefore, it remains confusing that I did not observe an enrichment of SATIII in the centromeric RNA-ChIP, while Blastopia is enriched 5-fold. Especially considering that SATIII RNA is usually present as a single focus in cells, which, compared to the many foci of Blastopia, would result in less RNA to be pulled down in the pan-chromatin H3 RNA-ChIP control and thus a stronger enrichment.

There are several possible explanations for this result. First, SATIII RNA is only expressed in 21 % of the cells, which is significantly less than Blastopia, with 90 % of the cells. Consequently, in absolute numbers, more cells have Blastopia RNA at a centromeric focus than SATIII RNA. Furthermore, a slightly higher fraction of the CENP-C foci has an overlap with Blastopia RNA compared to SATIII RNA. Nevertheless, this cannot explain the full discrepancies observed between the RNA-ChIP data and the IF-smRNA-FISH experiment.

Second, it is possible that unspecific chromatin was pulled down with the CID RNA-ChIP. The control western blot showed a successful enrichment of CID, but also Lamin was present in the CID RNA-ChIP sample. This could mean that the putative centromere-associated RNAs I identified, including Blastopia, are not centromere-associated after all. However, such an unspecific pull down should be random, and therefore have a different composition each time. The different replicates of the RNA-

ChIP, even between irradiated and control samples, consistently pulled down the candidates I identified, especially Blastopia. Making it unlikely that Blastopia is an unspecific candidate.
Third, there is a possibility that the designed smRNA-FISH probe set is not specific enough for Blastopia, and cross-reacts with related TEs. The potential additional foci from other TEs could inflate the non-centromeric signal observed. However, since the strong expression pattern in the anterior ring has been previously described for Blastopia (FROMMER ET AL., 1994), at least the majority of the RNA-FISH signal observed should originate from Blastopia RNA. Still, it would be important to verify the probe specificity for Blastopia in tissue or cells where Blastopia is knocked down.
Last, it is possible that differences between the RNA-ChIP and the IF-smRNA-FISH experimental set-ups produced these contradicting results, making both findings correct in their specific conditions. For example, I performed the IF-smRNA-FISH colocalization experiment in stage 5 embryos, while the RNA-ChIP was performed with chromatin from embryos collected 2-8 hours after laying, which roughly corresponds to stages 4 to 12. During this period, there are many changes in the cell cycle, which could affect centromere-associated RNAs. The first embryonic nuclear divisions occur very rapidly, due to the omission of both G phases during interphase (FARRELL & O'FARRELL, 2014). In stage 5 embryos, the cellularization prolongs the S phase, after which, during the start of gastrulation at stage 6, embryonic cells experience their first G2 phase (FARRELL & O'FARRELL, 2014). Only several divisions later, interphase will also include a G1 phase (FARRELL & O'FARRELL, 2014). Therefore, if SATIII or Blastopia differentially localize to centromeres during different cell cycle phases, this could be missed in the stage 5 embryos used in the IF-smRNA-FISH experiment, but will alter their relative abundance in the centromeric RNA-ChIP data. Furthermore, the RNA-ChIP pulls down chromatin bound RNA, while the smRNA-FISH labels all cellular Blastopia and SATIII RNA. Centromeric Blastopia, and not the other Blastopia or SATIII, potentially is more abundant in the chromatin bound fraction, and thus significantly enriched in the CID RNA-ChIP data. These could be plausible explanations for the discrepancies observed for SATIII and Blastopia between the different experiments. However, the data must still be interpreted with caution and requires further validation.

3.1.2 Blastopia knockdown in embryos not successful

In order to assess the function of Blastopia, in particular at the centromere, I attempted to knockdown the RNA to observe potential defects. I wanted to perform this experiment in embryos, as the centromeric enrichment was initially detected there. Inducing a knockdown in early embryos is efficient by starting such a knockdown already in ovaries and thus affecting the maternally deposited RNAs. Furthermore, mitotic defects would be easily detected due to the synchronous syncytial divisions. However, with smRNA-FISH I observed that Blastopia RNA is only expressed starting from the blastoderm stage, meaning no detectable levels of Blastopia RNA are deposited maternally, making

such an early knockdown uninformative. For this reason, I attempted to induce a ubiquitous knockdown during later stages of embryonic development.

I designed and created TRiP fly lines, which can express a 21 bp short hairpin RNA targeting Blastopia under the control of the UAS/Gal4 inducible system. By crossing these UAS-TRiP fly lines with a ubiquitous embryonic Gal4 driver line, Armadillo-Gal4, the knockdown should be induced starting shortly after activation of zygotic transcription in all embryonic cells. However, inducing a knockdown in embryonic tissue in notoriously difficult due to the fast development. Therefore, it was not surprising I was unable to observe a specific reduction of Blastopia RNA levels with qPCR in embryos laid by these crosses. Even the positive control, the UAS-TRiP fly line which targets the mRNA of CG32344 capable of inducing a strong knockdown in ovaries, gave only a moderate knockdown in embryos. However, there was some specific decrease in CG32344 RNA levels, while Blastopia showed no specific decrease at all. Which would suggest additional efficiency issues with the Blastopia TRiP fly line, besides the difficulty of embryonic knockdowns. The reason for this is unclear, but may have something to do with the unique nature of TEs. For example, different transgenic fly lines might have different copy numbers of each TE, since they have a slightly different genetic background and retrotransposons can increase their copy number by transposition (MÉREL ET AL., 2020). This could affect the basal expression levels of TEs, as more copies could result in higher expression. Blastopia RNA levels vary between the different control crosses I used, which supports the hypothesis that different fly lines have different copy numbers or different regulation. However, the TRiP-Blastopia #2 and TRiP-G2/Jockey-3 transgenic fly lines were created using the exact same genetic background, and thus should have the same copy number and basal expression level. Since there is no difference in Blastopia RNA levels between those two crosses, it can be concluded that the Blastopia TRiP construct is ineffective. Another difficulty with TEs are the potential mutations between the different genomic copies of Blastopia, compared to the consensus sequence used to design the short hairpin. Therefore, a 21 bp short hairpin RNA might be too short to be able to target all RNAs from these different copies. Altogether, TRiP fly lines might not be the best system to induce a knockdown of TEs.

As an alternative, I induced a knockdown of Blastopia in S2 cells using longer dsRNA, which can then be processed and used by the cell's own RNAi machinery. Each dsRNA is 400 bp long and can therefore target a larger portion of the transcript. Different Blastopia copies, with potentially slight sequence variations, are therefore more likely all targeted. Indeed, although this result is preliminary, the knockdown of Blastopia with longer dsRNA in S2 cells is efficient, reducing transcript levels to only 22 % of the control knockdown. Unfortunately, I was unable to observe an increase in mitotic defects after Blastopia RNA knockdown, potentially due to the high basal rate of mitotic defects in S2 cells.

Nevertheless, S2 cells are frequently used to study mitosis, therefore additional replicates are needed to clarify the effect of the Blastopia knockdown in S2 cells.

However, using dsRNA instead of the short hairpin RNAs of the TRiP fly lines, would be a promising strategy in flies as well. Either by injecting dsRNA into embryos or by creating a UAS-dsRNA transgenic fly line, it would potentially be possible to induce a strong Blastopia knockdown in a whole organism. By combining the UAS-dsRNA with UAS-Dicer2, the RNAi pathway is enhanced, which can thus create an even stronger knockdown (DIETZL ET AL., 2007). Successful knockdown experiments would be imperative to gain understanding of the function of Blastopia RNA at the centromere.

3.1.3 Stress induces a depletion of centromere-associated RNAs

Since different stressors elevated levels of centromeric satellite RNAs in mice and human (BOUZINBA-SEGARD ET AL., 2006; HÉDOUIN ET AL., 2017; JOLLY ET AL., 2004; VALGARDSDOTTIR ET AL., 2005, 2008), I performed the centromeric RNA-ChIP with chromatin from both control and γ-irradiated embryos. This allowed me to assess whether there were any stress-specific centromere-associated RNAs in *Drosophila*. I was unable to observe elevated levels of centromere-associated RNAs or unique stress-specific centromere-associated RNAs after by γ-irradiation induced DNA damage. In contrary, there seem to have been a general depletion of centromere-associated RNAs in embryos exposed to γ-irradiation.

Embryos were exposed to a dose of γ-irradiation which induced a strong DNA damage response without a severe lethality rate. They were aged one hour after the irradiation in order to allow time for the putative stress-specific centromere-associated RNAs to be transcribed and/or recruited. Indeed, this experimental set-up induced the DNA damage response, with a clear upregulation of stress-specific genes and elevated cellular levels of DNA damage marker γH2Av. Surprisingly, besides genes involved in the DNA damage response, several muscle development transcripts were also elevated in the stress-treated embryos. Potentially this is due to a halted development after exposure to γ-irradiation, which could shift expression waves of genes involved in specific developmental stages.

Though it is possible that stress does not affect expression levels or composition of centromere-associated RNAs in *Drosophila*, limitations in the experimental set-up might have restricted detection of changes in centromere-associated RNAs after stress.

First of all, embryos were only aged for one hour after exposure to γ-irradiation. While this maturation time was enough to activate the stress-specific genes of the DNA damage response, upregulation of stress-specific centromere-associated RNAs might require more time. In human HeLa cells, it was observed that different stressors elevated HSATIII transcripts after different recovery times (VALGARDSDOTTIR ET AL., 2008). For example, after heat shock the highest HSATIII RNA levels were

observed within 2 to 4 hours, while with UV-C the peak was only observed more than 8 hours after the exposure. Even with two DNA damaging drugs, MMS and etoposide, the peak HSATIII expression differed, at 2 to 4 hours and 6 to 8 hours after treatment, respectively (VALGARDSDOTTIR ET AL., 2008). Therefore, it is possible that aging the embryos for a longer period after γ-irradiation would have changed the results.

Second, in both mouse and human cells, different types of stress treatments increased satellite transcripts to different extends (BOUZINBA-SEGARD ET AL., 2006; VALGARDSDOTTIR ET AL., 2008). Exposing human cells to a heat shock elevated HSATIII expression more than 10-fold more than DNA damaging agents MMS and etoposide, while hypoxia and thymidine treatment did not induce a higher HSATIII expression (VALGARDSDOTTIR ET AL., 2008). Suggesting a need to explore the effect of different stress inducing treatments and recovery times on centromere-associated RNAs.

Last, it is possible that in *Drosophila* stress-specific centromere-associated RNAs are not specific to centromeres and therefore difficult to specifically identify in this experimental set-up. Repetitive elements that are enriched in the centromeric DNA, are not unique to the centromere in *Drosophila*, all of them are also found at chromosome arms (CHANG ET AL., 2019). It is plausible that, if stress de-represses repetitive elements throughout the genome, an increase of repetitive RNAs at centromeres is not detectable as a specific enrichment at centromeric chromatin compared to genome wide chromatin.

As mentioned, I did observe a significant depletion of several centromere-associated TE RNAs after exposure to γ-irradiation. Although the reductions are small, the quantity of transcripts that show this trend is interesting. This could suggest that upon DNA damage centromeric transcription is repressed or centromeric transcripts are degraded or relocated. Centromeric transcription might be scaled down in order to prioritize transcription of genes involved in the DNA damage response. Alternatively, by reducing centromeric transcripts, centromere functioning is disrupted, which contributes to the cell cycle arrest that is required for repair after DNA damage. Furthermore, centromeric transcript reduction might be a side effect from the DNA damage repair pathway, as severe chromatin remodeling affecting transcription at and near damaged sites are known to be induced (FERRAND ET AL., 2021). It could also be that cells accumulated in a specific cell cycle stage after DNA damage induction. Since fluctuations in centromeric transcript levels during different cell cycle phases have been observed in yeast, mouse and human (FERRI ET AL., 2009; HEDOUIN ET AL., 2022; HOYT ET AL., 2022), this would result in slightly altered centromeric RNA levels in γ-irradiated embryos compared to the more randomly cycling cells in the control embryos. Last, one must consider that RNA sequencing is a relative quantification method. After DNA damage induction, a larger portion of the reads is occupied by stress-

specific RNAs, leaving fewer reads available for centromere-associated RNAs. This could artificially result in a slight general depletion of centromere-associated RNAs in the stressed datasets.

3.2 TE expression upregulated with age and in specific tissues

With smRNA-FISH, I observed Blastopia RNA expressed in specific cell types across embryonic development. The modENCODE consortium has performed extensive analysis of expression patterns of coding and regular non-coding RNAs in *Drosophila* (BROWN ET AL., 2014; DUFF ET AL., 2015). They used deep sequencing to systematically sequence rRNA-depleted total RNA from 30 developmental stages and a multitude of dissected tissues. This dataset was used to provide the *Drosophila* research community an important resource, describing expression patterns of all genes. However, like most studies, they omitted TEs and repeats from their analysis, as repetitive sequences are difficult to analyze. Meaning that information on the expression patterns of TEs and repeats is still not publicly available. In order to assess the general expression pattern of Blastopia and other TEs and repeats, I re-analyzed the published modENCODE RNA-seq dataset (DUFF ET AL., 2015). Although centromere-associated RNAs discovered to date have all been transcribed by RNA polymerase II, and thus containing a poly-A tail, it is plausible not all RNA copies of satellite repeats or TEs have this tail. Therefore, it is beneficial the modENCODE consortium used rRNA-depleted total RNA instead of poly-A selected RNA, preventing detection biases. I fully reprocessed this large dataset, normalized the counts to sequencing depth (reads per million) and transformed these normalized counts to Z-scores in order to visualize expression patterns for each TE or repeat. Comparison of transcript levels between different TEs or repeats are difficult, as these differences could be caused by the unknown copy number differences in the incomplete genome.

As expected, most TEs showed increased expression levels in aging adults. Aging is associated with severe epigenetic remodeling, loss of heterochromatin marks and a resulting increase in TE transcription (GORBUNOVA ET AL., 2021; LÓPEZ-OTÍN ET AL., 2013). For example in flies, repressive marks, such as H3K9me3 and HP1, are lost at all heterochromatin loci with age (WOOD ET AL., 2010). The associated increase in TE expression are suggested to aggravate the aging process, as dietary restrictions and genetic interventions which maintained heterochromatin, delayed TE upregulation and increased the life span of the animal (WOOD ET AL., 2016).

Tissues with a general higher TE expression are pupae fat bodies and the central nervous system. The fat body, which is equivalent of mammalian liver and adipose tissue, has been connected to an upregulation of TE expression in older adult flies (H. CHEN ET AL., 2016; WOOD ET AL., 2016). This increase was linked to an age-dependent reduction of Lamin B, which leads to the loss of H3K9me3 and increased DNA damage (H. CHEN ET AL., 2016). The high TE expression levels observed in pupae fat bodies are independent of aging, but might be linked to fat body remodeling which occurs during metamorphosis (ZHENG ET AL., 2016). Metamorphosis is a period without feeding, energy stored within fat body cells has to be released which induces programmed cell death (ZHENG ET AL., 2016).

In both mammalian and *Drosophila* brains, TE expression is associated with several neurodegenerative diseases and psychiatric disorders (AHMADI ET AL., 2020; REILLY ET AL., 2013). However, recent advances has linked TE expression to healthy brain development (AHMADI ET AL., 2020; TREIBER & WADDELL, 2020). Brain complexity benefits from an increase in neural cell diversity. To generate more different cells, brains make use of a phenomenon called somatic mosaicism, meaning the presence of slightly different genomes in different cells of the same individual. TE activity is one of the mechanisms with which the brain induces mutations to diversify the genome. The exact mechanism remains unclear, however in fly, individual TE expression in specific neurons has been linked to alternative splicing of genes containing those TEs in their introns (AHMADI ET AL., 2020; TREIBER & WADDELL, 2020). The higher expression levels of many TEs in neuronal tissues throughout development I observe in the modENCODE data, supports this hypothesis.

3.2.1 Blastopia RNA in the brain

Blastopia follows a similar expression pattern observed for most TEs, with higher expression in fat body and neuronal tissue. Especially in brains of older adults, Blastopia expression spikes. With RNA-FISH I was able to observe Blastopia RNA expressed in a small subset of neurons of 3rd instar larval brains. Suggesting that Blastopia expression might be one of the TEs contributing to neuronal mosaicism, diversifying neurons by expressing different TEs in different neurons. Treiber & Waddell (2020) showed that a Blastopia copy inserted into an intron of *cacophony*, coding for a voltage-gated calcium channel, truncates and thus inactivates this protein in a small portion of neurons. Thereby changing currents present in these neurons. Although this is a very interesting function of TEs like Blastopia, it is unlikely that Blastopia's neuronal expression pattern is related to its potential centromeric function, especially since neurons are post mitotic.

3.3 RNA-binding centromeric proteins

The second aim of my thesis was to assess and understand whether the proteins involved in CID loading, namely CID, Cal1 and CENP-C, directly interact with RNA. All of these proteins, or their homologs in other species, have been suggested to have an interaction with cenRNAs (DU ET AL., 2010; GRENFELL ET AL., 2016; MCNULTY ET AL., 2017; QUÉNET & DALAL, 2014; ROŠIĆ ET AL., 2014; WONG ET AL., 2007). Due to the intrinsically disordered regions of Cal1 and CENP-C, both are strong candidates to directly bind cenRNAs (CALABRETTA & RICHARD, 2015; MEDINA-PRITCHARD ET AL., 2020). Centromeric CID levels are affected by knockdown of SATIII and inhibition of transcription, suggesting cenRNAs could impact the functioning of this complex (ROŠIĆ ET AL., 2014). Knowing which of these proteins is responsible for the binding the RNA component of the complex, and subsequently which RNA this protein is interacting with, could further our understanding of this process.

I made use of a relatively new isolation method for RNA-protein complexes, called XRNAX, and adapted the method to Drosophila S2 cells (TRENDEL ET AL., 2019). UV crosslinked RNA-protein complexes separate into the interphase of TRIzol isolation, making it easy to select and purify. By analyzing the XRNAX extract with western blot, comparing the relative levels of the proteins of interest in the XRNAX extract to the levels in whole cell lysate, I could determine which proteins directly bind RNA. I was able to confirm that Cal1 is indeed an RNA binding protein. For CENP-C there was no enrichment in the XRNAX extract, suggesting this protein do not directly interact with RNA in S2 cells. Although not previously described as a direct RNA binding protein, also CID was enriched in the XRNAX extract.

3.3.1 CENP-C

Especially for CENP-C, it is surprising to not find it enriched in the XRNAX extract. *Drosophila* CENP-C was able to pull down SATIII RNA in an RNA-IP (ROŠIĆ ET AL., 2014). Also in other species, CENP-C was shown to interact with cenRNAs with IPs or electrophoretic mobility shift assays (EMSA) (DU ET AL., 2010; GRENFELL ET AL., 2016; MCNULTY ET AL., 2017; WONG ET AL., 2007). Next to it intrinsically disordered domains, CENP-C contains two AT-hook motifs and an Arginine-rich motif responsible for its interaction with DNA, which are also able to bind RNA (BAYER ET AL., 2005; FILARSKY ET AL., 2015; HEEGER ET AL., 2005). Despite these previous findings, I was consistently unable to observe CENP-C enriched in the XRNAX extract. This could suggest that those previous described interactions are indirect or not conserved in Drosophila. Alternatively, the interaction of CENP-C with RNA is transient or only during specific phases of the cell cycle, making that the interaction is not detected in asynchronously cycling S2 cells. Potentially there are also technical limitation with the XRNAX extraction method, which could limit the enrichment of CENP-C. While the size of CENP-C often limits its isolation or detection, XRNAX extraction has no difficulty with high mass proteins. The original paper was able to identify 246 proteins

with a larger molecular weight than CENP-C as RNA-binding proteins, with the largest protein being 3816 kDa (TRENDEL ET AL., 2019). What could be a potential limitation is the lack of the specific amino acid types in the RNA-protein interface of CENP-C, which are capable to be covalently linked by UV crosslinking. UV crosslinking activates the photo-reactivity of closely interacting nucleotides, specifically pyrimidines, and certain amino acids, most frequently Cysteine, Lysine, Phenylalanine, Tryptophan or Tyrosine (ULE ET AL., 2005). RNA-protein interactions are often mediated by hydrogen bonds or hydrophobic interactions, in which those UV sensitive amino acids can play a big role (CORLEY ET AL., 2020). While, AT-hook and Arginine-rich motifs by themselves do not contain any UV photo-reactive amino acids, there are several Lysine amino acids surrounding these motifs in CENP-C. Furthermore, CENP-C consists of a similar percentage of UV photo-reactive amino acids (12.4 %) as Cal1 (14.8 %) and CID (9.3 %). Therefore, although possible, it is unlikely CENP-C is missing from the XRNAX extract due to a missing UV reacting substrate. Meaning that CENP-C is most likely not directly RNA binding in *Drosophila* S2 cells.

3.3.3 Cal1 and CID

Cal1 showed a clear enrichment in the XRNAX extract, confirming for the first time that it is an RNA-binding protein. The intrinsically disordered regions of Cal1 indicated its capacity to interact with RNAs, as most RNA-binding proteins contain them (CORLEY ET AL., 2020). Also CID was significantly enriched in the XRNAX extract. The human orthologue of Cal1, HJURP was shown to bind RNAs while in a soluble complex with CENP-A, prior to CENP-A loading (QUÉNET & DALAL, 2014). I was able to confirm HJURP as an RNA-binding protein with XRNAX. It is possible that also in *Drosophila* a soluble, pre-loading complex of Cal1 and CID interacts with RNA.

This observation may support the hypothesis that RNA is the missing link it the meiotic drive in *Drosophila*. Both CENP-A/CID and centromeric DNA is rapidly evolving (ROSIN & MELLONE, 2017). Centromeric DNA is constantly mutating and expanding, which would also expand the core centromere domain and consequently stronger mitotic attachments. Chromosomes with expanded centromeres are selected for during female meiosis, as the strongest of the two homologous chromosomes migrates towards the egg. In male meiosis on the other hand, this centromere imbalance between homologous chromosomes could reduce fertility due to meiotic defects. It has been hypothesized that rapid CENP-A/CID evolution compensates centromeric DNA evolution and expansion, although it is still unclear how such mutations in CENP-A/CID would specifically favor incorporation at the weaker centromeric DNA. Furthermore, it is even more surprising that in *Drosophila*, even though Cal1 evolves slowly, there is an incompatibility between Cal1 and CID proteins of related species (ROSIN & MELLONE, 2017). Centromeric RNA could bind at the interphase of Cal1 and CID, bridging the divergently evolving proteins. The protein-protein interface between Cal1 and CID potentially adjusts to the secondary

structure of the newly evolved cenRNAs, allowing Cal1 to evolve slower as long as it is able to interact with RNA. Additionally, the bound cenRNAs could aid the distribution of CID across all centromeres, by tether the complex to its site of transcription. Upregulating transcription at smaller centromeres, could recruit more CID loading complex and strengthen its centromere.

3.3.4 Copia RNA putative interaction partner of Cal1

In order to establish which RNA interacts with Cal1, I performed a pulldown of Cal1 from the UV cross-linked XRNAX extract followed by deep sequencing of the attached RNA (XRNAX-CLIP-seq). There are several advantages for using the XRNAX extract instead of UV cross-linked whole cell lysate as an input for the CLIP. First and foremost, during XRNAX isolation the samples are treated with DNase, removing the genomic DNA. This prevents contamination, which is especially advantageous in the case of repetitive sequences. Furthermore, for low abundant proteins like Cal1, this additional enrichment step benefits its detection.

I was able to specifically pull down Cal1, as was shown with the control western blots. Among the differentially enriched RNAs in the Cal1 CLIP compared to both the size-matched input and the IgG CLIP controls, were several mRNAs and ncRNAs. However, the most significantly enriched RNA was Copia, both the internal part and its LTR are at least 2-fold enriched. There are 70 full length copies of Copia present in the genome and the re-analyzed modENCODE data showed exceptionally high expression levels of Copia throughout the *Drosophila* life cycle and in most tissues. Importantly, there is also a truncated copy of Copia found in the core centromeric DNA of the Y chromosome (CHANG ET AL., 2019). Furthermore, Copia RNA is also one of the candidate centromere-associated RNAs I identified with the CID RNA-ChIP. Altogether, this makes Copia an interesting candidate RNA for interacting with Cal1.

To get an initial understanding of the localization of Copia, I performed smRNA-FISH in combination with CENP-C IF in stage 5 embryos. Copia RNA is present in all cells of these embryos in multiple foci throughout the nuclei. One-third of the cells have at least one overlapping Copia RNA-FISH focus with a CENP-C IF focus, which is comparable to Blastopia and SATIII RNA. However, it is plausible that Cal1 interact with Copia RNA not only at the centromere but also elsewhere in the nucleus. A Bachelor student in the lab, David Klein, performed an electrophoretic mobility shift assay (EMSA), an *in vitro* binding assay, with purified middle part of Cal1 and *in vitro* transcribed Copia, SATIII and ncRNA Hsromega. As there were difficulties cloning the full length of Copia, he used two pieces of Copia which I had identified as most enriched in the Cal1 XRNAX-CLIP-seq data, and are most likely to be involved in the direct interaction with Cal1. Although the middle part of Cal1 seems to be a promiscuous RNA binder, lower concentrations of this protein are required to induce interaction with Copia RNA

compared to SATIII or Hsromega RNA. Altogether, these results give an initial indication Copia is indeed associated with centromere where it can interact with Cal1.

3.4 Increased SATIII RNA without heterochromatin formation

During early embryogenesis, heterochromatin has to be newly established in order to repress expression of potentially damaging repeats. In a collaboration with the Iovino lab of the Max-Planck Institute of Immunobiology and Epigenetics in Freiburg, I explored how SATIII and G2/Jockey-3 expression was affected in embryos which lack functional copies of all three H3K9 methyltransferases, Su(var)3-9, Eggless and G9a (SELLER ET AL., 2019). RT-qPCR showed elevated expression levels of both repeats in stage 5 embryos. Surprisingly, RNA-FISH experiments found that SATIII sense and antisense were affected differently by the lack of heterochromatin formation. While SATIII sense RNA was robustly upregulated throughout the stage 5 embryos, SATIII antisense RNA was significantly more elevated at the posterior pole of the embryo. It is unclear why SATIII antisense behaves differently. However, these results confirm that H3K9 methyltransferases are required for suppressing (peri)centromeric repeat expression. The Iovino lab showed that these embryos die shortly after activation of zygotic transcription. This is not surprising, as high levels of expression of (peri)centromeric repeats probably induce mitotic defects and TE activity can have detrimental effects on the genome (ARUNKUMAR & MELTERS, 2020; BOURQUE ET AL., 2018).

3.5 Conclusion and perspectives

In this thesis, I showed that CID and Cal1 are RNA binding proteins. Furthermore, I identified several novel centromere-associated RNAs, Blastopia and Copia, of which Copia directly interacts with Cal1 is S2 cells. The results of this study strengthen the idea that RNAs are an integral component of the centromere, with several RNA binding proteins at the heart of the centromere. Furthermore, these data suggest that more of the repetitive elements present at the (peri)centromeric DNA, not just SATIII, are transcribed and present at centromeric chromatin.

The study was only able to validate the centromeric localization of Blastopia and Copia to a limited extend. For both RNAs there the co-localization with centromeres was only observed in some cells and at some centromeres. Knockdown experiments of Blastopia in embryos were unsuccessful and

although the knockdown in cells was efficient, initial observations did not show an increase in mitotic defects. Therefore, it remains elusive what the function is of Blastopia RNA at centromeres. Considerably more work will need to be done to validate the centromeric localization and function of both Blastopia and Copia.

First, reciprocal RNA pull downs can be performed to confirm the interaction with centromeric proteins, and for Copia RNA Cal1 specifically. Either by incubating cell lysate with *in vitro* transcribed RNA containing streptavidin-binding S1M-loops, followed by western blot on the eluted proteins (LEPPEK & STOECKLIN, 2014). Alternatively, by performing the same affinity purification *in vivo*, using transgenic cell or fly lines with an S1M-loop tagged version of Blastopia and Copia. This would capture the interactions as they actually occur in the cells, especially when these interactions are first stabilized by UV crosslinking.

Additional exploration of the function of Blastopia and Copia should be conducted, either with successful knockdown experiments or by overexpressing the RNAs. Since the knockdown with longer dsRNA seem to be a promising method to use for potentially variable TE copies, this can be used in both S2 cells and flies. Systematic analysis of cell cycle progression and CID levels can disclose potential functions of these novel cenRNAs. Furthermore, the knockdown cells can confirm the specificity of the RNA-FISH probe sets used. Since elevated levels of several cenRNAs have detrimental impact on centromere functioning (BOUZINBA-SEGARD ET AL., 2006; D. Y. L. CHAN ET AL., 2017; Q. ZHU ET AL., 2011), overexpressing Blastopia and Copia could also further our understanding of their impact on the centromere.

For Copia specifically, it will be important to repeat the IF-RNA-FISH colocalization experiments with Cal1 instead of CENP-C. Although Cal1, similarly to CENP-C, is localized to the centromere throughout the cell cycle, in interphase Cal1 is also observed at the nucleolus (SCHITTENHELM ET AL., 2010). Unfortunately, soluble, pre-loading complexes of Cal1 and CID might be difficult to observe as their signal might be too diffuse.

As both mono-ubiquitinated CID and Cal1 are RNA-binders, it is conceivable that Copia or other cenRNAs are a structural component of the CID loading complex, as was proposed in human (QUÉNET & DALAL, 2014). Future research in our lab, performed by Vojtěch Dolejš, aims to resolve the entire structure of the CID loading complex, including its putative RNA component. He plans to use hydrogen-deuterium exchange Mass Spectrometry to map the interactions of the proteins with RNA and each other. It will be interesting to determine which domains of Cal1 and CID are responsible for their RNA binding capacity. For this, truncated and mutated versions of the proteins can be used in EMSAs.

I also tried to identify stress-specific centromere-associated RNAs. Although I did not observe any RNAs specifically recruited to centromeres one hour after exposure to γ-irradiation, there was a general

depletion of centromere-associated RNAs. An implication of this is the possibility that in *Drosophila* centromere-associated RNAs are degraded or displaced in order to limit mitotic progression. As discussed, there are several limitations to this experimental set-up, making it plausible that under different conditions specific RNAs are recruited to the centromere. Without additional genome-wide assays, it will be difficult to assess whether different stress treatments and/or recovery times would change the centromere-associated RNA composition. However, it is possible to explore the effects of these treatments on the by this study newly identified centromere-associated RNAs. After exposing cells or embryos to different stress inducing conditions, the abundance of different centromere-associated RNAs can be assess with RT-qPCR and changes in transcript size can be explored with northern blot.

Last, I was able to provide the expression patterns of repetitive elements in *Drosophila* development and tissues. By re-analyzing the modENCODE RNA-seq data for satellite repeats and TEs, I was able to confirm previous findings which implicated TE upregulation with age and in brain and fat body. For Blastopia, I determined a specific expression pattern in a subset of neurons of 3rd instar larval brains. Suggesting that Blastopia has additional functions outside the centromere, being one of the TEs contributing to the somatic mosaicism stimulating post-mitotic neuronal cell diversity.

In conclusion, the results in this thesis have extended the list of cenRNAs and broadened our understanding of them. I have demonstrated cenRNAs are indeed a structural component of centromeres, binding directly to CID and Cal1 *in vivo*. I have identified several novel putative centromere-associated RNAs. This will aid future research on centromere-associated RNAs in *Drosophila*.

