Protein-RNA interactions in centromeric chromatin

Vojtěch Dolejš

2024

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

Mgr. Vojtěch Dolejš

born in Praha, Czechoslovakia

Oral examination: July 19th, 2024

Protein-RNA interactions in centromeric chromatin

Referees:

Prof. Dr. Sylvia Erhardt

Prof. Dr. Matthias Mayer

Summary

Chromosome segregation is a complex and tightly regulated process that is necessary for proper cell division. Chromosomes are aligned at the equatorial plate and transported to nascent daughter cells using motor proteins and microtubules. They are attached to the microtubules by a multiprotein complex, the kinetochore. Kinetochores are anchored to the centromeric regions of the chromosome. These regions consist of repetitive noncoding DNA which forms the centromeric heterochromatin and is embedded in larger regions of noncoding pericentromeric heterochromatin. Centromeres are not defined by an underlying DNA sequence, but rather by the epigenetic marker Cenp-A, a histone H3 variant specific for centromeres.

Cenp-A, or Cid in *Drosophila melanogaster* (fruit fly), together with its loading factor Cal1 and the Cenp-C protein form the basis of the inner kinetochore complex. The outer kinetochore then connects to the spindle microtubule. This process requires the presence of noncoding RNAs. These RNAs are transcribed from the pericentromeric regions and colocalise with the inner parts of the kinetochore. The absence of these RNAs leads to chromosome segregation defects, but their exact function as well as their interaction partners are not well described. These defects are observed not just in *Drosophila melanogaster*, but also in every other organism that has been researched to date.

In this work, I investigated the interaction of centromeric proteins of *Drosophila melanogaster* with centromeric RNA using biochemical assays (EMSA) and mass spectrometry (HDX-MS). I overexpressed and purified proteins as well as *in vitro* transcribed RNAs. I observed the interaction of kinetochore proteins and centromeric RNA with focus on Cal1 and satellite RNA. I demonstrated that Cal1 fragments are RNA-binding using EMSA. However, HDX-MS was not sufficient to measure the details of the interaction, as addition of RNA did not significantly change the deuteration pattern of the protein fragments. I observed that Cal1 is unfolded and the fragment I worked with is a promiscuous RNA-binder.

Zusammenfassung

Die Chromosomentrennung ist ein komplexer und streng regulierter Prozess, der für eine ordnungsgemäße Zellteilung erforderlich ist. Die Chromosomen werden in der Äquatorialebene ausgerichtet und mit Hilfe von Motorproteinen und Mikrotubuli zu den sich formenden Tochterzellen transportiert. Sie sind durch einen Multiproteinkomplex, dem Kinetochor, an die Mikrotubuli befestigt. Kinetochore sind in den zentromerischen Regionen des Chromosoms verankert. Diese Regionen bestehen aus repetitiver, nicht kodierender DNA, die das zentromerische Heterochromatin bilden und in die größeren Regionen des nicht kodierenden perizentromerischen Heterochromatins eingebettet sind. Zentromere werden nicht durch die zentromerischen DNA-Sequenzen definiert, sondern durch den epigenetischen Marker Cenp-A, eine für Zentromere spezifische Histon-H3-Variante.

Cenp-A, oder Cid in *Drosophila melanogaster* (Fruchtfliege), bildet zusammen mit seinem Ladefaktor Cal1 und dem Cenp-C-Protein die Grundlage des inneren Kinetochor-Komplexes. Das äußere Kinetochor verbindet diesen Komplex mit den Spindelmikrotubuli. Dieser Prozess erfordert nichtcodierende RNAs. Diese RNAs werden in den perizentromerischen Regionen transkribiert und kolokalisieren mit den inneren Teilen des Kinetochors. Das Fehlen dieser RNAs führt zu Defekten bei der Chromosomensegregation, aber ihre genaue Funktion sowie ihre Interaktionspartner sind nicht genau bekannt. Diese Defekte werden nicht nur in *Drosophila melanogaster* beobachtet, sondern auch in allen anderen Organismen, die bislang erforscht wurden.

In dieser Arbeit untersuchte ich die Interaktion zentromerischer Proteine von *Drosophila melanogaster* mit zentromerischer RNA mithilfe biochemischer Assays (EMSA) und Massenspektrometrie (HDX-MS). Ich überexprimierte und reinigte Proteine sowie *in vitro* transkribierte RNAs auf. Ich beobachtete die Interaktion von Kinetochor-Proteinen und zentromerischer RNA mit Schwerpunkt auf Cal1 und Satelliten-RNA. Mit Hilfe von EMSA konnte ich nachweisen, dass Cal1-Fragmente RNA-bindend sind. HDX-MS reichte jedoch nicht aus, um die Details der Interaktion zu messen, da die Zugabe von RNA das Deuterationsmuster der Proteinfragmente nicht wesentlich veränderte. Ich habe festgestellt, dass Cal1 entfaltet ist und das Fragment, mit dem ich gearbeitet habe, ein promiskuitiver RNA-Binder ist.

Contents

1. Introduction

The cell is generally recognised as the smallest living part of any organism (Koonin, 2014). Cell division is a carefully orchestrated process that requires the cooperation of a large number of cellular components. Failure to divide leads to cell stagnation and apoptosis, while uncontrolled division leads to defects and diseases, such as cancer (Liu et al., 2022). Most importantly, the genetic information itself must be evenly distributed into daughter cells. To ensure that DNA is split evenly, each chromosome is condensed and aligned at the metaphase plate and then equally segregated to the daughter cells. The regions on every chromosome responsible for this process are the centromeres, which are visible at the primary constriction sites of mitotic chromosomes. They form the basis of the kinetochore complex, which in turn facilitates attachment to spindle microtubules. Centromeres are embedded in larger regions of centromeric and pericentromeric chromatin (Talbert & Henikoff, 2020; Kyriacou & Heun, 2023).

1.1. Chromatin structure

DNA, the carrier of genetic information in every known organism, is stored in the nucleus of each eukaryotic cell in the form of chromatin. Chromatin consists of the DNA, RNA and proteins, with both structural and functional properties. Structural proteins, such as histones, help to condense long polymers of DNA molecules into compact structures by neutralising their negative charge and forming 'beads on a string' (Klug et al., 1979). DNA is wrapped around nucleosomes consisting of two pairs of four histone proteins: H2A, H2B, H3 and H4 [***Figure 1***]. When DNA is wrapped around them, the two parts are locked together by histone H1. The length of DNA wrapped around the histone particle is 146-147 basepairs (bp) but DNA between two nucleosomes has various length (Davey et al., 2002). Histones are among the most conserved proteins, especially their core structures. Although there are differences between species, the structures are very similar in all organisms that have histones even though their secondary functions and N-terminal parts have been observed to evolve (Malik & Henikoff, 2003). Additional to five canonical histones which build the canonical nucleosome, there are several non-canonical histones, that have specific functions at different time points of cell cycle and differ from their canonical counterparts only by a few amino acids (Hake & Allis, 2006). For instance, histone H3 has two variants: H3.1 and H3.2 which are considered canonical,

as well as non-canonical H3.3 that, among other functions such as recruitment of chromatin remodelling proteins (P. Chen et al., 2013), also functions as a placeholder for the deposition of the centromeric variant of H3 histone, or Cenp-A (Dunleavy et al., 2011). Its name varies depending on the species. It is called centromeric protein A (CENP-A) in humans, centromere alignment defect (Cid) in fruit fly and chromosome segregation protein 4 (Cse4) in fission yeast.

N-terminal tails of histones reach out of the nucleosome core particle and facilitate interactions with chromatin remodelling proteins and other interaction partners [***Figure 1***]. Histones also allow DNA to slide in both directions, which is essential for its proper function, because it allows access of modifier proteins to DNA (Felsenfeld, 1978; Morrison & Thakur, 2021). In addition to the structural proteins, functional chromatin-associated proteins read, copy, and cleave DNA, processes that are essential for the correct function of a cell. This is necessary for the transcription of DNA and subsequent proteosynthesis. It is also important for the functional organisation of chromatin (Morrison & Thakur, 2021). Chromatin is heterogeneous and exhibits different properties based on the specific requirements of the cells. Gene rich regions of the genome, that are transcriptionally active are referred to as euchromatin. Euchromatin carries specific histone modifications such as di- and trimethylation on lysine 4 of histone H3, abbreviated as H3K4me2/3. Methylation of lysine changes the charge state of the chromatin, which is recognised by the remodelling enzymes, that carry out the opening or closing of the chromatin, as a specific histone code (Jenuwein & Allis, 2001). H3K4me2/3 promotes opening of the chromatin and therefore its activation by recruiting the nucleosome remodelling factor (NURF) (Wysocka et al., 2006). This type of chromatin is typical for actively transcribed gene rich regions and their enhancer sequences (Liang et al., 2004).

Transcriptionally silent chromatin, or heterochromatin, is divided into two groups. Facultative heterochromatin is repressed but can be activated under certain conditions. It carries H3K9me2 and H3K27me3 modifications. Methylation of lysines 9 and 27 leads to recruitment of polycomb proteins and heterochromatin protein 1 (HP1) that actively work on closing the chromatin, as well as deacetylases, that help to mitigate the charge and assist with the packing process (Kouzarides, 2007). This also means that heterochromatin is hypoacetylated compared to the euchromatin (Richards & Elgin, 2002). The other type is constitutive heterochromatin, which is condensed and remains

inactive. It carries H3K9me3 and H4K20me3 modifications and is mostly located at the ends of the chromosomes (telomeres) or around the centromere (pericentromeric chromatin) (Richards & Elgin, 2002).

Centromeric chromatin itself is enriched for H3K4me2 and depleted for H3K4me3 and H3K9me2/3 (Sullivan & Karpen, 2004). This makes it a unique type of chromatin, that differs from both euchromatin and heterochromatin. It carries histone modification typical for both and therefore has unique properties. Even though it is heterechromatic, it bears signs of active transcription, most importantly the active form of RNA polymerase II (Saffery et al., 2012; Rošić et al., 2014). Many transcripts originating from pericentromeres have been identified to date (Corless et al., 2020). These RNAs are important for the proper centromere function and Cenp-A loading (Rošić et al., 2014; Bobkov et al., 2018).

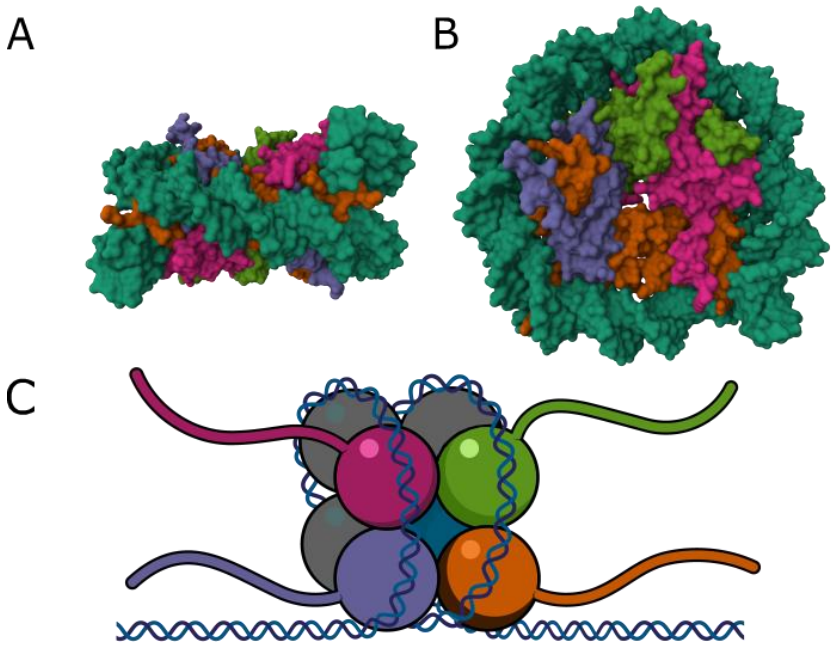


Figure 1. **Schematic representation of a nucleosome core particle**

Histones forming a core particle of the nucleosome with DNA wrapped around it. **(A)** View from below **(B)** view from the front **(C)** schematic view. DNA is represented with dark green **(A, B)** or blue **(C)**. Histone H2A is pink, H2B is light green, H3 orange and H4 purple. The entire nucleosome core particle contains two copies of each, assembled into two adjacent heterotetramers. DNA is then wrapped around the resulting octamer. They are not visible in the crystal structure due to their flexibility. Structure obtained at RCSB PDB (2NQB) and BioRender (https://www.biorender.com).

1.2. Centromeres and centromeric sequences

Centromeres are specialised regions on chromosomes that are essential for accurate chromosome segregation during cell division. They are identified by the presence of centromeric proteins rather than the underlying DNA sequence (Carroll & Straight, 2006). Centromeric chromatin is defined by the presence of the histone H3 variant Cenp-A. Centromeric DNA sequence is not conserved and differs significantly between organisms (Henikoff & Dalal, 2005). There are no known protein coding genes present in centromeric and pericentromeric regions. Centromeres can be a single point (one Cenp-A molecule, i.e. in budding yeast) or a small region (several neighbouring Cenp-A molecules, i.e. in fruit fly) [***Figure 2***]. These are so called monocentric centromeres. As opposed to that, some organisms are holocentric (i.e. *C. elegans*) and have their Cenp-A and therefore centromeres diffused along the length of the entire chromosome (Allshire & Karpen, 2008) [***Figure 2***]. *Drosophila* chromosomes are monocentric, meaning that they have one centromere per chromosome. The centromeres are embedded in large regions of repetitive noncoding heterochromatin. The region can be generally separated into two parts: the core centromere and pericentromere. The core centromere contains Cid nucleosomes and forms the base of the kinetochore complex formation, while the pericentromere flanks this region on both sides [***Figure 3***]. This region does not contain Cid and is generally formed by large numbers of repeating sequences, termed satellite DNA. Satellite sequences are repetitive sequences that form regions of heterochromatin that are up to several megabases long. These sequences form higher order repeats (HOR) that can contain thousands of repeat units, in either head-to-head or head-to-tail orientation (Waye & Willard, 1985; Corless et al., 2020). The repetitive nature of these regions has resulted in their absence from genome assemblies, which has notably impaired the understanding of their function. Only recently there has been a breakthrough and the full sequence of human centromere was published by the T2T consortium (Nurk et al., 2022). Currently, the full sequence of *Drosophila* centromere is yet to be published, but the core centromere sequences are available (Chang et al., 2019a). Comparison of the *Drosophila* and human centromere sequences can be made based on this information. While human centromeres are embedded in satellite sequences, *Drosophila* centromeres are surrounded by islands of mobile elements, or transposons, with lower number of satellites interspersed between them. Each chromosome has different numbers and ratios

of transposons except G2/Jockey-3, which is present on all chromosomes (Chang et al., 2019a; Hartley & O'Neill, 2019).

DNA in (peri)centromeric regions are subject to high mutagenic pressure leading to the diversity in the centromere sequences and kinetochore proteins even among closely related organisms (Malik, 2009; Sullivan et al., 2011; Kursel & Malik, 2018). It also has implications for meiotic drive, where some sequences are preferentially selected by an oocyte (Chmátal et al., 2014). It has been suggested that the rapid evolution of the centromeric proteins mitigates this effect (Henikoff et al., 2001). This leads to a dynamic equilibrium and coevolution of centromeric sequences together with the associated proteins. Side effects of this competition are the increasing size of centromeric as well as pericentromeric regions and higher mutation rate compared to the rest of the genome (Malik & Henikoff, 2009). It was also suggested as a possible mechanism for reproductive isolation of nascent species (Henikoff et al., 2001). Although noncoding centromeric sequences may appear to be parasitic DNA, they play a crucial role in maintaining chromosome stability and kinetochore formation. Deleting centromeres can result in chromosome instability, segregation defects, and the formation of new centromeres, or neocentromeres, in different regions of the chromosome (Talbert & Henikoff, 2020; Murillo-Pineda et al., 2021). Especially the neocentromeres have been observed in various organisms, and while the specific phenotypes may differ, the outcome generally remains identical (Thakur & Sanyal, 2013). The neocentromeres are formed through CENP-A deposition, rather than the presence of satellite DNA or any other repetitive DNA, as observed in yeast and other fungi (Ishii et al., 2008; Ketel et al., 2009; Schotanus & Heitman, 2020), chicken (Shang et al., 2013) and humans (Murillo-Pineda et al., 2021). Furthermore, the neocentromeres prefer repetitive regions, although they are not strictly necessary (Alonso et al., 2010; Logsdon et al., 2019). This information is in line with the fact that centromeres are defined epigenetically and do not depend on the underlying sequence.