4. Materials

4.1 Chemicals

Chemical	Provider
Acrylamide/Bisacrlamide(30:0.8 %)	AppliChem
Agar bacteriology grade	AppliChem
Agarose Ultra Pure	Sigma-Aldrich
Ammonium persulfate (APS)	AppliChem
Ampicillin Sodium salt	AppliChem
Aprotinin	AppliChem
Bromphenol blue	AppliChem
BSA (Bovine serum albumin)	Th. Geyer
Chloroform	Roth/Sigma-Aldrich
Copper sulphate (CuSO4)	Th. Geyer
DAPI	Sigma-Aldrich
DEPC	AppliChem
Dextran sulfate salt	AppliChem
Dimethyl sulfoxide (DMSO)	AppliChem
Dithiothreitol (DTT)	Sigma-Aldrich
Egtazic acid (EGTA)	AppliChem
Ethanol absolute (EtOH)	Sigma-Aldrich
Ethidium Bromide (EtBr)	AppliChem
Ethylenediaminetetraacetic acid (EDTA)	Roth
Ficoll 400	Sigma-Aldrich
Formaldehyde 16 % ampules	Thermo Fisher Scientific
Formaldehyde 37 % (methanol balanced)	Sigma-Aldrich
Glucose	AppliChem
Glycerol	Roth
Glycine	Sigma-Aldrich
Grape juice	Rewe
HEPES	Sigma-Aldrich
Heptane	Sigma-Aldrich
Isopropanol	Sigma-Aldrich
Leupeptin-Hemisulfate	AppliChem
Lithium chloride (LiCl)	Sigma-Aldrich
Magnesium chloride (MgCl2)	AppliChem
Methanol	ZMBH
NIPAGIN	Sigma-Aldrich
N-Lauroylsarcosine	Sigma-Aldrich
Nonidet P-40 (NP-40)	AppliChem
Paraformaldehyde (PFA)	AppliChem
Pepstatin	AppliChem

Chemical	Provider
Phenol-chloroform-isoamyl alcohol 25:24:1 (PCI)	AppliChem
Phenol-chloroform-isoamyl alcohol 25:24:1, pH 6.6	Sigma-Aldrich
Phenylmethylsulfonylfluoride (PMSF)	Sigma-Aldrich
Polyvinylpyrrolidone (PVP)	Sigma-Aldrich
Ponçeau S solution	AppliChem
Potassium chloride (KCl)	AppliChem
potassium dihydrogenphosphate (KH2PO4)	AppliChem
RNaseZAP	Sigma-Aldrich
Skim milk powder	Sigma-Aldrich
Sodium acetate (NaOAc)	Sigma-Aldrich
Sodium chloride (NaCl)	Sigma-Aldrich
Sodium citrate	Sigma-Aldrich
sodium deoxycholate (DOC)	AppliChem
Sodium dodecyl sulfate pellets (SDS)	Serva
Sodium hydroxide pellets (NaOH)	AppliChem
Sodium hypochlorite	Roth
Sodium phosphate, dibasic (Na2HPO4)	AppliChem
Sodium phosphate, monobasic (NaH2PO4)	Sigma-Aldrich
Sucrose	Sigma-Aldrich
Tetramethylethylenediamine (TEMED)	AppliChem
Tris base	AppliChem
Tris-HCl	AppliChem
Triton X-100	Merck
Trypan Blue	Sigma-Aldrich
Tween-20	AppliChem
Urea	Sigma-Aldrich
Yeast cube	Rewe
β-Mercaptoethanol	AppliChem

4.2 Reagents/Enzymes/Kits

Reagent/Enzyme/Kit	Provider
1 kb Plus DNA ladder	NEB
100 bp DNA ladder	NEB
4-12 % Bis-Tris Criterion XT Midi gel, 12+2 well	Bio-Rad
Ampure XP beads	NEB
ATP (Rnase-free)	Invitrogen
Bioanalyzer High Sensitivity DNA kit	Agilent
Cellfectin II	Gibco
Complete EDTA-free Protease Inhibitor cocktail	Roche
Direc-Zol RNA microprep kit	Zymo Research
Direct-Zol RNA miniprep kit	Zymo Research
Distilled water, Dnase/Rnase-free	Gibco/Invitrogen

Reagent/Enzyme/Kit	Provider
DNaseI (Rnase-free)	NEB
dNTPs	NEB
ECL Femto Maximum Chemiluminescent Substrate	Thermo Fisher Scientific
ECL SuperSignal West Pico PLUS	Thermo Fisher Scientific
FastAP	Thermo Fisher Scientific
Fetal bovine serum (FBS)	PAN
Fragment analyzer HS NGS kit	Agilent
Fragment analyzer HS RNA kit	Agilent
Gel Loading Dye, purple (6X)	NEB
Gibson Assembly mix	lab-made
GlycoBlue	Invitrogen
HuluFISH hybridization buffer	PixelBiotech
HuluFISH probes	PixelBiotech
HuluFISH wash buffer	PixelBiotech
LightCycler 480 SYBR Green I Master	Roche
MaXtract High Density tube	Qiagen
MEGAscript RNAi Kit	Thermo Fisher Scientific
Mounting medium aqua/polymount	Polysciences
NextSeq 500/550 high output kit v2.5 with 150 cycles	Illumina
NextSeq 500/550 high output kit v2.5 with 75 cycles	Illumina
NucleoSpin Gel and PCR Clean-up kit	Machery-Nagel
Nucleospin plasmid kit	Machery-Nagel
NuPAGE MOPS SDS Running Buffer	Bio-Rad
NuPAGE transfer buffer	Bio-Rad
Penicillin/Streptomycin	Capricorn Scientific
PNK	Thermo Fisher Scientific
Protein A Dynabeads	Thermo Fisher Scientific
Proteinase K	lab-made
Proteinase K	Thermo Fisher Scientific
Puromycin	Sigma-Aldrich
Q5 Polymerase	NEB
QuantiTect Reverse Transcription Kit	Qiagen
Qubit dsDNA HS Assay Kit	Invitrogen
Qubit RNA HS Assay Kit	Invitrogen
RedTaq Polymerase mix	Jena Bioscience
Restriction enzymes	NEB
Ribonucleoside Vanadyl Complex (RVC)	NEB
RNA clean and concentrator kit-5	Zymo Research
RNA Loading Dye (2X)	NEB
RNase A	AppliChem
RNase A	Sigma-Aldrich
RNasin Plus Ribonuclease Inhibitor	Promega
Schneider’s Drosophila Medium	Gibco
SMARTer RNA unique Dual index kit – 24U	TAKARA

Reagent/Enzyme/Kit	Provider
SMARTer smRNA-Seq kit for Illumina	TAKARA
Sodium acetate (3 M NaOAc, Rnase-free, pH 5)	Thermo Fisher Scientific
ssRNA ladder	NEB
TRIsure	Bioline
TRIzol	Invitrogen
TURBO DNase	Invitrogen
Universal RNA-seq kit with NuQuant and Drosophila specific rRNA depletion	TECAN

4.3 Equipment and consumables

Equipment	Provider
Agarose gel electrophoresis tank, trays and combs	ZMBH workshop
Balance	Sartorius, Kern AG
Binocular microscope	Zeiss
Bioanalyzer 2100	Agilent
Biorupter Plus	Diagenode
Blotting paper	Roth
Cell counting system, LUNA	Logos Biosystems
Cell culture 6-well plate	Greiner Bio-One
Cell culture flasks	TPP
Cell culture incubator	Heraeus
Centrifuges	Eppendorf
Covaris m220 Focussed-ultrasonicator	Covaris
Coverslips	Neolab
Criterion vertical electrophoresis	Bio-Rad
Deltavision microscope	Olympus/GE Healthcare
Fly cages	ZMBH workshop
Fly vials	Stein, Greiner
Fragment Analyzer 5200	Agilent
Gel Doc XR+ System	Bio-Rad
LAS-4000	Fujifilm Life Science
LightCycler 480 384-well system and plates	Roche
Magnetic stand	ZMBH workshop, GE Healthcare
Microscopy slides (polylysine/Superfrost Plus)	Thermo Fisher Scientific
Mini Trans-Blot cell	Bio-Rad
Mini-PROTEAN tetra handcast system	Bio-Rad
Mini-PROTEAN tetra vertical electrophoresis	Bio-Rad
Mr. Frosty	Thermo Fisher Scientific
Nanodrop A260	Nanodrop
NextSeq550 sequencer	Illumina
Nitrocellulose membrane	Amersham Biosciences
Nutating mixer	VWR

Equipment	Provider
Petri dishes	Greiner Bio-One
pH meter	Satorius, Kern AG
Pipette tips	Sarstedt, TipOne, Avant Guard
PowerPac power supply	Bio-Rad
Qubit 3 Fluorometer	Invitrogen
Rotation wheel	Labinco BV
Thermocycler	Bio-Rad, Nippon Genetics
Thermomixer	Eppendorf
Trans-Blot cell, plate electrodes	Bio-Rad
Tube roller	IDL
Tubes 0.2 µl – 50 ml	Sarstedt, Eppendorf
UV crosslinker	Bio-Link
Zeiss LSM 900	Zeiss

4.4 Buffers

Buffer	Components
10X Phosphate Buffered Saline (PBS)	1.37 M NaCl 27 mM KCl 100 mM Na_2HPO_4 18 mM KH_2PO_4
1X SDS running buffer	25 mM Tris-Base 192 mM Glycine 0.1 % SDS
20X SSC	175.3 g/l NaCl 88.2 g/l sodium citrate-2H2O adjusted to pH 7.0 with HCl
4X Laemmli sample buffer	50 mM Tris-HCl (pH 6.8) 10 % Glycerol 2 % SDS 0.5 % β -Mercaptoethanol 0.02 % Bromphenolblue
50X Tris-acetate-EDTA (TAE)	242 g/l Tris-HCl 18.6 g/l EDTA pH 7.7 adjusted with acetic acid
Denhardt's solution	1 % (w/v) BSA 1 % (w/v) Ficoll 400 1 % (w/v) Polyvinylpyrrolidone (PVP)

Buffer	Components
Fly food (by ZMBH kitchen)	7.2% (w/v) maize 2.4% molasses 7.2% (w/v) malt 0.88% (w/v) soya 1.464% (w/v) yeast acid mix (1% propionic acid + 0.064% orthophosphoric acid)
Hulu Hybridization buffer (cells)	2X SSC 2 M Urea 10 % Dextran sulfate sodium salt 5X Denhardt's solution
Hulu Hybridization buffer (tissue)	2X SSC 2 M Urea 10 % Dextran sulfate sodium salt 5X Denhardt's solution 0.1 % Tween-20
Hulu Wash buffer (cells)	2X SSC 2 M Urea
Hulu Wash buffer (tissue)	2X SSC 2 M Urea 0.1 % Tween-20
PBST	1X PBS 0.1 % Tween-20
Protease inhibitor solution A	1 mg/ml Aprotinin 1 mg/ml Leupeptin-Hemisulfate
Protease inhibitor solution B	0,5 g/ml Pepstatin in EtOH
RIPA buffer	50 mM Tris-HCl (pH 7.5) 500 mM NaCl 1 % NP-40 0.5 % sodium deoxycholate (DOC, 10 % stock, in plastic, stored dark) 0.5 % SDS 2 mM PMSF Proteinase Inhibitor A (1:1000) Proteinase Inhibitor B (1:500)
RNA-ChIP in embryos Cell lysis buffer	5 mM HEPES (pH 8.0) 85 mM KCl 0.5 % NP-40 1X Roche protease inhibitors cocktail Ribonucleoside Vanadyl Complex 1:20

Buffer	Components
RNA-ChIP in embryos Dilution buffer	4 % glycerol 10 mM Tris-HCl (pH 8) 1 mM EDTA 0.5 mM EGTA 1X Roche protease inhibitors cocktail Ribonucleoside Vanadyl Complex 1:20
RNA-ChIP in embryos Elution buffer	10 mM Tris-HCl (pH 8) 1 mM EDTA Ribonucleoside Vanadyl Complex 1:20
RNA-ChIP in embryos Fixing solution	50 mM HEPES (pH 8) 1 mM EDTA (pH 8) 0.5 mM EGTA (pH 8) 10 mM NaCl 1.6 % formaldehyde (methanol-free ampules, 16 % stock)
RNA-ChIP in embryos High-salt wash	500 mM NaCl 10 mM Tris-HCl (pH 8) 1 mM EDTA 0.1 % SDS 1 % Triton X-100 0.1 % sodium deoxycholate (DOC) 1 mM PMSF (Ribonucleoside Vanadyl Complex 1:50)
RNA-ChIP in embryos LiCl wash	0.25 M LiCl 10 mM Tris-HCl (pH 8) 1 mM EDTA 0.5 % NP-40 0.5 % sodium deoxycholate (DOC) (Ribonucleoside Vanadyl Complex 1:50)
RNA-ChIP in embryos Low-salt wash	140 mM NaCl 10 mM Tris-HCl (pH 8) 1 mM EDTA 0.1 % SDS 1 % Triton X-100 0.1 % sodium deoxycholate (DOC) 1 mM PMSF (Ribonucleoside Vanadyl Complex 1:50)
RNA-ChIP in embryos Nuclear lysis buffer	50 mM HEPES (pH 8.0) 10 mM EDTA (pH 8.0) 0.5 % (w/v) N-Lauroylsarcosine 1X Roche protease inhibitors cocktail Ribonucleoside Vanadyl Complex 1:20
RNA-ChIP in embryos PBST (+ inhibitors)	1X PBS 0.1 % Tween-20 (1X Roche protease inhibitors cocktail) (Ribonucleoside Vanadyl Complex 1:20)

Buffer	Components
RNA-ChIP in embryos Stop solution	1X PBS 0.1 % Tween-20 125 mM glycine
RNA-ChIP in embryos TE wash	10 mM Tris-HCl (pH 8) 1 mM EDTA (Ribonucleoside Vanadyl Complex 1:50)
Separation gel	375 mM Tris-HCl (pH 8.8) 8-12 % Acrylamide/Bisacrylamide 30:0.8 % 0.1 % SDS 0.05 % APS 0.05 % TEMED
Stacking gel	123 mM Tris-HCl (pH 6.8) 4.4 % Acrylamide/Bisacrylamide 30:0.8% 0.1 % SDS 0.03 % APS 0.1 % TEMED
TBST	20 mM Tris-HCl (pH 7.5) 150 mM NaCl 0.1 % Tween-20
Transfer buffer	25 mM Tris-Base 192 mM Glycine 20 % Methanol
XRNAX Proteinase K buffer	50 mM Tris-HCl (pH 8) 10 mM EDTA 150 mM NaCl 1 % SDS
XRNAX TE	10 mM Tris-HCl (pH 7.5) 1 mM EDTA
XRNAX-CLIP 2X IP buffer	100 mM Tris-HCl (pH 7.5) 200 mM NaCl 2 % NP-40 0.2 % SDS 1 % sodium deoxycholate (DOC) 2X Roche Proteinase Inhibitor cocktail
XRNAX-CLIP 5X SDS loading buffer	250 mM Tris-HCl (pH 6.8) 10 % SDS 0.02 % Bromphenol blue 30 % Glycerol
XRNAX-CLIP TBST	50 mM Tris-HCl (pH 7.5) 150 mM NaCl 0.1 % Tween-20

4.5 Antibodies

4.4.1 Primary antibodies

Antibody	Dilution	Species	Company
Actin	WB 1:10.000	Mouse	Millipore (MAB1501)
Cal1	WB 1:3000	Rabbit	Erhardt lab
CENP-C	IF 1:800	Rat	Erhardt lab
CID	WB 1:2000	Rabbit	Active Motif (39713)
CID	IF 1:500 RNA-ChIP 6 µl	Rabbit	Active Motif (39719)
Elav	IF 1:10	Rat	DSHB (7E8A10)
H3	WB 1:10.000 RNA-ChIP 3 µl	Rabbit	Abcam (ab1791)
H3phosphoSer10	IF 1:5000	Rabbit	Abcam (ab5176)
HJURP	WB 1:50	Rabbit	Abcam (ab100800)
hnRNPm	WB 1:1000	Mouse	Santa Cruz (sc-20001)
IgG	CLIP 5 µg	Rabbit	Sigma (I-8140-10MG)
Lamin	WB 1:1000	Mouse	DSHB (ADL84.12-c)
Rumpelstiltskin	WB 1:2000	Mouse	DSHB (10C3)
Tubulin	IF 1:2000	Mouse	Sigma (T5168)
Tubulin	WB 1:5000	Mouse	Sigma (T9026)
V5	IF 1:1000 WB 1:5000	Mouse	Invitrogen (R960-25)
V5	CLIP 5 µg	Rabbit	Abcam (ab15828)
γH2Av	WB 1:1000	Rabbit	Rockland (600-401-914)

4.4.2 Secondary antibodies

Antibody	Dilution	Species	Company
AF-488 α-mouse IgG	1:500	Goat	Thermo Fisher Scientific
AF-488 α-rat IgG	1:500	Goat	Thermo Fisher Scientific
AF-488 α-rabbit IgG	1:500	Goat	Thermo Fisher Scientific
α-mouse IgG HRP	1:10.000	Rabbit	Sigma (A9044)
α-rabbit IgG HRP	1:5000	Goat	Sigma (A0545)
TrueBlot α-rabbit IgG HRP	1:1000	Mouse	Biomol

4.6 Primers

Primer	Sequence
Cal1_KpnI Fw	atctagatcggggtacATGGCGAATGCGGTGGTG
Cal1_XhoI Rv	aagggccctctagactcCTTGTCACCGGAATTATTCTCGAG
Blastopia_shRNA_top	ctagcagtGTGCTAAAGTCGAAGAGAAATtagttatattcaagcataATTTCTCT TCGACTTTAGCACgcg

Primer	Sequence
Blastopia_shRNA_bottom	aattcgcGTGCTAAAGTCGAAGAGAAATtatgcttgaatataactaATTTCTCTT CGACTTTAGCACactg
G2_shRNA_top	ctagcagtCAACAGATGCTTCTCACTAGGtagttatattcaagcataCCTAGTGA GAAGCATCTGTTGgcg
G2_shRNA_bottom	aattcgcCAACAGATGCTTCTCACTAGGtatgcttgaatataactaCCTAGTGAG AAGCATCTGTTGactg
qPCR Blastopia Fw	ATGAATACTCTAACCCAGAG
qPCR Blastopia Rv	CATCACAGTCACTTACAACT
qPCR Cen4-G2 Fw	CCAATTGCGATGCGGTTCC
qPCR Cen4-G2 Rv	CTGCAGAGGTGGATGTTGTG
qPCR CenX-G2 Fw	TCCAACGTCCATCCCTACCA
qPCR CenX-G2 Rv	TAGTCAAGAGGGTTGACGGC
qPCR CenX-Doc2 Fw	CTGCACTCGTTATTGCAGTG
qPCR CenX-Doc2 Rv	GAGCACCCTCACCTTCAACT
qPCR SATIII Fw	AATGGAAATTAAATTTTTTGGCC
qPCR SATIII Rv	GTTTTGAGCAGCTAATTACC
qPCR CG32344 Fw	CGTCCTGGTGGTACCGTTTT
qPCR CG32344 Rv	CCTCGTCTATGTCCTCTTCTGC
qPCR Actin Fw	TGGCACCGTCGACCATGAAGATC
qPCR Actin Rv	TTAGAAGCACTTGCGGTGCAC
qPCR GAPDH Fw	GCTCCGGGAAAAGGAAAA
qPCR GAPDH Rv	TCCGTTAATTCCGATCTTCG
qPCR rp49/RpL32 Fw	CGGATCGATATGCTAAGCTGT
qPCR rp49/RpL32 Rv	GCGCTTGTTCGATCCGTA
IVT T7 Blastopia #1 Fw	*TAATACGACTCACTATAGG*TTTGTCTTTTAACCGGTGGC
IVT T7 Blastopia #1 Rv	*TAATACGACTCACTATAGG*TATGCCTCGGCATTTTTCTC
IVT T7 Blastopia #2 Fw	*TAATACGACTCACTATAGG*TGCTTCAACTGTGGTTCGAG
IVT T7 Blastopia #2 Rv	*TAATACGACTCACTATAGG*CCCACTATTCCATCGCAACT
IVT T7 Blastopia #3 Fw	*TAATACGACTCACTATAGG*ACAGTATTGCGAGGTTTGGG
IVT T7 Blastopia #3 Rv	*TAATACGACTCACTATAGG*GAAAGGGTTTAATCACGCCA
IVT T7 Blastopia #4 Fw	*TAATACGACTCACTATAGG*AGTCCCTTTCAAGCTCGTCA
IVT T7 Blastopia #4 Rv	*TAATACGACTCACTATAGG*TGCTATGCATCGTCTTTTGC
IVT T7 G2/Jockey-3 Fw	*TAATACGACTCACTATAGG*CGATACTGTTCGGCCGTATT
IVT T7 G2/Jockey-3 Rv	*TAATACGACTCACTATAGG*TGATCATGATGTTCCAGCGT
IVT T7 GFP Fw	*TAATACGACTCACTATAGG*GCTTCAGCCGCTACCCC
IVT T7 GFP Rv	*TAATACGACTCACTATAGG*TGGTTGTCGGGCAGCAGC

4.7 HuluFISH probe sets

HuluFISH probe	Sequences	
Blastopia	tgcctgcgatcgtccttc	ggctttgttcgtgcactatca
	ttccactcgcctctactgc	tctgatcctgtgtccaccaga
	tcgactgactgcaacgctgac	cccaaacctcgcaatactgtt
	ccgaacgactgctgctgagaa	ccactattccatcgcaactga
	gccccccatgtaagattgttt	tcttctcggtattgtggtgca
	actatacccacgcacacaaca	ggtggtgtttcgtatgttctc
	ccgacttttaccgacttgtgt	tcagatagtcgggtgtgtcc
	cagttctctctagcagctttc	cctctacctgcttctttacag
	gacaatgttctaacggcgcga	gattgactgctcgacccatt
	gaacattccgcgacaaggcac	tcaagcacctcctccattatg
	ggcgccaacttttgatcttct	cttcttctacggccacatgaa
	gcactagcgtctttattctgc	ctggggtagcggcatatacaa
	cttcttcttcctgtcttgca	ccgtcttttccatgcattct
	ctccttttttaatgtcgaggcc	ctgcagctctcttaagtacca
	gccctcttgattacagctgaa	ccttcgataatatgtcccagg
	tccggacacttgcttgagata	gttttctcttttccagggca
Copia	tggactgggcccataacc	cagcatcccacaacttaa
	gccttgtccatttttcac	caacatctcttgcgacaa
	acttctcgccatcaaacgg	cctactgtcattcggaaa
	tcttgctcggctaaaaga	ctcattcgggaattctgt
	cctcgttaggcattaaacc	gttgtcgcattccttact
	ccttttttccaggagtcatc	cctactgtcattcggaaa
	aacgagtcgcttaggtac	ctcattcgggaattctgt
	gccgtaatgtcgcttgttg	gttgtcgcattccttact
	ttctcaagaatctgacgc	ggtgatcatctcgctttct
	tcataaacggcgtccaaa	gcctgatcccttactttca
	cgacgccaaacttttttcgt	ccctactctcattcgggtt
	tttcgcagcgccagttg	gctctgctgtttcacttt
	cttcagagaaagcaaacg	ggatatctgaggcttagtc
	tgacatctcactcgatag	tgatggcttcttcccaaga
	gctgccaacaattcactt	cccatctgctatctacaa
	ggcaatgtgatcagtaga	gcaaccaatctagctttg
	cacaaacgccaatgtcaa	ggagcaaatgtctcttca
	gcttgtatcgttgtggtca	cgtgccatttaagaaagc
	cgatcgcgttcataactt	gagtccgtaaattgcctt
	gccttctctgccacagtg	ctctttcaatgcttgctc
	cgtgtgatgttgcagtttg	agcgatcaactgaagagt
	gttgtccatcactgaagt	cctgtagctataaccaca
	ccagaatcaaggacaaac	ggtcagtcatcctaaact
	ggtatacagcgactcatca	tcctgcatctctatccta
	gcacaacctccacactgtc	gggtattgcagtcttcatc
	cttgcttggccactgc	catcctatgaggctacggc

HuluFISH probe	Sequences	
(Copia continued)	ggacaataccacgcttagt	gggcgtgtacaaagcatta
	ccagtgtaatctcatggtc	ctgccataattcggagtt
	gcagcttccttacaaaag	cgatagtgcccttcaaat
	cctcttggagacgctttac	gcccaatcagaatccacat
	ttcgatcgacattcctg	gctactgagttctgtctct
	acaccgcttttgtcaa	ctcagcttcagttgatga
	cctctcatgccataaacga	gcttcaaatagggccata
	catcgcttatatggccaa	gccttgattgtcttcgta
	gtcttgcctgtttaccat	ggggttgtttgctatgcta
	taggcccacagacatct	gctctctggcaaaatgat
	gcttcactcttggctacaa	ccagttgattctctgtagg
	gtatgtggcactgttaag	ccacaaatctcgcagcagg
	cgagctttttccgtaatg	cccaatttgtctcgtaac
	ctttgcaccactaaccatg	tggtcgtcttgcagcaaa
	cgccccaaaagcttttatc	cccgatcagatcattgta
	ggggtctttgaactatca	catttgctgcgaggatac
	cttattgtgccacatctc	gggaagtgttaactgatcc
	gcaccaaacactctcaaa	tgtgggtggtgtgcattct
	ggttcatagcccacaaaa	ccccctcaaaaataacgtc
SATIII-sense	ccgaaacttggaaaaacgga	cggatttcgcccaaaagttga
	gttgtggcaaaagaactg	cggcattccatattcagaca
	ggcattccatattcagac	tattacgagctcaacgagg
	acgagctcaactaggtat	tgggaggatatggccaaaaa
	gggtcataagggtcatcaaa	gggttatcagggtcatcaaa
	gcgctaacaatccctatcact	gcgctaaagatccctatcat
SATIII-antisense	gcactggtaattagctgctca	cctccttacgaaaaatgcga
	cccctccatacagaaaatacg	gcaaattttgatgacccct
	cctcactgagctcgtaataa	cctcactgagctcgtaataa
	gcagagtctgtttttcca	gcagagtctgtttttcca
	gcactggtaattagctgctca	gtgacccttttttagccaagt
	cccctccatacagaaaatacg	gcactggtaattagctgctca
	cctcactgagctcgtaataa	gacccccattcttacaaa
	gcagagtctgtttttcca	cctcactgagctcgtaataa
	gcactggtaattagctgctca	gcagagtctgtttttcca

4.8 Plasmids

Plasmid	Source
pMT-Cal1-V5-His-Puro	This study
pMT-CENP-C-V5-His-Puro	This study
pUASz-TRiP-Blastopia	This study

Plasmid	Source
pUASz-TRiP-G2/Jockey-3	This study

4.9 Organisms

E. coli	Genotype	Source
DH5α	F- Phi80dlacZ DeltaM15 Delta(lacZYAargF) U169 deoR recA1 endA1 hsdR17(rKmK+) phoA supE44 lambda- thi-1	Erhardt lab

S2 cell line	Source
Wild type	Erhardt lab
pMT-Cal1-V5-His-Puro	This study
pMT-CENP-C-V5-His-Puro	This study

Fly line	Genotype	Source
OregonR		Erhardt lab
W1118		Erhardt lab
Arm-Gal4	W[*]; P{w[+mW.hs]=Gal4-arm.S}11	Bloomington 1560
MTD-Gal4	P{w[+mC]=otu-GAL4::VP16.R}1, w[*]; P{w[+mC]=GAL4-nos.NGT} 40; P{w[+mC]=GAL4::VP16-nos.UTR} CG6325[MVD1]	Bloomington 31777
$Elav^{C155}$-Gal4	P{W[+mW.hs]=GawB}elav[C155]	Bloomington 458
Elav-Gal4/cyo	P{w[+mC]=GAL4-elav.L}2/Cyo	Bloomington 8765
GFP dsRNA	W[1118]; P{w[+mC]=UAS-GFP.dsRNA.R}142	Bloomington 9330
TRiP-mCherry	Y[1] sc[*] v[1] sev[21]; P{y[+t7.7] v[+t1.8]=Valium20-mCherry}attP2	Bloomington 35785
TRiP-Blastopia #1	P{ y[+t7.7]=pUASz-TRiP-Blastopia}attP2	This study
TRiP-Blastopia #2	PBac{y[+]=pUASz-TRiP-Blastopia}attP-3B VK00033	This study
TRiP-G2/Jockey-3	P{ y[+t7.7]=pUASz-TRiP-G2/Jockey-3}attP2	This study
TRiP-CG32344	y1 v1; P{TRiP.HMJ23354}attP40	Bloomington 61865
UAS-NLS-GFP		Erhardt lab

4.10 Published datasets

These datasets are published in Duff et al., 2015 for modENCODE(DUFF ET AL., 2015).

SRA accession	Sample name	strain
SRR1197370	Embryo 0-2 h rep1	y1; cn bw1 sp1
SRR1197337	Embryo 0-2 h rep2	y1; cn bw1 sp1
SRR1197368	Embryo 2-4 h rep1	y1; cn bw1 sp1
SRR1197336	Embryo 2-4 h rep2	y1; cn bw1 sp1

SRA accession	Sample name	strain
SRR767609	Embryo 4-6 h rep1	y1; cn bw1 sp1
SRR1197338	Embryo 4-6 h rep2	y1; cn bw1 sp1
SRR767610	Embryo 6-8 h rep1	y1; cn bw1 sp1
SRR1197333	Embryo 6-8 h rep2	y1; cn bw1 sp1
SRR767615	Embryo 8-10 h rep1	y1; cn bw1 sp1
SRR1197335	Embryo 8-10 h rep2	y1; cn bw1 sp1
SRR1197367	Embryo 10-12 h rep1	y1; cn bw1 sp1
SRR1197334	Embryo 10-12 h rep2	y1; cn bw1 sp1
SRR1197369	Embryo 12-14 h rep1	y1; cn bw1 sp1
SRR1197332	Embryo 12-14 h rep2	y1; cn bw1 sp1
SRR767618	Embryo 14-16 h rep1	y1; cn bw1 sp1
SRR1197331	Embryo 14-16 h rep2	y1; cn bw1 sp1
SRR767605	Embryo 16-18 h rep1	y1; cn bw1 sp1
SRR1197330	Embryo 16-18 h rep2	y1; cn bw1 sp1
SRR1197363	Embryo 18-20 h rep1	y1; cn bw1 sp1
SRR1197327	Embryo 18-20 h rep2	y1; cn bw1 sp1
SRR767620	Embryo 20-22 h rep1	y1; cn bw1 sp1
SRR1197329	Embryo 20-22 h rep2	y1; cn bw1 sp1
SRR1197366	Embryo 22-24 h rep1	y1; cn bw1 sp1
SRR1197328	Embryo 22-24 h rep2	y1; cn bw1 sp1
SRR1197426	Larvae st1 rep1	y1; cn bw1 sp1
SRR1197325	Larvae st1 rep2	y1; cn bw1 sp1
SRR1197425	Larvae st2 rep1	y1; cn bw1 sp1
SRR1197324	Larvae st2 rep2	y1; cn bw1 sp1
SRR1197424	Larvae st3 rep1	y1; cn bw1 sp1
SRR1197326	Larvae st3 rep2	y1; cn bw1 sp1
SRR1197423	White prepupae (WPP) rep1	y1; cn bw1 sp1
SRR1197290	WPP rep2	y1; cn bw1 sp1
SRR1197422	WPP +12h rep1	y1; cn bw1 sp1
SRR1197289	WPP +12h rep2	y1; cn bw1 sp1
SRR1197421	WPP +24h rep1	y1; cn bw1 sp1
SRR1197288	WPP +24h rep2	y1; cn bw1 sp1
SRR1197420	WPP +2d rep1	y1; cn bw1 sp1
SRR1197287	WPP +2d rep2	y1; cn bw1 sp1
SRR1197419	WPP +3d rep1	y1; cn bw1 sp1
SRR1197285	WPP +3d rep2	y1; cn bw1 sp1
SRR1197416	WPP +4d rep1	y1; cn bw1 sp1
SRR1197286	WPP +4d rep2	y1; cn bw1 sp1
SRR1197415	male eclosion +1d rep1	y1; cn bw1 sp1
SRR1197315	male eclosion +1d rep2	y1; cn bw1 sp1
SRR1197418	female eclosion +1d rep1	y1; cn bw1 sp1
SRR1197317	female eclosion +1d rep2	y1; cn bw1 sp1
SRR1197417	male eclosion +5d rep1	y1; cn bw1 sp1
SRR1197316	male eclosion +5d rep2	y1; cn bw1 sp1

SRA accession	Sample name	strain
SRR1197414	female eclosion +5d rep1	y1; cn bw1 sp1
SRR1197313	female eclosion +5d rep2	y1; cn bw1 sp1
SRR1197413	male eclosion +30d rep1	y1; cn bw1 sp1
SRR1197311	male eclosion +30d rep2	y1; cn bw1 sp1
SRR1197411	female eclosion +30d rep1	y1; cn bw1 sp1
SRR1197314	female eclosion +30d rep2	y1; cn bw1 sp1
SRR1197357	Larvae st3 fat body rep1	OregonR
SRR1197352	Larvae st3 fat body rep2	OregonR
SRR1197295	Larvae st3 imaginal discs rep1	OregonR
SRR1197298	Larvae st3 imaginal discs rep2	OregonR
SRR1197351	Larvae st3 salivary glands rep1	OregonR
SRR1197463	Larvae st3 salivary glands rep2	OregonR
SRR1197353	Larvae st3 digestive system rep1	OregonR
SRR1197354	Larvae st3 digestive system rep2	OregonR
SRR1197350	Larvae st3 carcas rep1	OregonR
SRR1197349	Larvae st3 carcas rep2	OregonR
SRR1197348	Larvae st3 CNS rep1	OregonR
SRR1197291	Larvae st3 CNS rep2	OregonR
SRR1197347	WPP salivary gland rep1	OregonR
SRR1197345	WPP salivary gland rep2	OregonR
SRR1197344	WPP fat body rep1	OregonR
SRR1197343	WPP fat body rep2	OregonR
SRR1197346	WPP +2d CNS rep1	OregonR
SRR1197460	WPP +2d CNS rep2	OregonR
SRR1197462	WPP +2d fat body rep1	OregonR
SRR1197461	WPP +2d fat body rep2	OregonR
SRR1197279	virgin eclosion +1d head rep1	OregonR
SRR1197275	virgin eclosion +1d head rep2	OregonR
SRR1197273	female eclosion +1d head rep1	OregonR
SRR1197274	female eclosion +1d head rep2	OregonR
SRR1197276	male eclosion +1d head rep1	OregonR
SRR1197272	male eclosion +1d head rep2	OregonR
SRR1197271	adult eclosion +1d digestive system rep1	OregonR
SRR1197270	adult eclosion +1d digestive system rep2	OregonR
SRR1197269	adult eclosion +1d carcass rep1	OregonR
SRR1197487	adult eclosion +1d carcass rep2	OregonR
SRR1197294	virgin eclosion +4d head rep1	OregonR
SRR1197486	virgin eclosion +4d head rep2	OregonR
SRR1197296	virgin eclosion +4d ovary rep1	OregonR
SRR1197480	virgin eclosion +4d ovary rep2	OregonR
SRR1197485	female eclosion +4d head rep1	OregonR
SRR1197484	female eclosion +4d head rep2	OregonR
SRR1197483	female eclosion +4d ovary rep1	OregonR
SRR1197482	female eclosion +4d ovary rep2	OregonR

SRA accession	Sample name	strain
SRR1197481	male eclosion +4d head rep1	OregonR
SRR1197293	male eclosion +4d head rep2	OregonR
SRR1197297	male eclosion +4d testes rep1	OregonR
SRR1197301	male eclosion +4d testes rep2	OregonR
SRR1197479	male eclosion +4d accessory glands rep1	OregonR
SRR1197478	male eclosion +4d accessory glands rep2	OregonR
SRR1197476	adult eclosion +4d digestive system rep1	OregonR
SRR1197477	adult eclosion +4d digestive system rep2	OregonR
SRR1197475	adult eclosion +4d carcass rep1	OregonR
SRR1197292	adult eclosion +4d carcass rep2	OregonR
SRR1197472	virgin eclosion +20d head rep1	OregonR
SRR1197473	virgin eclosion +20d head rep2	OregonR
SRR1197474	female eclosion +20d head rep1	OregonR
SRR1197362	female eclosion +20d head rep2	OregonR
SRR1197361	male eclosion +20d head rep1	OregonR
SRR1197359	male eclosion +20d head rep2	OregonR
SRR1197360	adult eclosion +20d digestive system rep1	OregonR
SRR1197358	adult eclosion +20d digestive system rep2	OregonR
SRR1197355	adult eclosion +20d carcass rep1	OregonR
SRR1197356	adult eclosion +20d carcass rep2	OregonR

5. Methods

All methods listed here are standard protocols used in the Erhardt lab unless otherwise specified.

5.1 Cell biology techniques

5.1.1 Cell culture

Drosophila Schneider 2 (S2) cells were cultured at 25°C in Schneider's Drosophila medium with 10 % Fetal Bovine Serum (FBS) (v/v) and 200 µg/ml penicillin and streptomycin. The cells were split twice a week and diluted depending to their confluence (usually 1:10). The medium of cells containing a transfected plasmid was supplemented with 50 µg/ml puromycin to maintain the selection.

For storage cells were frozen and kept in liquid nitrogen. Cells were grown to confluency in a T75 cell culture flask. They were released from the flasks surface by pipetting and pelleted in a 15 ml falcon tube by centrifuging 2 min at 400 x g. The supernatant was removed until roughly 2.25 ml conditioned medium was left, and cells were resuspended with 2.25 ml fresh medium. 500 µl DMSO was added, after which the cell suspension was quickly distributed in 1 ml aliquotes in cryotubes. The cell suspension was frozen to -80°C using a Mr Frosty (Thermo Fisher) freezing container filled with isopropanol, after which the cryotubes were switched to liquid nitrogen.

Cells were quickly thawed by pipetting 1 ml fresh medium over the frozen cells. The cells were transferred to a 15 ml falcon tube containing 4 ml fresh medium and centrifuged for 2 min at 400 x g. Supernatant was removed and the cells were resuspended in 3 ml fresh medium and transferred to a T25 cell culture flask. Cells were allowed to recover for at least 2 weeks before using them for experiments.

5.1.2 Transfections

To create stable cell lines, plasmid DNA was transfected into S2 cells using Cellfectin II reagent (Gibco). A day before the transfection, 1.5*10^6 cells were seeded per well of a 6-well plate in 2 ml medium. The transfection solution was prepared in 2 steps. First 5 µg plasmid was mixed with 300 µl serum free medium (SFM) and separately 5 µl Cellfectin II was mixed with 300 µl SFM. These two mixtures were combined and incubated for 30 min at room temperature (RT). During this incubation time, the cells were washed twice with 1 ml SFM. SFM was added up to 1 ml, after which this mixture was added to the cells. Cells were incubated at 25°C for 4 h, before 1 ml medium containing 20 % FBS was added. To

avoid medium evaporation, the 6-well plates were sealed with parafilm. Cells were incubated at 25°C until confluent and then transferred to a T25 flask with the appropriate selection medium. As a control, untransfected S2 cells were grown in the same selection medium. The selection of transfected cells was considered complete, when all untransfected cells had died.

5.1.3 Induction of pMT-V5-His expression

Cells containing pMT-V5-His plasmids with various protein genes were grown until confluent. Expression of the genes under the pMT promoter was induced by supplementing the medium with various concentrations of copper sulfate (ranging from 20 - 500 µM) overnight.

5.1.4 DNA damage induction in S2 cells

Cells were treated increasing concentrations of Neocarzinostatin for 1 h to induce double strand breaks in the DNA. After this treatment, cells were harvested for western blot analysis or used for RNA-ChIP experiments. Alternatively, fresh medium was added to the cells for continued culture.

5.1.5 RNAi in S2 cells

$1.5*10^6$ cells were seeded in a 6-well plate in 2 ml medium the evening before the treatment. The cells were washed twice with 1 ml SFM. In sterile 1.5 ml tubes, 20 µg dsRNA was mixed with 1 ml SFM and incubated for 15 min. The last wash was removed from the cells and replaced with the dsRNA/SFM mixture. After 1 h incubation at 25°C, 1 ml medium containing 20% FBS was added. The cells were harvested for RNA isolation and IF after 4 days.

dsRNA synthesis

A DNA template with on either side a T7 promoter was generated with PCR reaction using drosophila OregonR embryo cDNA or plasmid DNA as input. The PCR reaction combined 2.5 µl cDNA or plasmid DNA with 5 µl of both forward and reverse T7 primers (10 µM), 20 µl RedTaq polymerase mix and 67.5 µl H_2O. The following program was used on a thermal cycler: 2 min at 94°C, followed by 5 cycles of 30 sec at 94°C, 30 sec at 50°C and 30 sec at 72°C, followed by 25 cycles of 30 sec at 94°C, 30 sec at 72°C and 30 sec at 72°C, followed by 2 min at 72°C and cooling. The product was run on a 1 % agarose gel, cut out and purified with the NucleoSpin Gel and PCR Clean-up kit (Machery-Nagel) according to manufacturer's protocol. The purified template was used for in vitro transcription with the MegaScript RNAi kit (Invitrogen) according to the manufacturer's protocol with an overnight incubation.

5.1.6 Immunofluorescence (IF) on settled S2 cells

Cells were released from their cell culture flasks and washed once in PBS, spun down 500 x g for 7 min. $4*10^5$ cells in 100 µl PBS were settled for 10 min on a Polysine slide (Thermo Fisher Scientific) marked with a PAPpen. The cells were fixed in 120 µl 4 % PFA in PBS for 10 min at RT, the PFA was heated to

65°C before use to break up the polymers. Subsequently, cells were washed 3X 5 min with PBS, permeabilized with 0.5 % Triton-X-100 in PBS for 5 min, before washing again 3X 5 min with PBS. Blocking of cells was done with 150 µl 1 % BSA in PBS for 30 min at RT. The primary antibodies were diluted in 50 µl blocking solution. Primary antibody incubation was done in a humid chamber either overnight at 4°C or for 1 h at RT. Cells were washed 3X 10 min with PBS before incubating with the corresponding fluorescent secondary antibody diluted 1:500 in blocking solution for 1 h at RT in a humid chamber protected from light. Cells were washed 3X 10 min with PBS protected from light. Cells were counterstained with 1 µg/ml DAPI in PBS for 5 min and washed 3X 10 min with PBS, before mounting with Aqua-Poly/Mount (Polysciences) and coverslips (Neolab). Finished slides were incubated 2 h at RT and stored at 4°C until imaging.

5.1.7 IF-RNA-FISH on settled S2 cells

For RNA-FISH HuluFISH from PixelBiotech were used. They provide a custom design probe set of 32 enzymatically labelled probes covering the length a gene for high signal to noise single molecule FISH analysis. All probes used for this study are labeled with atto647. RNA-FISH was performed based on the manufacturer's protocol as described below.

All reagents and buffers used were prepared RNase-free. Let $5*10^5$ cells in 100 µl PBS settle for 15 min on a round 1 mm thick coverslip with an 18 mm diameter. The cells were fixed in 100 µl 4 % PFA in PBS for 10 min at RT, the PFA was heated to 65°C before use to break up the polymers. Subsequently, cells were washed in a 12-well plate 3X 10 min with 500 µl PBS, permeabilized with 0.5 % Triton-X-100 in PBS for 5 min, before washing again 3X 5 min with PBS. Blocking of cells was done with 100 µl 1 % BSA in PBS for 20 min at RT. The primary antibodies were diluted in 100 µl blocking solution, placed as a drop on a piece of parafilm on which the coverslip was placed. Primary antibody incubation was done in a humid chamber for 30 min at RT. Cells were washed 3X 10 min with PBS before incubating with the corresponding fluorescent secondary antibody diluted 1:500 in blocking solution for 30 min at RT in a humid chamber protected from light in the same manner as the primary antibody. Cells were washed 2X 10 min with PBS protected from light.

After the IF was completed, the cells were washed 2X 10 min with 600 µl Hulu Wash buffer. Next, 0.75 µl Hulu probe was mixed with 75 µl Hulu Hybridization buffer. This mixture was places as a drop on parafilm, on which the coverslip was placed in a dark humid chamber. This was incubated overnight at 30°C. The coverslips were washed again 3X 10 min with 600 µl Hulu Wash buffer, incubated with 1 µg/ml DAPI in Hulu Wash buffer for 5 min and washed 1X 5 min with Hulu Wash buffer. The coverslips were mounted on slides with Aqua-Poly/Mount (Polysciences). Finished slides were incubated 2 h at RT and stored at 4°C until imaging.

5.2 Drosophila techniques

5.2.1 Fly husbandry

Flies in use were kept at RT, while stocks were kept at 18°C with vials exchange every 3-4 weeks. Fly food consisted of 0.72 % (w/v) agar, 7.2 % (w/v) maize, 2.4 % molasses, 7.2 % (w/v) malt, 0.88 % (w/v) soya, 1.464 % (w/v) yeast and acid mix (1 % propionic acid + 0.064 % orthophosphoric acid).

Virgins needed for crosses were collected either at pupae stage or directly after hatching. Pupae were removed from the vial and placed in H_2O in a dish to be observed under the light microscope. Females were identified by the absence of black sex combs and kept in separate vials for them to hatch. To collect virgins directly after hatching, the vials were cleared of all flies. Every 3 h newly hatched flies were sorted under the light microscope and separated into new vials based on their sex.

Crosses were set up with a ratio of 2:1 virgins and male flies and kept at 25°C or 29°C.

5.2.2 Generation of transgenic flies

For in vivo fly RNAi (TRiP) (NI ET AL., 2008; PERKINS ET AL., 2015), transgenic flies were generated for both Blastopia and G2/Jockey-3 in which the expression of a short hairpin could be induced with the UAS-Gal4 system (DUFFY, 2002). A short hairpin was designed using the online tool Designer of Small Interfering RNA (DSIR), which predicts siRNA efficiency (VERT ET AL., 2006). Short hairpins with a minimal DSIR score of 90 were blasted against the reference transcriptome (blastn SS sequence to refseq_rna), to select those with a maximum of 15 nucleotides overlap with other RNAs. These were blasted against a custom blast database created for the *Drosophila melanogaster* transposable element and complex repeat consensus sequences (File_S1 (CHANG ET AL., 2019)), again selecting those with a maximum of 15 nucleotide overlap (ZHANG ET AL., 2000). One short hairpin sequence for each transposable element was selected and cloned into the pUASz-MIR vector (DGRC #1432) (DELUCA & SPRADLING, 2018). The plasmid was sent to the fly facility of the Cambridge University for PhiC31 mediated injection into the nos-int; ; attP2 embryos using their plan 5. The plasmid containing the short hairpin targeting Blastopia was also injected into VK33 embryos by Sandra Müller of AG Teleman at the DKFZ. Resulting flies were genotyped by PCR and crossed until a homozygous stock was obtained with the UASz-TRiP integrated on the 3rd chromosome.

5.2.3 Dissections 3rd instar larvae brain for staining

To dissect 3rd instar larvae brains, larvae that were crawling up the wall of the vial were selected. In PBS, organs in the head were exposed by pulling with tweezers at the mouth hooks while constraining the middle of the larvae. From there the brain lobes, ventral nerve cord and potentially attached discs were separated from surrounding tissue and moved for a 0.5 ml tube in PBS on ice. Once enough

material was collected, the brains were fixed in 4 % PFA for 30 min at RT, the PFA was heated to 60°C before use to break up the polymers. Afterwards, the brains were washed with PBS twice 5 min on a tilting mixer to remove the fixative. When brains were used for RNA-FISH, all the reagents were prepared RNase-free and the brains were stored in 70 % Ethanol at 4°C overnight for permeabilization.

5.2.4 Collecting and fixing embryos for staining

Flies were kept in cages with grape juice plates and fresh yeast paste, plates were changed daily. Before collecting embryos, the plate was changed for at least an hour to allow females to lay already aging embryos that had accumulated. Embryos were laid and aged for the time needed for that experiment. The embryos were washed from the plate with water and collected in a fine mesh strainer. They were dechorionated in 50 % bleach for 4 min and rinsed generously with water. Embryos were transferred to a glass vial containing 2 ml PBS and 2 ml heptane. To fix 425 µl 36 % formaldehyde was added (final concentration 4 %), incubating for 20 min at RT shaking vigorously. The lower PBS layer was replaced twice with 2 ml fresh PBS to wash away the formaldehyde. To remove the vitelline membrane the PBS was replaced with 2 ml Methanol and the vial was vortexed for 30 sec. The embryos were collected from the bottom of the glass vial and transferred to an 1.5 ml tube. They were washed twice with 1 ml Methanol before storing them in 500 µl Methanol at -20°C. The embryos were stored at least overnight before continuing with the staining. When brains were used for RNA-FISH, all the reagents were prepared RNase-free.

Grape juice plates

3.5 g Agar were dissolved in 150 ml H2O by heating it up in the microwave. Subsequently, 50 ml of biological grape juice (REWE) were added. Optional: 350 mg of NIPAGIN was diluted in 2 ml 100 % ethanol and added to the slightly cooled solution (approximately 50°C). The solution was mixed by swirling and poured into 5 cm or 15 cm round plastic dishes up to a height of approximately 5 mm. The plates were stored at 4°C.

5.2.5 IF on larval brain and embryos

The staining of larval brains was done in 200 µl PCR tubes, washes were 200 µl. The staining of embryos was done in 1.5 ml tubes, washes were 800 µl. All washes and incubations were done on a tilting mixer. After the above described fixing, samples were rinsed twice and washed 3X 10 min in PBST (0.1 % Tween-20 in PBS). The samples were blocked in 100 µl (brains) or 200 µl (embryos) 5 % BSA in PBS for 1 h at RT, followed by incubation with the primary antibody in 50 µl (brains) or 100 µl (embryos) 5 % BSA in PBS either overnight at 4°C or for 2 h at RT. Unbound primary antibody was washed away by washing 4X 15 min with PBST. Samples were incubated for 2 h at RT with the fluorescent secondary antibody in 50 µl (brains) or 100 µl (embryos) 5 % BSA in PBS protected from light. Unbound antibody

was again washed away, 4X 15 min with PBST. The samples were counterstained with DAPI (1:1000) in PBST for 5 min and rinsed once more with PBST. The tissue was pipetted onto a coverslip, excess PBST was removed and replaced by Aqua-Poly/Mount (Polysciences) mounting medium. The tissue was repositioned under the microscope if needed and the mounting medium was allowed to dry for 20 min before mounting the coverslips on slides. Finished slides were incubated 2 h at RT and stored at 4°C until imaging.

5.2.6 RNA-FISH on drosophila embryos and larval brain

For RNA-FISH HuluFISH from PixelBiotech were used. They provide a custom design probe set of 32 enzymatically labelled probes covering the length a gene for high signal to noise single molecule FISH analysis. All probes used for this study are labeled with atto647, except SATIII-antisense, which is labelled with atto565. RNA-FISH was performed based on the manufacturer's protocol as described below.

The staining of larval brains was done in 200 µl PCR tubes, washes were 200 µl. The staining of embryos was done in 1.5 ml tubes, washes were 800 µl. All washes and incubations were done on a tilting mixer unless otherwise specified. All reagents were prepared RNase-free. After the above described fixing and Ethanol or Methanol incubation, the samples were rinsed twice and washed 3X 10 min with Hulu Wash buffer, before adding 50 µl Hulu Hybridization buffer with 0.5 µl Hulu probe. The PCR tubes with brains were place in a 30°C water bath and the 1.5 ml tubes with embryos were placed in a 30°C heat block shaking 400 RPM. The samples were incubated at 30°C overnight in the dark. The next day, the samples were rinsed twice and washed 4X 10 min with Hulu Wash buffer. The samples were counterstained with DAPI (1:1000 from 1mg/ml stock) in Hulu Wash buffer for 5 min and rinsed once more with Hulu Wash buffer. The tissue was pipetted onto a coverslip, excess buffer was removed and replaced by Aqua-Poly/Mount (Polysciences) mounting medium. The tissue was repositioned under the microscope if needed and the mounting medium was allowed to dry for 20 min before mounting the coverslips on slides. Finished slides were incubated 2 h at RT and stored at 4°C until imaging.

As a control, the RNase A treatment was performed just before the staining steps, after the Ethanol or Methanol step. The tissues were rinsed twice and washed twice for 10 min in PBST. The samples were incubated with 0.02 µg/µl RNase A (Sigma-Aldrich, 1:2000 from 40 mg/ml stock) in PBST for 1.5 h at 37°C.

5.2.7 IF-RNA-FISH on larval brain and embryos

To combine IF with RNA-FISH, the samples were first stained using the IF protocol described above using only 2 h antibody incubation times and with RNase-free buffers. After the washes to remove unbound secondary antibody, the samples were fixed again with 4 % PFA in PBS for 10 min at RT. The

formaldehyde was quenched with 135 mM Glycine in PBS for 10 min at RT and washed with PBS. After this the above described RNA-FISH protocol was continued.

5.3 Molecular biology techniques

5.3.1 Cloning

Coding regions of genes were amplified by PCR with the Q5 polymerase (NEB) from S2 cell cDNA or other plasmids containing the gene of interest using primers which would add sequences complementary to restriction enzyme cut sites. The primers used for synthesizing the insert were designed with the NEBuilder Assembly Tool (https:// nebuilder.neb.com/). The product was run on an agarose gel to clean-up the correct size fragment with the Nucleospin gel and PCR clean-up kit (Machery-Nagel). This fragment was cloned into vectors using GIBSON assembly ligation with a restriction enzyme digested backbone plasmid. Basically, a 15 µl reaction with 7.5 µl Gibson Assembly Master Mix (GIBSON ET AL., 2009), 0.03 pmol linearized plasmid and 0.06 pmol insert were incubated at 50°C for 15 min. This ligated plasmid was transformed into chemically competent *E. Coli* (DH5α) by mixing the plasmid with 50 µl thawed cells and incubating it on ice for 25 min. After a shock for 45 sec at 42°C, the mixture was placed on ice for another 2 min and supplemented with 250 µl LB medium. The *E. Coli* cells were incubated for 1 h on a shaker at 37°C, centrifuged for 2 min at 400 x g to pellet and remove excess medium before plating on agar plates with appropriate antibiotic. A colony PCR of several colonies was performed using RedTaq polymerase mix (Jena Bioscience) to briefly assess the colonies that grew overnight at 37°C. Positive colonies were used to grow a 5 ml miniprep culture overnight at 37°C in LB medium supplemented with appropriate antibiotic. Plasmid DNA was isolated with the Nucleospin plasmid kit (Machery-Nagel). All plasmid sequences were verified using Sanger sequencing.

Colony PCR

A colony was picked with a pipette tip and transferred to 10 µl H_2O in a PCR tube. This cell suspension was pipetted vigorously in order to release the DNA content from some of the cells. The remaining intact cells were pelleted by centrifuging 2 min at 400 x g. 1 µl of the plasmid DNA containing supernatant was used for the colony PCR, the remaining cells and liquid was kept on ice for later culture. The PCR reaction combined 1 µl plasmid DNA containing supernatant with 1 µl of both forward and reverse primers (10 µM), 4 µl RedTaq polymerase mix (Jena Bioscience) and 13 µl H_2O. The following program was used on a thermal cycler: 2 min at 94°C, followed by 30 cycles of 30 sec at 94°C,

30 sec at an annealing temp and 30 sec at 72°C, followed by 2 min at 72°C and cooling. The amplified reaction was run on a 1 % Agarose gel to assess for positive colonies.

The following backbone plasmids were used:
pMTpuro-V5-His (Amp)
pUASz-MIR (Amp)

5.3.2 RNA extraction

From S2 cells: harvested S2 cells were spun down and resuspended in 1 ml of TRIzol (Invitrogen) or TRIsure (Bioline).

From Drosophila: whole flies or embryos were snapfrozen in liquid nitrogen and sometimes stored at -80°C. The tissue was homogenized in 50 µl TRIzol (Invitrogen) with a pistel in an 1.5 ml tube. More TRIzol was added (up to 1ml) and the samples were centrifuged 10 min at 12.000 x g at 4°C to remove debris. The supernatant was transferred to a new tube.

RNA was extracted according to the manufacturer's protocol. 1 µl GlycoBlue (Invitrogen) was added to aid precipitation and pellet recognition. RNA pellets were resuspended in RNase-free water according to the pellet size.

5.3.3 TURBO DNase treatment of RNA

To remove gDNA traces from RNA which would interfere with the accurate quantification of repeat transcript levels with RT-qPCR, RNA was treated with TURBO DNase (Invitrogen). This was especially needed for RNA extracted from Drosophila tissue. In a 50 µl reaction 10 µg RNA was mixed with 5 µl 10x TURBO DNase buffer, 1 µl TURBO DNase and 0.5 µl RNasin Plus Ribonuclease Inhibitor (Promega). This reaction was incubated at 37°C for 30 min. The RNA was isolated again using 500 µl TRIsure or TRIzol according the manufacturer's protocol or with Phenol-Chloroform-Isoamyl alcohol (P/C/I) extraction as follows.

RNase-free water was added to the reaction to a volume of 200 µl and 20 µl 3 M NaOAc (pH 5). 1 Volume of P/C/I was added, the reaction was vortexed for 10 sec to mix. Centrifuge for 5 min at 16000 x g and transfer the upper phase to a new 1.5 ml tube. Add 1 Volume of chloroform, vortex and centrifuge, and transfer the upper phase to a new 1.5 ml tube. Precipitate the RNA with 1 Volume isopropanol and 1 µl GlycoBlue (Invitrogen), mix and incubate at least 1 h at -80°C. Centrifuge for 30 min at 18000 x g at 4°C. Remove the supernatant and wash the pellet twice with 1 ml 75 % Ethanol. Centrifuge to remove the wash for 5 min at 18000 x g at 4°C. Air dry the pellet before resuspending in 20 µl RNase-free water.

5.3.4 Reverse transcription and qPCR

For reverse transcription the QuantiTect Reverse Transcription Kit (Qiagen) was used according to manufacturer's instructions as follows. The RNA for both the + and – RT reaction was treated together with gDNA Whipeout Buffer, 2 µg RNA in a 28 µl gDNA digestion reaction, and incubated for 7 min at 42°C. This reaction was cooled quickly on ice and split in two (14 µl each) for the + and – RT reaction. The total volume of the reverse transcription reaction is 20 µl, in the – RT reaction the Reverse Transcriptase was replaced with water. All reactions were incubated for 15 min at 42°C and 3 min at 95°C. If limited RNA was available only half of the reaction was used. On occasion, when many primer pairs were used in the following qPCR, the cDNA was diluted 1:2 with water.

qPCRs were performed using LightCycler® 480 SYBR Green I Master Mix (2X) (Roche) with one reaction containing: 7.5 µl 2X SYBR Green, 1 µl (diluted) cDNA, 0.75 µl each of forward and reverse primers (10 µM) and 5 µl H_2O. Each sample was measured in triplicate, one mastermix was made which was divided over 3 wells of the 384-well plate (Roche). The following program was used on the LightCycler® 480 (Roche): 10 min at 95°C, followed by 40 cycles of 10 sec at 95°C, 10 sec at an annealing temperature and 10 sec at 72°C, followed by a melting curve.

5.4 Biochemical techniques

5.4.1 SDS-PAGE and Western Blot analysis

Cells were washed once with PBS and lysed in 20 µl RIPA buffer per $1*10^6$ cells. To improve release of chromatin associated proteins, lysates were sonicated 5 cycles with the Bioruptor (Diagenode), 30 sec on, 30 sec off, with the high power setting, in an ice water bath. Samples were spun down 18000 x g for 10 minutes at 4°C, taking the supernatant to a fresh tube to which 4X Laemmli sample buffer was added. Before loading, samples were boiled for 5 min at 95°C.

The denatured protein samples were separated on SDS-PAGE gels (between 8 % and 12 %, 1.5 mm) with the Biorad Tetracell system, running at 0.02 Amp per gel in the tank. Separated proteins were transferred onto a nitrocellulose membrane (0.45 µm) with wet/tank transfer in methanol containing transfer buffer. The gel and the membrane were sandwiched between foam pads and Whatman paper (cathode, 1 foam pad, 1 Whatman paper, gel, membrane, 1 Whatman paper, 1 foam pad, anode) before applying an electrical field of 100 V for 2 h at 4°C with an ice pack and a stirring fish in the tank. To check the success of transfer, the membrane was stained with Ponceau and if needed the membrane was cut. Destaining was done with TBST after which the membrane was blocked for 1 h in 5 % skimmed dry milk powder in TBST. Primary antibodies were diluted in 3 ml 5 % milk and incubated

overnight at 4°C in a 50 ml Falcon tube while rotating. The membrane was washed 5X in TBST for 10 min with agitation to remove residual primary antibody. The corresponding horseradish peroxidase (HRP) conjugated secondary antibody was also diluted in 5 % milk and incubated similarly to the primary antibody but for 1 h at RT. Again the membrane was washed 5X in TBST for 10 min with agitation. Membranes were incubated with SuperSignal West Pico PLUS or Femto Maximum Chemiluminescent Substrate (ECL, Thermo Fisher Scientific) mix for 5 min. Signal was detected on the Luminescent Image Analyzer LAS-3000 (Fujifilm Life Science).

5.4.2 RNA-ChIP-seq in embryos

This protocol was adapted and optimized from the embryo chromatin immunoprecipitation (ChIP) protocols by Domsch et al., 2019 and Sandmann et al., 2006, and an RNA-ChIP protocol by Sun at al., 2006 (DOMSCH ET AL., 2019; SANDMANN ET AL., 2007; SUN ET AL., 2006). All buffers were prepared RNase-free.

Embryo collection

Large cages with OregonR flies were set-up. The flies were allowed to lay on fresh agar plates with yeast for 1 hour to clear already aging embryos from females. The plates were switched again with fresh ones for a 6 h collection, and were aged another 2 h to get 2-8 h old embryos for the experiment. DNA damage was induced by exposing half of the embryo containing plates to 5 Gray γ-irradiation at the IKTZ Heidelberg, Blutbank, when the embryos were 1-7 h old.

Embryos were collected and dechorionated for 2 min in 5 % sodium hypochlorite. The embryos were rinsed generously with dH_2O. Embryos were transferred to a glass vial containing 4.5 ml fresh fixing solution without formaldehyde and 15 ml heptane. To fix, 500 µl 16 % formaldehyde was added, incubating for 15 min at RT shaking vigorously. Embryos were transferred to a 15 ml falcon tube with a glass Pasteur pipette, and the fixing solution and heptane was aspirated. 10 ml stop solution was added to quench the formaldehyde, the tube was shaken for 1 min at RT and placed on ice. Embryos were pelleted by centrifugation 1000 x g for 2 min at 4°C and washed twice with 10 ml cold PBST. The liquid was removed and embryos were flash frozen in liquid nitrogen and stored at -80°C.

Chromatin preparation

To prepare fragmented chromatin for the ChIP, a total of 500 mg of embryos from different collections was combined. A 7 ml douncer on ice was used to homogenize embryos in 5 ml cold PBST with inhibitors into a cell suspension with 20 strokes with the loose fitting pestle. Embryo debris was pelleted in a 15 ml falcon tube by centrifugation 400 x g for 1 min at 4°C. The supernatant was centrifuged again to pellet the cells at 1100 x g for 10 min at 4 °C, the supernatant was discarded. The cell pellet was resuspended in 5 ml cold cell lysis buffer and transferred to the douncer. Cells were

disrupted to isolate the nuclei by applying 20 strokes with the tight fitting pestle on ice. Nuclei were pelleted by centrifugation 2000 x g for 4 min at 4°C, discard supernatant. Nuclei were lysed in 800 µl cold nuclear lysis buffer and incubated on ice for 45 min. The sample was divided in three 1.5 ml tubes, 267 µl each, and the chromatin was sheared into 100-300 bp fragments with the Bioruptor (Diagenode). Shearing was done with 15 cycles of 30 sec on, 30 sec off, with the high power setting, in an ice water bath, the ice was replaced every 5 cycles. The sample was centrifuged at 18000 x g for 10 min at 4°C to remove insoluble debris. The three supernatants were pooled and an additional 1600 µl cold nuclear lysis buffer was added. 50 µl chromatin was removed to assess the shearing quality, the remainder was distributed in 200 µl aliquots, flash frozen and stored at -80°C.