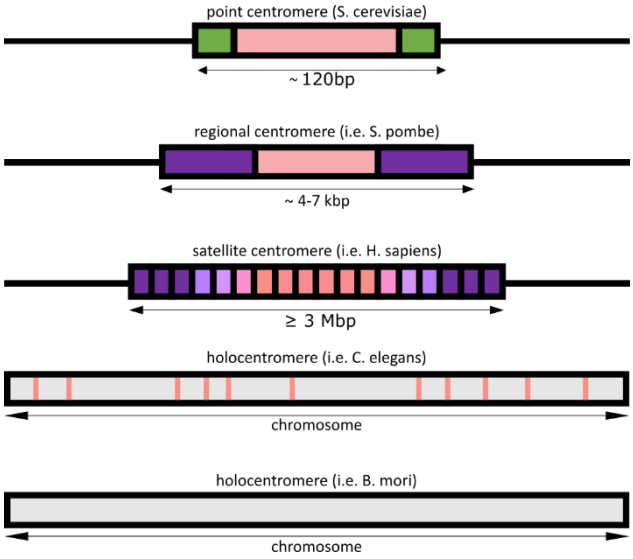


Figure 2. **Representation of mono- and holocentromeres in various organisms**

The simplest centromere (pink), or point centromere, contains only one Cenp-A nucleosome in 120bp sequence. Regional centromere is several kbp long and contain several Cenp-A nucleosomes surrounded by pericentromeric sequences. Holocentromeres have Cenp-A nucleosomes dispersed along the entire chromosome arms, either in groups (*C. elegans* in this picture) or completely dispersed (*B. mori* in this picture). Adapted from (Steiner & Henikoff, 2015)

1.3. Centromeric RNAs

For a long time, it was believed that centromeres were transcriptionally silent. However, recent studies have shown that this is not the case. There is an active form of RNApolII present in the centromeric regions, although there are no active protein coding genes (Saffery et al., 2012). Instead, transcripts of the satellite regions form noncoding RNAs of various lengths, which are generally referred to as long noncoding RNA (lncRNA) or specifically as centromeric RNA (cenRNA), such as satellite III (SatIII) I used for most of this work. The transcripts were found to colocalise with inner kinetochore proteins (Rieder, 1979) and chromatin in general (Huang & Bonner, 1965; Holmes et al., 1972). Despite the fact that the transcripts have been discovered already, their origin was unclear. Recently it was shown that RNAs play a role in both structure and function of the chromatin as it recruits chromatin remodelling enzymes as well as structural and kinetochore proteins (Sawyer & Dundr, 2017; Thakur & Henikoff, 2020). Additionally, they are necessary for proper kinetochore function, namely chromosome segregation and regulation of heterochromatin (Johnson & Straight, 2017). The absence of these transcripts leads to chromosome segregation defects (Rošić et al., 2014; Ling & Yuen,

2019). The ongoing discussion in the field concerns whether the transcripts themselves have a function or if it is solely the act of transcription itself and the resulting chromatin rearrangements that are crucial for proper kinetochore function. Some studies suggest that cenRNA is required for the recruitment of kinetochore proteins, but a clear mechanism of the recruitment has not yet been described (Wong et al., 2007; Quénet & Dalal, 2014; Rošić et al., 2014; Blower, 2016; McNulty et al., 2017). However, it is clear that the centromeric transcription is necessary for Cenp-A deposition (Bobkov et al., 2018). Additionally, it was demonstrated that cenRNA can partially restore the centromere function of another chromosome after it has been impaired by a knock down of centromeric proteins or the original cenRNA (Wong et al., 2007; Rošić et al., 2014). The levels of cenRNA expression change upon various stimuli. For example, it has shown that stress increases the production of cenRNAs (Valgardsdottir et al., 2005, 2008; Hédouin et al., 2017). Similarly, some cancers also facilitate increased cenRNA expression (Ting et al., 2011; Bersani et al., 2015).

Various possible functions of cenRNA have been proposed, including signalling, scaffolding, guiding, tethering and phase separation (Corless et al., 2020). Moreover, there is increasing evidence suggesting that cenRNA carries various RNA modifications, such as 6-methylation of adenine (Ninomiya et al., 2021) or capping with 7-methylguanine (Choi et al., 2011), that are important for its proper function (Arunkumar & Melters, 2020; Ninomiya et al., 2023). These modifications are known to regulate heat shock stress response (Ninomiya et al., 2021), but may also be responsible for other functions, such as regulation of the interaction with kinetochore proteins.

Due to their repetitive nature and low natural abundance, working with cenRNAs can be challenging. One of the difficulties is that cenRNA sequences vary not only between different organisms, but also between different chromosomes of a single organism (Ideue & Tani, 2020). Even closely related species, such as *Drosophila* genus, can have very different centromeric sequences (Talbert et al., 2018). Recent developments in single molecule sequencing methods have begun to provide further insights, but much remains unknown (Leger et al., 2021; Ohshiro et al., 2021).

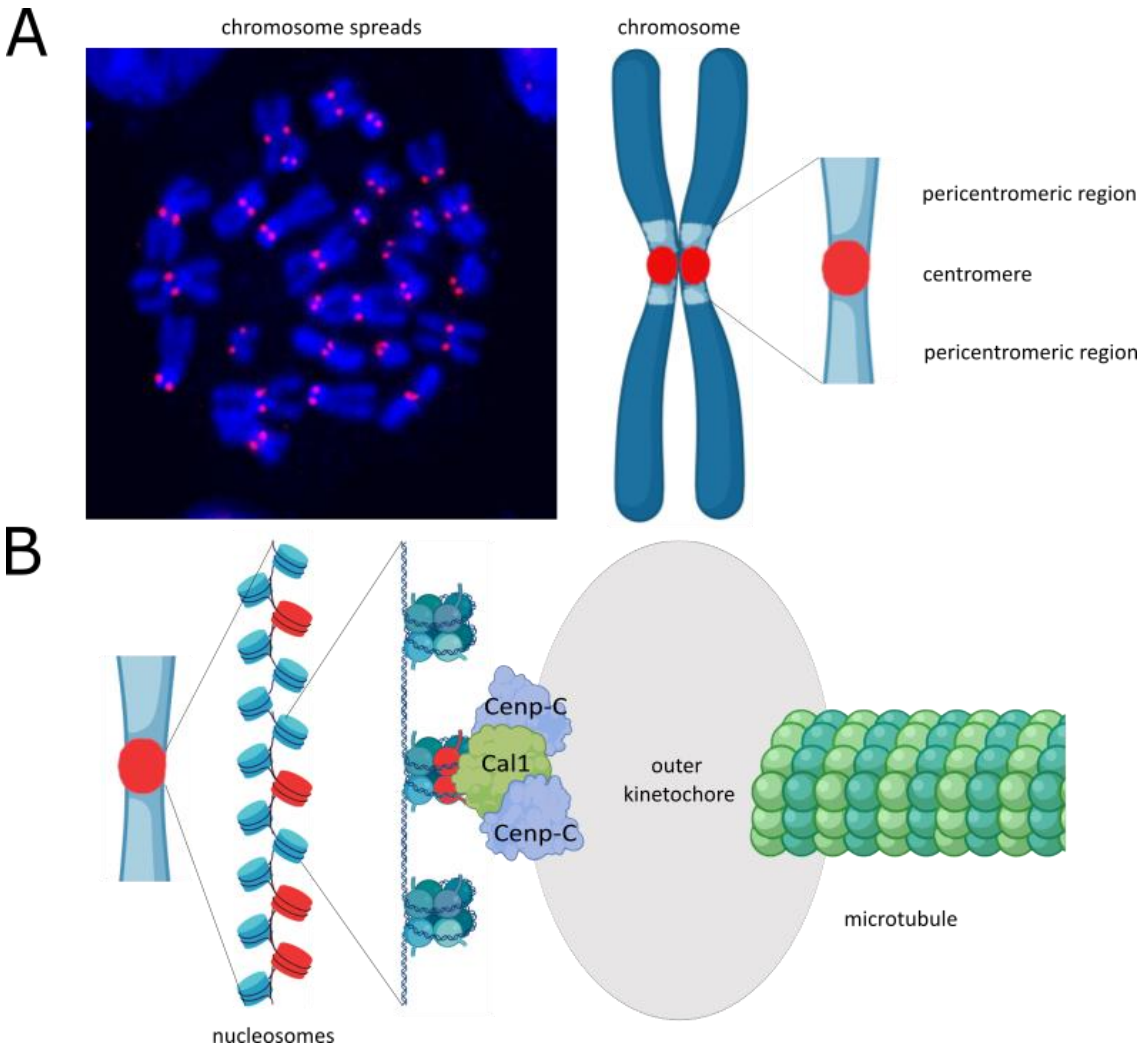


Figure 3. **Schematic representation of *D. melanogaster* chromosomes and a kinetochore.**

(A) Left: a micrograph of mitotic chromosome spreads from *Drosophila* Schneider 2 (S2) cells. DNA in blue, Cid in red. Right: Cid is an indicator of centromeric chromatin.
(B) Left: a close-up schematic of the centromeric region of a fruit fly chromosome. The Cid-containing centromeric region is embedded within a wider pericentromeric region. Cid-nucleosomes are exclusively present in the centromeric region and are interspersed with classical H3 nucleosomes. They serve as a foundation for kinetochore formation by binding to Cal1 and Cenp-C proteins, which in turn bind the outer kinetochore proteins responsible for microtubule binding.
Adapted from (Chang et al., 2019b; Medina-Pritchard et al., 2020), a micrograph of chromosome spreads was kindly provided by Sylvia Erhardt.
Created using BioRender (https://www.biorender.com).

1.4. Inner kinetochore proteins of *Drosophila melanogaster*

The process of aligning and segregating chromosomes during cell division is facilitated by motor proteins and the microtubule cytoskeleton network. The gap between the chromosome and the microtubule is physically bridged by a multiprotein complex called the kinetochore. Kinetochore proteins recognise centromeric regions and assemble the complex that connects them to the microtubules (Chan et al., 2005; Cheeseman, 2014).

The kinetochore complex can be structurally divided into two sub-complexes: the core (inner) kinetochore complex and the outer kinetochore complex, also called the KMN (Knl1-Mis12-Ndc80) network [***Figure 4 A***]. While both sub-complexes are compositionally simpler in *Drosophila* than in human, the major difference lies in the inner kinetochore complex. In *Drosophila*, it consists of only three proteins, Cid, Cal1 and Cenp-C**.** While Cid marks the position of the centromere, Cenp-C and Cal1 maintain the integrity of the inner kinetochore complex (Kyriacou & Heun, 2023). Even in this minimalistic setup, the function of the kinetochore is preserved, making *Drosophila* a suitable model organism for centromere studies. All three of these proteins have orthologues in other species, even though the sequence similarity can be very divergent. It has been proposed that the divergence may be driven by the rapid evolution of the underlying satellite sequences in the centromeric chromatin (Cooper & Henikoff, 2004; Malik, 2009). Loading of CENP-A onto the chromosome is also simpler in flies. In humans, CENP-A deposition occurs in the early G1 phase and requires the MIS18 complex in cooperation with the HJURP chaperone (Jansen et al., 2007; Black & Cleveland, 2011). On the other hand, Cid loading only requires the chaperone Cal1 and Cenp-C [***Figure 4 B***] (Mellone et al., 2011).

The simplicity of the *Drosophila* kinetochore is further underlined by the lack of centromere-associated proteins that are present in human cells. The consecutive centromere associated network (CCAN), absent in *Drosophila*, contains numerous centromeric proteins and their associated complexes (CENP-B to CENP-W). These proteins connect the chromosome with the outer kinetochore KMN network, which in turn binds to the spindle microtubules (Kyriacou & Heun, 2023) [***Figure 4 C, D***].

There are published structures of most of the CCAN proteins in humans (Pesenti et al., 2022), including a cryoEM structure of the entire complex (Yan et al., 2019), whereas no

such resources exist for *Drosophila*. This makes it an interesting problem, and this thesis aims to further our understanding of kinetochore assembly in *Drosophila*.

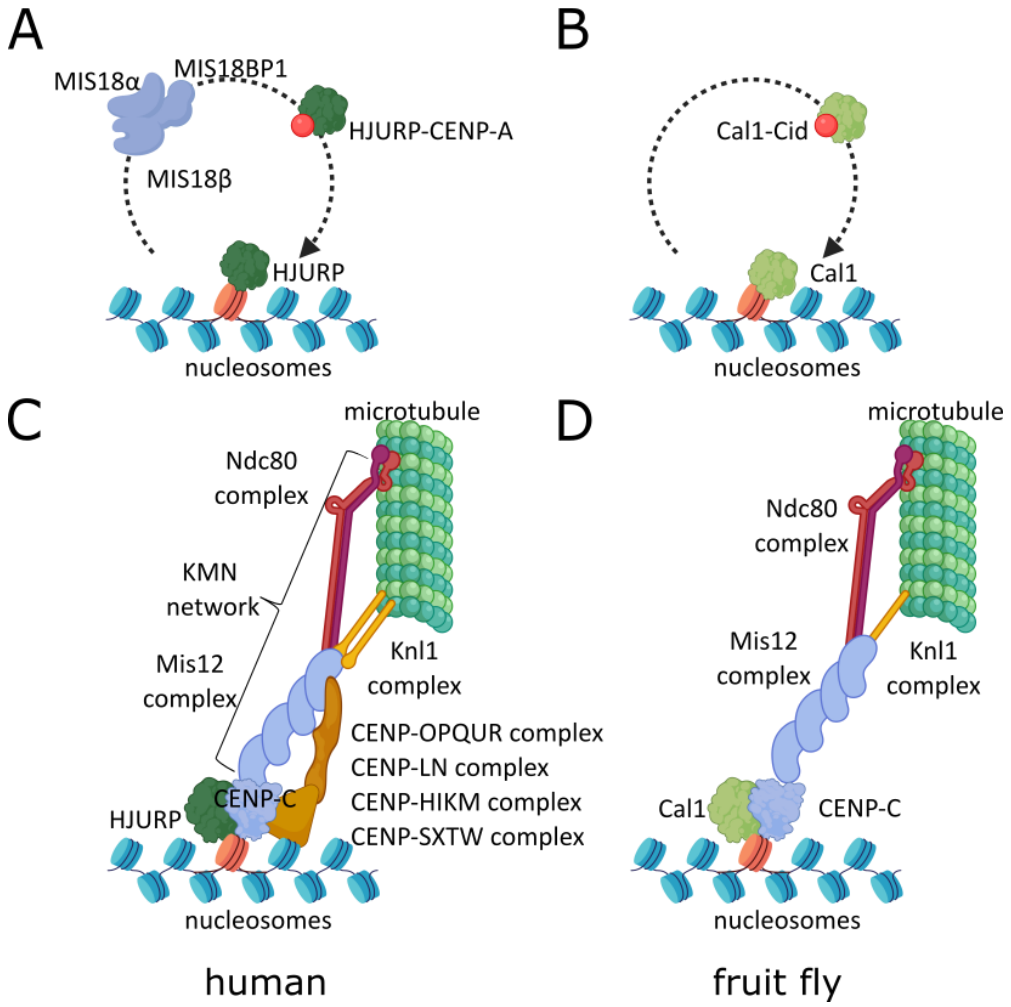


Figure 4. **Comparison of human and *Drosophila* kinetochore complexes**

(A) CENP-A loading in human cells. The MIS18 complex, in cooperation with HJURP, deposits CENP-A onto the nucleosome.
(B) Cid loading in fly cells. Cal1 deposits Cid onto the nucleosome in cooperation with Cenp-C. There is no equivalent of the MIS18 complex required.
(C) A schematic representation of the human kinetochore complex. Inner kinetochore proteins, known as CENP, followed by a one letter code, form the consecutive centromere associated network. This multiprotein complex binds to the outer kinetochore KMN network, which in turn binds the microtubules.
(D) A schematic representation of the fly kinetochore complex. In contrast to the human kinetochore, most of the inner kinetochore complex proteins are absent. The KMN connecting function is maintained only through Cal1 and Cenp-C.
Adapted from (Kyriacou & Heun, 2023), created using BioRender (https://www.biorender.com)

1.4.1. Cid

Cid, also known as centromere identifier in *Drosophila*, is a histone H3 variant that is specific to centromeric chromatin (Henikoff et al., 2000). This highly conserved protein has homologues in all screened eukaryotic organisms. The general name for all of them is Cenp-A, or centromeric variant of histone H3, although they have species specific names as well, such as Cid in *D. melanogaster*. Most commonly it is referred to as CENP-A, or centromeric protein A (Stellfox et al., 2013). The structural similarity between species is highly conserved, although the sequence can vary greatly. For example, the Cenp-A protein in flies and humans share 42% sequence identity but have a RMSD of 1.221 Å. RMSD stands for the root square of standard deviation of positions of atoms in compared structures, the lower the score the higher the similarity. RMSD score below 2 Å is considered very similar (Kufareva & Abagyan, 2012). Comparably, the Cenp-A protein in flies and mice share 35% sequence identity, but have a RMSD of 1.482 Å, and between flies and yeast, the sequence identity is 39% with a RMSD of 1.184 Å (PDBeFold). Structurally, Cid consists of a conserved C-terminal histone-fold domain and a long unstructured N-terminal chain [***Figure 5***]. The N-terminal chain reaches out of the nucleosome complex and is responsible for directly binding Cal1 and Cenp-C (described in next two chapters) (Carroll et al., 2010). While structures of CENP-A nucleosomes in humans, mice and *Xenopus* have been solved, only parts of Cid have been published (Tachiwana et al., 2011; Zhou et al., 2019; Boopathi et al., 2020; Medina-Pritchard et al., 2020). Notably, the N-terminal part of the protein lacks a predicted structure and is absent from the solved structures [***Figure 6***].