Shearing quality control

The shearing quality was assessed by running the DNA on an agarose gel after digesting away the crosslinked RNA and protein. 50 µl TE was added to the 50 µl sheared chromatin. To digest away the RNA, 50 µg/ml RNase A (0.5 µl of 10 mg/ml stock, AppliChem) was added and the sample was incubated 30 min at 37°C, shaking 700 RPM. To digest the proteins, 2.5 µl of 20 % SDS and 0.5 µg/µl Proteinase K (10 µl of 5 mg/ml stock) was added and the sample was incubated 45 min at 42°C, shaking 700 RPM. After this, the sample was incubated for 1 h at 65°C to reverse the formaldehyde crosslink, shaking 700 RPM. The DNA was purified with P:C:I extraction with isopropanol precipitation as described above (section 5.3.3), and dissolved in 12 µl H_2O. The concentration was measured on nanodrop, to guide the quantity used for the later ChIP. Mix 5 µl of isolated DNA with 1 µl 6X Loading dye and run on a 1 % Agarose gel.

RNA chromatin Immunoprecipitation (RNA-ChIP)

For the ChIP, sheared chromatin equaling 10 µg isolated DNA was used per IP (α-CID, α-H3 and no antibody mock). This chromatin was diluted in cold Dilution buffer to a total of 500 µl for each IP in 1.5 ml low binding DNA tubes (Eppendorf). The following was added in this order, while mixing in between: 100 µl 10 % Triton X-100, 100 µl 1 % Sodium Deoxycholate, 100 µl 1 % SDS, 100 µl 1.4 M NaCl, 20 µl Protease Inhibitors cocktail (stock 25X, Roche) and 25 µl Ribonucleoside Vanadyl Complex (stock 20X, NEB). As input, 15 µl from all three ChIP reactions were combined and split in two 22.5 µl samples, one for the western blot and one for RNA isolation. 7.5 µl 4X SDS Laemmli buffer was added to the input for the western blot, and stored at -20°C. The input for RNA isolation was kept at 4°C for later isolation together with the ChIP sample. To each ChIP reaction the corresponding antibodies were added, no antibody to the mock ChIP, 3 µl α-H3 antibody (ab1791, Rb, Abcam) or 6 µl α-CID antibody (39720, Rb, Active Motif). These reactions were incubated overnight at 4°C under rotation. The next day, 25 µl Protein A Dynabeads (Invitrogen) were prepared per ChIP reaction in 1.5 ml low binding DNA tubes (Eppendorf) by washing them once with 1 ml PBS and once with 1 ml low salt wash. The ChIP reaction

was added to the beads and incubated for 3.5 h at 4°C under rotation. The beads with the attached antibody-chromatin were separated from the rest of the reaction using a magnetic rack, it is important not to let the beads dry out while exchanging washes. 60 µl supernatant was kept as flow through for the western blot, 20 µl 4X SDS Laemmli buffer was added and stored at -20°C. The beads were washed 5 times with 1 ml washes for 10-15 min each at 4°C under rotation with the following washes: low-salt wash, high-salt wash, LiCl wash and two times with TE (pH 8). In the last wash, the beads are split into 2, one half for the western blot and the other half for RNA isolation.

For western blot, the proteins were eluted of the beads by adding 20 µl 1X SDS Laemmli buffer, which was diluted from 4X with low-salt wash. This was incubated at RT for 20 min, before the beads were removed and the eluate was transferred to a new tube. This was stored at -20°C. Before loading on the SDS-PAGE gel the input, flow through and ChIP samples were all incubated for 1 h at 65°C to reverse the crosslink and for 5 min at 95°C to denature. The quality control western blot was performed as all western blots, see section 5.4.1.

For RNA isolation, 100 µl TE was added to the beads and 77.5 µl TE to the input samples. To this 10 µl 10X DNaseI buffer (NEB), 1 µl DNaseI (NEB) and 1 µl RNasin Plus Ribonuclease Inhibitor (Promega) was added, this was incubated for 30 min at 37°C shaking 700 RPM. To digest the proteins, 2.8 µl 20 % SDS and 10 µl of 5 mg/ml proteinase K (final concentration 0.5 µg/µl) was added, and incubated for 45 min at 42°C shaking 700 RPM. The beads were removed and the eluate was transferred to a new low binding DNA tube (Eppendorf). To reverse the formaldehyde crosslink, the sample was incubated for 45 min at 65°C, shaking 700 RPM. 375 µl TRIzol (Invitrogen) was added to the samples, and stored at -80°C. RNA was isolated with the Direc-Zol RNA microprep kit (Zymo Research) including the DNaseI step according to the manufacturer's manual and eluted in 10 µl RNase-free H_2O in a low binding DNA tube (Eppendorf). RNA concentrations were determined with the Qubit using the RNA High Sensitivity kit (Invitrogen), using 3 µl of the RNA.

Sequencing

For sequencing, libraries were prepared with the Universal RNA-seq kit with NuQuant and Drosophila specific rRNA depletion (nr. 0364-32, lot 1806358, Nugen/Tecan). The libraries were prepared by me in the lab of David Ibberson of the Bioquant sequencing facility, according to the manufacturer's manual v2.1 omitting the extra DNase step. For each sample either 10 µg or a maximum of 5 µl was used. For the RNA fragmentation, the samples were transferred to a microTUBE-130 AFA Fiber Snap-Cap tube (#520045, Covaris) and placed in a Covaris s220 Focussed-ultrasonicator with the following settings. For replicate 1, 10% Duty, Intensity 4, 200 cycles, 120 sec. For replicates 2-4, 10% Duty, Intensity 5, 200 cycles, 180 sec. The libraries had similar fragments sizes in the end. The PCR amplification cycle number was determined as described in the protocol with qPCR, replicate 1 was

amplified with 19 cycles, replicates 2-4 with 18 cycles. The size of the libraries was assessed on a Bioanalyzer with the HS DNA chip (Agilent), libraries diluted 1:10. The concentration was measured on the Qubit (Invitrogen) with the NuQuant protocol of the library kit. Replicate 1 was pooled together with replicated 2, and replicate 3 was pooled together with replicate 4, in a way that should result in an equal distribution of the number of reads across the libraries. The library pool concentrations were measure again with NuQuant on the Qubit, and the library pools were diluted with 10 mM Tris-HCl (pH 8.5) for sequencing. The sequencing was done on the NextSeq 550 in the Bukau lab by Kai Fenzl using the NextSeq 500/550 mid output kit v2.5 with 150 cycles (Illumina): 75 bp paired end, 8 nt barcode. For the pool of replicate 1/2 2 pM was loaded, for the pool of replicate 3/4 2.2 pM was loaded.

Fixing solution

50 mM HEPES (pH 8)
1 mM EDTA (pH 8)
0.5 mM EGTA (pH 8)
10 mM NaCl
1.6 % formaldehyde (methanol-free ampules, 16 % stock)

Stop solution

PBS
0.1 % Tween-20
125 mM glycine

PBST (+ inhibitors)

PBS
0.1 % Tween-20
(1X Roche protease inhibitors cocktail)
(Ribonucleoside Vanadyl Complex 1:20)

Cell lysis buffer

5 mM HEPES (pH 8.0)
85 mM KCl
0.5 % NP-40
1X Roche protease inhibitors cocktail
Ribonucleoside Vanadyl Complex 1:20

Nuclear lysis buffer

50 mM HEPES (pH 8.0)
10 mM EDTA (pH 8.0)
0.5 % N-Lauroylsarcosine (w/v)
1X Roche protease inhibitors cocktail
Ribonucleoside Vanadyl Complex 1:20

Dilution Buffer

4 % glycerol
10 mM Tris-HCl (pH 8)
1 mM EDTA
0.5 mM EGTA

1X Roche protease inhibitors cocktail
Ribonucleoside Vanadyl Complex 1:20

Low-salt wash

140 mM NaCl
10 mM Tris-HCl (pH 8)
1 mM EDTA
0.1 % SDS
1 % Triton X-100
0.1 % sodium deoxycholate (DOC, 10 % stock, in plastic, stored dark)
1 mM PMSF
(Ribonucleoside Vanadyl Complex 1:50)

High-salt wash

500 mM NaCl
10 mM Tris-HCl (pH 8)
1 mM EDTA
0.1 % SDS
1 % Triton X-100
0.1 % sodium deoxycholate (DOC)
1 mM PMSF
(Ribonucleoside Vanadyl Complex 1:50)

LiCl wash

0.25 M LiCl
10 mM Tris-HCl (pH 8)
1 mM EDTA
0.5 % NP-40
0.5 % sodium deoxycholate (DOC
(Ribonucleoside Vanadyl Complex 1:50)

TE wash

10 mM Tris-HCl (pH 8)
1 mM EDTA
(Ribonucleoside Vanadyl Complex 1:50)

Elution Buffer

10 mM Tris-HCl (pH 8)
1 mM EDTA
Ribonucleoside Vanadyl Complex 1:20

5.4.3 XRNAX extraction

This protocol was adapted to drosophila S2 cells and optimized from the XNRAX protocol published by Trendel et al. for human cells (TRENDEL ET AL., 2019). It makes use of the by UV induced direct and short distance link between directly interacting nucleotides (pyrimidines) and amino acids, such as Cys, Lys, Phe, Trp, and Tyr. UV does not create links between proteins and at 254nm with the intensity used also not between DNA and protein (TRENDEL ET AL., 2019; ULE ET AL., 2005). During TRIzol (Invitrogen) extraction these RNA-protein complexes get trapped in the interphase between the RNA containing

aqueous phase and the protein containing organic phase (TRENDEL ET AL., 2019). All buffers used are prepared RNase-free.

UV crosslinking and TRIzol lysis

XRNAX was done with drosophila S2 stable cell lines containing plasmids with V5 tagged Cal1 or CENP-C under a copper inducible pMT promotor. To induce expression of the tagged proteins, 400 µM $CuSO_4$ was added to the medium the night before harvesting the cells. For each XRNAX extraction $5*10^8$ cells were plated distributed over 6 cell culture dishes of 15 cm in 15-18 ml. Cells were allowed to settle for 4 h. The 6 dishes were combined into 1 lysate, so they were handled one after the other. The cells were carefully washed with 10 ml PBS without releasing the cells from the surface. The cells were allowed to settle 10 min before removing the PBS wash completely. The remaining PBS was drained by tilting the dish on its side for 1 min. The dish was placed on another dish filled with ice and UV crosslinked at 200 mJ/cm^2 with 254nm UV with the Bio-Link UV crosslinker (Vilber). To lyse the cells, 8 ml TRIzol (Invitrogen) was added and incubated for 10 min at RT. During this incubation time, the next dish was drained from PBS and UV crosslinked and to the dish after the PBS wash was added. The TRIzol was pipetted over the dish to disrupt and collect the UV crosslinked cells and transferred to the next dish. Since TRIzol is viscous, this was always slightly less than 8 ml. To fill up to 8 ml, a few hundred microliters of TRIzol was used to rinse the remaining lysate from the now previous dish. After the last dish, the lysate was collected in a 15 ml tube. The lysate was stored at -80°C until extraction.

From the same induced cells, $5*10^6$ cells were lysed in RIPA buffer as described in section 5.4.1 as a control lysate to assess the enrichment during the SDS-Page western blot analysis. Of this control lysate, 20 µl was loaded after denaturing the sample.

For the XRNAX with Human HEK293T cells, another PhD student in the lab Vojtěch Dolejš grew the cells, UV crosslinked them and lysed the crosslinked cells in TRIzol. I performed all the following steps for that experiment.

XRNAX extraction

The TRIzol lysate was thawed on ice. 1.6 ml Chloroform was added, mixed by inverting the tube several times and incubated at RT for 5 min. Phases were separated by centrifugation 3220 x g for 30 min at 4°C. The upper aqueous phase containing loose RNA was removed and discarded. The sticky interphase clump was transferred to a 2 ml tube. The interphase was washed twice with 1 ml TE + 0.1 % SDS, inverting several times to clear away more loose RNA and protein. The interphase clump was disrupted and dissolved in 5 steps of 1 ml each, first 2 steps with TE + 0.1 % SDS and the last 3 steps with TE + 0.5 % SDS. Each time the clump or resulting smaller flakes were pipetted at least 30 times to aid dissolving. The remaining interphase flakes were pelleted by centrifuging 5000 x g for 2 min at RT, after which the supernatant was transferred to a new 2 ml tube. The 5 resulting tubes with 1 ml dissolved interphase

was precipitated by adding 60 µl of 5 M NaCl, 1 µl GlycoBlue (Invitrogen) and 1 ml cold isopropanol, and mixed. The precipitate was pelleted by centrifuging 18000 x g for 15 min at 4°C, discard supernatant. The pellets were collected into 1 of the 5 2 ml tubes using 1 ml cold 70 % Ethanol, another 1 ml 70 % Ethanol was used to collect residual pieces of pellet. The pellets were combined by centrifuging 18000 x g for 2 min at 4°C, the Ethanol was discarded and removed with additional short spins. The pellet was dissolved in 1800 µl H_2O by hydrating on ice for 1 h while occasionally inverting the tube. After this hydration step, the pellet was dissolved by pipetting and if needed incubated longer on ice. The extract is very viscous due to the DNA that is still present, so the pipetting has to be done carefully and slowly. Next, the DNA is digested away with DNaseI. For this, the extract is split over 2 2ml tubes. To each 100 µl 10X DNaseI Buffer (NEB), 1 µl RNasin Plus Ribonuclease Inhibitor (Promega) and 50 µl DNaseI (NEB) was added, mixed and incubated for 90 min at 37°C while shaking 700 RPM. The RNA-protein complexes were precipitated again by adding 60 µl 5 M NaCl and 1 ml cold isopropanol, mixed and centrifuged 18000 x g for 15 min at 4°C. The pellets were collected into 1 of the 2 2 ml tubes using 1 ml cold 70 % Ethanol, another 1 ml 70 % Ethanol was used to collect residual pieces of pellet. The pellets were combined by centrifuging 18000 x g for 2 min at 4°C, the Ethanol was discarded and removed with additional short spins. The XRNAX extract was dissolved in 1 ml H_2O, again by hydrating for 1 hour on ice, occasionally mixing and pipetting after. If needed this hydrating step can also be done overnight on ice in the cold room. Remove the volume needed for measuring the concentration on the nanodrop, the QC digestions and western blot. The rest of the XRNAX extract can be stored at -80°C if needed.

Digestions for QC gel and western blot

The concentration of the XRNAX extract was measured with nanodrop, diluted 1:5 with H_2O.

In order to confirm the success of the XRNAX extraction, two control digestions were performed and compared to undigested extract on an Agarose gel. For the Proteinase K digest, 1 µl XRNAX extract was mixed with 10 µl 2X Proteinase K buffer, 5 µl H_2O and 4 µl Proteinase K (5 mg/ml stock, Thermo Fisher Scientific). This reaction was incubated at 60°C for 30 min while shaking 700 RPM. For the RNase A digest, 1 µl XRNAX extract was mixed with 18 µl 10 mM Tris-HCl (pH 8) and 1 µl or 0.25 µl RNase A (stocks respectively 10 mg/ml, AppliChem, and 40 mg/ml, Sigma-Aldrich). This reaction was incubated at 37°C for 30 min shaking 700 RPM. For the undigested control, 1 µl XRNAX extract was mixed with 19 µl 10 mM Tris-HCl (pH 8) and kept on ice. To each tube 20 µl 2X RNA loading dye (NEB) was added, 20 µl of this mix was loaded on a 1 % Agarose gel and separated for 45 min at 70 Volt. In case of a successful XRNAX extract, the undigested RNA-protein complexes are trapped in the well of the gel, while the RNA is able to migrate into the gel after the Proteinase K digest and no RNA is visible after in either well of gel after the RNase A digest.

The XRNAX extract was analyzed with SDS-Page followed by western blot to assess whether our proteins of interest were enriched in the XRNAX extract and therefore likely RNA-binding proteins. Before running the extract on SDS-Page gels, the RNA was digested away to get clear bands to quantify the enrichment. For this digestion 60 µl XRNAX extract was mixed with 1.5 µl 1 M Tris-HCl (pH 8) and 6 µl or 2 µl RNase A (stocks respectively 10 mg/ml, AppliChem, and 40 mg/ml, Sigma-Aldrich), and incubated at 37°C for 30 min while shaking 700 RPM. To this 22 µl 4X SDS Laemmli buffer was added, 40 µl was loaded on the SDS-Page gel after denaturing by heating. For SDS-Page and western blot protocol see section 5.4.1.

TE

10 mM Tris-HCl (pH 7.5)
1 mM EDTA

Proteinase K buffer

50 mM Tris-HCl (pH 8)
10 mM EDTA
150 mM NaCl
1 % SDS

5.4.4 XRNAX-CLIP-seq

The XRNAX extract was used as input for enhanced crosslinking and immunoprecipitation (eCLIP)-seq. The protocol was adapted from the XRNAX-CLIP-seq method described by Trendel et al. (2019) and the eCLIP-seq methods described by Van Nostrand et al. (2016) (TRENDEL ET AL., 2019; VAN NOSTRAND ET AL., 2016). The advantages of using the XRNAX extract instead of regular cell lysate is that the XRNAX extract has an enrichment of RNA-binding proteins and the DNA is already removed. For low abundant proteins like our proteins of interest this extra enrichment is beneficial. However, the removal of DNA is especially essential, given that our proteins of interest potentially interact with RNA derived from repetitive sequences. This also allows to switch to RNA fragmentation using sonication instead of diluted RNase A digestion, which is a difficult to optimize step of CLIP protocols.

RNA fragmentation

The XRNAX extract was thawed on ice and 6 µl 1 M Tris-HCl (pH 7.5) and 6 µl 0.5 M EDTA was added. Sonication was done with the Covaris m220 Focussed-ultrasonicator in a microTUBE-130 AFA Fiber Snap-Cap tube (#520045, Covaris) in 130 µl batches, re-using the sonication tube. The sonication settings were 240 sec, 50 W peak intensity power, 20 % duty factor, 200 cycles per burst. The sonicated extract was collected in a new 2 ml tube. If needed H_2O was added to a total volume of 900 µl.

CLIP

Two input samples were taken, one for the control western blot and the other for the size-matched input control (SMI-control). For the input of the control western blot, 18 µl sonicated extract was taken

and treated with 1 µl RNase A (Sigma-Aldrich) for 30 min at 37°C. After this treatment 6 µl 5X SDS loading dye and 6 µl 1 M DTT was added, and the sample was stored on ice until further processing. For the SMI-control, 5 µl sonicated extract was taken and the RNA fragments were repaired. First 30 µl H_2O, 5 µl 10X FastAP buffer and 10 µl FastAP (Thermo Fisher Scientific) was added, mixed and incubated for 15 min at 37°C. The enzyme was inactivated for 5 min at 80°C and after which the sample was placed on ice. Next 5 µl 10X PNK buffer A, 10 µl 10 mM ATP (Thermo Fisher Scientific), 10 µl H_2O and 10 µl PNK (Thermo Fisher Scientific) was added, mixed and incubated 15 min at 37°C. 15 µl of this sample was combined with 5 µl 5X SDS loading dye and 5 µl 1 M DTT as the SMI-control and stored on ice until further processing.

To the remaining sonicated extract 900 µl 2X IP buffer and 5 µl α-V5 antibody (polyclonal, Rb, Abcam ab15828, Lot GR3265659-1, concentration 1 mg/ml) or 0.5 µl IgG antibody (control IgG, Rb, Sigma-Aldrich I-8140-10MG, Lot 029H9125, concentration 11 mg/ml) was added. The antibodies were allowed to bind their targets during a 4 h incubation at 4°C on a rotating wheel. Meanwhile, 50 µl Protein A Dynabeads (10002D, Lot 00825220, 30 mg/ml, Thermo Fisher Scientific) were washed 3 times with 1 ml 1X IP buffer and resuspended in 100 µl 1X IP buffer. The washed beads were added to the IP sample and incubated overnight at 4°C on a rotating wheel.

Beads were separated on a magnetic rack to take a 38 µl flow through sample for the control western blot. This sample was treated with 1 µl RNase A (Sigma-Aldrich) for 30 min at 37°C, after which 12.7 µl 5X SDS loading dye and 12.7 µl 1 M DTT was added. The flow through sample was stored on ice until further processing. The remaining supernatant with unbound RNA-protein complexes was removed and the beads were washed 3 times with 1 ml cold 1X IP buffer for 5 min at 4°C on a rotating wheel. Next the beads were washed 2 times with 1 ml cold TBST, of which the first for 5 min at 4°C on a rotating wheel and the second while tubes remain on the magnetic rack. The RNA fragments of the IP-ed RNA-protein complexes have to be repaired for the library preparation. For the first end repair, the dephosphorylation, 80 µl H_2O, 10 µl 10X FastAP buffer, 8 µl FastAP (Thermo Fisher Scientific) and 2 µl RNasin Plus Ribonuclease Inhibitor (Promega) was added to the beads and incubated for 15 min at 37°C while shaking 1000 RPM. The beads were washed 2 times with 1 ml cold TBST while the tubes remain on the magnetic rack. For the second end repair, adding a phosphate group back to the 5'-end, 70 µl H_2O, 10 µl 10 mM ATP (Thermo Fisher Scientific), 10 µl 10X PNK buffer A, 8 µl PNK (Thermo Fisher Scientific) and 2 µl RNasin Plus Ribonuclease Inhibitor (Promega) was added and incubated for 10 min at 37°C while shaking 1000 RPM. The beads were resuspended in 1 ml cold TBST, from which 200 µl was taken for the control western blot. The beads for the control western blot were resuspended in 14 µl H_2O with 1 µl RNase A (Sigma-Aldrich) and incubated for 30 min at 37°C. To this 5 µl 5X SDS loading dye and 5 µl 1 M DTT was added, the sample was kept on ice until further processing. The remaining beads of the IP was eluted in 5 µl 5X SDS loading dye, 5 µl 1 M DTT and 15 µl H_2O by

incubating for 15 min at 70°C while shaking 1000 RPM. At this step also the SMI-control was heated the same. The beads were separated from the IP sample and the eluate was transferred to a new tube ready for loading on the SDS-PAGE gel.

SDS-PAGE

To separate unspecific RNA-protein complexes of another molecular weight to the one of interest, the CLIP and SMI-control loaded on an SDS-PAGE gel. 25 µl of the IP eluate and 25 µl of the SMI-control was loaded on a 4-12 % Bis-Tris Criterion XT Midi gel, 12+2 well (Bio-Rad). For easier excision, one well was kept empty between samples and ladders. The sample was separated at 150 Volt in NuPAGE MOPS SDS Running Buffer (Bio-Rad) until the dye front reached the bottom of the gel. MOPS buffer was used as it allows for better separation of larger proteins. Separated RNA-protein complexes were transferred to a 0.45 µm nitrocellulose membrane in NuPAGE transfer buffer (Bio-Rad) supplemented with 20 % methanol for 1 h at 100 Volt in the cold room with an ice pack.

RNA release and clean-up

Using a fresh scalpel, an area of the membrane was excised for both CLIP and SMI-control corresponding to the molecular weight of the protein of interest and above where the RNA-protein complexes should run. This piece of membrane was cut into smaller slices and transferred to a 1.5 ml low binding DNA tube (Eppendorf). To release the RNA from the membrane, the protein had to be digested away. For this 200 µl Proteinase K buffer and 50 µl Proteinase K (Thermo Fisher Scientific) was added to the membrane slices and incubated for 30 min at 55°C. 150 µl 6 M Urea in Proteinase K buffer was added to aid the protein digestion and incubated for another 10 min at 37°C. To isolate the RNA from this reaction, 400 µl acid phenol/chloroform/isoamyl alcohol (Sigma-Aldrich) was added, mixed by shaking and incubated for 5 min at 37°C while shaking 1200 RPM. The tubes were quickly centrifuged to settle the membrane pieces and the supernatant was transferred to a pre-spun 1.5 ml MaXtract High Density tube (Qiagen). The samples were incubated for another 5 min at 37°C while shaking 1200 RPM before inducing phase separation by centrifuging at 13000 x g for 15 min. The upper phase was transferred to a 5 ml tube. The RNA was isolated further using the RNA clean and concentrator kit from Zymo Research according to manufacturer's protocol. In short, 2 Volumes RNA binding buffer and 1 Volume 100 % Ethanol was added. After mixing, it was loaded on the columns in 750 µl batches. The column was washed subsequently with 400 µl RNA prep buffer, 700 µl RNA wash buffer and 400 µl RNA wash buffer. The columns were centrifuged to dry before eluting the RNA in 10 µl H_2O. The RNA was stored at -80°C until library preparation.

Control western blot

The samples for the control western blot were heated for 5 min at 95°C before loading the entire

volume of the Input, Flow Through and IP sample on a 4-12 % Bis-Tris Criterion XT Midi gel, 12+2 well (Bio-Rad). SDS-PAGE and transfer was performed in parallel to the SDS-PAGE of the CLIP with the same buffers and settings. After the transfer, the membrane was blocked, incubated with antibodies and developed as described in section 5.4.1.

Sequencing

For sequencing, libraries were prepared with the SMARTer smRNA-Seq kit for Illumina (cat 635032, lot 1909875A, TAKARA), using the indexes from the SMARTer RNA unique Dual index kit – 24U (cat 634451, lot 2007495A, TAKARA), indexes U25-U36. The libraries were prepared by me in our own lab, according to the manufacturer's protocol, and also including a H_2O negative control. For each sample 7 µl RNA was used, no extra ATP was added during the polyadenylation step. The PCR amplification cycle number was 11 cycles. Before bead size selection, the quantity of the library was measure with a Qubit using the dsDNA High Sensitivity kit (Invitrogen) by me, and the quality of the libraries was assessed on a Fragment Analyzer using the HS NGS kit (Agilent) in the Kaessmann lab. The concentration was a bit lower than expected, I could have used more PCR cycles for amplifying the libraries. The negative control and IgG IP libraries were almost undetected by the Fragment Analyzer. There were adapter dimers present in the libraries. Library size selection was performed according to the SMARTer smRNA-Seq kit protocol (TAKARA) using AMPURE XP beads (A63881, Beckman Coulter). Again the quantity of the library was measure with a Qubit using the dsDNA High Sensitivity kit (Invitrogen) by me, and the quality of the libraries was assessed on a Fragment Analyzer using the HS NGS kit (Agilent) in the Kaessmann lab. Adapter dimers were removed successfully. The IgG Input and IP samples of replicate 1 and 3 were excluded from pooling due to too low concentrations in the IP samples. The remaining libraries were pooled together in a way that should result in an equal distribution of the number of reads across the libraries. The library pool concentrations were measure again on the Qubit with dsDNA High Sensitivity kit (Invitrogen), and the library pools were diluted with 10 mM Tris-HCl (pH 8.5) for sequencing. The sequencing was done on the NextSeq 550 in the Bukau lab by Anja Schubert and Jaro Schmitt using the NextSeq 500/550 high output kit v2.5 with 75 cycles (Illumina): 75 bp single end, 8 nt barcode. The final library concentration loaded onto the sequencing chip was 1.6 pM with a 30% phiX spike-in according to the TAKARA manual instructions.

2X IP buffer

100 mM Tris-HCl (pH 7.5)
200 mM NaCl
2 % NP-40
0.2 % SDS
1 % sodium deoxycholate (DOC, 10 % stock, in plastic, stored dark)
2X Roche Proteinase Inhibitor cocktail

TBST

50 mM Tris-HCl (pH 7.5)
150 mM NaCl
0.1 % Tween-20

5X SDS loading buffer

250 mM Tris-HCl (pH 6.8)
10 % SDS
0.02 % Bromphenol blue
30 % Glycerol

Proteinase K buffer

50 mM Tris-HCl (pH 8)
10 mM EDTA
150 mM NaCl
1 % SDS

5.5 Microscopy

5.5.1 Light microscope

Dissections and fly sorting were performed on the Stereomicroscope from Zeiss with external light source.

5.5.2 Deltavision microscope

S2 cells with IF or FISH stainings were imaged with a wide field fluorescence microscope with a DeltaVision Core system (Applied Precision) and a charge-coupled device camera (CoolSNAP HQ2, Photometrics) in combination with the softWoRx v5.5 suite operating software. Either the 100x UPlan-SApochromat (NA 1.4) or 60x Plan-Apochromat N (NA 1.42) objective lenses with a binning of 1x1 or 2x2 were used for image acquisition. To image the entire cell, z-stacks were made with 0.2 µm slices.

The images were deconvolved (conservative ratio, 10 cycles), maximum projected or kept as a z-stack, cropped and exported as TIFF images using the softWoRx 6.5.2 suite software. Further image processing and analysis was done in Fiji.

5.5.3 Zeiss LSM 900 laser scanning microscope

Drosophila tissues were imaged using the Zeiss LSM 900 laser scanning microscope with either the LD LCI Plan-Apochromat 40x or the C Plan-Apochromat 63x objective lenses and the ZEN 3.0 software by Zeiss. If higher resolution was required, the Airyscan mode was used.

5.5.4 Image analysis in Fiji

Microscopy and western blot images were processed and quantified in Fiji (SCHINDELIN ET AL., 2019). The background was subtracted using the rolling ball radius tool with the radius set to the radius of the largest object in the image, for example a western blot band or a cell nucleus.

Centromeric CID foci intensity in images of S2 cells was measured with a custom plugin developed by the Nikon Imaging Centre for the Erhardt lab. In the DAPI channel, nuclei were marked and saved as ROIs. Foci in the CID channel were enhanced using the DoG spot enhancer plugin and the threshold was set with the Li method. Mean centromeric CID foci intensity per nucleus was determined with the ROI particle analyzer plugin.

SATIII foci intensity in microscopy images was measured using a custom python plugin made by Patrick van Nierop Y Sanchez from Ingrid Lohmann's lab at COS. To capture signals with a wide range of intensity during thresholding, the histograms of all analyzed images were stretched to one reference image using the stack contrast adjustment plugin (MICHALEK ET AL., N.D.). Now that all images have similar brightness histograms, auto thresholding with the Moments method was used to select ROIs to measure on the original images.

Foci colocalization was either identified manually or by using the Fiji plugin Just Another Colocalization Plugin (JACoP) (BOLTE & CORDELIERES, 2006). With JACoP, the Manders' overlap coefficients M1 and M2 were calculated (MANDERS ET AL., 1993). M1 is the ratio of the "summed intensities of pixels from image A for which the intensity in image B is above zero" to the "total intensity in the image A" and M2 vice versa. These coefficients go from 0 to 1, with 0 being no overlap and 1 perfect overlap. For this analysis, a threshold is required, which was set manually. Alternatively, also with JACoP, the Costes' overlap analysis was performed. This approach calculates the Pearson's coefficient to describe the correlation between the pixel intensities of two channels. Next, JaCOP creates 500 randomized images of one of the two channels and computes the range of Pearson's coefficients to this set of randomized images, and whether the original Pearson's coefficient is significantly different.

The signal intensity of western blot bands was quantified by measuring the "volume" of the band made up by its length, width and pixel intensity value. Individual bands were outlined with a rectangle and selected as lanes (Analyze>Gels>Select First/Next Lane). The profiles were plotted (Analyze>Gels>Plot Lanes), the peak was closed to be able to measure the area of the peak which corresponds with the intensity of the band.

Images were prepared for publication by adjusting the brightness and contrast on concatenated images, cropping and adding scale bars when necessary. All figures in this thesis where prepared using Gimp and Inkscape.

5.6 Bioinformatic analysis

Initial sequencing data processing, mapping and counting were performed in the Baden-Württemberg's high performance computing cluster (bwForCluster MLS&WISO Production) using UNIX-shell scripts. Further analysis and data visualization was done in R.