Cid is responsible for determining the position of the centromere. Overexpression of Cid leads to ectopic binding and formation of neocentromeres along the chromosome arms, resulting in increased stress, chromosome damage, and segregation defects (Heun et al., 2006; Fukagawa & Earnshaw, 2014). This process can be partially mitigated if there is insufficient amount of endogenous Cal1, as it is necessary for its deposition (Schittenhelm et al., 2010).

Cid carries several post-translational modifications forming a part of the so-called histone code (Jenuwein & Allis, 2001), such as monoubiquitination (Bade et al., 2014). This functions as an epigenetic marker that helps to define the centromere independently of

the underlying DNA sequence and facilitates interaction with the other centromeric proteins, such as Cenp-C.

Deposition of Cid depends on the presence of the other two *Drosophila* centromeric proteins, Cal1 and Cenp-C (C.-C. Chen et al., 2014; Erhardt et al., 2008). Prior to its deposition, the canonical histones H3.1 and H3.3 occupy its position. They are deposited there during the S phase (Sullivan & Karpen, 2004; Dunleavy et al., 2011). Cid remains associated with the centromere throughout the cell cycle, but it needs to be replenished after each division, because it gets diluted twice by each replication (Vafa & Sullivan, 1997). The timing of Cid deposition depends on the cell type. In Schneider 2 (S2) cells, new Cid is deposited during metaphase, while in embryos it is deposited during anaphase (Goshima et al., 2007; Jansen et al., 2007). Deposition of Cid to the nucleosomes occurs during G1 phase, while it is synthesised during G2 phase. Therefore, it is independent of DNA replication (Mellone et al., 2011).

1.4.2. Cal1

In *Drosophila* Cid is loaded onto the centromere by a loading factor called Cal1, or chromosome alignment defect 1 (Phansalkar et al., 2012). It is an ortholog of HJURP, or Holliday junction recognition protein, which loads the CENP-A in humans, and Scm3, or Suppressor of chromosome missegregation, which fulfils the same function in budding yeast (Hayashi et al., 2004; Shivaraju et al., 2011). Despite having two identified functional domains, Cal1 is largely unstructured [***Figure 5***]. Cal1 binds Cid with its N-terminal Scm3-like domain and Cenp-C with its C-terminal domain (C.-C. Chen et al., 2014; Unhavaithaya & Orr-Weaver, 2013). The Scm3-like domain is predicted to have a conserved amino acid composition similar to human HJURP and budding yeast Scm3 by structure prediction software such as Phyre2 (Kelley et al., 2015). It has been proposed that Cal1 is the product of convergent evolution because there is very little sequence similarity between Cal1 and HJURP. Only small parts of the protein have been crystallised in complex with the Cid-nucleosome complex or with the Cenp-C cupin domain (Medina-Pritchard et al., 2020) [***Figure 6***].

In humans, CENP-A is targeted to centromeres by the HJURP and MIS18 complex. *Drosophila* species lack these proteins and instead rely on the Cal1-Cenp-C complex for their function (Phansalkar et al., 2012). Cal1, although not structurally similar to HJURP,

recognises Cid structurally and forms a complex with it. When a new Cid is produced, it is stored in a complex with H4 and Cal1 in the cytoplasm until it is needed after DNA replication. After transporting the Cal1-Cid-H4 complex into the nucleus, Cal1, in cooperation with Cenp-C, deposits Cid onto the chromosome. Together with Cid-containing nucleosomes and Cenp-C, Cal1 forms the basis of the inner kinetochore complex. Cal1 has the capacity to form oligomers which suggests a possible mechanism for bridging several Cid nucleosomes and forming a wider network that connects the chromosome to spindle microtubules (Roure et al., 2019). Cid is also monoubiquitylated in Cal1 dependent manner. This is necessary for the stability of the Cal1-Cid complex and deletion of the responsible SUMO ligase leads to severe segregation defects (Bade et al., 2014).

1.4.3. Cenp-C

Cenp-C, also known as centromeric protein C, is the core protein of the inner kinetochore complex in fruit flies (Heeger et al., 2005). It is stably associated with centromeric chromatin and connects other kinetochore proteins together to form the inner kinetochore complex (Mellone et al., 2011). In *Drosophila*, it is the only protein conserved from the entire constitutive centromere associated network complex (Orr & Sunkel, 2011; Rosin & Mellone, 2017). Cenp-C has orthologues in humans (Earnshaw & Rothfield, 1985; Saitoh et al., 1992), yeast (Brown, 1995; Meluh & Koshland, 1995), nematodes (Moore & Roth, 2001) and plants (Dawe et al., 1999). It binds the Cal1-Cid nucleosome complex on one side and the outer kinetochore proteins of the KMN complex on the other (Mellone et al., 2011). In addition to the Cal1 binding domain at the C-terminus, there are other recognized domains in Cenp-C. Close to the N-terminus there is an arginine rich region responsible for binding the outer kinetochore proteins. Moreover, it contains several conserved regions that are homologous to other *Drosophila* species and other organisms [***Figure 5***]. Furthermore, it contains two AT hook that are predicted to bind nucleic acids (Bayer et al., 2005; Filarsky et al., 2015; Medina-Pritchard et al., 2020). It has been demonstrated that this protein is associated with RNA in humans (Du et al., 2010a; McNulty et al., 2017; Quénet & Dalal, 2014), *Xenopus* (Grenfell et al., 2016) and *Drosophila* (Rošić et al., 2014). However, the only region of the protein with a known

structure is the cupin domain, responsible for its dimerisation (Chik et al., 2019). The remainder of the protein is predicted to be unstructured [***Figure 6***].

Depleting CENP-C has little effect on CENP-A localisation in humans. However, depleting its homologue has a detrimental effect on the Cid localisation in *Drosophila*, such that Cid is no longer deposited on the centromeres (Carroll et al., 2010; Erhardt et al., 2008).

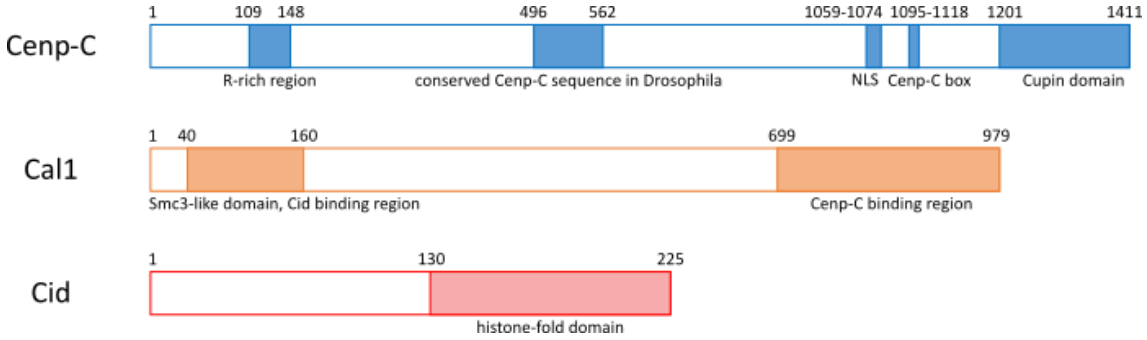


Figure 5. **Schematic view of *D. melanogaster* inner kinetochore proteins**

Cenp-C has several regions with identified functions. The arginine rich region is important for binding to outer kinetochore proteins. The conserved region in the middle is shared with other *Drosophila* species. NLS stands for nuclear localisation sequence. The Cenp-C box is a conserved region of Cenp-C proteins found in different species. The C-terminal cupin domain is responsible for dimerization of Cenp-C and binding to Cal1.
Cal1 has two identified regions of importance. The C-terminal part binds Cenp-C, while the N-terminal part binds Cid.
Cid is notably shorter than the other two proteins and is characterised by the histone fold domain at the C-terminus. The unstructured N-terminal tail is responsible for binding Cal1 and Cenp-C.
Adapted from (Medina-Pritchard et al., 2020)

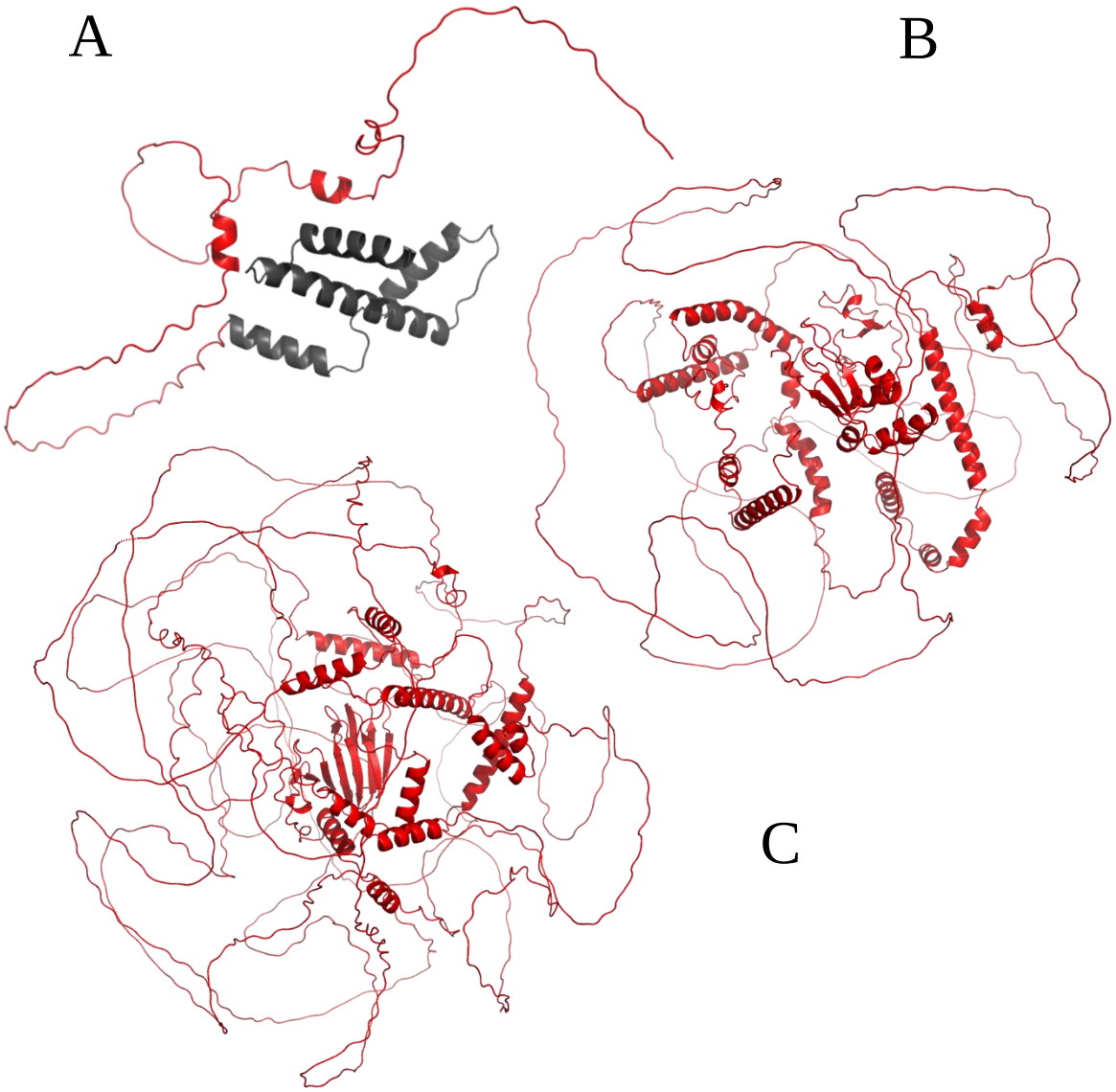


Figure 6. **Structure prediction of *Drosophila* inner kinetochore proteins by AlphaFold**

(A) Prediction of Cid structure (AlphaFold structure code Q9V6Q2). The histone-fold domain (shown in black) is a structurally conserved region of Cid, as well as other histones. This part of the protein has been predicted with a high degree of confidence (pLDDT > 90). The N-terminal tail domain (shown in red) does not have a predicted stable structure and the prediction confidence is accordingly low (pLDDT < 50) (Cooper & Henikoff, 2004).
(B) Prediction of Cal1 structure (AlphaFold structure code Q9VEN2). No stable domains are predicted, only a few secondary structure motifs in the center have higher prediction confidence (pLDDT > 70). Most of the protein does not have any predicted stable structure and has low prediction confidence (pLDDT < 50).
(C) Prediction of Cenp-C structure (AlphaFold structure code Q9VHP9). The algorithm shows that an even larger portion of the protein is predicted to be a random coil with low prediction confidence (pLDDT < 50). The only structured domain known in Cenp-C is a small β-sheet cupin domain located in the center of the structure. This domain is also the only part predicted with higher prediction confidence (pLDDT > 90) (Chik et al., 2019; Medina-Pritchard et al., 2020).
Structures were predicted using AlphaFold (Jumper et al., 2021) https://alphafold.ebi.ac.uk

1.5. Structure determination by mass spectrometry

Mass spectrometry is a reliable and fast method for chemical and biochemical analysis (Maher et al., 2015). It allows for very precise measurements across a wide range of masses, including biomolecules and enables identification of their chemical and biochemical modifications (Domon & Aebersold, 2006).

Hydrogen-deuterium exchange mass spectrometry (HDX MS), or more correctly proton-deuteron exchange ($^{1}H/^{2}H$), is a modification of this method for structural biology applications. Proteins that cannot be crystallised and that are too large or aggregate too fast for nuclear magnetic resonance spectroscopy and too small or too flexible for cryoelectron microscopy, have been challenging to analyse (Trivedi & Nagarajaram, 2022). HDX MS partially addresses this issue (Konermann et al., 2011; Ozohanics & Ambrus, 2020), as it provides insights into the protein structure, as well as the localisation of the interaction surface with binding partners [***Figure 7***]. Another important advantage of this method is that it can be used for examining intrinsically disordered proteins, making it uniquely suitable for investigating centromeric proteins (Karch et al., 2018; Mitra, 2021). Further developing this approach by adding RNA to the reaction mixture could be the ideal way how to solve the lack of structural data of intrinsically disordered centromeric proteins and to map the protein-RNA interactions there.

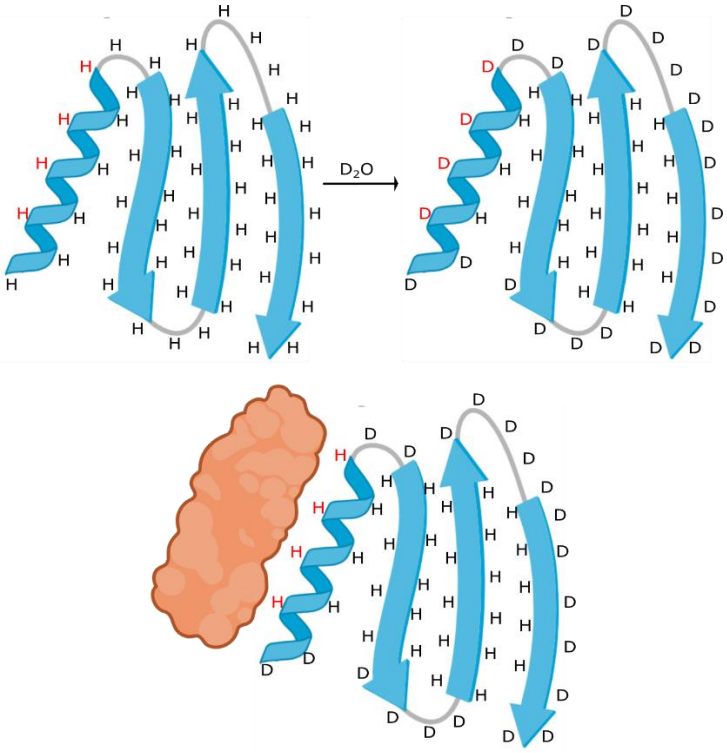


Figure 7. **A schematic representation of $^{1}H/^{2}H$ exchange mechanism**

A protein is placed in a buffer containing only heavy water (D_2O). The surface of the protein will freely exchange all acidic, basic, and amidic protons with the solvent. Acidic and basic protons exchange rapidly, while amidic protons exchange slower. The exchange rate of amidic protons can be significantly reduced by acidifying the environment. This enables analysis by mass spectrometry. Following $^{1}H/^{2}H$ exchange, the protein is digested by a specific protease and loaded onto the mass spectrometer. Deuterated peptides have a higher mass and therefore produce a different mass spectrum. By comparing non-deuterated and deuterated samples, it is possible to determine which parts of the protein form the solvent-accessible surface. If a binding partner blocks part of the surface, as shown in the lower part of the picture, $^{1}H/^{2}H$ exchange is also blocked. Comparing the blocked sample with the free one provides structural information about the interaction (Ozohanics & Ambrus, 2020).
The schematic was created using BioRender (https://www.biorender.com)

Aims of the thesis

It is evident that the proper function of Drosophila centromere depends on the transcripts of the pericentromeric DNA, or satellite RNAs, such as SatIII. However, the exact function and molecular structural details of the interaction between centromeric proteins and SatIII RNA have not been described in detail to date. Hence, the aim of my thesis is to shine more light on this interaction by expressing and purifying centromeric proteins from *Drosophila melanogaster* as well as *in vitro* transcribing the SatIII RNAs, and by performing *in vitro* structural experiments. *Drosophila* is the perfect model organism for this experiment, because the inner kinetochore complex consists of only three proteins – Cid, Cal1 and Cenp-C. This simplification in fruit fly inner kinetochore complex can enhance our understanding of the human counterparts and the general role of RNA in kinetochore function.