5.6.1 RNA-ChIP-seq data processing

The raw sequencing data quality was assessed with FastQC (version 0.11.5) (ANDREWS, N.D.). Next, reads were mapped to the reference genome using Spliced Transcripts Alignment to a Reference (STAR, version 2.6.1a) (DOBIN ET AL., 2013). As a reference genome, a new assembly from the Mellone and Larracuente labs was used (CHANG ET AL., 2019). This reference was generated from PacBio long-read sequences and CID DNA-ChIP derived centromeric DNA contigs, making it the first *Drosophila* reference to include the centromere and parts of the pericentromere. The reference was downloaded on 12.09.2019 from the Dryad repository of the Chang et al., 2019 paper, with File.S8 being the genome reference, File.S10 being the gene annotation, and File.S9 being the Transposable Element (TE) and repeat annotation. The reference was indexed with STAR, using the following command.

```
STAR \
      --runThreadN 8 \
      --runMode genomeGenerate \
      --genomeDir . \
      --genomeFastaFiles ./File.S8.Chang_et_al.fasta \
      --sjdbGTFfile ./File.S10.Chang_et_al.gtf \
      --sjdbOverhang 74
```

To align the reads to the reference, the following command was used, with ${1} being the sample name.

```
STAR \
      --runThreadN 8 \
      --genomeDir path_to_reference/fly/ \
      --genomeLoad LoadAndKeep \
      --readFilesIn \
            <(zcat ../raw_data/${1}_R1_001.fastq.gz) \
            <(zcat ../raw_data/${1}_R2_001.fastq.gz) \
      --outFileNamePrefix ./${1}_ \
      --outFilterMultimapNmax 100 \
      --winAnchorMultimapNmax 100 \
      --outSAMtype BAM Unsorted
```

This mapping was optimized so that the mapped reads can be used with TEcount (TEToolkit, version 2.0.3) (JIN ET AL., 2015). TEcount is a read counter which quantifies to both genes and repeats or TEs, using both uniquely mapped and multimapped reads. This program needs a GTF file with the annotations where in the genome TEs and repeats are located. Chang et al., 2019 have provided this annotation in File.S9, however this file needed to be transformed slightly to fit the GTF format required by this tool. The following command was used to get the counts with TEcount, with ${1} being the sample name.

```
TEcount -b path_to_alignments/${1}_Aligned.out.bam \
      --GTF path_to_reference/fly/File.S10.Chang_et_al.gtf \
      --TE ./File_S9_TE_fixed.txt \
      --project ${1}_tecount_out
```

Finally, to be able to view the mapped reads with the Integrative Genomics Viewer (IGV) (ROBINSON ET AL., 2011), the BAM files that were the output of STAR had to be sorted and indexed with SAMtools (DANECEK ET AL., 2021). These were the commands, with ${1} being the sample name.

```
samtools sort \
      -@ 8 \
      -O BAM \
      -o ${1}_sorted.bam \
      -T ${TMPDIR}/${1}_tempsort \
      ${1}_Aligned.out.bam

samtools index ${1}_sorted.bam
```

Further analysis of the count tables were performed with R programming language version 4.1.0 (R CORE TEAM, 2018) in RStudio (RSTUDIO TEAM, 2020). Differential enrichment analysis was done with DESeq2 (LOVE ET AL., 2014), which provides Fold Changes and p-values that are adjusted for multiple testing with the Benjamini-Hochberg correction. ClusterProfiler was used for GO term enrichment analysis, using the enrichGO function for the top differentially enriched genes (G. YU ET AL., 2012). Tidyverse and its data visualization tool ggplot2 was used to plot the data (WICKHAM ET AL., 2019).

5.6.2 XRNAX-CLIP-seq data processing

The first step of the CLIP-seq data processing is adapter trimming by Cutadapt (version 3.4) which requires Python (version 3.8.6) (MARTIN, 2011). The following parameters were used for adapter trimming, with ${1} being the sample name.

```
cutadapt \
      -m 15 \
      -O 5 \
      -a AAAAAAAAAA \
      -o ${1}_out.fastq \
      ../raw_data/${1}_R1_001.fastq.gz
```

These trimmed reads were aligned to the reference genome from Chang et al. (2019) paper mentioned above. Again STAR (version 2.6.1a) was used for mapping the reads to the reference, allowing for multi mappers to be able to count reads for repetitive sequences as well (DOBIN ET AL., 2013). The indexed Mellone and Larracuente labs genome reference was still available from previous analysis. The following command was used, with ${1} being the sample name.

```
STAR \
      --runThreadN 8 \
      --genomeDir path_to_reference/fly/ \
      --genomeLoad LoadAndKeep \
      --readFilesIn ../cutadapt/${1}_out.fastq \
      --outFileNamePrefix ./${1}_ \
      --outFilterMultimapNmax 100 \
      --winAnchorMultimapNmax 100 \
      --outFilterMismatchNmax 20 \
      --outSAMtype BAM Unsorted
```

Next the BAM files were sorted in order to remove PCR duplicates which are very abundant in CLIP dataset due to their low complexity (UHL ET AL., 2017). There is some sequence bias in the PCR amplification during library preparation, potentially inflating counts of certain genes. SAMtools (version 1.9) (DANECEK ET AL., 2021) was used to sort the BAM files and Picard (version 2.20.0) (BROAD INSTITUTE, 2019) was used to remove duplicates as follows, with ${1} being the sample name.

```
samtools sort \
      -@ 8 \
      -O BAM \
      -o ${1}_sorted.bam \
      -T ${TMPDIR}/${1}_tempsort \
      ${1}_Aligned.out.bam

java -jar /opt/bwhpc/common/bio/picard/2.20.0-java_jdk-
1.8.0/lib/picard.jar \
      MarkDuplicates \
      I=${1}_sorted.bam \
      REMOVE_DUPLICATES=true \
      O=${1}_sorted_dupRm.bam \
      METRICS_FILE=${1}_dupRm.txt
```

As above TEcounts (TEToolkit, version 2.2.1), together with Python (version 3.8.6) was used to count reads mapping to coding genes, non-coding genes, TEs and repeats (JIN ET AL., 2015). The following script was used, with ${1} being the sample name.

```
TEcount -b path_to_alignments/${1}_sorted_dupRm.bam \
      --sortByPos \
      --GTF path_to_reference/fly/File.S10.Chang_et_al.gtf \
      --TE ./File_S9_TE_fixed3.gtf \
      --project ${1}_tecount_out
```

Also for this dataset the sorted BAM files were indexed to be able to see the read coverage with IGV (ROBINSON ET AL., 2011). The following commands from SAMtools (DANECEK ET AL., 2021) were used, with ${1} being the sample name.

```
samtools index ${1}_sorted_dupRm.bam
```

Further analysis of the count tables were performed with R programming language version 4.1.0 (R CORE TEAM, 2018) in RStudio (RSTUDIO TEAM, 2020). Differential enrichment analysis was done with DESeq2 (LOVE ET AL., 2014), which provides Fold Changes and p-values that are adjusted for multiple testing with the Benjamini-Hochberg correction. Tidyverse and its data visualization tool ggplot2 was used to plot the data (WICKHAM ET AL., 2019).

5.6.3 Data mining development and tissue

To assess the levels of repeats and TEs in drosophila across development and in different tissues, the modENCODE RNA-seq datasets published by Duff et al., 2015 were reprocessed (DUFF ET AL., 2015). Originally multi mappers and reads mapping to repetitive regions were discarded during the analysis. With the current genome reference and tools it is possible to get information on the relative expression across development and tissues.

These datasets from modENCODE are publicly available on the NCBI SRA repository. The SRA accession numbers of the datasets used can be found in section 4.9. The fastq files were downloaded using the SRA toolkit (version 2.11.0) (SRA TOOLKIT DEVELOPMENT TEAM, N.D.) with the following commands, with ${1} being the SRA accession numbers.

```
prefetch \
      --output-directory . \
      ${1}

fastq-dump \
      --outdir . \
      --gzip \
      --skip-technical \
      --read-filter pass \
      --dumpbase \
      --split-e \
      --clip \
      ./${1}.sra
```

Again STAR (version 2.6.1a) was used for mapping the reads to the reference, allowing for multi mappers to be able to count reads for repetitive sequences as well (DOBIN ET AL., 2013). The indexed Mellone and Larracuente labs genome reference was still available from previous analysis. The following command was used, with ${1} being the sample name.

```
STAR \
      --runThreadN 8 \
      --genomeDir path_to_reference/fly/ \
      --genomeLoad LoadAndKeep \
      --readFilesIn \
            <(zcat ../raw_data/${1}_pass_1.fastq.gz) \
            <(zcat ../raw_data/${1}_pass_2.fastq.gz) \
      --outFileNamePrefix ./${1}_ \
      --outFilterMultimapNmax 100 \
      --winAnchorMultimapNmax 100 \
      --outFilterMismatchNmax 20 \
      --outSAMtype BAM Unsorted
```

As above TEcounts (TEToolkit, version 2.2.1), together with Python (version 3.8.6) was used to count reads mapping to coding genes, non-coding genes, TEs and repeats (JIN ET AL., 2015). The following script was used, with ${1} being the sample name.

```
TEcount -b path_to_alignments/${1}_Aligned.out.bam \
      --GTF path_to_reference/fly/File.S10.Chang_et_al.gtf \
      --TE ./File_S9_TE_fixed2.gtf \
      --project ${1}_tecount_out
```

Further analysis of the count tables were performed with R programming language version 4.1.0 (R CORE TEAM, 2018) in RStudio (RSTUDIO TEAM, 2020). Counts were normalized to reads per million (RPM), to account for sequencing depth differences between samples. In order to visualize expression trends of all elements without them being overshadowed by the highly expressed elements, Z-score normalization was applied to the RPM counts. Tidyverse and its data visualization tool ggplot2 was used to plot the data (WICKHAM ET AL., 2019).

5.6.4 BLAST search against custom BLAST database in command line

To assess the specificity of designed shRNAs for specific TEs, a custom BLAST database was created containing all TE consensus sequences. The TE consensus sequences file (File.S1) was downloaded on 12.09.2019 from the Dryad repository of the Chang et al., 2019 paper. The command-line tool from the NCBI, BLAST+, was used to create a custom BLAST database for this file with the following script (CAMACHO ET AL., 2009). The *Drosophila melanogaster* taxid is 7227.

```
makeblastdb \
      -in File.S1.Chang_et_al.fasta \
      -dbtype nucl \
      -parse_seqids \
      -taxid 7227 \
      -title "Blastdb_TE_Drosophila" \
      -out ./Blastdb_TE_Drosophila.fasta
```

Sequences were blasted against this custom database with the following script.

```
blastn
      -query dsRNA_1.fasta \
      -db Blastdb_TE_Drosophila \
      -word_size 10 \
```

Bibliography

Abad, J. P., Agudo, M., Molina, I., Losada, A., Ripoll, P., & Villasante, A. (2000). Pericentromeric regions containing 1.688 satellite DNA sequences show anti-kinetochore antibody staining in prometaphase chromosomes of Drosophila melanogaster. *Molecular Genetics and Genomics*, *264*(4), 371–377. https://doi.org/10.1007/s004380000331

Ahmadi, A., De Toma, I., Vilor-Tejedor, N., Eftekhariyan Ghamsari, M. R., & Sadeghi, I. (2020). Transposable elements in brain health and disease. *Ageing Research Reviews*, *64*, 101153. https://doi.org/10.1016/j.arr.2020.101153

Akdemir, F., Christich, A., Sogame, N., Chapo, J., & Abrams, J. M. (2007). p53 directs focused genomic responses in Drosophila. *Oncogene*, *26*(36), 5184–5193. https://doi.org/10.1038/sj.onc.1210328

Allshire, R. C., & Karpen, G. H. (2008). Epigenetic regulation of centromeric chromatin: old dogs, new tricks? *Nature Reviews Genetics*, *9*(12), 923–937. https://doi.org/10.1038/nrg2466

Alonso, A., Hasson, D., Cheung, F., & Warburton, P. E. (2010). A paucity of heterochromatin at functional human neocentromeres. *Epigenetics & Chromatin*, *3*(1), 6. https://doi.org/10.1186/1756-8935-3-6

Altemose, N., Logsdon, G. A., Bzikadze, A. V., Sidhwani, P., Langley, S. A., Caldas, G. V., Hoyt, S. J., Uralsky, L., Ryabov, F. D., Shew, C. J., Sauria, M. E. G., Borchers, M., Gershman, A., Mikheenko, A., Shepelev, V. A., Dvorkina, T., Kunyavskaya, O., Vollger, M. R., Rhie, A., … Miga, K. H. (2022). Complete genomic and epigenetic maps of human centromeres. *Science*, *376*(6588). https://doi.org/10.1126/science.abl4178

Andrews, S. (n.d.). *FastQC*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Anwar, S. L., Wulaningsih, W., & Lehmann, U. (2017). Transposable Elements in Human Cancer: Causes and Consequences of Deregulation. *International Journal of Molecular Sciences 2017, Vol. 18, Page 974*, *18*(5), 974. https://doi.org/10.3390/IJMS18050974

Arunkumar, G., & Melters, D. P. (2020). Centromeric Transcription: A Conserved Swiss-Army Knife. *Genes*, *11*(8), 911. https://doi.org/10.3390/genes11080911

Batki, J., Schnabl, J., Wang, J., Handler, D., Andreev, V. I., Stieger, C. E., Novatchkova, M., Lampersberger, L., Kauneckaite, K., Xie, W., Mechtler, K., Patel, D. J., & Brennecke, J. (2019). The nascent RNA binding complex SFiNX licenses piRNA-guided heterochromatin formation. *Nature Structural & Molecular Biology*, *26*(8), 720–731. https://doi.org/10.1038/s41594-019-0270-6

Bayer, T. S., Booth, L. N., Knudsen, S. M., & Ellington, A. D. (2005). Arginine-rich motifs present multiple interfaces for specific binding by RNA. *RNA*, *11*(12), 1848. https://doi.org/10.1261/RNA.2167605

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Bergmann, J. H., Rodríguez, M. G., Martins, N. M. C., Kimura, H., Kelly, D. A., Masumoto, H., Larionov, V., Jansen, L. E. T., & Earnshaw, W. C. (2011). Epigenetic engineering shows H3K4me2 is required for HJURP targeting and CENP-A assembly on a synthetic human kinetochore. *The EMBO Journal*, *30*(2), 328–340. https://doi.org/10.1038/emboj.2010.329

Black, B. E., & Bassett, E. A. (2008). The histone variant CENP-A and centromere specification. *Current*

Opinion in Cell Biology, *20*(1), 91–100. https://doi.org/10.1016/j.ceb.2007.11.007

Blower, M. D. (2016). Centromeric Transcription Regulates Aurora-B Localization and Activation. *Cell Reports*, *15*(8), 1624–1633. https://doi.org/10.1016/j.celrep.2016.04.054

Blower, M. D., Daigle, T., Kaufman, T., & Karpen, G. H. (2006). Drosophila CENP-A Mutations Cause a BubR1- Dependent Early Mitotic Delay without Normal Localization of Kinetochore Components. *PLoS Genetics*, *2*(7), e110. https://doi.org/10.1371/journal.pgen.0020110

Blower, M. D., Sullivan, B. A., & Karpen, G. H. (2002). Conserved Organization of Centromeric Chromatin in Flies and Humans. *Developmental Cell*, *2*(3), 319–330. https://doi.org/10.1016/S1534-5807(02)00135-1

Bobkov, G. O. M., Gilbert, N., & Heun, P. (2018). Centromere transcription allows CENP-A to transit from chromatin association to stable incorporation. *Journal of Cell Biology*, *217*(6), 1957–1972. https://doi.org/10.1083/jcb.201611087

Bodor, D. L., Mata, J. F., Sergeev, M., David, A. F., Salimian, K. J., Panchenko, T., Cleveland, D. W., Black, B. E., Shah, J. V, & Jansen, L. E. (2014). The quantitative architecture of centromeric chromatin. *ELife*, *3*. https://doi.org/10.7554/eLife.02137

Bolte, S., & Cordelieres, F. P. (2006). A guided tour into subcellular colocalization analysis in light microscopy. *Journal of Microscopy*, *224*, 213–232.

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology*, *19*(1), 199. https://doi.org/10.1186/s13059-018-1577-z

Bouzinba-Segard, H., Guais, A., & Francastel, C. (2006). Accumulation of small murine minor satellite transcripts leads to impaired centromeric architecture and function. *Proceedings of the National Academy of Sciences*, *103*(23), 8709–8714. https://doi.org/10.1073/pnas.0508006103

Brand, A. H., & Perrimon, N. (1993). Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development*, *118*(2), 401–415. https://doi.org/10.1242/DEV.118.2.401

Broad Institute, G. repository. (2019). *Picard toolkit*. Broad Institute. https://broadinstitute.github.io/picard/

Brodsky, M. H., Weinert, B. T., Tsang, G., Rong, Y. S., McGinnis, N. M., Golic, K. G., Rio, D. C., & Rubin, G. M. (2004). Drosophila melanogaster MNK/Chk2 and p53 Regulate Multiple DNA Repair and Apoptotic Pathways following DNA Damage . *Molecular and Cellular Biology*, *24*(3), 1219–1231. https://doi.org/10.1128/mcb.24.3.1219-1231.2004

Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., Booth, B. W., Wen, J., Park, S., Suzuki, A. M., Wan, K. H., Yu, C., Zhang, D., Carlson, J. W., Cherbas, L., Eads, B. D., Miller, D., Mockaitis, K., Roberts, J., … Celniker, S. E. (2014). Diversity and dynamics of the Drosophila transcriptome. *Nature*, *512*(7515), 393–399. https://doi.org/10.1038/nature12962

Bunch, T. A., Grinblat, Y., & Goldstein, L. S. B. (1988). Characterization and use of the Drosophila metallothionein promoter in cultured Drosophila melanogaster cells. *Nucleic Acids Research*, *16*(3), 1043–1061. https://doi.org/10.1093/nar/16.3.1043

Calabretta, S., & Richard, S. (2015). Emerging Roles of Disordered Sequences in RNA-Binding Proteins. *Trends in Biochemical Sciences*, *40*(11), 662–672. https://doi.org/10.1016/J.TIBS.2015.08.012

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009).

BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 1–9. https://doi.org/10.1186/1471-2105-10-421/FIGURES/4

Carmena, M., Wheelock, M., Funabiki, H., & Earnshaw, W. C. (2012). The chromosomal passenger complex (CPC): from easy rider to the godfather of mitosis. *Nature Reviews Molecular Cell Biology*, *13*(12), 789–803. https://doi.org/10.1038/nrm3474

Catania, S., Pidoux, A. L., & Allshire, R. C. (2015). Sequence features and transcriptional stalling within centromere DNA promote establishment of CENP-A chromatin. *PLoS Genetics*, *11*(3), e1004986. https://doi.org/10.1371/journal.pgen.1004986

Chakraborty, R., Li, Y., Zhou, L., & Golic, K. G. (2015). Corp Regulates P53 in Drosophila melanogaster via a Negative Feedback Loop. *PLoS Genet*, *11*(7), 1005400. https://doi.org/10.1371/journal.pgen.1005400

Chan, D. Y. L., Moralli, D., Khoja, S., & Monaco, Z. L. (2017). Noncoding Centromeric RNA Expression Impairs Chromosome Stability in Human and Murine Stem Cells. *Disease Markers*, *2017*. https://doi.org/10.1155/2017/7506976

Chan, F. L., Marshall, O. J., Saffery, R., Won Kim, B., Earle, E., Choo, K. H. A., & Wong, L. H. (2012). Active transcription and essential role of RNA polymerase II at the centromere during mitosis. *Proceedings of the National Academy of Sciences*, *109*(6), 1979–1984. https://doi.org/10.1073/pnas.1108705109

Chang, C.-H., Chavan, A., Palladino, J., Wei, X., Martins, N. M. C., Santinello, B., Chen, C.-C., Erceg, J., Beliveau, B. J., Wu, C.-T., Larracuente, A. M., & Mellone, B. G. (2019). Islands of retroelements are major components of Drosophila centromeres. *PLOS Biology*, *17*(5), e3000241. https://doi.org/10.1371/journal.pbio.3000241

Cheeseman, I. M. (2014). The Kinetochore. *Cold Spring Harbor Perspectives in Biology*, *6*(7), a015826–a015826. https://doi.org/10.1101/cshperspect.a015826

Chen, C.-C., Bowers, S., Lipinszki, Z., Palladino, J., Trusiak, S., Bettini, E., Rosin, L., Przewloka, M. R., Glover, D. M., O'Neill, R. J., & Mellone, B. G. (2015). Establishment of Centromeric Chromatin by the CENP-A Assembly Factor CAL1 Requires FACT-Mediated Transcription. *Developmental Cell*, *34*(1), 73–84. https://doi.org/10.1016/j.devcel.2015.05.012

Chen, C.-C., Dechassa, M. L., Bettini, E., Ledoux, M. B., Belisario, C., Heun, P., Luger, K., & Mellone, B. G. (2014). CAL1 is the Drosophila CENP-A assembly factor. *Journal of Cell Biology*, *204*(3), 313–329. https://doi.org/10.1083/jcb.201305036

Chen, H., Zheng, X., Xiao, D., & Zheng, Y. (2016). Age-associated de-repression of retrotransposons in the Drosophila fat body, its potential cause and consequence. *Aging Cell*, *15*(3), 542–552. https://doi.org/10.1111/acel.12465

Chen, Y., Zhang, Q., Teng, Z., & Liu, H. (2021). Centromeric transcription maintains centromeric cohesion in human cells. *The Journal of Cell Biology*, *220*(7). https://doi.org/10.1083/jcb.202008146

Choi, E. S., Strålfors, A., Castillo, A. G., Durand-Dubief, M., Ekwall, K., & Allshire, R. C. (2011). Identification of Noncoding Transcripts from within CENP-A Chromatin at Fission Yeast Centromeres. *Journal of Biological Chemistry*, *286*(26), 23600–23607. https://doi.org/10.1074/jbc.M111.228510

Christich, A., Kauppila, S., Chen, P., Sogame, N., Ho, S. I., & Abrams, J. M. (2002). The Damage-Responsive Drosophila Gene sickle Encodes a Novel IAP Binding Protein Similar to but Distinct from reaper, grim, and hid. *Current Biology*, *12*(2), 137–140. https://doi.org/10.1016/S0960-

9822(01)00658-3

Chueh, A. C., Northrop, E. L., Brettingham-Moore, K. H., Choo, K. H. A., & Wong, L. H. (2009). LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin. *PLoS Genetics*, *5*(1), e1000354. https://doi.org/10.1371/journal.pgen.1000354

Coffin JM, Hughes SH, & Varmus HE. (1997). Structural Classes of Retroelements and Replication Strategies. In *Retroviruses*. Cold Spring Harbor Laboratory Press. https://www.ncbi.nlm.nih.gov/books/NBK19412/

Corless, S., Höcker, S., & Erhardt, S. (2020). Centromeric RNA and Its Function at and Beyond Centromeric Chromatin. *Journal of Molecular Biology*, *432*(15), 4257–4269. https://doi.org/10.1016/j.jmb.2020.03.027

Corley, M., Burns, M. C., & Yeo, G. W. (2020). How RNA binding proteins interact with RNA: molecules and mechanisms. *Molecular Cell*, *78*(1), 9. https://doi.org/10.1016/J.MOLCEL.2020.03.011

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

De Cecco, M., Criscione, S. W., Peterson, A. L., Neretti, N., Sedivy, J. M., & Kreiling, J. A. (2013). Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. *Aging*, *5*(12), 867–883. https://doi.org/10.18632/aging.100621

DeBose-Scarlett, E. M., & Sullivan, B. A. (2021). Genomic and Epigenetic Foundations of Neocentromere Formation. *Annual Review of Genetics*, *55*(1), 331–348. https://doi.org/10.1146/annurev-genet-071719-020924

Deloger, M., Cavalli, F. M. G., Lerat, E., Biémont, C., Sagot, M. F., & Vieira, C. (2009). Identification of expressed transposable element insertions in the sequenced genome of Drosophila melanogaster. *Gene*, *439*(1–2), 55–62. https://doi.org/10.1016/j.gene.2009.03.015

Deluca, S. Z., & Spradling, A. C. (2018). Efficient expression of genes in the drosophila germline using a uas promoter free of interference by hsp70 pirnas. *Genetics*, *209*(2), 381–387. https://doi.org/10.1534/genetics.118.300874

Dietzl, G., Chen, D., Schnorrer, F., Su, K.-C., Barinova, Y., Fellner, M., Gasser, B., Kinsey, K., Oppel, S., Scheiblauer, S., Couto, A., Marra, V., Keleman, K., & Dickson, B. J. (2007). A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. *Nature*, *448*(7150), 151–156. https://doi.org/10.1038/nature05954

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Domsch, K., Carnesecchi, J., Disela, V., Friedrich, J., Trost, N., Ermakova, O., Polychronidou, M., & Lohmann, I. (2019). The Hox transcription factor Ubx stabilizes lineage commitment by suppressing cellular plasticity in Drosophila. *ELife*, *8*, e42675. https://doi.org/10.7554/eLife.42675.001

Drinnenberg, I. A., Henikoff, S., & Malik, H. S. (2016). Evolutionary Turnover of Kinetochore Proteins: A Ship of Theseus? *Trends in Cell Biology*, *26*(7), 498–510. https://doi.org/10.1016/j.tcb.2016.01.005

Du, Y., Topp, C. N., & Dawe, R. K. (2010). DNA binding of centromere protein C (CENPC) is stabilized by

single-stranded RNA. *PLoS Genetics*, *6*(2), e1000835. https://doi.org/10.1371/journal.pgen.1000835

Dubridge, R. B., Tang, P., Chao Hsia, H., Leong, P.-M., Miller, J. H., & Calos1, M. P. (1987). Analysis of Mutation in Human Cells by Using an Epstein-Barr Virus Shuttle System. *MOLECULAR AND CELLULAR BIOLOGY*, *7*(1), 379–387.

Duff, M. O., Olson, S., Wei, X., Garrett, S. C., Osman, A., Bolisetty, M., Plocik, A., Celniker, S. E., & Graveley, B. R. (2015). Genome-wide identification of zero nucleotide recursive splicing in Drosophila. *Nature*, *521*(7552), 376–379. https://doi.org/10.1038/nature14475

Duffy, J. B. (2002). GAL4 system in Drosophila: A fly geneticist's Swiss army knife. In *Genesis* (Vol. 34, Issues 1–2, pp. 1–15). Wiley-Liss Inc. https://doi.org/10.1002/gene.10150

Dunleavy, E. M., Almouzni, G., & Karpen, G. H. (2011). H3.3 is deposited at centromeres in S phase as a placeholder for newly assembled CENP-A in G1 phase. *Nucleus*, *2*(2), 146–157. https://doi.org/10.4161/nucl.2.2.15211

Dunleavy, E. M., Roche, D., Tagami, H., Lacoste, N., Ray-Gallet, D., Nakamura, Y., Daigo, Y., Nakatani, Y., & Almouzni-Pettinotti, G. (2009). HJURP Is a Cell-Cycle-Dependent Maintenance and Deposition Factor of CENP-A at Centromeres. *Cell*, *137*(3), 485–497. https://doi.org/10.1016/j.cell.2009.02.040

Erhardt, S., Mellone, B. G., Betts, C. M., Zhang, W., Karpen, G. H., & Straight, A. F. (2008). Genome-wide analysis reveals a cell cycle–dependent mechanism controlling centromere propagation. *Journal of Cell Biology*, *183*(5), 805–818. https://doi.org/10.1083/jcb.200806038

Fabry, M. H., Falconio, F. A., Joud, F., Lythgoe, E. K., Czech, B., & Hannon, G. J. (2021). Maternally inherited piRNAs direct transient heterochromatin formation at active transposons during early Drosophila embryogenesis. *ELife*, *10*. https://doi.org/10.7554/eLife.68573

Farrell, J. A., & O'Farrell, P. H. (2014). From Egg to Gastrula: How the Cell Cycle Is Remodeled During the Drosophila Mid-Blastula Transition. *Annual Review of Genetics*, *48*(1), 269–294. https://doi.org/10.1146/annurev-genet-111212-133531

Ferrand, J., Plessier, A., & Polo, S. E. (2021). Control of the chromatin response to DNA damage: Histone proteins pull the strings. *Seminars in Cell & Developmental Biology*, *113*, 75–87. https://doi.org/10.1016/J.SEMCDB.2020.07.002

Ferri, F., Bouzinba-Segard, H., Velasco, G., Hubé, F., & Francastel, C. (2009). Non-coding murine centromeric transcripts associate with and potentiate Aurora B kinase. *Nucleic Acids Research*, *37*(15), 5071–5080. https://doi.org/10.1093/nar/gkp529

Filarsky, M., Zillner, K., Araya, I., Villar-Garea, A., Merkl, R., Längst, G., & Németh, A. (2015). The extended AT-hook is a novel RNA binding motif. *RNA Biology*, *12*(8), 864–876. https://doi.org/10.1080/15476286.2015.1060394/SUPPL_FILE/KRNB_A_1060394_SM1916.ZIP

Fisher, B., Weiszmann, R., Frise, E., Hammonds, A., Tomancak, P., Beaton, A., Berman, B., Quan, E., Shu, S., Lewis, S., Rubin, G., Barale, C., Laguertas, E., Quinn, J., Ghosh, A., Hartenstein, V., Ashburner, M., & Celniker, S. (2012). *BDGP insitu homepage*. http://insitu.fruitfly.org/cgi-bin/ex/insitu.pl

Flockhart, I. T., Booker, M., Hu, Y., McElvany, B., Gilly, Q., Mathey-Prevot, B., Perrimon, N., & Mohr, S. E. (2012). FlyRNAi.org - The database of the Drosophila RNAi screening center: 2012 update. *Nucleic Acids Research*, *40*(D1). https://doi.org/10.1093/nar/gkr953

Foltz, D. R., Jansen, L. E. T., Bailey, A. O., Yates, J. R., Bassett, E. A., Wood, S., Black, B. E., & Cleveland, D. W. (2009). Centromere-Specific Assembly of CENP-A Nucleosomes Is Mediated by HJURP. *Cell*,

137(3), 472–484. https://doi.org/10.1016/j.cell.2009.02.039

Francis, M. J., Roche, S., Cho, M. J., Beall, E., Min, B., Panganiban, R. P., & Rio, D. C. (2016). Drosophila IRBP bZIP heterodimer binds P-element DNA and affects hybrid dysgenesis. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(46), 13003–13008. https://doi.org/10.1073/pnas.1613508113

Frommer, G., Schuh, R., & Jäckle, H. (1994). Localized expression of a novel micropia-like element in the blastoderm of Drosophila melanogaster is dependent on the anterior morphogen bicoid. *Chromosoma*, *103*(2), 82–89. https://doi.org/10.1007/BF00352316

Fujita, Y., Hayashi, T., Kiyomitsu, T., Toyoda, Y., Kokubu, A., Obuse, C., & Yanagida, M. (2007). Priming of Centromere for CENP-A Recruitment by Human hMis18α, hMis18β, and M18BP1. *Developmental Cell*, *12*(1), 17–30. https://doi.org/10.1016/j.devcel.2006.11.002

Fukagawa, T., & Earnshaw, W. C. (2014). The Centromere: Chromatin Foundation for the Kinetochore Machinery. *Developmental Cell*, *30*, 496–508. https://doi.org/10.1016/j.devcel.2014.08.016

Gamba, R., & Fachinetti, D. (2020). From evolution to function: Two sides of the same CENP-B coin? *Experimental Cell Research*, *390*(2), 111959. https://doi.org/10.1016/j.yexcr.2020.111959

Gascoigne, K. E., & Cheeseman, I. M. (2011). Kinetochore assembly: if you build it, they will come. *Current Opinion in Cell Biology*, *23*(1), 102–108. https://doi.org/10.1016/j.ceb.2010.07.007

Gheisari, E., Aakhte, M., & Müller, H. A. J. (2020). Gastrulation in Drosophila melanogaster: Genetic control, cellular basis and biomechanics. *Mechanisms of Development*, *163*, 103629. https://doi.org/10.1016/J.MOD.2020.103629

Gibson, D. G., Young, L., Chuang, R. Y., Venter, J. C., Hutchison, C. A., & Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods 2009 6:5*, *6*(5), 343–345. https://doi.org/10.1038/nmeth.1318

Gorbunova, V., Seluanov, A., Mita, P., McKerrow, W., Fenyö, D., Boeke, J. D., Linker, S. B., Gage, F. H., Kreiling, J. A., Petrashen, A. P., Woodham, T. A., Taylor, J. R., Helfand, S. L., & Sedivy, J. M. (2021). The role of retrotransposable elements in ageing and age-associated diseases. *Nature*, *596*(7870), 43–53. https://doi.org/10.1038/s41586-021-03542-y

Grenfell, A. W., Heald, R., & Strzelecka, M. (2016). Mitotic noncoding RNA processing promotes kinetochore and spindle assembly in Xenopus. *The Journal of Cell Biology*, *214*(2), 133–141. https://doi.org/10.1083/jcb.201604029

Grimaldi, G., & Di Nocera, P. P. (1988). Multiple repeated units in Drosophila melanogaster ribosomal DNA spacer stimulate rRNA precursor transcription. *Proceedings of the National Academy of Sciences*, *85*(15), 5502–5506. https://doi.org/10.1073/pnas.85.15.5502

Groth, A. C., Fish, M., Nusse, R., & Calos, M. P. (2004). Construction of transgenic Drosophila by using the site-specific integrase from phage phiC31. *Genetics*, *166*(4), 1775. https://doi.org/10.1534/GENETICS.166.4.1775

Guzzardo, P. M., Muerdter, F., & Hannon, G. J. (2013). The piRNA pathway in flies: highlights and future directions. *Current Opinion in Genetics & Development*, *23*(1), 44–52. https://doi.org/10.1016/j.gde.2012.12.003

Hall, L. L., Byron, M., Carone, D. M., Whitfield, T. W., Pouliot, G. P., Fischer, A., Jones, P., & Lawrence, J. B. (2017). Demethylated HSATII DNA and HSATII RNA Foci Sequester PRC1 and MeCP2 into Cancer-Specific Nuclear Bodies. *Cell Reports*, *18*(12), 2943–2956. https://doi.org/10.1016/j.celrep.2017.02.072

Hancks, D. C., & Kazazian, H. H. (2016). Roles for retrotransposon insertions in human disease. *Mobile DNA*, *7*(1), 9. https://doi.org/10.1186/s13100-016-0065-9

Hartenstein, V. (1993). *Atlas of Drosophila development*. Cold Spring Harbor Laboratory Press.