Since *Drosophila* possesses various centromeric RNAs including SatIII, I also aimed at investigating other possible interaction partners, such as other satellite RNAs (Valent, 2022). The method applied for the *in vitro* analysis of protein-RNA interactions of centromeric proteins and centromeric RNAs is $^{1}H/^{2}H$ exchange mass spectrometry, because it has no theoretical limitations for work with unstructured proteins, like other common structural methods.

The overall aim of my thesis is to give an experimental description of cenRNA-protein interaction in the centromeric chromatin of *Drosophila melanogaster*.

2. Results

2.1. Cloning of expression constructs

I selected three proteins, Cid, Cal1 and Cenp-C, for this project based on previous work in our group, which found out that Cid loading is dependent on SatIII RNA (Rošić et al., 2014). Together they form the inner kinetochore complex in *D. melanogaster* and are essential for proper chromosome segregation during anaphase and kinetochore assembly. In *Drosophila*, they are also the only inner kinetochore proteins, which makes the use of this model system perfect for research of kinetochore properities. Plasmids containing sequences of all three proteins were available in the laboratory, but they needed to be transformed into bacterial or insect expression vectors. Fortunately, both insect and bacterial codon-optimized versions were available, and no further optimization was necessary.

Both Cal1 and Cenp-C are large intrinsically disordered proteins, which complicates their expression, purification, and further work [***Figure 6***]. It soon became apparent that constructing the appropriate expression vectors will be more challenging than initially anticipated. While Cid, being a short protein, was not problematic, Cal1 and Cenp-C, being significantly larger, proved difficult to subclone. I used several available molecular cloning techniques - Gibson assembly (NEB) for Cid, PCR amplification cloning (CPEC) (Quan & Tian, 2009) for full-length constructs of Cal1 and Cenp-C, and HiFi assembly (NEB) for insect cells expression constructs of Cal1 and Cenp-C.

I also designed and created shorter constructs when expression of the full-length proteins was difficult, based on Medina-Pritchard et al., (2020). I used restriction cloning for all the fragments with restriction enzymes depending on the particular fragment. The information about the constructs is listed in ***Table 1***. Cal1 does not have any predicted structured domains, while Cenp-C has only one C-terminal domain and Cid has a histone-fold domain as predicted by Psipred (http://bioinf.cs.ucl.ac.uk/psipred/), Phyre (http://www.sbg.bio.ic.ac.uk/phyre2), and Alphafold (https://alphafold.ebi.ac.uk) [***Figure 5, Figure 6, Appendix 1***]. Therefore, splitting the proteins into smaller parts is unlikely to damage any obvious structure-function relationship.

Bacterial expression and purification of truncated sequences proved to be feasible, and I eventually obtained nearly all of the prepared constructs in quantities and purities that enabled further research [***Table 1***].

Table 1: **Overview of used constructs.**

Name	Length of the gene (bp)	Selected region (AA)	Molecular weight (kDa)	pI	Successfully purified	Concentration (mg/ml)
Cid	675	1-225	26.0	10.20	yes	9.5
Cal1	2940	1-979	109.5	8.30	no	/
Cenp-C	4233	1-1411	159.3	8.60	no	/
Cid ΔN	378	101-225	14.6	10.31	no	/
Cid N-terminus	297	1-100	11.4	9.50	yes	1.4
Cal1N	1224	1-407	47.8	6.06	no	/
Cal1 first 100AA	300	1-100	11	3.84	yes	2.2
Cal1M	993	392-722	36.5	9.56	yes	2-4
Cal1C	846	699-979	30.2	6.99	yes	2.5
Cenp-C1	1692	1-575	63.7	9.67	yes	1.6
Cenp-C2	903	558-857	34.3	5.29	yes	0.6
Cenp-C3	753	857-1106	28.5	8.32	no	/
Cenp-C4	957	1106-1411	36.0	6.87	yes	1.5-2.2

Table 2: **Constructs and optimized conditions of expression tests in *E. coli*.**

Construct	Bacterial strain	IPTG concentration	Temperature	Expression time
SUMO-Cid	BL21DE3	0,1M	16 °C	3 h
SUMO-Cal1	BL21DE3 codon+		20 °C	5 h
SUMO-Cenp-C	Rosetta	1M		
GST-Cid	RIL		25 °C	O/N
MBP-Cal1	pLysS			

2.2. Protein expression in *E. coli*

E. coli is the most commonly usored expression system for the production of recombinant proteins due to the availability of well-established protocols. After I prepared bacterial expression vectors containing my proteins of interest in full-length, expression conditions were tested by transforming BL21DE3 strain of *E. coli.* Cid was successfully expressed under these conditions [***Figure 8 A***]. However, Cal1 and Cenp-C showed poor expression due to their large sizes and low structural organisation [***Figure 6***]. I needed to optimise the purification process by testing different strains of *E. coli* and a wide range of expression conditions. I tested BL21DE3 codon plus, Rosetta, RIL and pLysS expression strains [***Figure 9, Table 2, Appendix 6***.]. I also tested expression at three different temperatures (16, 20 and 25 °C) and two concentrations of IPTG [***Table 2***]. I took samples for further analysis after 3, 5 and 24 h of expression [***Appendix 4, Appendix 5***.]. None of the above-mentioned conditions led to successful expression of these proteins, which is why I switched to the insect expression system.

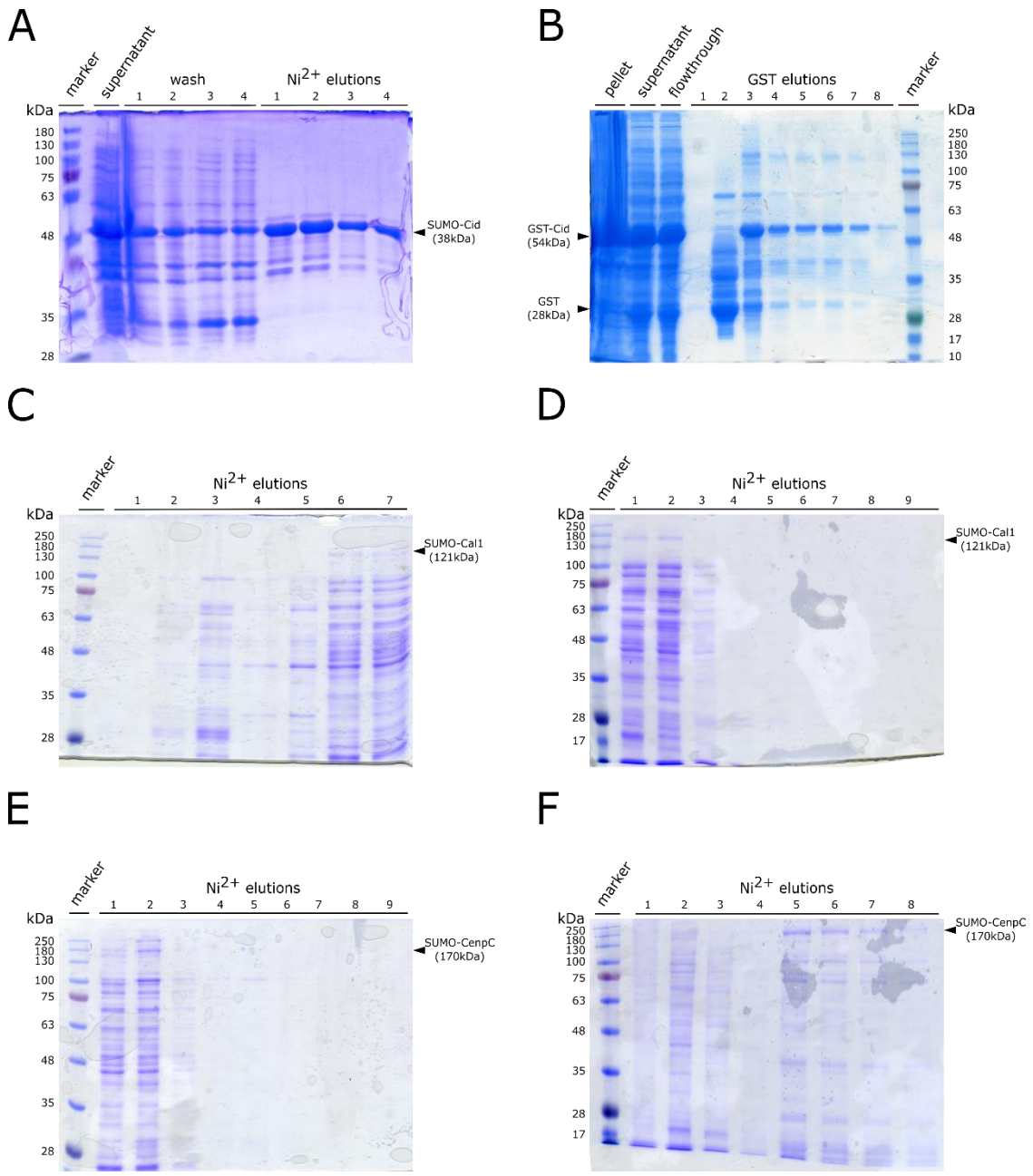


Figure 8. **Centromeric proteins do not express well in E. coli**

All proteins were expressed in the BL21DE3 strain. After harvesting, the samples were lysed and purified using an ÄKTA liquid chromatography system (see materials and methods).
(A) Purified SUMO-Cid (38 kDa).
(B) Purified GST-Cid (54 kDa), with GST alone at 28 kDa.
(C) Purified SUMO-Cal1 (121 kDa), the band corresponding to the protein is not visible.
(D) Purified SUMO-Cal1 (121 kDa), using 3% EtOH in the growth medium. The band corresponding to the protein is not visible.
(E) Purified SUMO-Cenp-C (170kDa). The band corresponding to the protein is not visible.
(F) Purified SUMO-Cenp-C (170kDa), using 3% EtOH in the growth medium. The band corresponding to the protein is not visible.

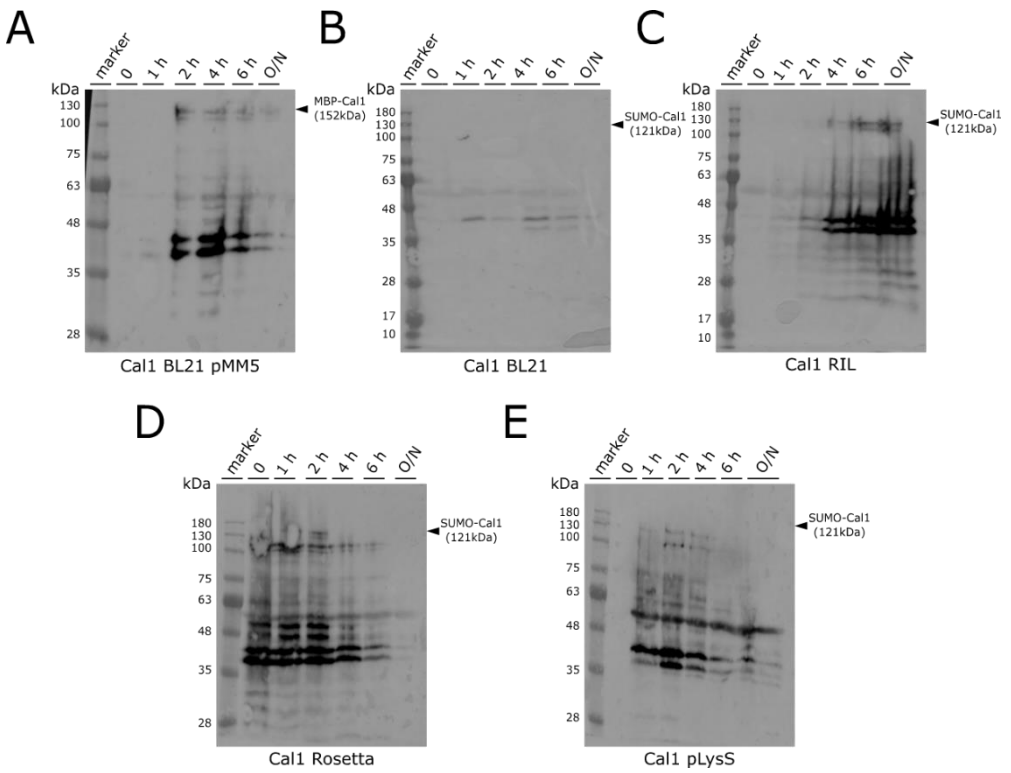


Figure 9. **Cal1 does not express well in various bacterial strains**

Cal1 protein was expressed in several expression strains of bacteria. After harvesting, the samples were lysed and purified using an ÄKTA liquid chromatography system (see materials and methods) and analysed by Western blot (WB).
(A) WB of MBP-Cal1 (152kDa) expressed in BL21DE3 strain over time.
(B) WB of SUMO-Cal1 (121kDa) expressed in BL21DE3 codon plus strain over time.
(C) WB of SUMO-Cal1 (121kDa) expressed in RIL strain over time.
(D) WB of SUMO-Cal1 (121kDa) expressed in Rosetta strain over time.
(E) WB of SUMO-Cal1 (121kDa) expressed in the pLysS strain over time.

2.3. Cal1 expression in insect cells

Since I work with insect proteins and insect expression system is reported to work better for larger and less structured proteins, I expressed Cal1 in Sf21 and High Five insect cell lines which are also evolutionary closer to *D. melanogaster*. Sf21 cell line grows in a monolayer while High Five cells grow in a suspension, resulting in higher protein yields.

The preparation of construct is more complex than for bacterial expression and requires several steps. First, I inserted Cal1 gene, codon-optimised for insect cell expression in pFastBac plasmid (Invitrogen), and used it to transform the packaging strain of *E. coli*, DH10Bac (Invitrogen) with it. These cells produce bacmid DNA (large plasmid containing viral genome) and contain an apparatus to transpose the gene of interest from pFastBac into the bacmid. The bacmid can be purified and directly used for lipofectamine transfection of the expression insect cells. Subsequently, infected insect cells produce authentic baculoviral particles within 72 hours post-transfection. I obtained viral particles from the medium by filtering it through 0.2μm filter. Resulting viral stock can be stored at -80 °C and used to infect insect cells. The protocol I used was developed in Melchior group and based on Scholz & Suppmann, (2017). I used the P1 baculoviral stock to infect a grown culture of insect cells at 60-80% confluency. I harvested the infected cells after 24 h, collected them by centrifugation and lysed them by freezing at -80 °C.

The protein was degrading too quickly during nickel chelation chromatography using cell lysate, so I employed a denaturing purification protocol instead. I used a column with nickel chelating matrix and lysis buffer containing 6M guanidinium hydrochloride to denature all proteins in the solution. After allowing the lysate to flow through the column, I washed it with a gradient of lysis buffer containing urea, ranging from 8M to 0. This step should enable the denatured protein to refold while still being bound to the column. Following extensive washing with the lysis buffer, I eluted the protein by elution buffer containing 300mM imidazole. I measured the concentration of the protein in the eluate using a nanodrop and confirmed the protein quality by SDS-PAGE.

The yield of the purification was low, but the resulting protein was pure. However, the storage conditions were problematic. I flash-froze the protein samples in liquid nitrogen and stored them at -80 °C for several days. (Simpson, 2010) Upon thawing, the protein quality was poor, and it rapidly degraded [***Figure 10***].

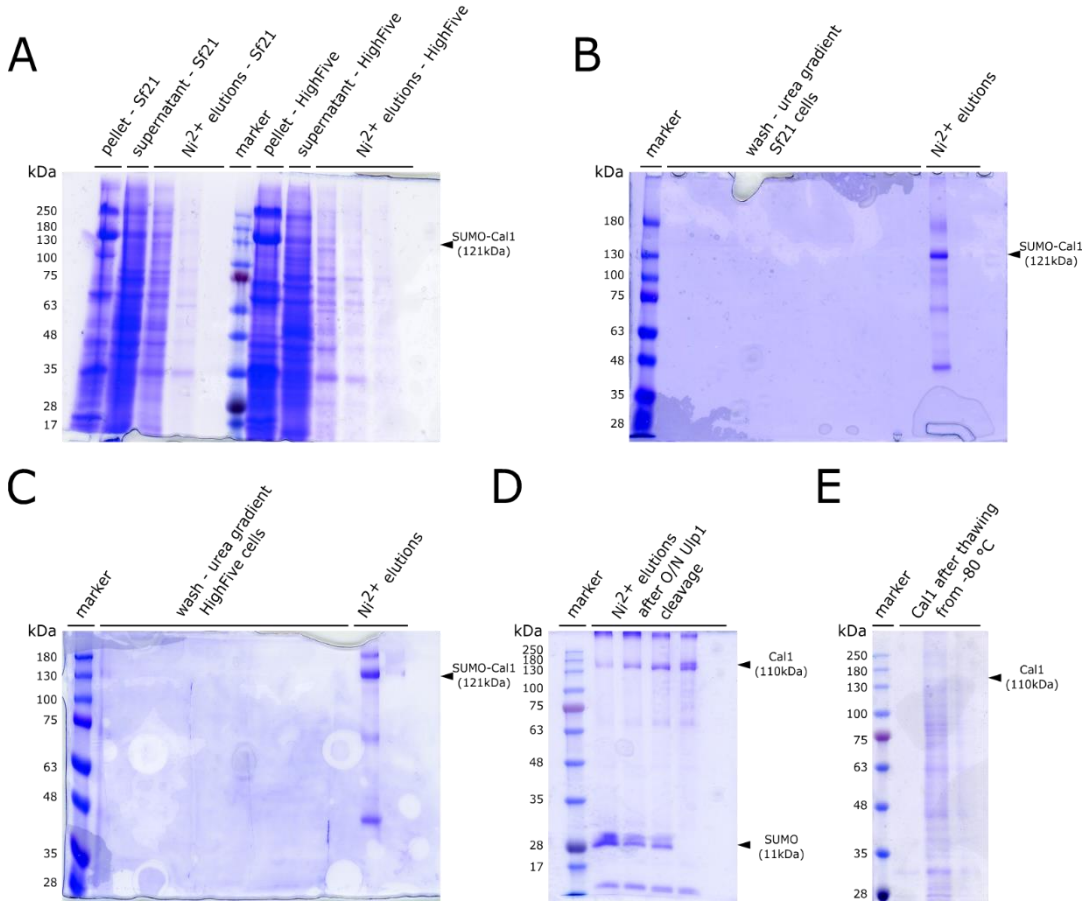


Figure 10. **Cal1 can be expressed and purified from insect cells**

SUMO-Cal1 protein was expressed in Sf21 and HighFive insect cells. After harvesting, the samples were lysed and purified using an ÄKTA liquid chromatography system (see materials and methods).

(A) SUMO-Cal1 (121kDa) purified from Sf21 and HighFive cells using standard protocol.

(B) SUMO-Cal1 (121kDa) purified from Sf21 cells using denaturing conditions.

(C) SUMO-Cal1 (121kDa) purified from HighFive cells using denaturing conditions.

(D) Cal1 (110kDa) after overnight Ulp1 SUMO protease treatment of the purified SUMO-Cal1.

(E) The purified Cal1 after flash-freezing in N_2(l) and subsequent thawing.

2.4. Purification of centromeric proteins and protein fragments

Following initial expression tests, I finalised a functional expression and purification protocol. Full-length Cenp-C and Cal1 expression proved to be problematic, resulting in poor yields. Therefore, I decided to split both proteins into smaller parts and express and purify those instead. I created constructs based on the paper by Medina-Pritchard et al., (2020). The expression of these parts mostly succeeded in *E. coli*.

Bacterial cultures of the BL21DE3 codon plus strain, transformed with the expression plasmid containing the protein of interest, were grown in TB medium until reaching an OD_{600} of 0,8-1 at 37 °C. The temperature was then lowered to the desired expression level of 16 °C and the expression was induced by adding IPTG. The bacterial culture was left to produce the protein under these conditions overnight and harvested by centrifugation in the morning.

I used ÄKTA liquid chromatography system for the purification steps. Initially, nickel chelation affinity chromatography was employed to bind the tagged protein. Following extensive washing with at least 5 column volumes of buffer, the protein was released from the column using a gradient of imidazole, up to 300mM. The eluate from this step was then subjected to buffer exchange to remove the imidazole. After this step, the affinity tag was cleaved by Ulp1 SUMO protease (1µl/ml), and another round of nickel chelation affinity chromatography was performed. The flowthrough was collected this time. I obtained protein of sufficient purity for the subsequent experiments, as shown in [***Figure 11*, *Figure 12***].

I measured the quality and quantity of the purified protein using several complementary methods. Throughout the purification protocol I collected samples for SDS-PAGE to assess protein purity and size.

I used the Cal1M fragment for most of the experiments described in this work. It was due to its stability after purification and apparent affinity for nucleic acids. A representative purification is shown in [***Figure 13***]. AlphaFold confirmed that this part of Cal1 does not have any predicted structure [***Figure 14***]. However, it does possess a net positive charge in biological conditions (18^{+}), which suggests it may be the RNA-binding component of the protein. Cal1N is net charge-neutral and Cal1C is net negative (13^{-}). The properties of the protein parts were predicted using the Expasy ProtParam tool (https://web.expasy.org/protparam/)

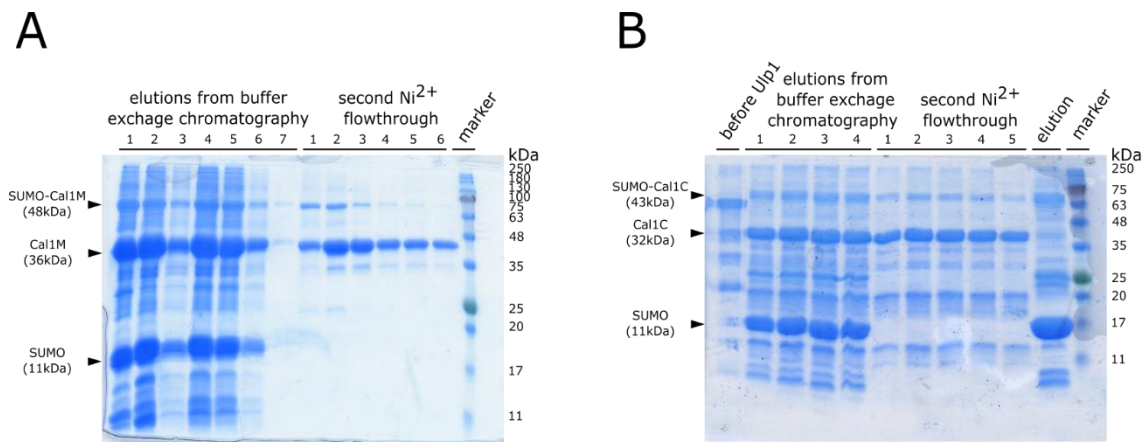


Figure 11. **SUMO-Cal1 fragments can be expressed in and purified from bacteria**

SUMO-Cal1 parts were expressed in BL21DE3 strain. After harvesting, the samples were lysed and purified using an ÄKTA liquid chromatography system (see materials and methods).
(A) SUMO-Cal1M before (48kDa) and after (36kDa) Ulp1 cleavage.
(B) SUMO-Cal1C before (43kDa) and after (32kDa) Ulp1 cleavage.

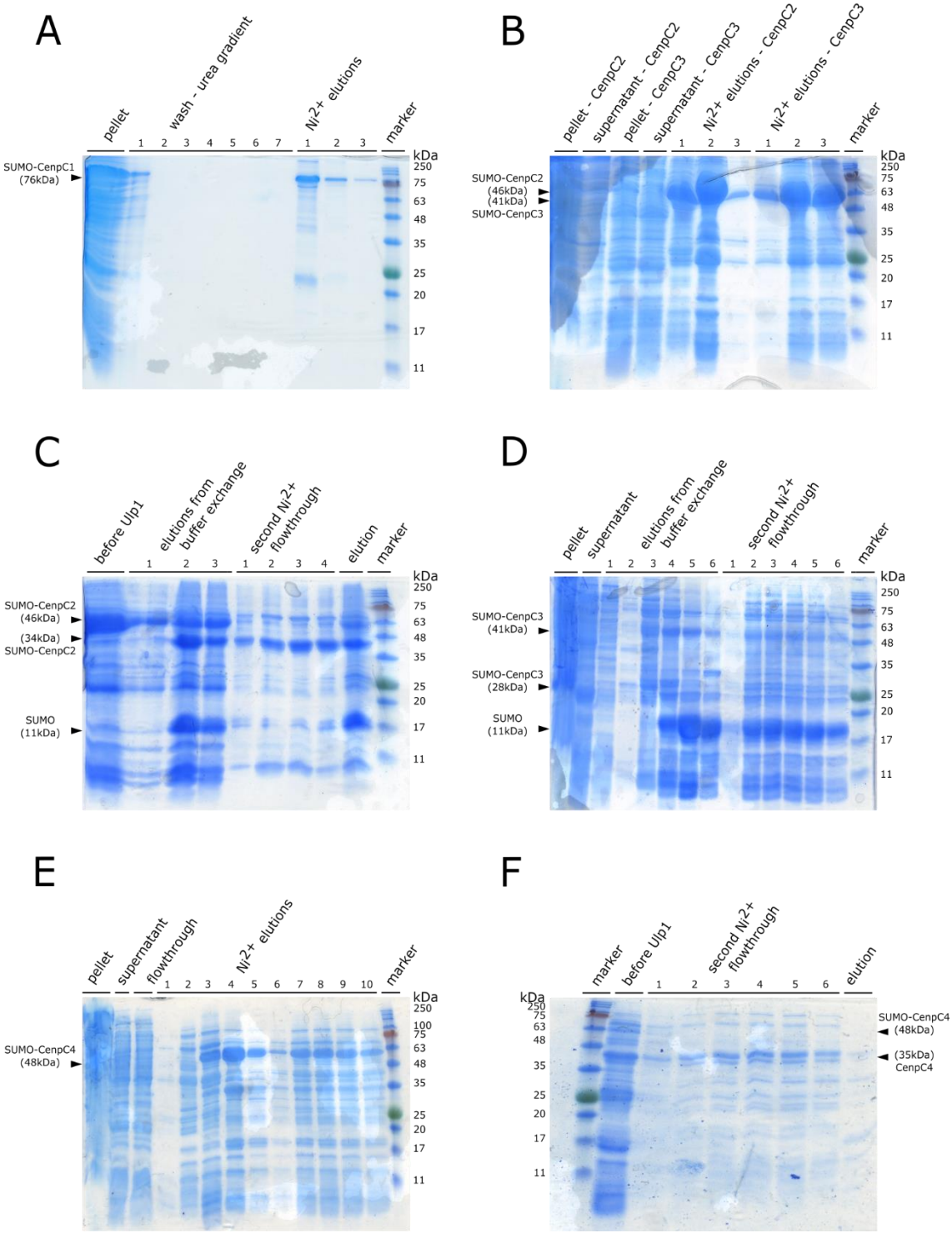

A
pellet
wash - urea gradient
Ni2+ elutions
marker
1 2 3 4 5 6 7
1 2 3
kDa
250
75
63
48
35
25
20
17
11
SUMO-CenpC1
(76kDa)
B
pellet - CenpC2
supernatant - CenpC2
pellet - CenpC3
supernatant - CenpC3
Ni2+ elutions - CenpC2
Ni2+ elutions - CenpC3
marker
1 2 3
1 2 3
SUMO-CenpC2
(46kDa)
(41kDa)
SUMO-CenpC3
C
before Ulp1
elutions from buffer exchange
second Ni2+ flowthrough
elution
marker
1 2 3
1 2 3 4
SUMO-CenpC2
(46kDa)
(34kDa)
SUMO-CenpC2
SUMO
(11kDa)
D
pellet
supernatant
elutions from buffer exchange
second Ni2+ flowthrough
marker
1 2 3 4 5 6
1 2 3 4 5 6
SUMO-CenpC3
(41kDa)
SUMO-CenpC3
(28kDa)
SUMO
(11kDa)
E
pellet
supernatant
flowthrough
Ni2+ elutions
marker
1 2 3 4 5 6 7 8 9 10
kDa
250
100
75
63
48
35
25
20
17
11
SUMO-CenpC4
(48kDa)
F
marker
before Ulp1
second Ni2+ flowthrough
elution
kDa
250
75
63
48
35
25
20
17
11
1 2 3 4 5 6
SUMO-CenpC4
(48kDa)
(35kDa)
CenpC4

Figure 12. **Expression of SUMO-Cenp-C fragments**

SUMO-Cenp-C parts were expressed in BL21DE3 strain. After harvesting, the samples were lysed and purified using an ÄKTA liquid chromatography system (see materials and methods).
(A) SUMO-Cenp-C1 (76kDa) purified using denaturing conditions.
(B) The first step of SUMO-Cenp-C2 (46kDa) and SUMO-Cenp-C3 (41kDa) purification.
(C) The second step of SUMO-Cenp-C2 (46kDa) purification.
(D) The second step of SUMO-Cenp-C3 (41kDa) purification.
(E) The first step of SUMO-Cenp-C4 (48kDa) purification.
(F) The second step of SUMO-Cenp-C4 (48kDa) purification. Cenp-C4 (35kDa) after Ulp1 cleavage.

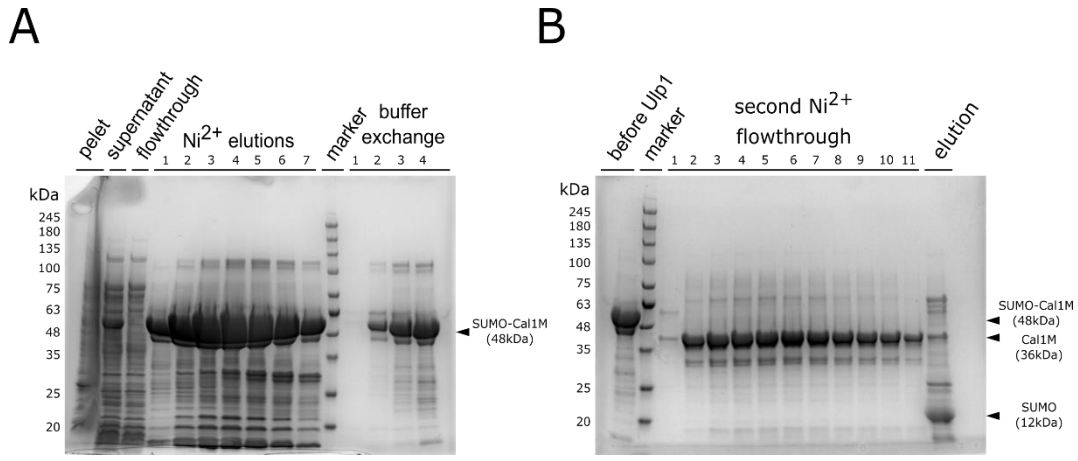


Figure 13. **Purification of SUMO-Cal1M fragment**

SUMO-Cal1M parts were expressed in BL21DE3 strain. After harvesting, the samples were lysed and purified using an ÄKTA liquid chromatography system (see materials and methods).
(A) The SUMO-Cal1M (48kDa) purification.
(B) The Cal1M (36kDa) purification.

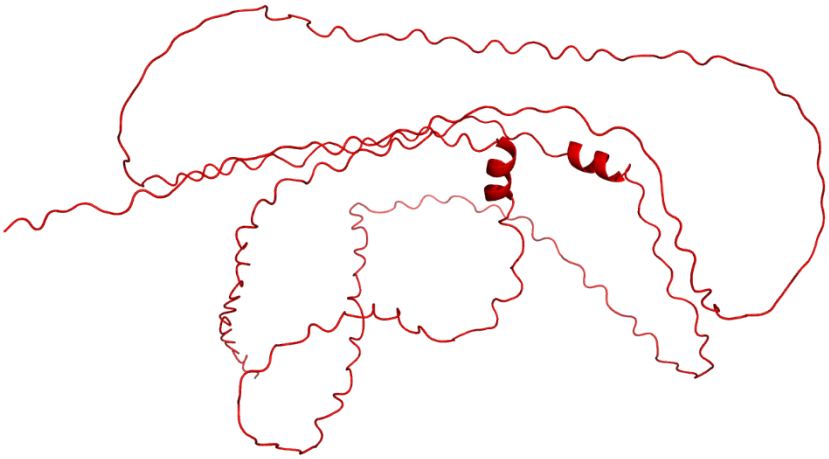


Figure 14. **Cal1M does not have a stable predicted structure**

As predicted by the previous structure analysis (Phyre, PsiPred) Cal1M lacks a stable structure. The protein structure was predicted using AlphaFold, https://alphafold.ebi.ac.uk (Jumper et al., 2021).