Hartley, G., & O'Neill, R. (2019). Centromere Repeats: Hidden Gems of the Genome. *Genes*, *10*(3), 223. https://doi.org/10.3390/genes10030223

Hassold, T., & Hunt, P. (2001). To err (meiotically) is human: the genesis of human aneuploidy. *Nature Reviews Genetics*, *2*(4), 280–291. https://doi.org/10.1038/35066065

Hédouin, S., Grillo, G., Ivkovic, I., Velasco, G., & Francastel, C. (2017). CENP-A chromatin disassembly in stressed and senescent murine cells. *Scientific Reports*, *7*(1), 42520. https://doi.org/10.1038/srep42520

Hedouin, S., Logsdon, G. A., Underwood, J. G., & Biggins, S. (2022). A transcriptional roadblock protects yeast centromeres. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkac117

Heeger, S., Leismann, O., Schittenhelm, R., Schraidt, O., Heidmann, S., & Lehner, C. F. (2005). Genetic interactions of separase regulatory subunits reveal the diverged Drosophila Cenp-C homolog. *Genes & Development*, *19*(17), 2041–2053. https://doi.org/10.1101/GAD.347805

Hengeveld, R. C. C., Vromans, M. J. M., Vleugel, M., Hadders, M. A., & Lens, S. M. A. (2017). Inner centromere localization of the CPC maintains centromere cohesion and allows mitotic checkpoint silencing. *Nature Communications*, *8*(1), 15542. https://doi.org/10.1038/ncomms15542

Heun, P., Erhardt, S., Blower, M. D., Weiss, S., Skora, A. D., & Karpen, G. H. (2006). Mislocalization of the Drosophila Centromere-Specific Histone CID Promotes Formation of Functional Ectopic Kinetochores. *Developmental Cell*, *10*(3), 303–315. https://doi.org/10.1016/j.devcel.2006.01.014

Ho, J. S. Y., Di Tullio, F., Schwarz, M., Low, D., Incarnato, D., Gay, F., Tabaglio, T., Zhang, J., Wollman, H., Chen, L., An, O., Chan, T. H. M., Hickman, A. H., Zheng, S., Roudko, V., Chen, S., Karz, A., Ahmed, M., He, H. H., … Guccione, E. (2021). Hnrnpm controls circrna biogenesis and splicing fidelity to sustain cancer cell fitness. *ELife*, *10*. https://doi.org/10.7554/ELIFE.59654

Hoffman, E. A., Frey, B. L., Smith, L. M., & Auble, D. T. (2015). Formaldehyde Crosslinking: A Tool for the Study of Chromatin Complexes. *Journal of Biological Chemistry*, *290*(44), 26404–26411. https://doi.org/10.1074/jbc.R115.651679

Howman, E. V., Fowler, K. J., Newson, A. J., Redward, S., MacDonald, A. C., Kalitsis, P., & Choo, K. H. A. (2000). Early disruption of centromeric chromatin organization in centromere protein A (CENP-A) null mice. *Proceedings of the National Academy of Sciences*, *97*(3), 1148–1153. https://doi.org/10.1073/pnas.97.3.1148

Hoyt, S. J., Storer, J. M., Hartley, G. A., Grady, P. G. S., Gershman, A., de Lima, L. G., Limouse, C., Halabian, R., Wojenski, L., Rodriguez, M., Altemose, N., Rhie, A., Core, L. J., Gerton, J. L., Makalowski, W., Olson, D., Rosen, J., Smit, A. F. A., Straight, A. F., … O'Neill, R. J. (2022). From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science*, *376*(6588). https://doi.org/10.1126/science.abk3112

Hsieh, T.-S., & Brutlag, D. (1979). Sequence and sequence variation within the 1.688 g/cm3 satellite DNA of Drosophila melanogaster. *Journal of Molecular Biology*, *135*(2), 465–481. https://doi.org/10.1016/0022-2836(79)90447-9

Ichida, K., Suzuki, K., Fukui, T., Takayama, Y., Kakizawa, N., Watanabe, F., Ishikawa, H., Muto, Y., Kato, T., Saito, M., Futsuhara, K., Miyakura, Y., Noda, H., Ohmori, T., Konishi, F., & Rikiyama, T. (2018). Overexpression of satellite alpha transcripts leads to chromosomal instability via segregation

errors at specific chromosomes. *International Journal of Oncology*, *52*(5), 1685–1693. https://doi.org/10.3892/ijo.2018.4321

Ishikura, S., Nakabayashi, K., Nagai, M., Tsunoda, T., & Shirasawa, S. (2020). ZFAT binds to centromeres to control noncoding RNA transcription through the KAT2B–H4K8ac–BRD4 axis. *Nucleic Acids Research*, *48*(19), 10848–10866. https://doi.org/10.1093/nar/gkaa815

Ismail, I. H., & Hendzel, M. J. (2008). The γ-H2A.X: Is it just a surrogate marker of double-strand breaks or much more? *Environmental and Molecular Mutagenesis*, *49*(1), 73–82. https://doi.org/10.1002/EM.20358

Ito, K., Awano, W., Suzuki, K., Hiromi, Y., & Yamamoto, D. (1997). The Drosophila mushroom body is a quadruple structure of clonal units each of which contains a virtually identical set of neurones and glial cells. *Development*, *124*(4), 761–771. https://doi.org/10.1242/DEV.124.4.761

Iwata-Otsubo, A., Dawicki-McKenna, J. M., Akera, T., Falk, S. J., Chmátal, L., Yang, K., Sullivan, B. A., Schultz, R. M., Lampson, M. A., & Black, B. E. (2017). Expanded Satellite Repeats Amplify a Discrete CENP-A Nucleosome Assembly Site on Chromosomes that Drive in Female Meiosis. *Current Biology*, *27*(15), 2365-2373.e8. https://doi.org/10.1016/j.cub.2017.06.069

Jäger, H., Rauch, M., & Heidmann, S. (2005). The Drosophila melanogaster condensin subunit Cap-G interacts with the centromere-specific histone H3 variant CID. *Chromosoma*, *113*(7), 350–361. https://doi.org/10.1007/s00412-004-0322-4

Jain, M., Olsen, H. E., Turner, D. J., Stoddart, D., Bulazel, K. V, Paten, B., Haussler, D., Willard, H. F., Akeson, M., & Miga, K. H. (2018). Linear assembly of a human centromere on the Y chromosome. *Nature Biotechnology*, *36*(4), 321–323. https://doi.org/10.1038/nbt.4109

Jansen, L. E. T., Black, B. E., Foltz, D. R., & Cleveland, D. W. (2007). Propagation of centromeric chromatin requires exit from mitosis. *Journal of Cell Biology*, *176*(6), 795–805. https://doi.org/10.1083/jcb.200701066

Jin, Y., Tam, O. H., Paniagua, E., & Hammell, M. (2015). TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics*, *31*(22), 3593–3599. https://doi.org/10.1093/bioinformatics/btv422

Jolly, C., Metz, A., Govin, J., Vigneron, M., Turner, B. M., Khochbin, S., & Vourc'h, C. (2004). Stress-induced transcription of satellite III repeats. *Journal of Cell Biology*, *164*(1), 25–33. https://doi.org/10.1083/jcb.200306104

Kaessmann, H., Vinckenbosch, N., & Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. *Nature Reviews. Genetics*, *10*(1), 19–31. https://doi.org/10.1038/nrg2487

Kaminker, J. S., Bergman, C. M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D. A., Lewis, S. E., Rubin, G. M., Ashburner, M., & Celniker, S. E. (2002). The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biology*, *3*(12), RESEARCH0084. https://doi.org/10.1186/GB-2002-3-12-RESEARCH0084

Kanne, J., Hussong, M., Isensee, J., Muñoz-López, Á., Wolffgramm, J., Heß, F., Grimm, C., Bessonov, S., Meder, L., Wang, J., Reinhardt, H. C., Odenthal, M., Hucho, T., Büttner, R., Summerer, D., & Schweiger, M. R. (2021). Pericentromeric Satellite III transcripts induce etoposide resistance. *Cell Death & Disease*, *12*(6), 530. https://doi.org/10.1038/s41419-021-03810-9

Kazazian, H. H. (2004). Mobile Elements: Drivers of Genome Evolution. *Science*, *303*(5664), 1626–1632. https://doi.org/10.1126/SCIENCE.1089670/ASSET/B44E3268-2216-42F8-8BFB-A56A4D8EDB7B/ASSETS/GRAPHIC/ZSE0100423600003.JPEG

Khost, D. E., Eickbush, D. G., & Larracuente, A. M. (2017). Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in Drosophila melanogaster. *Genome Research*, *27*(5), 709–721. https://doi.org/10.1101/gr.213512.116

King, M. R., Matzat, L. H., Dale, R. K., Lim, J. J., & Lei, E. P. (2014). The RNA-binding protein Rumpelstiltskin antagonizes gypsy chromatin insulator function in a tissue-specific manner. *Journal of Cell Science*, *127*(13), 2956–2966. https://doi.org/10.1242/jcs.151126

Klein, S. J., & O'Neill, R. J. (2018). Transposable elements: genome innovation, chromosome diversity, and centromere conflict. *Chromosome Research*, *26*(1–2), 5–23. https://doi.org/10.1007/s10577-017-9569-5

Kuhn, G. C. S., Küttler, H., Moreira-Filho, O., & Heslop-Harrison, J. S. (2012). The 1.688 Repetitive DNA of Drosophila: Concerted Evolution at Different Genomic Scales and Association with Genes. *Molecular Biology and Evolution*, *29*(1), 7–11. https://doi.org/10.1093/molbev/msr173

Landers, C. C., Rabeler, C. A., Ferrari, E. K., D'Alessandro, L. R., Kang, D. D., Malisa, J., Bashir, S. M., & Carone, D. M. (2021). Ectopic expression of pericentric HSATII RNA results in nuclear RNA accumulation, MeCP2 recruitment, and cell division defects. *Chromosoma*, *130*(1), 75–90. https://doi.org/10.1007/s00412-021-00753-0

Larracuente, A. M. (2014). The organization and evolution of the Responder satellite in species of the Drosophila melanogaster group: dynamic evolution of a target of meiotic drive. *BMC Evolutionary Biology*, *14*(1), 233. https://doi.org/10.1186/s12862-014-0233-9

Le Thomas, A., Rogers, A. K., Webster, A., Marinov, G. K., Liao, S. E., Perkins, E. M., Hur, J. K., Aravin, A. A., & Tóth, K. F. (2013). Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes & Development*, *27*(4), 390–399. https://doi.org/10.1101/gad.209841.112

Lee, F. C. Y., & Ule, J. (2018). Advances in CLIP Technologies for Studies of Protein-RNA Interactions. *Molecular Cell*, *69*(3), 354–369. https://doi.org/10.1016/j.molcel.2018.01.005

Leppek, K., & Stoecklin, G. (2014). An optimized streptavidin-binding RNA aptamer for purification of ribonucleoprotein complexes identifies novel ARE-binding proteins. *Nucleic Acids Research*, *42*(2), e13–e13. https://doi.org/10.1093/nar/gkt956

Lerat, E., Burlet, N., Biémont, C., & Vieira, C. (2011). Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. *Gene*, *473*(2), 100–109. https://doi.org/10.1016/j.gene.2010.11.009

Ling, Y. H., & Yuen, K. W. Y. (2019). Point centromere activity requires an optimal level of centromeric noncoding RNA. *Proceedings of the National Academy of Sciences*, *116*(13), 6270–6279. https://doi.org/10.1073/pnas.1821384116

Liu, H., Qu, Q., Warrington, R., Rice, A., Cheng, N., & Yu, H. (2015). Mitotic Transcription Installs Sgo1 at Centromeres to Coordinate Chromosome Segregation. *Molecular Cell*, *59*(3), 426–436. https://doi.org/10.1016/j.molcel.2015.06.018

Liu, Y., Su, H., Zhang, J., Liu, Y., Feng, C., & Han, F. (2020). Back-spliced RNA from retrotransposon binds to centromere and regulates centromeric chromatin loops in maize. *PLOS Biology*, *18*(1), e3000582. https://doi.org/10.1371/journal.pbio.3000582

Liu, Y., Zhou, J., & White, K. P. (2014). RNA-seq differential expression studies: More sequence or more replication? *Bioinformatics*, *30*(3), 301–304. https://doi.org/10.1093/bioinformatics/btt688

Logsdon, G. A., Gambogi, C. W., Liskovykh, M. A., Barrey, E. J., Larionov, V., Miga, K. H., Heun, P., &

Black, B. E. (2019). Human Artificial Chromosomes that Bypass Centromeric DNA. *Cell*, *178*(3), 624-639.e19. https://doi.org/10.1016/j.cell.2019.06.006

Lohe, A. R., & Brutlag, D. L. (1986). Multiplicity of satellite DNA sequences in Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, *83*(3), 696–700. https://doi.org/10.1073/pnas.83.3.696

Lohe, A. R., Hilliker, A. J., & Roberts, P. A. (1993). Mapping simple repeated DNA sequences in heterochromatin of Drosophila melanogaster. *Trends in Genetics*, *9*(11), 379. https://doi.org/10.1016/0168-9525(93)90135-5

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., & Kroemer, G. (2013). The hallmarks of aging. *Cell*, *153*(6), 1194–1217. https://doi.org/10.1016/j.cell.2013.05.039

Losada, A., & Villasante, A. (1996). Autosomal location of a new subtype of 1.688 satellite DNA of Drosophila melanogaster. *Chromosome Research*, *4*(5), 372–383. https://doi.org/10.1007/BF02257273

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12). https://doi.org/10.1186/s13059-014-0550-8

Mallm, J.-P., & Rippe, K. (2015). Aurora Kinase B Regulates Telomerase Activity via a Centromeric RNA in Stem Cells. *Cell Reports*, *11*(10), 1667–1678. https://doi.org/10.1016/j.celrep.2015.05.015

Manders, E. M. M., Verbeek, F. J., & Aten, J. A. (1993). Measurement of co-localization of objects in dual-colour confocal images. *Journal of Microscopy*, *169*, 375–382.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

Mazzalupo, S., & Cooley, L. (2006). Illuminating the role of caspases during Drosophila oogenesis. *Cell Death & Differentiation 2006 13:11*, *13*(11), 1950–1959. https://doi.org/10.1038/sj.cdd.4401892

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, *36*(6), 344–355. https://doi.org/10.1073/pnas.36.6.344

McCullers, T. J., & Steiniger, M. (2017). Transposable elements in Drosophila. *Mobile Genetic Elements*, *7*(3), 1–18. https://doi.org/10.1080/2159256X.2017.1318201

McNulty, S. M., Sullivan, L. L., & Sullivan, B. A. (2017). Human Centromeres Produce Chromosome-Specific and Array-Specific Alpha Satellite Transcripts that Are Complexed with CENP-A and CENP-C. *Developmental Cell*, *42*(3), 226-240.e6. https://doi.org/10.1016/j.devcel.2017.07.001

Medina-Pritchard, B., Lazou, V., Zou, J., Byron, O., Abad, M. A., Rappsilber, J., Heun, P., & Jeyaprakash, A. A. (2020). Structural basis for centromere maintenance by Drosophila CENP-A chaperone CAL1. *The EMBO Journal*, *39*(7). https://doi.org/10.15252/embj.2019103234

Mellone, B. G., Grive, K. J., Shteyn, V., Bowers, S. R., Oderberg, I., & Karpen, G. H. (2011). Assembly of Drosophila Centromeric Chromatin Proteins during Mitosis. *PLoS Genetics*, *7*(5), e1002068. https://doi.org/10.1371/journal.pgen.1002068

Mérel, V., Boulesteix, M., Fablet, M., & Vieira, C. (2020). Transposable elements in Drosophila. *Mobile DNA*, *11*(1), 23. https://doi.org/10.1186/s13100-020-00213-z

Michalek, J., Capek, M., & Janacek, J. (n.d.). *Stack Contrast Adjustment plugin*. https://imagej.nih.gov/ij/plugins/stack-contrast/index.htm

Molina, O., Vargiu, G., Abad, M. A., Zhiteneva, A., Jeyaprakash, A. A., Masumoto, H., Kouprina, N.,

Larionov, V., & Earnshaw, W. C. (2016). Epigenetic engineering reveals a balance between histone modifications and transcription in kinetochore maintenance. *Nature Communications*, *7*(1), 13334. https://doi.org/10.1038/ncomms13334

Moore, C. B., Guthrie, E. H., Huang, M. T.-H., & Taxman, D. J. (2010). Short Hairpin RNA (shRNA): Design, Delivery, and Assessment of Gene Knockdown. *Methods Mol Biol.*, *629*, 141–158. https://doi.org/10.1007/978-1-60761-657-3_10

Moran, E. C., Liu, L., Zasadzinska, E., Kestner, C. A., Sarkeshik, A., DeHoyos, H., Yates, J. R., Foltz, D., & Stukenberg, P. T. (2021). Mitotic R-loops direct Aurora B kinase to maintain centromeric cohesion. *BioRxiv*, 2021.01.14.426738. https://doi.org/https://doi.org/10.1101/2021.01.14.426738

Moree, B., Meyer, C. B., Fuller, C. J., & Straight, A. F. (2011). CENP-C recruits M18BP1 to centromeres to promote CENP-A chromatin assembly. *Journal of Cell Biology*, *194*(6), 855–871. https://doi.org/10.1083/jcb.201106079

Mount, S. M., Rubin, G. M., Mossie, K. G., Varmus, H. E., & Mol Biol, J. (1985). Complete nucleotide sequence of the Drosophila transposable element copia: homology between copia and retroviral proteins. *Molecular and Cellular Biology*, *5*(7), 1630–1638. https://doi.org/10.1128/MCB.5.7.1630-1638.1985

Mteirek, R., Gueguen, N., Jensen, S., Brasset, E., & Vaury, C. (2014). Drosophila heterochromatin: structure and function. *Current Opinion in Insect Science*, *1*, 19–24. https://doi.org/10.1016/j.cois.2014.04.003

Murillo-Pineda, M., Valente, L. P., Dumont, M., Mata, J. F., Fachinetti, D., & Jansen, L. E. T. (2021). Induction of spontaneous human neocentromere formation and long-term maturation. *Journal of Cell Biology*, *220*(3). https://doi.org/10.1083/JCB.202007210/211684

Musacchio, A., & Desai, A. (2017). A Molecular View of Kinetochore Assembly and Function. *Biology*, *6*(4), 5. https://doi.org/10.3390/biology6010005

Musacchio, A., & Salmon, E. D. (2007). The spindle-assembly checkpoint in space and time. *Nature Reviews Molecular Cell Biology*, *8*(5), 379–393. https://doi.org/10.1038/nrm2163

Nakano, M., Cardinale, S., Noskov, V. N., Gassmann, R., Vagnarelli, P., Kandels-Lewis, S., Larionov, V., Earnshaw, W. C., & Masumoto, H. (2008). Inactivation of a Human Kinetochore by Specific Targeting of Chromatin Modifiers. *Developmental Cell*, *14*(4), 507–522. https://doi.org/10.1016/j.devcel.2008.02.001

Naughton, C., Huidobro, C., Catacchio, C. R., Buckle, A., Grimes, G. R., Nozawa, R.-S., Purgato, S., Rocchi, M., & Gilbert, N. (2021). Human centromere formation activates transcription and opens chromatin fibre structure. *BioRxiv*, 2021.08.01.454615. https://doi.org/10.1101/2021.08.01.454615

Ni, J. Q., Markstein, M., Binari, R., Pfeiffer, B., Liu, L. P., Villalta, C., Booker, M., Perkins, L., & Perrimon, N. (2008). Vector and parameters for targeted transgenic RNA interference in Drosophila melanogaster. *Nature Methods*, *5*(1), 49–51. https://doi.org/10.1038/nmeth1146

Nicklas, R. B. (1997). How Cells Get the Right Chromosomes. *Science*, *275*(5300), 632–637. https://doi.org/10.1126/science.275.5300.632

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53.

https://doi.org/10.1126/science.abj6987

Ogienko, A. A., Andreyeva, E. N., Omelina, E. S., Oshchepkova, A. L., & Pindyurin, A. V. (2020). Molecular and cytological analysis of widely-used Gal4 driver lines for Drosophila neurobiology. *BMC Genetics*, *21*. https://doi.org/10.1186/s12863-020-00895-7

Ohkuni, K., & Kitagawa, K. (2011). Endogenous transcription at the centromere facilitates centromere activity in budding yeast. *Current Biology*, *21*(20), 1695–1703. https://doi.org/10.1016/j.cub.2011.08.056

Ohzeki, J., Larionov, V., Earnshaw, W. C., & Masumoto, H. (2019). De novo formation and epigenetic maintenance of centromere chromatin. *Current Opinion in Cell Biology*, *58*, 15–25. https://doi.org/10.1016/J.CEB.2018.12.004

Okamoto, Y., Nakano, M., Ohzeki, J., Larionov, V., & Masumoto, H. (2007). A minimal CENP-A core is required for nucleation and maintenance of a functional human centromere. *The EMBO Journal*, *26*(5), 1279–1291. https://doi.org/10.1038/sj.emboj.7601584

Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D., & Zamore, P. D. (2019). PIWI-interacting RNAs: small RNAs with big functions. *Nature Reviews Genetics*, *20*(2), 89–108. https://doi.org/10.1038/s41576-018-0073-3

Padeken, J., Mendiburo, M. J., Chlamydas, S., Schwarz, H. J., Kremmer, E., & Heun, P. (2013). The Nucleoplasmin Homolog NLP Mediates Centromere Clustering and Anchoring to the Nucleolus. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2013.03.002

Palozola, K. C., Donahue, G., Liu, H., Grant, G. R., Becker, J. S., Cote, A., Yu, H., Raj, A., & Zaret, K. S. (2017). Mitotic transcription and waves of gene reactivation during mitotic exit. *Science*, *358*(6359), 119–122. https://doi.org/10.1126/SCIENCE.AAL4671/SUPPL_FILE/AAL4671_SUPPLEMENTAL_TABLES.ZIP

Pauleau, A.-L., Bergner, A., Kajtez, J., & Erhardt, S. (2019). The checkpoint protein Zw10 connects CAL1-dependent CENP-A centromeric loading and mitosis duration in Drosophila cells. *PLOS Genetics*, *15*(9), e1008380. https://doi.org/10.1371/journal.pgen.1008380

Perkins, L. A., Holderbaum, L., Tao, R., Hu, Y., Sopko, R., McCall, K., Yang-Zhou, D., Flockhart, I., Binari, R., Shim, H. S., Miller, A., Housden, A., Foos, M., Randkelv, S., Kelley, C., Namgyal, P., Villalta, C., Liu, L. P., Jiang, X., … Perrimon, N. (2015). The transgenic RNAi project at Harvard medical school: Resources and validation. *Genetics*, *201*(3), 843–852. https://doi.org/10.1534/genetics.115.180208

Peters, J.-M. (2006). The anaphase promoting complex/cyclosome: a machine designed to destroy. *Nature Reviews Molecular Cell Biology*, *7*(9), 644–656. https://doi.org/10.1038/nrm1988

Phansalkar, R., Lapierre, P., & Mellone, B. G. (2012). Evolutionary insights into the role of the essential centromere protein CAL1 in Drosophila. *Chromosome Research*, *20*(5), 493–504. https://doi.org/10.1007/s10577-012-9299-7

Przewloka, M. R., Venkei, Z., Bolanos-Garcia, V. M., Debski, J., Dadlez, M., & Glover, D. M. (2011). CENP-C Is a Structural Platform for Kinetochore Assembly. *Current Biology*, *21*(5), 399–405. https://doi.org/10.1016/j.cub.2011.02.005

Quénet, D., & Dalal, Y. (2014). A long non-coding RNA is required for targeting centromeric protein A to the human centromere. *ELife*, *3*, e03254. https://doi.org/10.7554/eLife.03254

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Racca, C., Britton, S., Hédouin, S., Francastel, C., Calsou, P., & Larminat, F. (2021). BRCA1 prevents R-loop-associated centromeric instability. *Cell Death & Disease*, *12*(10), 896. https://doi.org/10.1038/s41419-021-04189-3

Ramanathan, M., Porter, D. F., & Khavari, P. A. (2019). Methods to study RNA–protein interactions. *Nature Methods*, *16*(3), 225–234. https://doi.org/10.1038/s41592-019-0330-1

Ravel-Godreuil, C., Znaidi, R., Bonnifet, T., Joshi, R. L., & Fuchs, J. (2021). Transposable elements as new players in neurodegenerative diseases. *FEBS Letters*, *595*(22), 2733–2755. https://doi.org/10.1002/1873-3468.14205

Reilly, M. T., Faulkner, G. J., Dubnau, J., Ponomarev, I., & Gage, F. H. (2013). The Role of Transposable Elements in Health and Diseases of the Central Nervous System. *Journal of Neuroscience*, *33*(45), 17577–17586. https://doi.org/10.1523/JNEUROSCI.3369-13.2013

Rizzi, N., Denegri, M., Chiodi, I., Corioni, M., Valgardsdottir, R., Cobianchi, F., Riva, S., & Biamonti, G. (2004). Transcriptional activation of a constitutive heterochromatic domain of the human genome in response to heat shock. *Molecular Biology of the Cell*, *15*(2), 543–551. https://doi.org/10.1091/mbc.e03-07-0487

Robinow, S., & White, K. (1991). Characterization and spatial distribution of the ELAV protein during Drosophila melanogaster development. *Journal of Neurobiology*, *22*(5), 443–461. https://doi.org/10.1002/NEU.480220503

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. In *Nature Biotechnology* (Vol. 29, Issue 1, pp. 24–26). https://doi.org/10.1038/nbt.1754

Rodríguez-Martín, C., Cidre, F., Fernández-Teijeiro, A., Gómez-Mariano, G., de la Vega, L., Ramos, P., Zaballos, Á., Monzón, S., & Alonso, J. (2016). Familial retinoblastoma due to intronic LINE-1 insertion causes aberrant and noncanonical mRNA splicing of the RB1 gene. *Journal of Human Genetics*, *61*(5), 463–466. https://doi.org/10.1038/jhg.2015.173

Rošić, S., Köhler, F., & Erhardt, S. (2014). Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. *Journal of Cell Biology*, *207*(3), 335–349. https://doi.org/10.1083/jcb.201404097

Rosin, L. F., & Mellone, B. G. (2017). Centromeres Drive a Hard Bargain. *Trends in Genetics*, *33*(2), 101–117. https://doi.org/10.1016/j.tig.2016.12.001

Roure, V., Medina-Pritchard, B., Lazou, V., Rago, L., Anselm, E., Venegas, D., Jeyaprakash, A. A., & Heun, P. (2019). Reconstituting Drosophila Centromere Identity in Human Cells. *Cell Reports*, *29*(2), 464-479.e5. https://doi.org/10.1016/j.celrep.2019.08.067

RStudio Team. (2020). *RStudio: Integrated Development for R.* http://www.rstudio.com/

Russo, J., Harrington, A. W., & Steiniger, M. (2016). Antisense transcription of retrotransposons in Drosophila: An origin of endogenous small interfering RNA precursors. *Genetics*, *202*(1), 107–121. https://doi.org/10.1534/GENETICS.115.177196/-/DC1

Sandmann, T., Jakobsen, J. S., & Furlong, E. E. M. (2007). ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in Drosophila melanogaster embryos. *Nature Protocols*, *1*(6), 2839–2855. https://doi.org/10.1038/nprot.2006.383

Schindelin, J., Arganda-carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., & Cardona, A. (2019). *Fiji : an open-source platform for biological-image analysis*. *9*(7).

https://doi.org/10.1038/nmeth.2019

Schittenhelm, R. B., Althoff, F., Heidmann, S., & Lehner, C. F. (2010). Detrimental incorporation of excess Cenp-A/Cid and Cenp-C into Drosophila centromeres is prevented by limiting amounts of the bridging factor Cal1. *Journal of Cell Science*, *123*(21), 3768–3779. https://doi.org/10.1242/jcs.067934

Schuh, M., Lehner, C. F., & Heidmann, S. (2007). Incorporation of Drosophila CID/CENP-A and CENP-C into Centromeres during Early Embryonic Anaphase. *Current Biology*, *17*(3), 237–243. https://doi.org/10.1016/j.cub.2006.11.051

Seller, C. A., Cho, C.-Y., & O'Farrell, P. H. (2019). Rapid embryonic cell cycles defer the establishment of heterochromatin by Eggless/SetDB1 in Drosophila. *Genes & Development*, *33*(7–8), 403–417. https://doi.org/10.1101/gad.321646.118

Shang, W.-H., Hori, T., Martins, N. M. C., Toyoda, A., Misu, S., Monma, N., Hiratani, I., Maeshima, K., Ikeo, K., Fujiyama, A., Kimura, H., Earnshaw, W. C., & Fukagawa, T. (2013). Chromosome Engineering Allows the Efficient Isolation of Vertebrate Neocentromeres. *Developmental Cell*, *24*(6), 635–648. https://doi.org/10.1016/j.devcel.2013.02.009

Shatskikh, A. S., Kotov, A. A., Adashev, V. E., Bazylev, S. S., & Olenina, L. V. (2020). Functional Significance of Satellite DNAs: Insights From Drosophila. *Frontiers in Cell and Developmental Biology*, *8*(May), 1–19. https://doi.org/10.3389/fcell.2020.00312

Shukla, M., Tong, P., White, S. A., Singh, P. P., Reid, A. M., Catania, S., Pidoux, A. L., & Allshire, R. C. (2018). Centromere DNA Destabilizes H3 Nucleosomes to Promote CENP-A Deposition during the Cell Cycle. *Current Biology*, *28*(24), 3924-3936.e4. https://doi.org/10.1016/j.cub.2018.10.049

Sienski, G., Batki, J., Senti, K.-A., Dönertas, D., Tirian, L., Meixner, K., & Brennecke, J. (2015). Silencio/CG9754 connects the Piwi-piRNA complex to the cellular heterochromatin machinery. *Genes & Development*, *29*(21), 2258–2271. https://doi.org/10.1101/gad.271908.115

Silva, M. C. C., Bodor, D. L., Stellfox, M. E., Martins, N. M. C., Hochegger, H., Foltz, D. R., & Jansen, L. E. T. (2012). Cdk Activity Couples Epigenetic Centromere Inheritance to Cell Cycle Progression. *Developmental Cell*, *22*(1), 52–63. https://doi.org/10.1016/j.devcel.2011.10.014

Singh, P. P., Shukla, M., White, S. A., Lafos, M., Tong, P., Auchynnikava, T., Spanos, C., Rappsilber, J., Pidoux, A. L., & Allshire, R. C. (2020). Hap2-Ino80-facilitated transcription promotes de novo establishment of CENP-A chromatin. *Genes & Development*, *34*(3–4), 226–238. https://doi.org/10.1101/gad.332536.119

Slotkin, R. K., & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, *8*(4), 272–285. https://doi.org/10.1038/nrg2072

SRA toolkit development team. (n.d.). *SRA toolkit*. Retrieved February 15, 2022, from https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

Stankovic, A., Guo, L. Y., Mata, J. F., Bodor, D. L., Cao, X.-J., Bailey, A. O., Shabanowitz, J., Hunt, D. F., Garcia, B. A., Black, B. E., & Jansen, L. E. T. (2017). A Dual Inhibitory Mechanism Sufficient to Maintain Cell-Cycle-Restricted CENP-A Assembly. *Molecular Cell*, *65*(2), 231–246. https://doi.org/10.1016/j.molcel.2016.11.021

Sudakin, V., Chan, G. K. T., & Yen, T. J. (2001). Checkpoint inhibition of the APC/C in HeLa cells is mediated by a complex of BUBR1, BUB3, CDC20, and MAD2. *Journal of Cell Biology*, *154*(5), 925–936. https://doi.org/10.1083/jcb.200102093

Sullivan, B. A., & Karpen, G. H. (2004). Centromeric chromatin exhibits a histone modification pattern

that is distinct from both euchromatin and heterochromatin. *Nature Structural & Molecular Biology*, *11*(11), 1076–1083. https://doi.org/10.1038/nsmb845

Sun, B. K., Deaton, A. M., & Lee, J. T. (2006). A transient heterochromatic state in Xist preempts X inactivation choice without RNA stabilization. *Molecular Cell*, *21*(5), 617–628. https://doi.org/10.1016/j.molcel.2006.01.028

Talbert, P. B., & Henikoff, S. (2018). Transcribing Centromeres: Noncoding RNAs and Kinetochore Assembly. *Trends in Genetics*, *34*(8), 587–599. https://doi.org/10.1016/j.tig.2018.05.001

Ting, D. T., Lipson, D., Paul, S., Brannigan, B. W., Akhavanfard, S., Coffman, E. J., Contino, G., Deshpande, V., Iafrate, A. J., Letovsky, S., Rivera, M. N., Bardeesy, N., Maheswaran, S., & Haber, D. A. (2011). Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science*, *331*(6017), 593–596. https://doi.org/10.1126/SCIENCE.1200801/SUPPL_FILE/TING.SOM.PDF

Topp, C. N., Zhong, C. X., & Dawe, R. K. (2004). Centromere-encoded RNAs are integral components of the maize kinetochore. *Proceedings of the National Academy of Sciences*, *101*(45), 15986–15991. https://doi.org/10.1073/pnas.0407154101

Treiber, C. D., & Waddell, S. (2020). Transposon expression in the Drosophila brain is driven by neighboring genes and diversifies the neural transcriptome. *Genome Research*, *30*(11), 1559–1569. https://doi.org/10.1101/gr.259200.119

Trendel, J., Schwarzl, T., Horos, R., Prakash, A., Bateman, A., Hentze, M. W., & Krijgsveld, J. (2019). The Human RNA-Binding Proteome and Its Dynamics during Translational Arrest. *Cell*, *176*(1–2), 391-403.e19. https://doi.org/10.1016/J.CELL.2018.11.004

Uhl, M., Houwaart, T., Corrado, G., Wright, P. R., & Backofen, R. (2017). Computational analysis of CLIP-seq data. In *Methods* (Vols. 118–119, pp. 60–72). Academic Press Inc. https://doi.org/10.1016/j.ymeth.2017.02.006

Ule, J., Jensen, K., Mele, A., & Darnell, R. B. (2005). CLIP: A method for identifying protein–RNA interaction sites in living cells. *Methods*, *37*(4), 376–386. https://doi.org/10.1016/j.ymeth.2005.07.018

Underwood, E. M., Caulton, J. H., Allis, C. D., & Mahowald, A. P. (1980). Developmental fate of pole cells in Drosophila melanogaster. *Developmental Biology*, *77*(2), 303–314. https://doi.org/10.1016/0012-1606(80)90476-5

Usakin, L., Abad, J., Vagin, V. V, de Pablos, B., Villasante, A., & Gvozdev, V. A. (2007). Transcription of the 1.688 Satellite DNA Family Is Under the Control of RNA Interference Machinery in *Drosophila melanogaster* Ovaries. *Genetics*, *176*(2), 1343–1349. https://doi.org/10.1534/genetics.107.071720

Valgardsdottir, R., Chiodi, I., Giordano, M., Cobianchi, F., Riva, S., & Biamonti, G. (2005). Structural and Functional Characterization of Noncoding Repetitive RNAs Transcribed in Stressed Human Cells. *Molecular Biology of the Cell*, *16*(6), 2597–2604. https://doi.org/10.1091/mbc.e04-12-1078

Valgardsdottir, R., Chiodi, I., Giordano, M., Rossi, A., Bazzini, S., Ghigna, C., Riva, S., & Biamonti, G. (2008). Transcription of Satellite III non-coding RNAs is a general stress response in human cells. *Nucleic Acids Research*, *36*(2), 423–434. https://doi.org/10.1093/nar/gkm1056

Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M., & Yeo, G. W. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, *13*(6), 508–514. https://doi.org/10.1038/nmeth.3810

Vert, J. P., Foveau, N., Lajaunie, C., & Vandenbrouck, Y. (2006). An accurate and interpretable model for siRNA efficacy prediction. *BMC Bioinformatics*, *7*. https://doi.org/10.1186/1471-2105-7-520

Vincent, J.-P., Girdham, C. H., & O'Farrell, P. H. (1994). A Cell-Autonomous, Ubiquitous Marker for the Analysis of Drosophila Genetic Mosaics. *Developmental Biology*, *164*(1), 328–331.

Walther, M., Schrahn, S., Krauss, V., Lein, S., Kessler, J., Jenuwein, T., & Reuter, G. (2020). Heterochromatin formation in Drosophila requires genome-wide histone deacetylation in cleavage chromatin before mid-blastula transition in early embryogenesis. *Chromosoma*, *129*(1), 83. https://doi.org/10.1007/S00412-020-00732-X

Weaver, B. A. A., & Cleveland, D. W. (2007). Aneuploidy: Instigator and Inhibitor of Tumorigenesis. *Cancer Research*, *67*(21), 10103–10105. https://doi.org/10.1158/0008-5472.CAN-07-2266

Wei, X., Eickbush, D. G., Speece, I., & Larracuente, A. M. (2021). Heterochromatin-dependent transcription of satellite DNAs in the Drosophila melanogaster female germline. *ELife*, *10*. https://doi.org/10.7554/eLife.62375

Wells, J. N., & Feschotte, C. (2020). A Field Guide to Eukaryotic Transposable Elements. *Annual Review of Genetics*, *54*(1), 539–561. https://doi.org/10.1146/annurev-genet-040620-022145

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Willard, H. F. (1985). Chromosome-specific organization of human alpha satellite DNA. *American Journal of Human Genetics*, *37*(3), 524. /pmc/articles/PMC1684601/?report=abstract

Wilson, E. B. (1925). *The cell in development and heredity* (3rd Edition). Macmillan. https://wellcomecollection.org/works/q2j4atgy

Wong, L. H., Brettingham-Moore, K. H., Chan, L., Quach, J. M., Anderson, M. A., Northrop, E. L., Hannan, R., Saffery, R., Shaw, M. L., Williams, E., & Choo, K. H. A. (2007). Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere. *Genome Research*, *17*(8), 1146–1160. https://doi.org/10.1101/gr.6022807

Wood, J. G., Hillenmeyer, S., Lawrence, C., Chang, C., Hosier, S., Lightfoot, W., Mukherjee, E., Jiang, N., Schorl, C., Brodsky, A. S., Neretti, N., & Helfand, S. L. (2010). Chromatin remodeling in the aging genome of Drosophila. *Aging Cell*, *9*(6), 971–978. https://doi.org/10.1111/J.1474-9726.2010.00624.X

Wood, J. G., Jones, B. C., Jiang, N., Chang, C., Hosier, S., Wickremesinghe, P., Garcia, M., Hartnett, D. A., Burhenn, L., Neretti, N., & Helfand, S. L. (2016). Chromatin-modifying genetic interventions suppress age-associated transposable element activation and extend life span in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(40), 11277–11282. https://doi.org/10.1073/PNAS.1604621113/SUPPL_FILE/PNAS.1604621113.SD01.XLS

Wu, T., Lane, S. I. R., Morgan, S. L., Tang, F., & Jones, K. T. (2021). Loss of centromeric RNA activates the spindle assembly checkpoint in mammalian female meiosis I. *Journal of Cell Biology*, *220*(10). https://doi.org/10.1083/jcb.202011153

Xiong, W. C., Okano, H., Patel, N. H., Blendy, J. A., & Montell, C. (1994). repo encodes a glial-specific homeo domain protein required in the Drosophila nervous system. *Genes & Development*, *8*(8), 981–994. https://doi.org/10.1101/GAD.8.8.981

Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). ClusterProfiler: An R package for comparing biological

themes among gene clusters. *OMICS A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Yu, Y., Gu, J., Jin, Y., Luo, Y., Preall, J. B., Ma, J., Czech, B., & Hannon, G. J. (2015). Panoramix enforces piRNA-dependent cotranscriptional silencing. *Science*, *350*(6258), 339–342. https://doi.org/10.1126/science.aab0700

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A Greedy Algorithm for Aligning DNA Sequences. In *JOURNAL OF COMPUTATIONAL BIOLOGY* (Vol. 7, Issue 2). Mary Ann Liebert, Inc. Pp. www.liebertpub.com

Zheng, H., Yang, X., & Xi, Y. (2016). Fat body remodeling and homeostasis control in Drosophila. *Life Sciences*, *167*, 22–31. https://doi.org/10.1016/j.lfs.2016.10.019

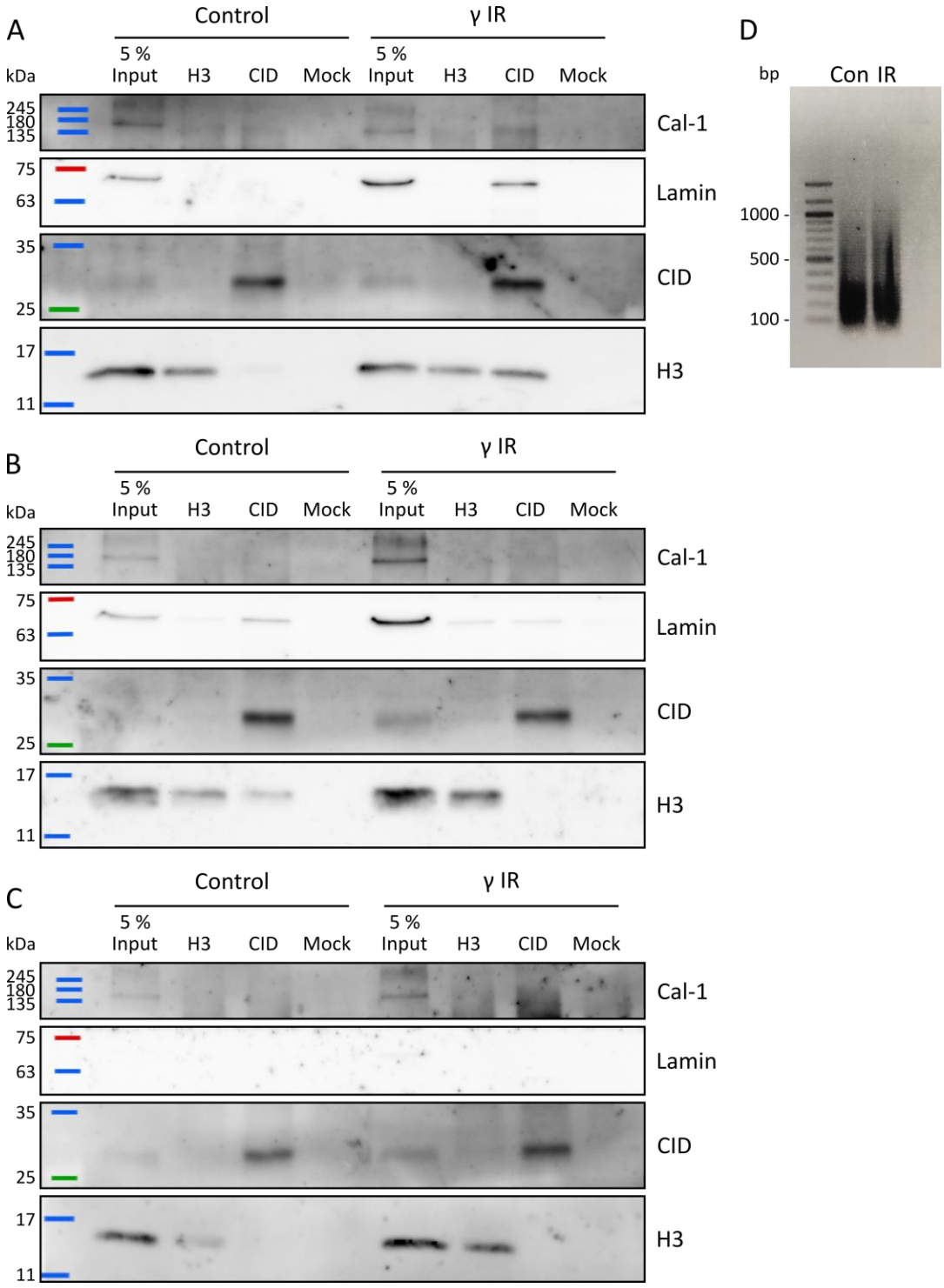
Zhu, A., Ibrahim, J. G., & Love, M. I. (2019). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*, *35*(12), 2084–2092. https://doi.org/10.1093/bioinformatics/bty895

Zhu, Q., Hoong, N., Aslanian, A., Hara, T., Benner, C., Heinz, S., Miga, K. H., Ke, E., Verma, S., Soroczynski, J., Yates, J. R., Hunter, T., & Verma, I. M. (2018). Heterochromatin-Encoded Satellite RNAs Induce Breast Cancer. *Molecular Cell*, *70*(5), 842-853.e7. https://doi.org/10.1016/j.molcel.2018.04.023

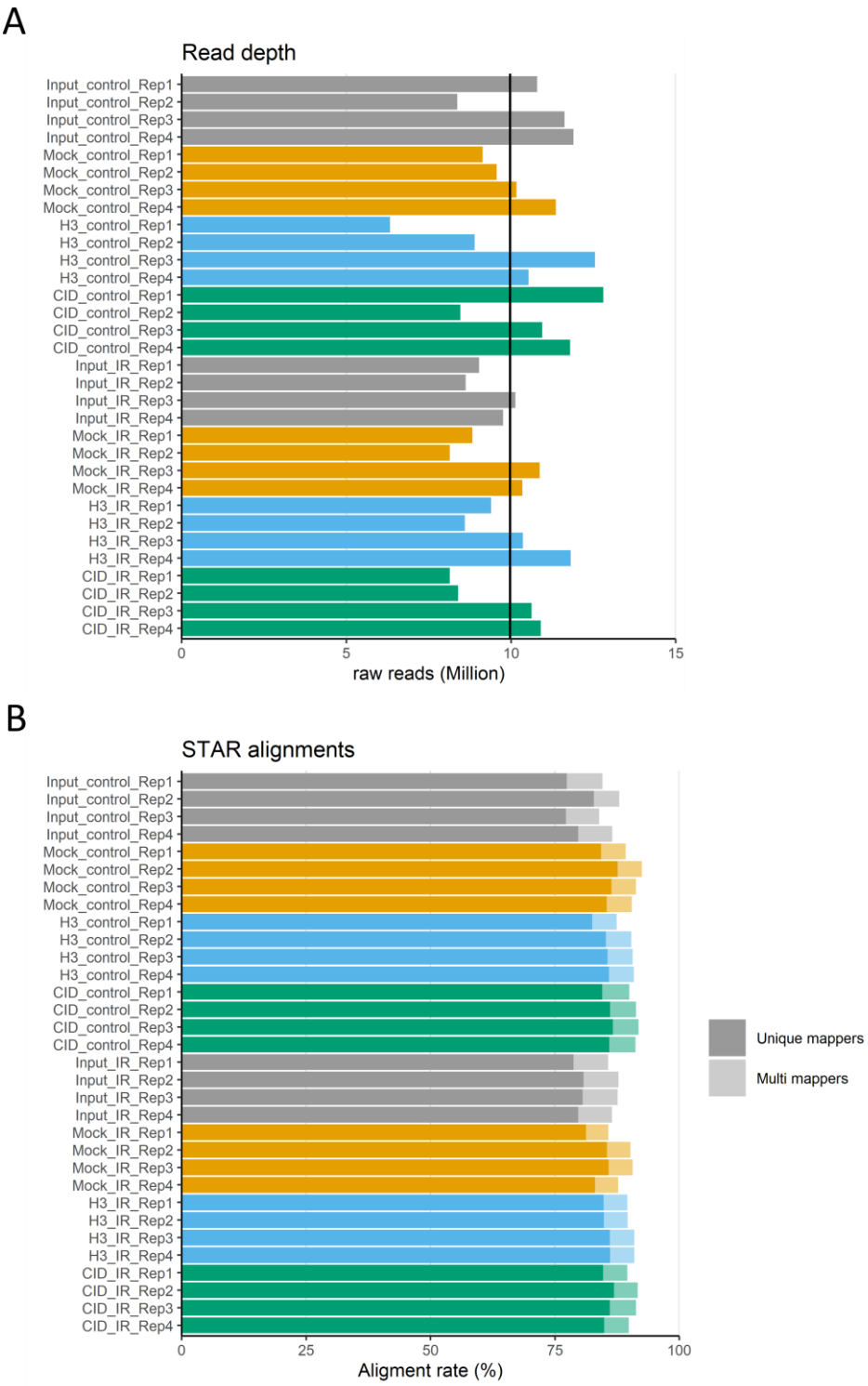
Zhu, Q., Pao, G. M., Huynh, A. M., Suh, H., Tonnu, N., Nederlof, P. M., Gage, F. H., & Verma, I. M. (2011). BRCA1 tumour suppression occurs via heterochromatin-mediated silencing. *Nature*, *477*(7363), 179–184. https://doi.org/10.1038/nature10371

Supplementary data

Supplementary Figures



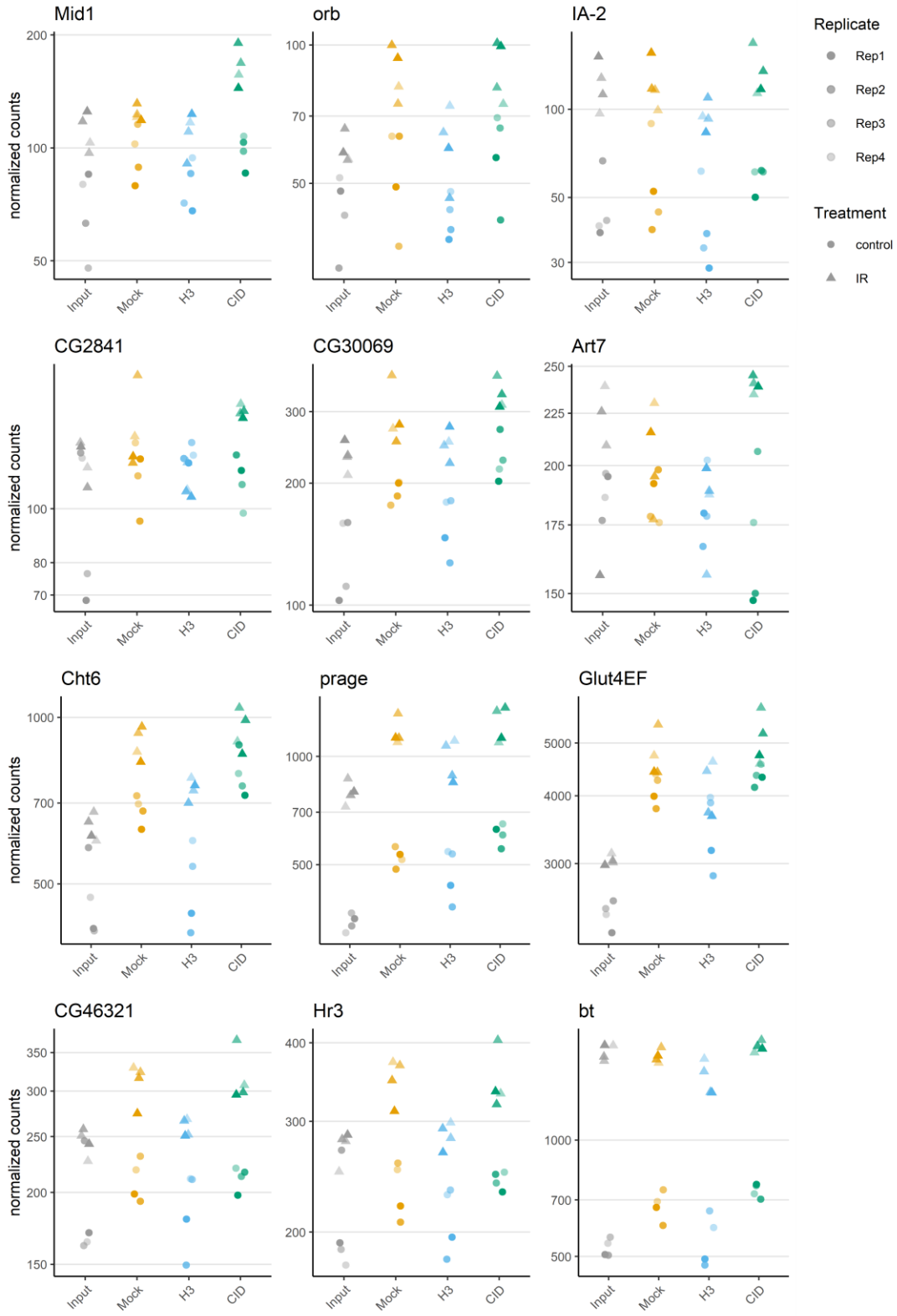
Supplementary Figure 1. Embryo RNA-ChIP control western blot of other replicates.
Input and indicated ChIP samples from both control and γ irradiated embryos were checked with WB using the indicated antibodies. A) Replicate 2. B) Replicate 3. C) Replicate 4.
D) Agarose gel of the shearing quality control of chromatin used for both control and γ irradiated ChIPs. IR = irradiated.



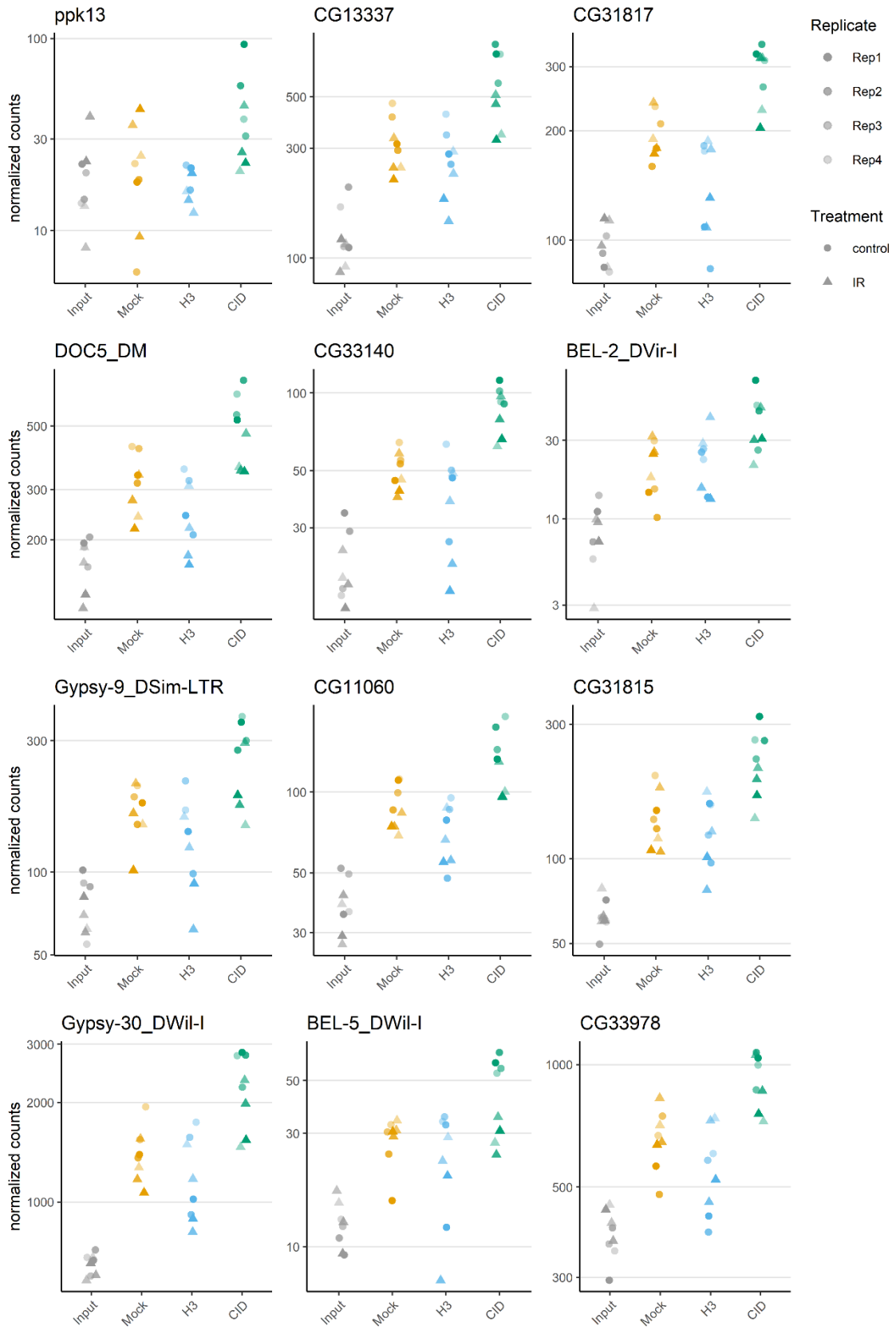
Supplementary Figure 2. Technical metrics of the embryo RNA-ChIP sequencing data.

A) Read depth in million raw reads per sample. Vertical black line indicates the average read depth of 9.97 million reads.

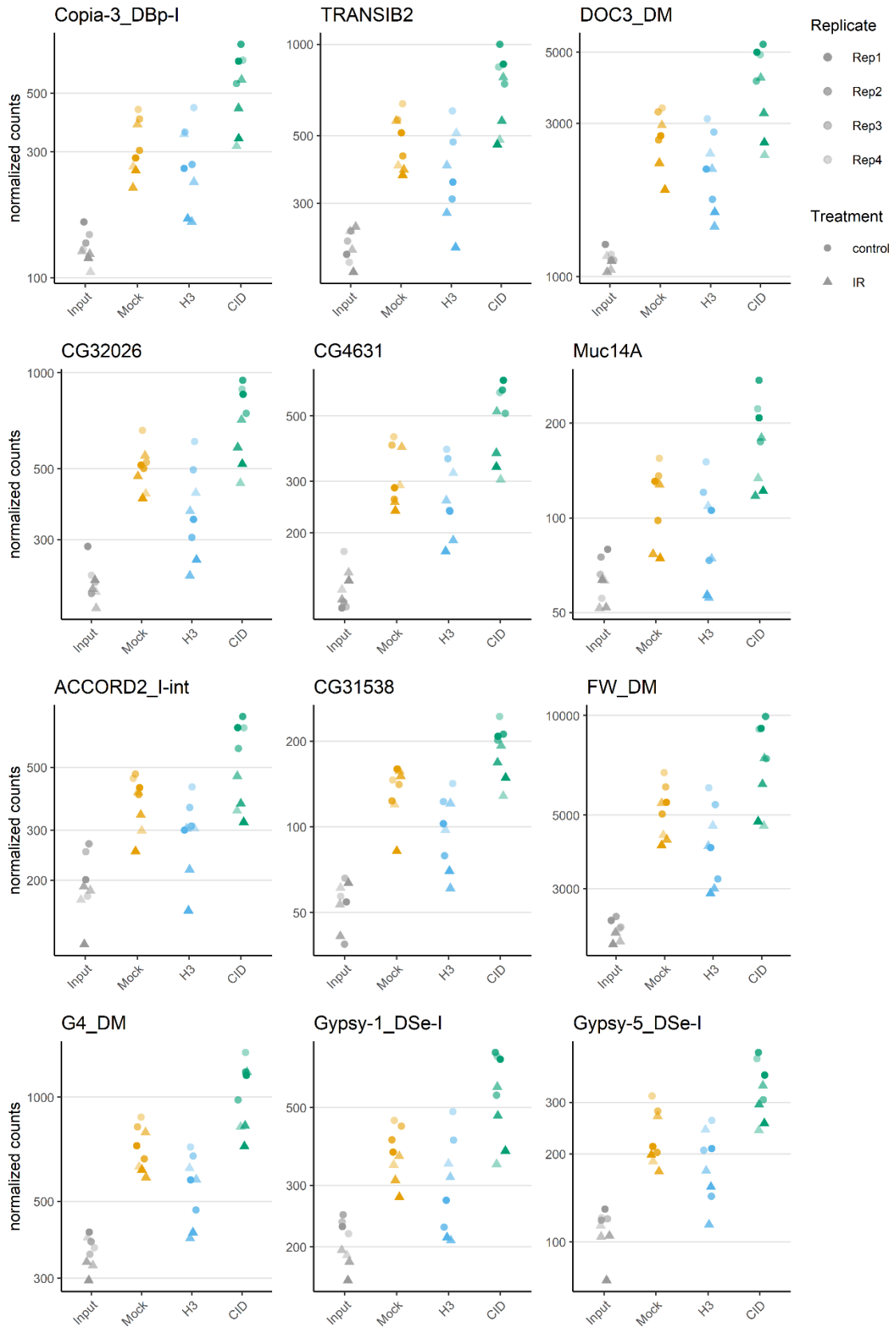
B) STAR alignment rate as percentage of the total raw reads per sample. Both unique and multi mappers are specified separately. IR = irradiated, Rep = replicate.