2.5. Analysis of the purified protein samples

In the following chapters I will focus at the Cal1M fragment. It was the fragment whose predicted physicochemical properties suggested that it may bind to nucleic acids. The fragment has a high pI value, resulting in a net positive charge in the physiological pH range. This is due to the high content of basic amino acid residues. Preliminary EMSA experiments have confirmed this prediction.

2.5.1. Spectrophotometry

I measured the concentration of the purified protein using spectrophotometry. At first, I used a simple nanodrop measurement at 280 nm, and then I used Bradford assay. The difference in measured concentrations between the two methods is due to the presence of non-negligible amounts of nucleic acids in the purified protein samples. The Bradford assay is specific for the proteins in the measured sample and therefore provides a more precise result, as listed in ***Table 1***.

2.5.2. Circular dichroism

After confirming the size and concentration of the purified proteins, the next step was to assess their structural stability and folding. Structure predictions for Cal1 and its parts indicate highly unstructured regions and a few α-helices, which I confirmed by circular dichroism (CD) spectrometry. Due to the high absorption of glycerol, I could not use a standard phosphate for circular dichroism. The sample was transferred into a 10mM phosphate buffer with a pH of 8.0, containing 500mM NaCl, 5mM imidazole, and 1mM DTT using a buffer exchange column.

The measurement confirmed the stability of the purified protein after freezing and thawing. I used it to investigate the quality of the purified protein and to confirm the presence of at least some minor structured segments [***Figure 15***]. Comparing the measured spectrum to standard curves provides information about the nature of the secondary structure of the Cal1M fragment. Most of the sequence is in a random coil conformation with minor α-helix components, which is in line with the AlphaFold predictions [***Figure 15***].

I also used CD to probe the secondary structure of the purified protein, and to measure its thermal stability. The protein sample was heated continuously from 10 to 85 °C. Thermal unfolding curve of the protein sample was measured at constant wavelength of 222 nm. This wavelength is commonly used because it is the peak absorbance of α-helices. The resulting data were fitted with a sigmoidal melting curve. The melting point of the Cal1M fragment, where half of the protein in the sample is unfolded, was determined by identifying the inflection point of the curve at 53.32 °C. The shape of the curve and realistic melting point suggest that there is some structure that can be disturbed by heating, and the protein is not entirely unstructured in solution [***Figure 16***]. These measurements formed the basis for subsequent analysis using mass spectrometry.

This analysis confirms that the protein sample I purified is of sufficient quality for the following experiments.

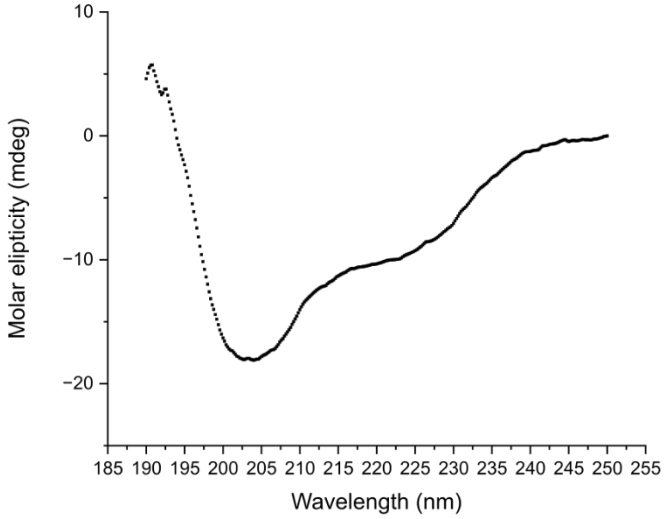


Figure 15. **Cal1M has measurable secondary structure motifs**

CD spectrum of the Cal1M fragment. The curve represents random coil with minor α-helix component.

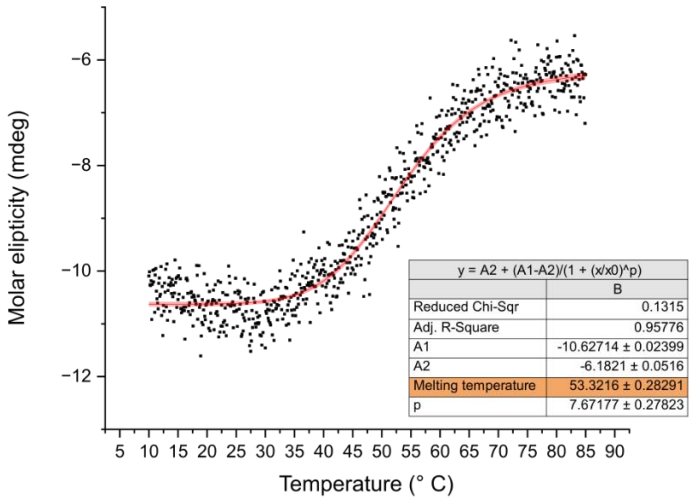


Figure 16. **Cal1M fragment has a structure that can be thermally unfolded**

Thermal unfolding curve of the Cal1M fragment measured at constant wavelength of 222 nm. Melting point was experimentally determined at 53.32 °C.

2.6. Analysing nucleic acid impurities in protein samples

During the initial quality control steps, I discovered that the purified Cal1M sample contains a detectable amounts of nucleic acids. This complicates quality analysis and further experiments. I have tried to remove the impurities during the purification, but without success. Therefore, I have decided to conduct an analysis of the impurities to determine the appropriate course of action. I used proteinase K, as well as DNAse I, RNAse A and RNAse H to investigate the nature of the impurities [***Figure 17 A, B***]. Proteinase K led to digestion of all protein in the sample and left only nucleic acids behind. DNAse I digests all DNA in the sample, but the band on the gel was still visible. RNAse H digests DNA-RNA hybrids and it did not change the band as well. Only RNAseA, digesting RNA, led to disappearance of the gel and proved, that the impurities are in fact RNA. Analysing the impurities by fragment analyser led me to believe it was in fact bacterial rRNA, which was supported by comparing the purified RNA with rRNA from bacteria on an agarose gel [***Figure 17 C, D***].

After confirming the identity of the nucleic acid impurities, I attempted to modify the purification protocol. Although the previous protocol already included nucleases, it was not effective enough. Therefore, I introduced an additional nuclease step into the protocol. By adding RNAse A after the buffer exchange chromatography during the tag cleavage process, I was able to slightly reduce the RNA content in the final purified sample. The issue was that it introduced RNAse A contamination to the sample that proved impossible to mitigate in the later stages of the protocol. This resulted in several failures of RNA-protein binding assays. Therefore, for the remainder of this work, I used the original samples of Cal1M containing rRNA-impurities.

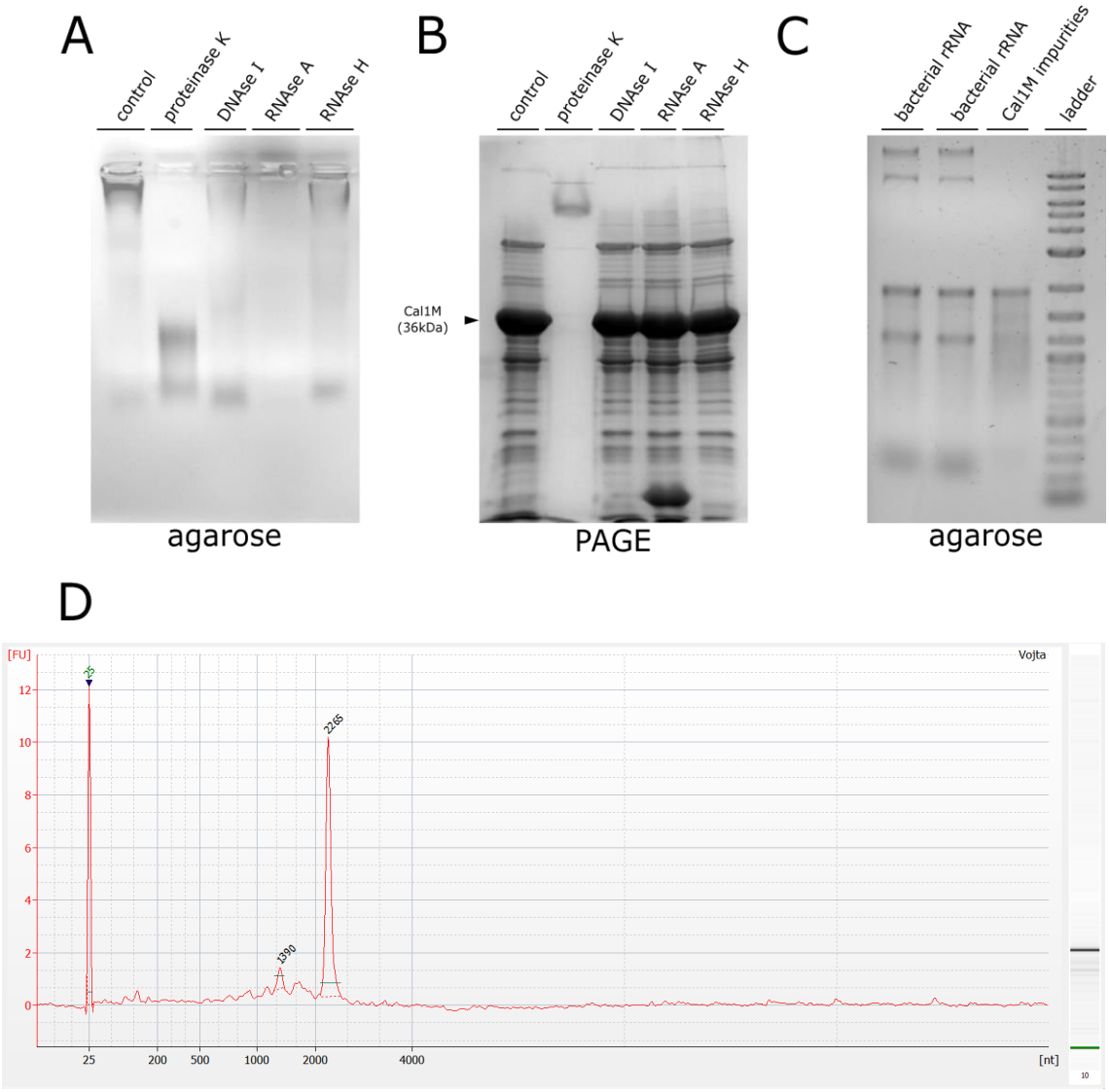


Figure 17. **Quality control of the Cal1M fragment identified RNA contamination**

(A) The agarose gel shows the purified protein sample after incubation at 37 °C with proteinase K, DNAseI, RNAseA, and RNAseH, respectively. The gel was stained with ethidium bromide to detect nucleic acids. Only the addition of RNAseA leads to complete digestion of nucleic acid impurities in the sample.
(B) The SDS-PAGE gel shows the purified protein sample after incubation at 37 °C with proteinase K, DNAseI, RNAseA, and RNAseH, respectively. This gel was stained with QuickStain to detect proteins. The protein remained stable throughout the experiment, and the digestion of nucleic acids did not affect its stability.
(C) The nucleic acid profile of the impurities from the Cal1M sample, purified by phenol/chloroform extraction and analysed by fragment analyser, revealed that the main source of impurities was bacterial rRNA. Despite the nuclease treatment of the bacterial lysate, this RNA remained intact throughout the purification process.

2.7. Investigating the protein-RNA interactions

I prepared all RNAs used in this work using in vitro transcription. The template DNA was prepared by cleaving a plasmid with a restriction endonuclease to create linear DNA, or by PCR amplification of plasmid DNA, or by reverse transcription of RNA extracted from S2 cells. A T7 promoter was attached to each DNA template using custom-made primers.

I prepared the cDNA by extracting RNA from the S2 insect cells using phenol-chloroform extraction, followed by reverse transcription using the Quantitect kit (Qiagen). The plasmid DNA used as a template was linearised through restriction cleavage and used for a reverse transcription reaction using the Quantitect kit. I extracted the rRNA from the bacteria using a protocol specifically designed for rRNA extraction. An overview of the RNAs used can be found in ***Table 3*** below, while the full sequences are listed in ***[Appendix 2]***. I selected several repetitive RNAs from Drosophila, based on the work of Valent, 2022, as well as rRNA and mRNA as controls.

Table 3**: Used RNAs**

name	length (bp)	molecular weight (kDa)	source	type
SatIII sense	718	221	plasmid	lncRNA
SatIII antisense	1436	442	plasmid	lncRNA
copia	4593	1479	PCR from cDNA	lncRNA
bacterial rRNA	1542, 2906	550, 990	bacteria	rRNA
Hsrω	1176	394	plasmid	lncRNA
Sat260	260	83	PCR from cDNA	lncRNA
Sat353	353	112	PCR from cDNA	lncRNA
αTub	1672	537	plasmid	mRNA
CR40469	214	69	PCR from cDNA	lncRNA

Investigating interactions between proteins and nucleic acids has been a long-standing goal of biochemistry due to their importance for regulation of cell processes (Kilchert et al., 2020). Various methods have been developed for this purpose, primarily for DNA, but they are also applicable to RNA with the caveat of ensuring RNA stability in experimental conditions (Popova et al., 2015; Ramanathan et al., 2019). Originally, I used microscopy to observe any changes the localisation of centromeric proteins upon RNAse treatment. This did not yield results, as the RNAse treatment led to no detectable changes in localisation of proteins and decrease in quality of the pictures, that made further analysis impossible. Subsequently I turned to two other basic approaches employed throughout this work. The aim of this study is to investigate the interaction of centromeric proteins with RNAs. Firstly, I used the electrophoretic mobility shift assay to directly observe the interaction between purified RNA and proteins. Secondly, I used various physical methods such as anisotropy of fluorescence and mass spectrometry, to measure the physicochemical parameters of the interaction.

2.7.1. Electrophoretic mobility shift assay identifies Cal1 as an RNA-binding protein

The electrophoretic mobility shift assay (EMSA) was the first method I employed to describe the protein-RNA interaction. This method takes advantage of the fact that particles of different size and charge move differently in a homogenous electric field (Righetti, 2005). For the experiments shown here I used agarose gels, after briefly testing the polyacrylamide gels. Agarose gels were easier to prepare, bands were sharper and they were easier to stain. I prepared the agarose gels in RNAse free environment and kept them in RNAse free conditions throughout the experiments. The gels were exposed to homogenous electric fields in a TAE buffer bath for 1-1,5 hours until the marker stain migrated to the lower part of the gel. Subsequently, they were stained in a TAE buffer bath containing 1μg/ml ethidium bromide for 10 minutes and visualised using GelDoc.

Initially, I tested the binding of the Cal1M fragment binds SatIII RNA and observed binding of control RNAs (Hsrω and αTub) as well [***Figure 18***]. Subsequently, I attempted to bind SatIII RNA to other proteins, that are not known RNA binders (lysozyme, BSA and GCP6) [***Figure 19***]. Only the Cal1M fragment exhibited RNA affinity, but it was not

specific to SatIII RNA. I further tested this with other cenRNA, based on the research of (Valent, 2022) [***Figure 20***]. Eventually I tested DNA as well [***Figure 21***].

These experiments led me to the realization that Cal1M is the RNA binding part of Cal1, and I further focused on it. Cal1M RNA interaction does not seem to be sequence specific, as it binds indiscriminately any RNA that is offered [***Figure 18***]. On the other hand, SatIII RNA does not bind any random protein [***Figure 19***]. Full length Cal1 seems to have smaller affinity to SatIII RNA than the M fragment, but it nevertheless still binds [***Figure 19 C***]. Based on these experiments, Cal1M was chosen for further analysis with mass spectrometry.

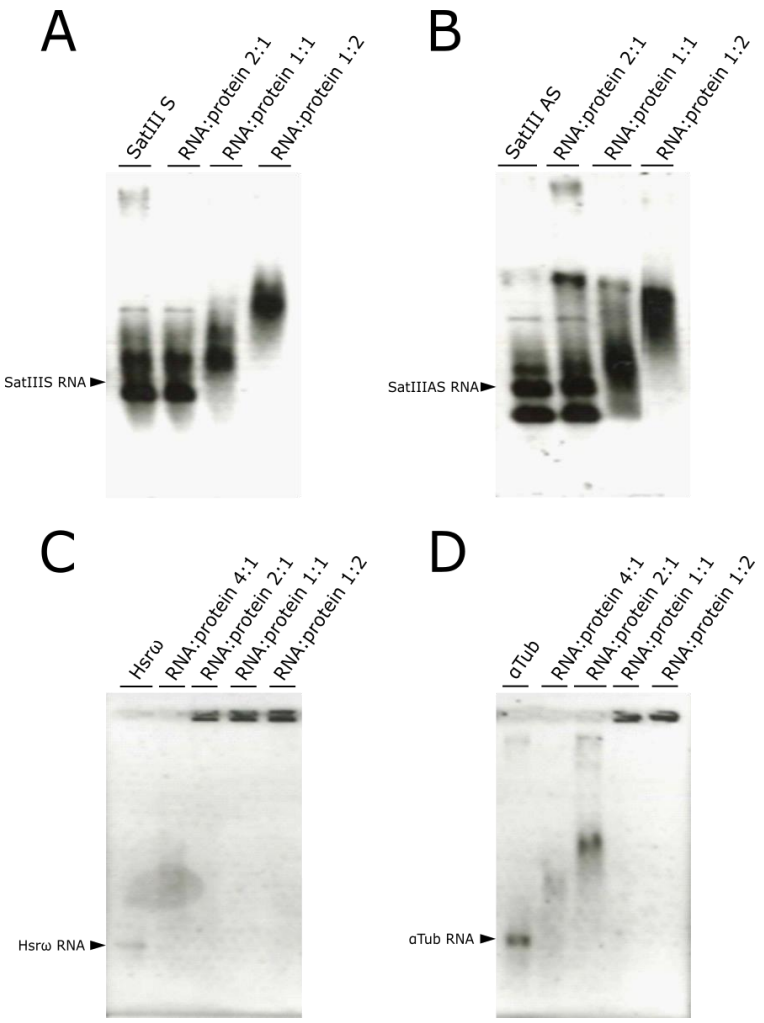


Figure 18. **Cal1M interacts with centromeric and other RNAs**

(A) Cal1M interaction with SatIII sense RNA. Each well contains 300 ng of RNA, or 1.36 pmol. The gel shown here is a chosen representative Increasing the protein:RNA molar ratio from left to right leads to a visible shift, indicating an interaction.
(B) Cal1M interaction with SatIII antisense RNA. Each well contains 300 ng of RNA, or 0,68 pmol. Increasing the protein:RNA molar ratio from left to right leads to a visible shift, indicating an interaction.
(C) Cal1M interaction with Hsrω RNA. Each well contains 200 ng of RNA, or 0.51 pmol. Increasing the protein:RNA molar ratio from left to right leads to a visible shift, but quickly results in precipitation in the well.
(D) Cal1M interaction with αTub mRNA. Each well contains 200 ng of RNA, or 0.37 pmol. Increasing the protein:RNA molar ratio from left to right leads to a visible shift, but also results in precipitation in the well.
Only one representative gel from each subfigure is shown here out of the three repeats.

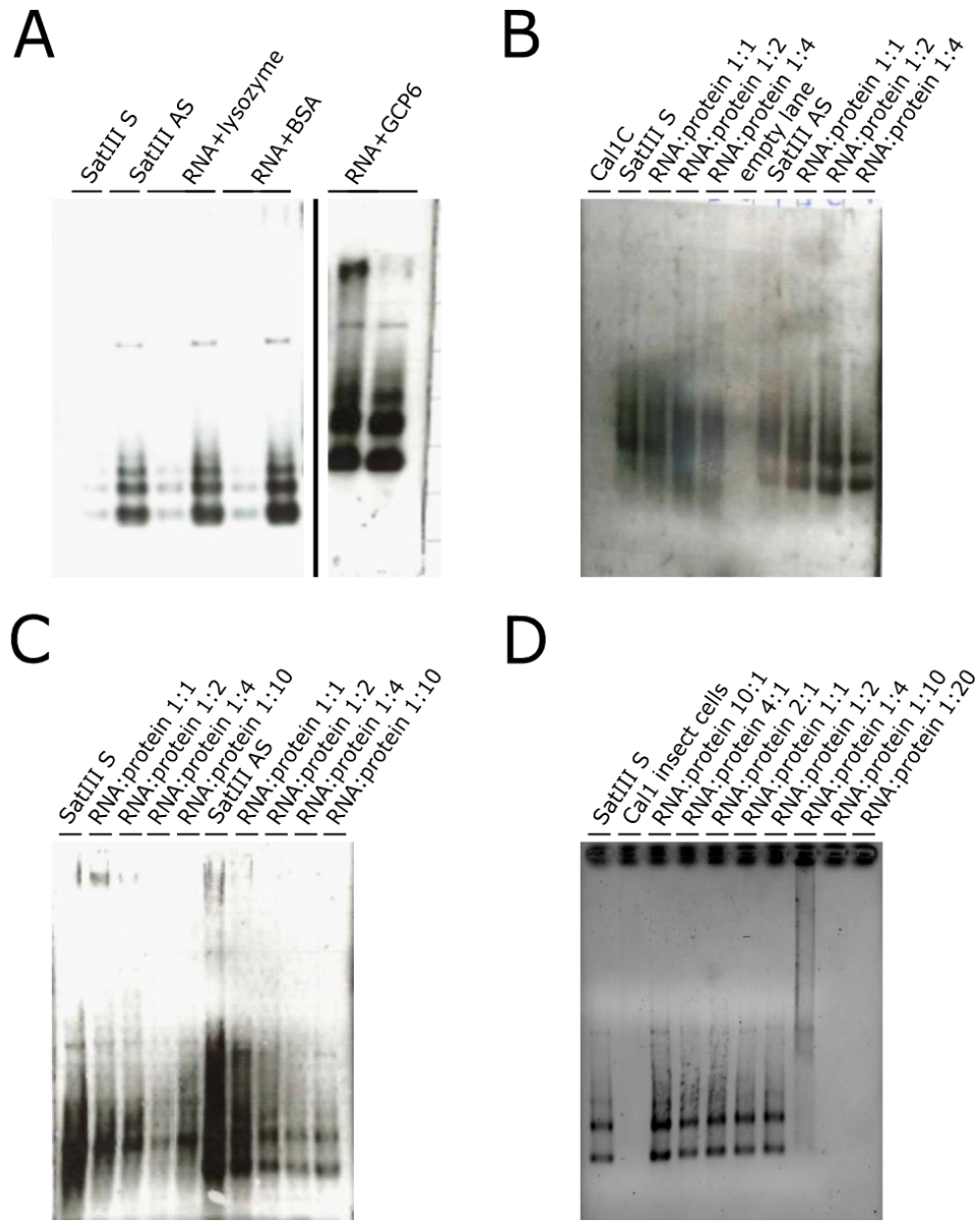


Figure 19. **SatIII RNA interaction with proteins is not unspecific**

(A) SatIII RNA interaction with lysozyme, BSA and GCP6.
(B) SatIII RNA interaction with Cal1C.
(C) SatIII RNA interaction with Cenp-C4.
(D) SatIII sense RNA interaction with full length Cal1. There is a visible shift with protein:RNA ratios increased above 2:1, but further increase leads to precipitation. Full length Cal1 is also not stable and quickly degrades, which prevents longer incubation times.
Only one representative gel from each subfigure is shown here out of the three repeats. Each well contains 200 ng of RNA, or 0.91 pmol for SatIIIS and 0.45 pmol for SatIIIAS.

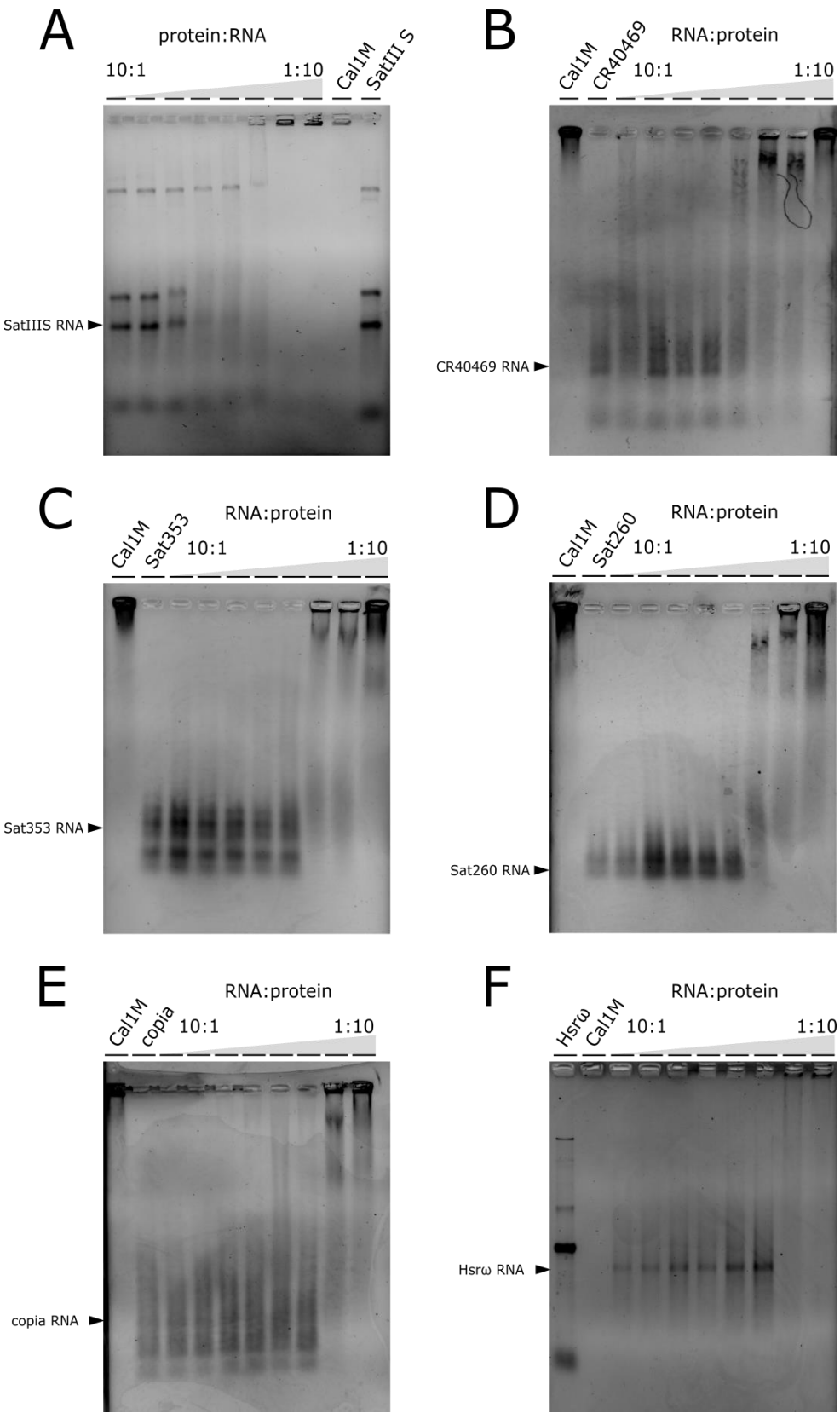

A
protein:RNA
10:1
1:10
Cal1M
SatIII S
SatIIIS RNA
B
RNA:protein
Cal1M
CR40469
10:1
1:10
CR40469 RNA
C
RNA:protein
Cal1M
Sat353
10:1
1:10
Sat353 RNA
D
RNA:protein
Cal1M
Sat260
10:1
1:10
Sat260 RNA
E
RNA:protein
Cal1M
copia
10:1
1:10
copia RNA
F
RNA:protein
Hsrω
Cal1M
10:1
1:10
Hsrω RNA

Figure 20. **Cal1M interacts with various centromeric RNAs**

(A) Interaction between Cal1M and SatIII sense RNA. Each well contains 200 ng of RNA, or 0.91 pmol. As the protein:RNA molar ratio increases from left to right the bands shift and disappear. At higher protein concentrations precipitation occurs in the wells.
(B) Interaction between Cal1M and CR40469 RNA. Each well contains 200 ng of RNA, or 2.89 pmol. As the protein:RNA molar ratio increases from left to right the bands shift and disappear. At higher protein concentrations precipitation occurs in the wells.
(C) Interaction between Cal1M and Sat353 RNA. Each well contains 200 ng of RNA, or 1.79 pmol. As the protein:RNA molar ratio increases from left to right the bands shift and disappear. At higher protein concentrations precipitation occurs in the wells.
(D) Interaction between Cal1M and Sat260 RNA. Each well contains 200 ng of RNA, or 2.41 pmol. As the protein:RNA molar ratio increases from left to right the bands shift and disappear. At higher protein concentrations precipitation occurs in the wells.
(E) Interaction between Cal1M and copia RNA. Each well contains 200 ng of RNA, or 0.13 pmol. As the protein:RNA molar ratio increases from left to right the bands do not shift but eventually precipitation in the wells occurs at higher protein concentrations.
(F) Interaction between Cal1M and Hsrω RNA. Each well contains 200 ng of RNA, or 0.51 pmol. As the protein:RNA molar ratio increases from left to right the bands do not shift but eventually precipitation in the wells occurs at higher protein concentrations.
Only one representative gel from each subfigure is shown here out of the three repeats.

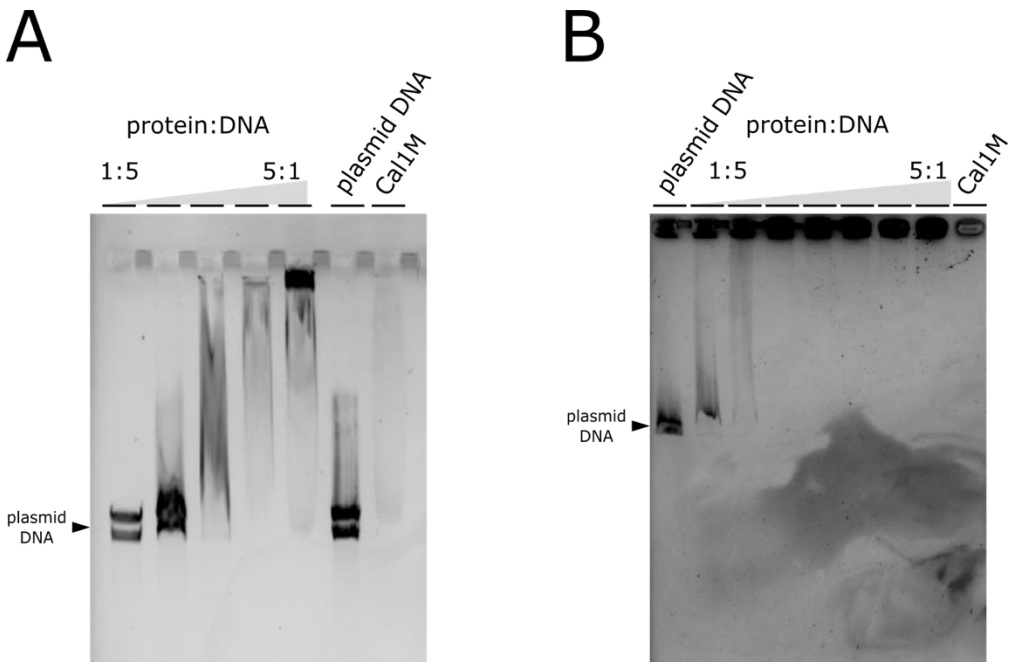


Figure 21. **Cal1M is binding DNA**

Interaction between Cal1M and linearised plasmid DNA: **(A)** with pCA528 Cal1 plasmid and **(B)** pHALO CENP-B plasmid. Each well contains 200 ng of linearised plasmid DNA, or 0.04 pmol in both cases. Only one representative gel out of the three repeats is shown here. There is a shift with linearised plasmid indicating binding of Cal1M to DNA.

2.7.2. Fluorescence anisotropy

In addition to EMSA, I used anisotropy of fluorescence (AF) to measure the interaction between satellite RNAs and Cal1M. This method measures the thermodynamic parameters of the interaction by taking advantage of the different rotational properties of molecules of different mass and diameter in solution (LiCata & Wowor, 2008). Proteins that are free rotate faster than those that are bound to RNA, which are more massive. Fluorophores are excited by linearly polarised light. The light emitted after relaxation is then measured under an angle to distinguish it from the original polarizing light. The plane of polarisation shifts based on the free rotation of molecules in the solutions, which varies for molecules of different molecular masses. These rotational properties can be probed by exciting the samples with linearly polarised light and measuring the emitted light.

I generated fluorescent molecules to measure the fluorescence anisotropy in my samples. I labelled satellite RNA using in vitro transcription with Alexa488 tagged uracil, replacing 25% of the supplied UTP with Alexa488-UTP. The polarisation value of Alexa 488 tagged RNA was set to 100 mP.

I fitted the experimental data with a sigmoid curve using OriginPro software. The inflection point of the resulting curve represents the K_D value of the interaction. Due to the precipitation of the interacting protein at higher concentrations, I was not able to obtain sufficient data points to fit the complete curve. At these low concentrations the interaction between RNA and Cal1M was observable, but impossible to quantify due to the missing part of the binding curve. Therefore, this method is not suitable for studying the interaction of RNA with Cal1M. The measured data nevertheless suggests that the interaction between satellite RNA and Cal1M is weak, and the corresponding K_D value will be large. Additionally, there is no difference between sense and antisense satellite RNA [***Figure 22***].