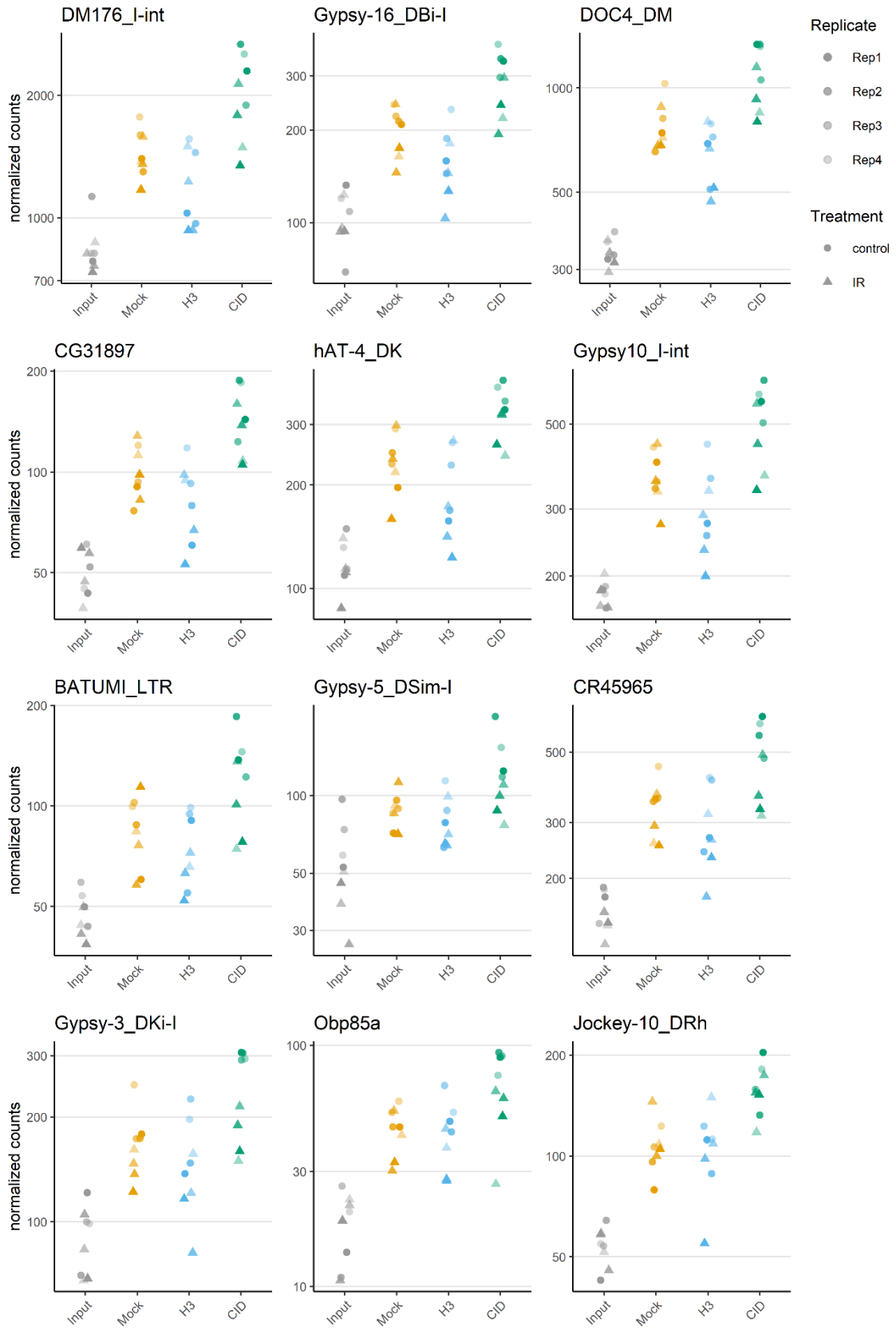
Supplementary Figure 3. Potential DNA damage specific centromere-associated RNAs.
Normalized counts plotted per sample for the top candidates indicated as significantly enriched in the irradiated CID RNA-ChIP over both control CID RNA-ChIP and irradiated H3 RNA-ChIP. IR = irradiated, Rep = replicate.



Supplementary Figure 4. Individual count data plotted for candidates 3 to 14.
Normalized counts plotted per sample for top significantly enriched genes in the untreated control CID RNA-ChIP. IR = irradiated, Rep = replicate.

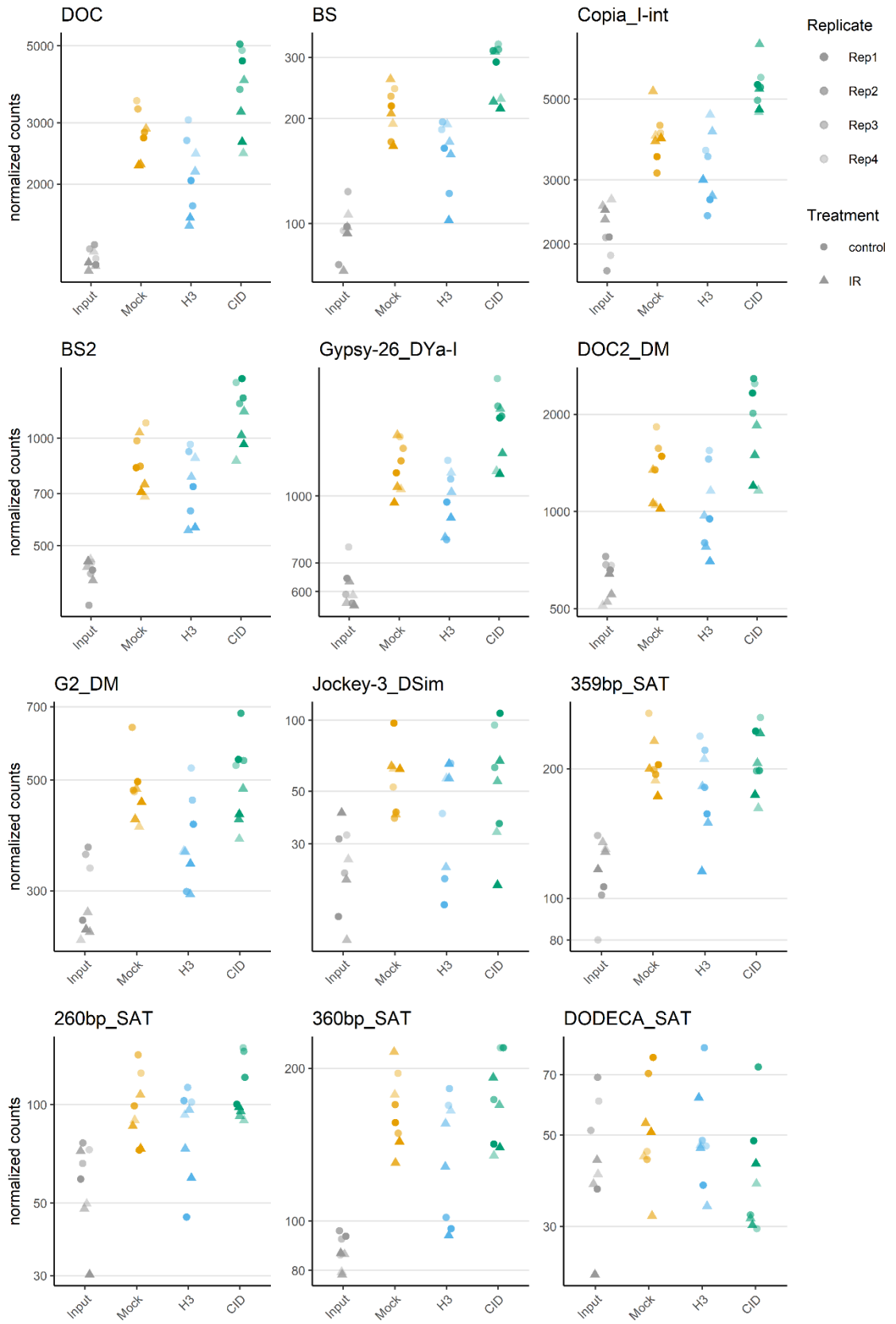


Supplementary Figure 5. Individual count data plotted for candidates 15 to 22, 24 to 26 and 28. Normalized counts plotted per sample for top significantly enriched genes in the untreated control CID RNA-ChIP. IR = irradiated, Rep = replicate.



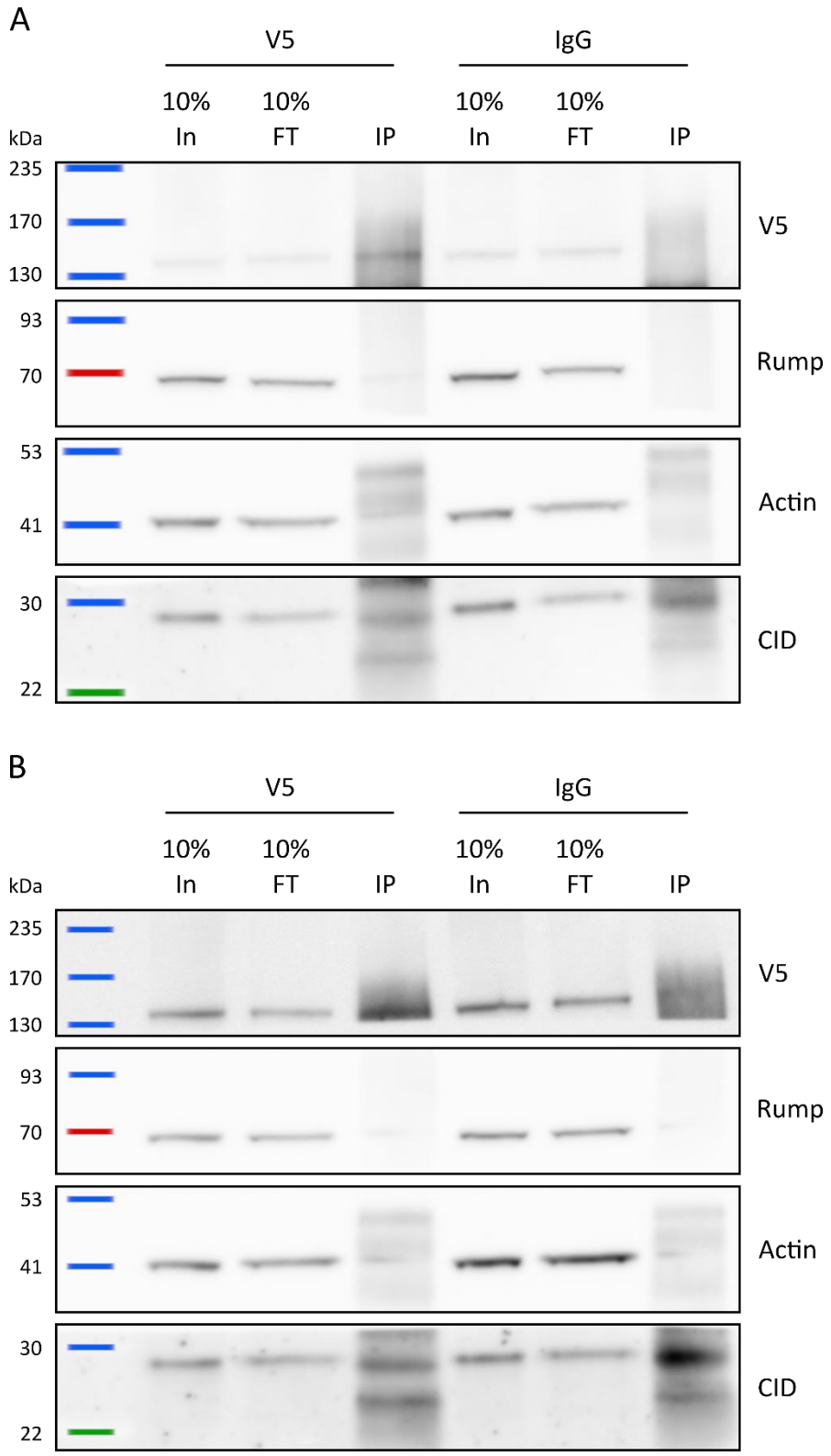
Supplementary Figure 6. Individual count data plotted for candidates 29 to 31, 34, 38, 40 to 42, 44, 49, 58 and 61.

Normalized counts plotted per sample for top significantly enriched genes in the untreated control CID RNA-ChIP. IR = irradiated, Rep = replicate.



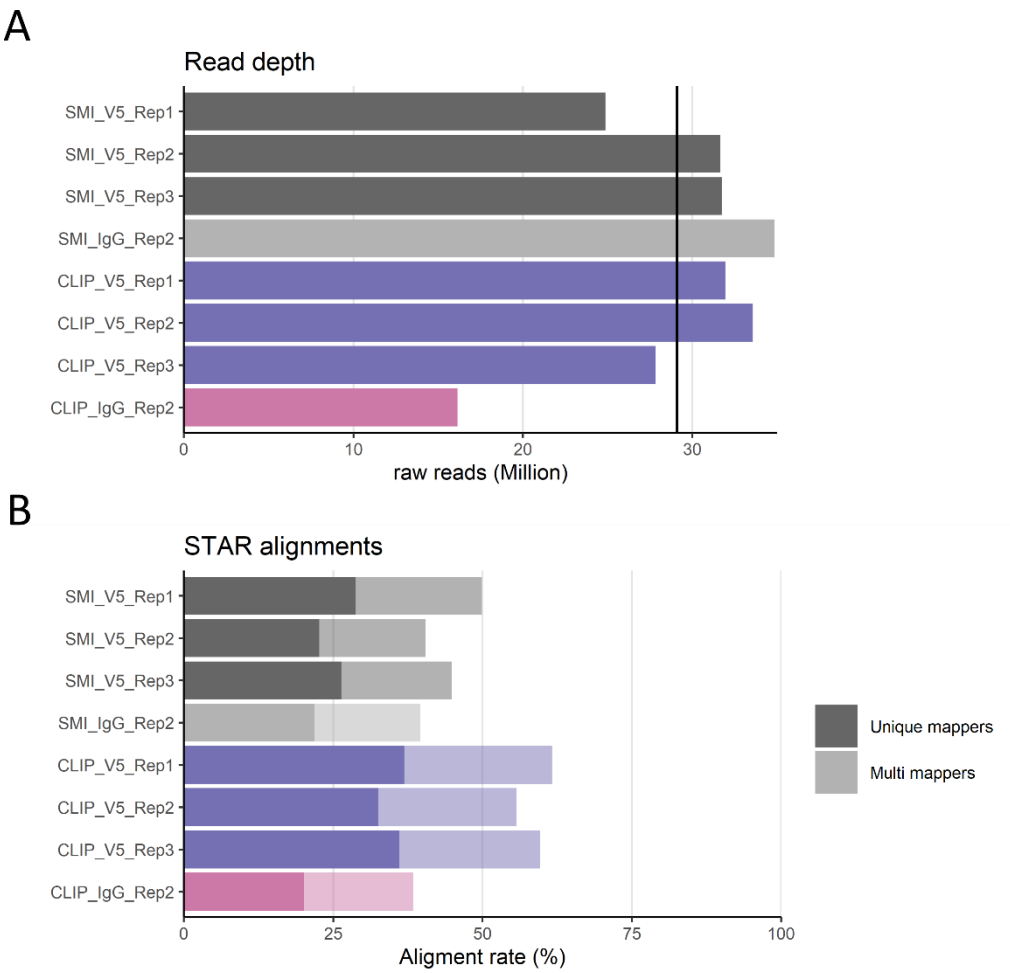
Supplementary Figure 7. Individual count data plotted for top cenTE candidates, G2_DM, Jockey-3_DSim and several satellite repeats.

Normalized counts plotted per sample for top significantly enriched genes in the untreated control CID RNA-ChIP. IR = irradiated, Rep = replicate.



Supplementary Figure 8. XRNAX-CLIP western blot shows specific enrichment of Cal1-V5 in the V5 CLIP.

XRNAX extract was used as input for a CLIP pulling down V5-tagged Cal1 with a V5 antibody. IgG antibody was used for a negative control pull down. Efficiency of the CLIP was checked with WB using the indicated antibodies. Western blots of replicate 1 (A) and replicate 3 (B). In = Input, FT = flow through, IP = CLIP, Rump = Rumpelstiltskin.

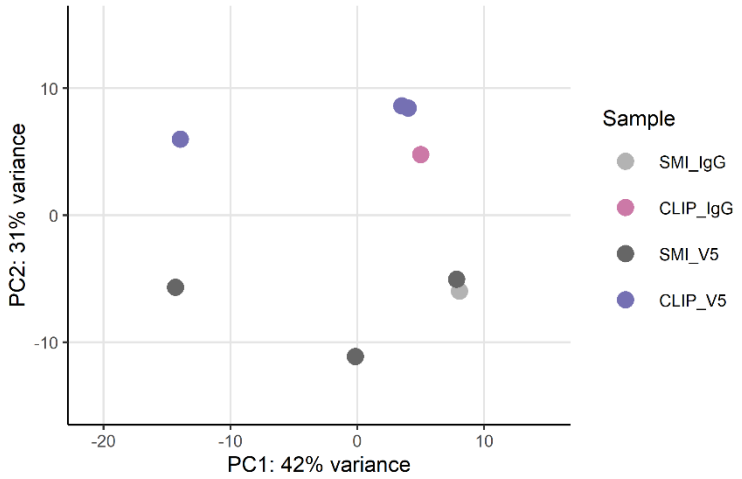


Supplementary Figure 9. Technical metrics of the Cal1-V5 XRNAX-CLIP sequencing data.

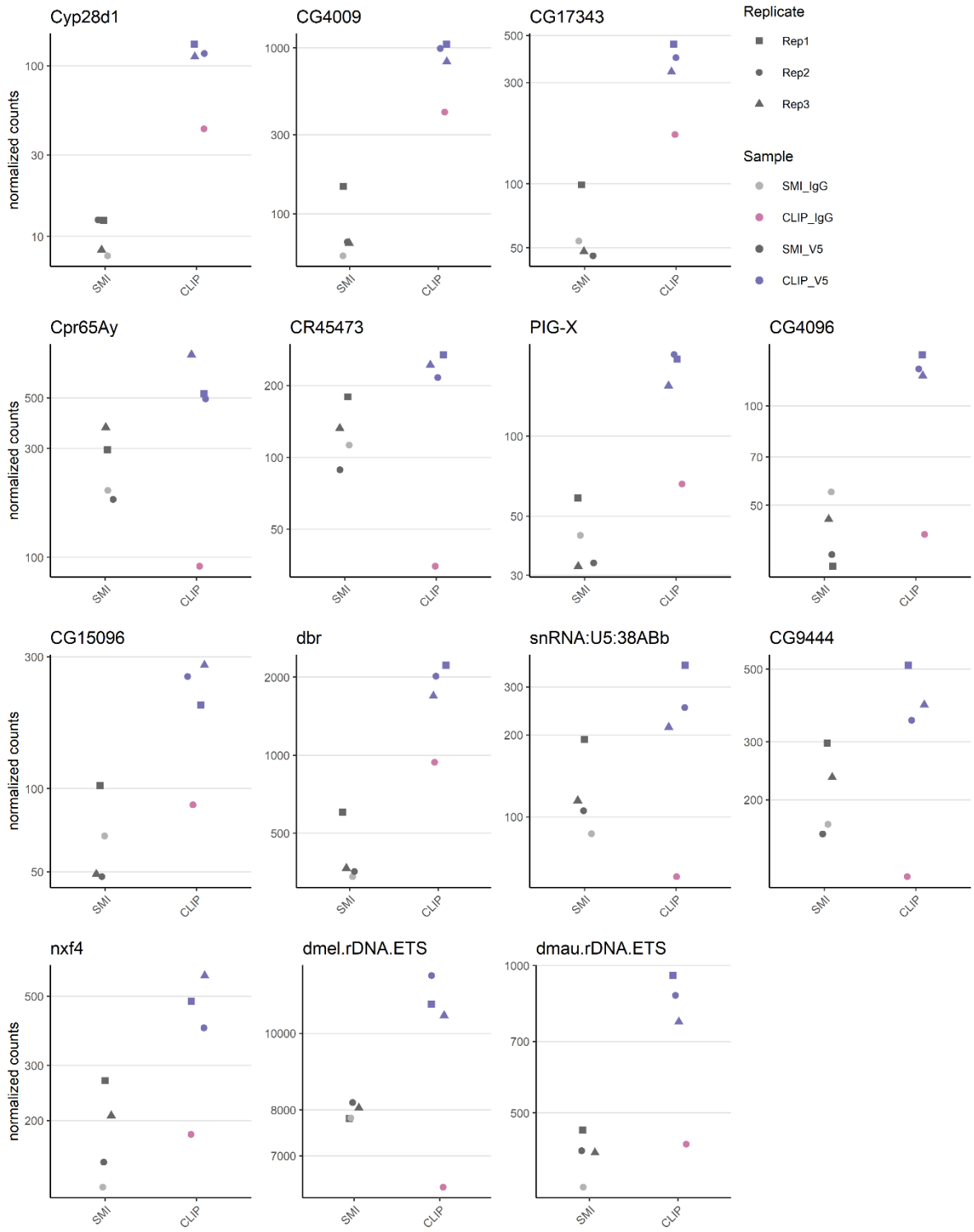
A) Read depth in million raw reads per sample. Vertical black line indicates the average read depth of 29.1 million reads.

B) STAR alignment rate as percentage of the total raw reads per sample. Both unique and multi mappers are specified separately.

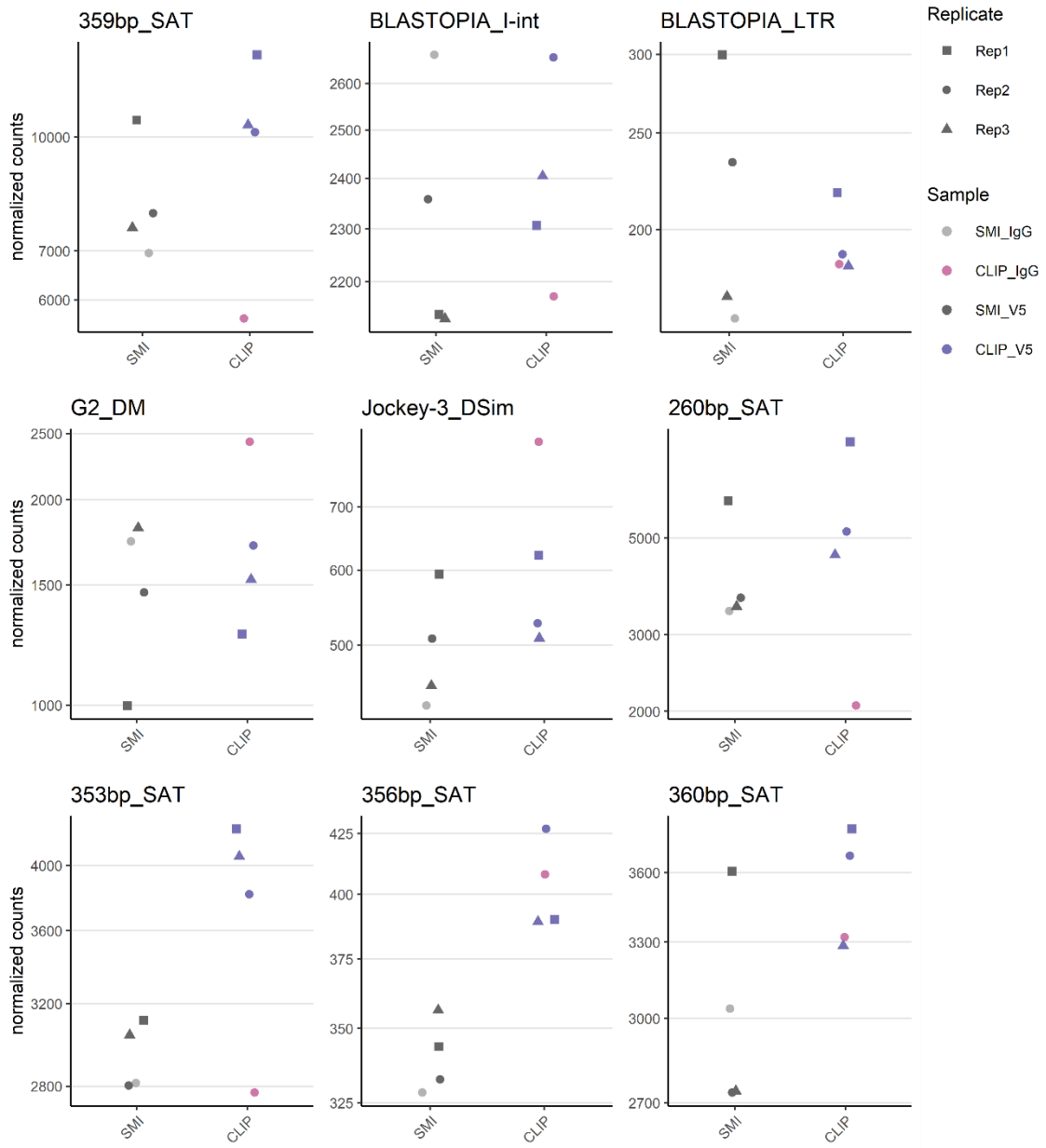
SMI = size-matched input control, Rep = replicate.



Supplementary Figure 10. PCA plot showing variation in the Cal1-V5 XRNAX-CLIP-seq dataset. Principal Component Analysis (PCA) was performed to assess the main causes of variation between the samples. The first two principal components were plotted which account for respectively 42 % and 31 % of the variation between the samples. Samples cluster based on their similarity. SMI = size-matched input control, PC = principle component.



Supplementary Figure 11. Individual count data plotted for the other candidates.
Normalized counts plotted per sample for top significantly enriched genes in the Cal1-V5 CLIP. SMI = size-matched input control, Rep = replicate.



Supplementary Figure 12. Individual count data plotted for centromere-associated RNAs and some other repeats.

Normalized counts plotted per sample for SATIII (359_bp), Blastopia, G2/Jockey-3 and other satellite repeats from the 1.688 family. SMI = size-matched input control, Rep = replicate.

Supplementary Tables

All Supplementary Tables containing the results of the embryo RNA-ChIP-seq and Cal1 XRNAX-CLIP-seq data analysis are stored on the attached compact disc storage device.

Supplementary Table 1. Differentially expressed RNAs in γ-irradiated embryos.
Significantly differentially expressed RNAs from the DESeq2 analysis of the embryo RNA-ChIP-seq dataset comparing all irradiated samples to all control samples.

Supplementary Table 2. Differentially enriched stress-specific centromere-associated RNAs.
Significantly differentially enriched RNAs from the DESeq2 analysis of RNA-ChIP-seq dataset comparing irradiated CID RNA-ChIP samples to both irradiated H3 RNA-ChIP and control CID RNA-ChIP samples.

Supplementary Table 3. Differentially depleted centromere-associated RNAs after stress.
Significantly differentially enriched RNAs from the DESeq2 analysis of RNA-ChIP-seq dataset comparing control CID RNA-ChIP samples to both control H3 RNA-ChIP and irradiated CID RNA-ChIP samples.

Supplementary Table 4. Differentially enriched centromere-associated RNAs.
Significantly differentially enriched RNAs from the DESeq2 analysis of RNA-ChIP-seq dataset comparing control CID RNA-ChIP samples to control H3 RNA-ChIP samples.

Supplementary Table 5. Putative centromere-associated RNAs.
Significantly differentially enriched RNAs from the DESeq2 analysis of RNA-ChIP-seq dataset comparing control CID RNA-ChIP samples to both control H3 RNA-ChIP and control Mock RNA-ChIP samples.

Supplementary Table 6. Putative Cal1 binding RNAs.
Significantly differentially enriched RNAs from the DESeq2 analysis of XRNAX-CLIP-seq dataset comparing Cal1 XRNAX-CLIP samples to both size-matched Input and IgG XRNAX-CLIP samples.

List of abbreviations

AED	After egg deposition
APC/C	Anaphase-promoting complex/cyclosome
Bp	Base pair
Cal1	Chromosome alignment defect 1
CCAN	Constitutive Centromere Associated Network
CDK	Cyclin-dependent kinase
cDNA	Copy DNA
CENP-A	Centromere protein A
CENP-B	Centromere protein B
CENP-C	Centromere protein C
cenRNA	Centromeric RNAs
cenTE	TE present in centromeric DNA
ChIP	Chromatin immunoprecipitation
CID	Centromere identifier (*Drosophila* CENP-A)
CLIP	UV crosslinking and immunoprecipitation
CNS	Central nervous system
CPC	Chromosomal passenger complex
DNA	Deoxyribonucleic acid
dsDNA	Double-stranded DNA
dsRNA	Double-stranded RNA
Endo-siRNA	Endogenous small interfering RNA
ERV	Endogenous retrovirus
GO	Gene ontology
H3	Histone 3
H3K4me2	Histone 3 lysine 4 di-methyl
H3K9me3	Histone 3 lysine 9 tri-methyl
H4	Histone 4
HJURP	Holliday junction recognition protein
hnRNPm	Heterogeneous nuclear ribonucleoprotein M
HOR	Higher-order repeat
HP1	Heterochromatin protein 1
IF	Immunofluorescence
INCENP	Inner centromere protein
IP	Immunoprecipitation
KNM network	Knl1/Spc105 complex, the Ndc80 complex and the Mis12 complex
LINE	Long interspersed nuclear element
LTR	Long terminal repeat
MCC	Mitotic checkpoint complex
ncRNA	Non-coding RNA
NLS	Nuclear localization signal
ORF	Open reading frames

piRNA	Piwi-interacting RNA
pMT	Metallothionein promoter
RISC	RNA-induced silencing complex
RNA	Ribonucleic acid
RNA-ChIP	RNA chromatin immunoprecipitation
RNAi	RNA interference
RNase	Ribonuclease
RPM	Reads per million
RT-qPCR	Reverse transcription quantitative PCR
S2 cells	*Drosophila* Schneider 2 cell line
SAC	Spindle assembly checkpoint
SDS-PAGE	Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis
shRNA	Short hairpin RNA
SINE	Short interspersed nuclear element
SMI-control	Size-matched input control
smRNA-FISH	Single molecule RNA fluorescence in situ hybridization
TE	Transposable element
TIR	Terminal inverted repeat
TRiP	Transgenic RNA interference project
tRNA	Transfer RNA
UAS	Upstream activating site
XRNAX	Protein-Xlinked RNA eXtraction

List of Figures

Acknowledgements

My PhD has been a journey which I could not have completed without the support of the Erhardt lab, the Lohmann lab, my friends and my family.

First of all, I would like to thank Prof. Dr. Erhardt for giving me the opportunity to join her lab as a PhD student. Dear Sylvia, thank you for always having an open door to discuss science and life. I appreciate your support through the lows and it was great celebrating with you through the highs.

Thank you to my TAC members, Prof. Dr. Georg Stoecklin and Prof. Dr. Frank Lyko, and my additional defense committee members, Prof. Dr. Ingrid Lohmann and Prof. Dr. Henrik Kaessmann. A special thanks to Prof. Dr. Georg Stoecklin for your moral support and for being the chair of my defense committee.

I am very grateful to Prof. Dr. Ingrid Lohmann for “adopting” me so warmly during the final years of my PhD. Without hesitation you gave me a place in your lab and all the facilities I needed to finish my PhD when the ZMBH abruptly closed down. This also gave me the opportunity to meet all the amazing Lohmann lab members, who immediately made me feel welcome and integrated me into their group. It was fun to join your lab meetings, lunches and dinners at Rafa’s place.

I would also like to thank David Ibberson, Simon Anders and the Bukau lab for their help and advice with sequencing and data analysis.

Many thanks to previous and present lab members of the Erhardt lab! You all created an amazing atmosphere in the lab, making the sometimes difficult times with failing experiments seem not so bad. It was great working with you, but even better to have breakfasts, lunches, dinners, retreats, parties, foosball and frisbee games, and game nights. Saskia and Elisa, thank you for being such great friends and being there for me throughout this journey!

Of course I want to thank my family for being so supportive and for visiting me often. You are always there for me. Last but definitely not least, I am so grateful for all your support and love Ioannis. I can’t wait to start our next adventure together.