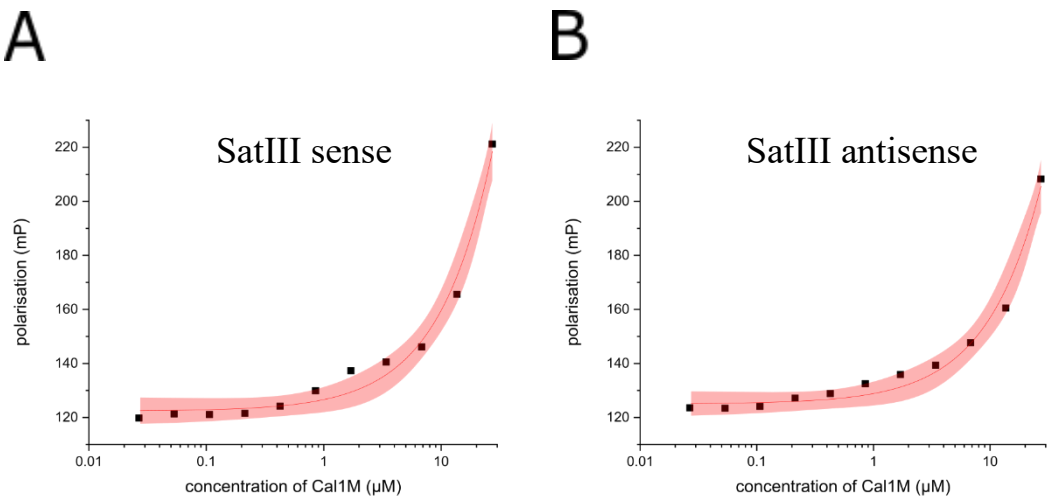


Figure 22. **Cal1M interaction with SatIII is weak**

Fluorescence polarisation curves demonstrate the binding of SatIII RNA to Cal1M. To determine the K_D of the interaction, the full sigmoidal curve must be measured. The absence of a portion of the curve is due to the precipitation of Cal1M at higher concentrations. The exact value of the K_D cannot be determined without the rest of the curve. These data suggest that Cal1M binds SatIII RNA weakly, and the K_D is large. The situation is very similar for both sense **(A)** and antisense **(B)** RNA.

2.8. Cal1M-RNA interaction was analysed by mass spectrometry

I initially selected the Cal1M and Cal1C fragments for subsequent mass spectrometry experiments based on previous experimental results. The Cal1C fragment was only used in the initial experiments as its RNA binding properties were not confirmed by other methods. All further descriptions pertain only to Cal1M unless otherwise specified.

After obtaining sufficient quantities of purified protein samples, it was necessary to confirm their stability under the conditions required for the amide hydrogen exchange experiments (Mitra, 2021). In order to do this experiment, protein sample is incubated with its binding partner, in this case RNA, in deuterated buffer, freely exchanging available protons for deuterons. After the incubation period the $^{1}H/^{2}H$ exchange is slowed down for analysis by quenching. To quench the $^{1}H/^{2}H$ exchange, the pH needed to be reduced to 2.0 by adding a large quantity of quench buffer. The purified protein samples were stored in the elution buffer from the last purification step, which had a pH of 8.0. The theoretical pI of 9.56 was predicted using the Expasy ProtParam tool. The experimental results suggest that the basic residues in the protein sequence are already almost fully protonated at pH 8.0 (pI=9). Decreasing pH to 2.0 will lead to protonation of histidine (pI=6-7) and acidic residues (pI=3-4), thus increasing the net charge and with it, the solubility. Therefore, this drastic change to pH 2.0 should not lead to insoluble aggregates. The problem was with the low solubility of RNA at lower pH. Because of this, I tested, whether the pH of the quench buffer could be increased without invalidating the experiment. This was confirmed in [***Figure 23***]. Proteolysis can still occur even if the pH does not drop all the way to 2.0, making the protein eligible for mass spectrometry even with RNA present. This experiment revealed two significant findings. Firstly, the protein can withstand drastic shifts in pH without precipitating. Secondly, pepsin remains active at higher pH levels than its recommended optimum of pH 2.0. This enabled me to design a protocol I later used for mass spectrometry experiments.

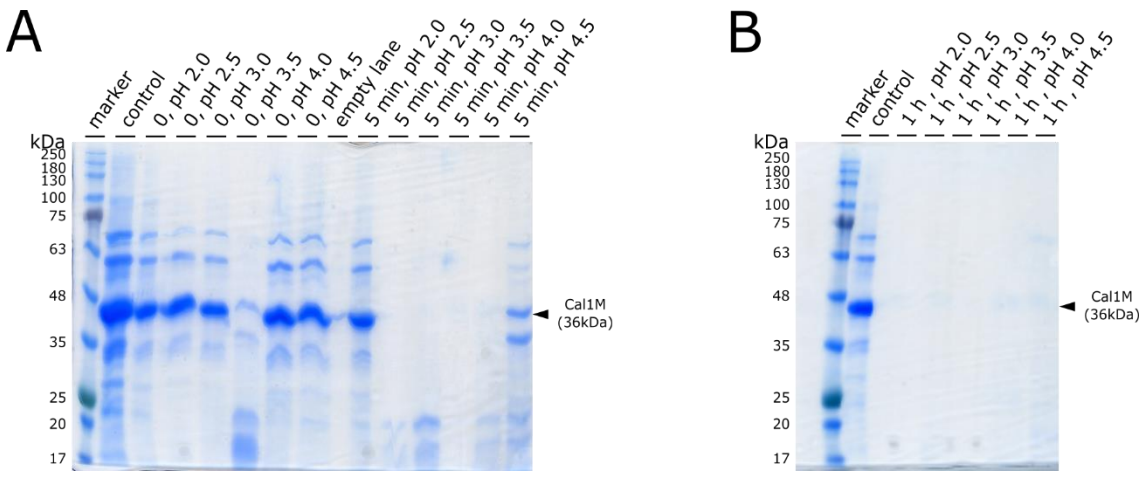


Figure 23. **Cal1M is stable in changing pH conditions and can undergo proteolysis by pepsin at higher than optimal pH**

The SDS PAGE of the Cal1M was conducted. Each sample was incubated under different pH conditions after quenching. After incubation, the sample was centrifuged to pellet the aggregates and the supernatant sample was mixed with the protease. Pepsin was used to degrade the protein in all conditions. The quenching solution consisted of 500mM NaCl, 50mM phosphate at a specific pH, 1mM DTT, 10% glycerol, and 6M guanidium-HCl.
(A) The gel shows the degradation efficiency of pepsin in different pH conditions immediately after acidification and 5 minutes later.
(B) The gel shows the degradation efficiency of pepsin in various pH conditions after 1 hour.

2.8.1. Mapping peptides

Once I had shown that the protein was stable enough to be loaded into the mass spectrometer, its proteolytic digestion and subsequent mass spectrum were measured. It was necessary to determine the optimal quenching buffer and protease column for this experiment. Two quenching buffers were used depending on the experimental setup. Initially, only the protein spectrum was measured, and the urea/thiourea quenching buffer was used (1M glycine-HCl, 6M urea, 2M thiourea, 0.4M TCEP and pH of 2.3). When RNA was also present, the use of urea quenching buffer resulted in fewer peptides being visible in the mass spectrum and guanidinium quench buffer was used instead (1 M glycine-HCl, 6M guanidine, 0.5M TCEP and pH of 2.3). All the details are shown in [***Appendix 9***.].

2.8.2. $^{1}H/^{2}H$ exchange mass spectrometry

Hydrogen-$^{1}H/^{2}H$-exchange mass spectrometry was used to evaluate the structural dynamics of Cal1M in the presence of satellite RNAs. This method is based on the phenomenon of solvent-solute hydrogen exchange, where the protein undergoes the exchange with the solvent (Ozohanics & Ambrus, 2020). Mixing the reaction with a low pH quench buffer slows down exchange of amide protons/deuterons sufficiently to conserve the current state for several seconds, providing enough time to inject the sample into a liquid chromatography system and pump it through a column with immobilized proteases. Short peptides are produced by cleaving the protein, which are then loaded into the mass spectrometer. Data analysis is performed using nondeuterated and 100% deuterated controls. This method enables mapping of the protein surface and the residues responsible that interact with some binding partner, in this case RNA. The deuteration level will decrease in the presence of RNA, which will be visible in the mass spectrum.

I used equimolar ratios of protein and RNA in all conducted experiments. $^{1}H/^{2}H$ exchange was performed at room temperature for a specific time and was stopped by adding equal volume of ice-cold quenching buffer. I performed the initial experiments in the mass spectrometry and proteomics core facility of ZMBH under supervision of Nicole Lübbehusen. Technical problems in the ZMBH mass spectrometry facility led me to the cooperation with the mass spectrometry core facility (CMS) in the Institute of

Biotechnology, BIOCEV in Vestec u Prahy, Czechia. Later experiments were performed on site by Dr. Pavla Vaňková.

The samples were prepared using the PAL DHR autosampler (CTC Analytics AG) for exchange steps longer than 20 s or by manual pipetting for shorter times. The quenched reaction mixture was injected into the Agilent Infinity II UPLC liquid chromatography system, which includes a proteolytic column, a trap column, and a desalting column.

The system was directly connected to the ESI source of the MaXis mass spectrometer (ZMBH) or timsTOF Pro mass spectrometer (BIOCEV, both Bruker Daltonics). To minimize D/H back-exchange, the entire chromatography system was cooled to 0 °C. LC-MS data were acquired, peak-picked, and exported using DataAnalysis (v. 5.3, Bruker Daltonics) and further processed using the DeutEx software (Trcka et al., 2019). I visualized the data using MSTools (Kavan & Man, 2011, http://peterslab.org/MSTools/index.php).

We initially tested several proteases to enhance the resolution of the protein mass spectrum. Eventually, the spectrum was sufficient to identify most of the Cal1M peptides [***Appendix 10***]. Once we could detect most of the peptides in the MS2 spectrum, we performed $^{1}H/^{2}H$. However, the peak intensity of the peptides in $^{1}H/^{2}H$ was lower than in the MS2 spectrum, resulting in some signal loss and decreased final resolution. Nevertheless, the majority of the peptides were visible in the spectrum, allowing us to proceed with RNA experiments. ***Figure 24*** demonstrates that the presence of SatIII RNA does not provide protection to the protein surface or any other present peptides from $^{1}H/^{2}H$ exchange. To ensure that we did not overlook any interactions, we conducted the experiment again in shorter time scales, limited only by our ability to pipette quickly, as depicted in ***Figure 25.*** Shortening the $^{1}H/^{2}H$ exchange times did not produce any different outcomes. All peptides visible in the spectrum remained fully exchanged and were therefore not protected by the interaction with the RNA. Following the previous work in our laboratory, I also tested copia RNA. ***Figure 26*** shows that copia did not provide any measurable protection. All peptides visible in the spectrum of copia remained fully exchanged.

Cal1M appears to be unstructured and fully exposed to solvent under the tested conditions, with no sign of RNA binding even in shortened timescales.

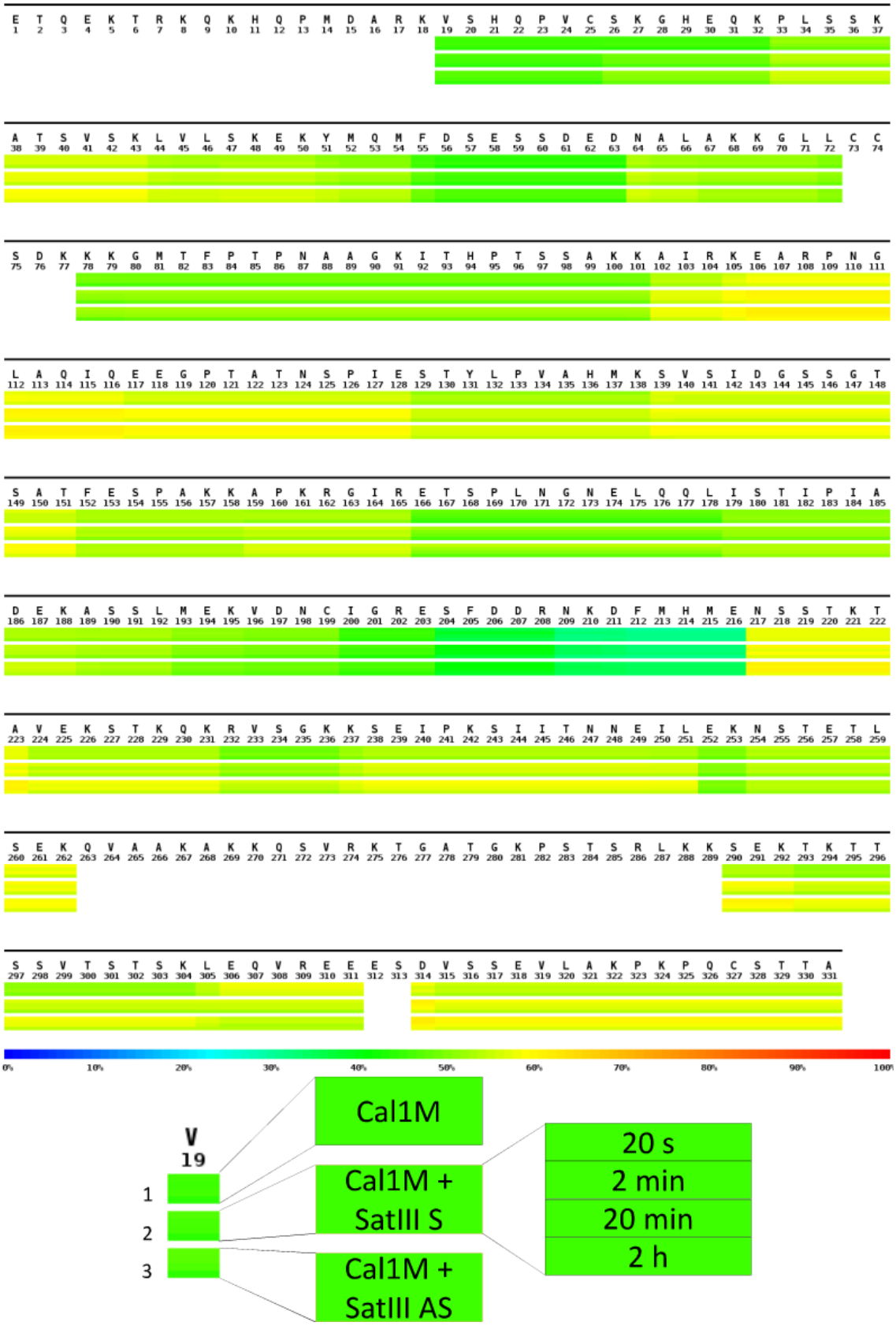

E T Q E K T R K Q K H Q P M D A R K V S H Q P V C S K G H E Q K P L S S K
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
A T S V S K L V L S K E K Y M Q M F D S E S S D E D N A L A K K G L L C C
38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74
S D K K K G M T F P T P N A A G K I T H P T S S A K K A I R K E A R P N G
75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111
L A Q I Q E E G P T A T N S P I E S T Y L P V A H M K S V S I D G S S G T
112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148
S A T F E S P A K K A P K R G I R E T S P L N G N E L Q Q L I S T I P I A
149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185
D E K A S S L M E K V D N C I G R E S F D D R N K D F M H M E N S S T K T
186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222
A V E K S T K Q K R V S G K K S E I P K S I I T N N E I L E K N S T E T L
223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259
S E K Q V A A K A K K Q S V R K T G A T G K P S T S R L K K S E K T K T T
260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296
S S V T S T S K L E Q V R E E E S D V S S E V L A K P K P Q C S T T A
297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
V
19
1
2
3
Cal1M
Cal1M +
SatIII S
Cal1M +
SatIII AS
20 s
2 min
20 min
2 h

Figure 24. **Presence of SatIII RNA does not protect Cal1M surface from $^{1}H/^{2}H$ exchange**

Heat maps under the sequence of Cal1M illustrate the different HDX experiments performed at room temperature. The scale for the heat maps is shown at the bottom of the figure. Each lane is further divided into smaller lanes showing different experiments with different exchange times, as schematically described in the diagram at the bottom of the figure.
The first lane shows HDX in the Cal1M sample in the absence of RNA. There was no significant difference between the different reaction times.
The second lane describes HDX in the Cal1M sample in the presence of a SatIII sense RNA fragment. The SatIII S RNA did not protect any of the Cal1M segments from exchange, resulting in a spectrum strikingly similar to the sample without RNA.
The third lane describes HDX in the Cal1M sample in the presence of a SatIII antisense RNA fragment. The SatIII AS RNA did not protect any of the peptides Cal1M segments from exchange, resulting in a spectrum strikingly similar to the sample without RNA.
The results are not corrected for back-exchange, resulting in an apparent total exchange of 50%.

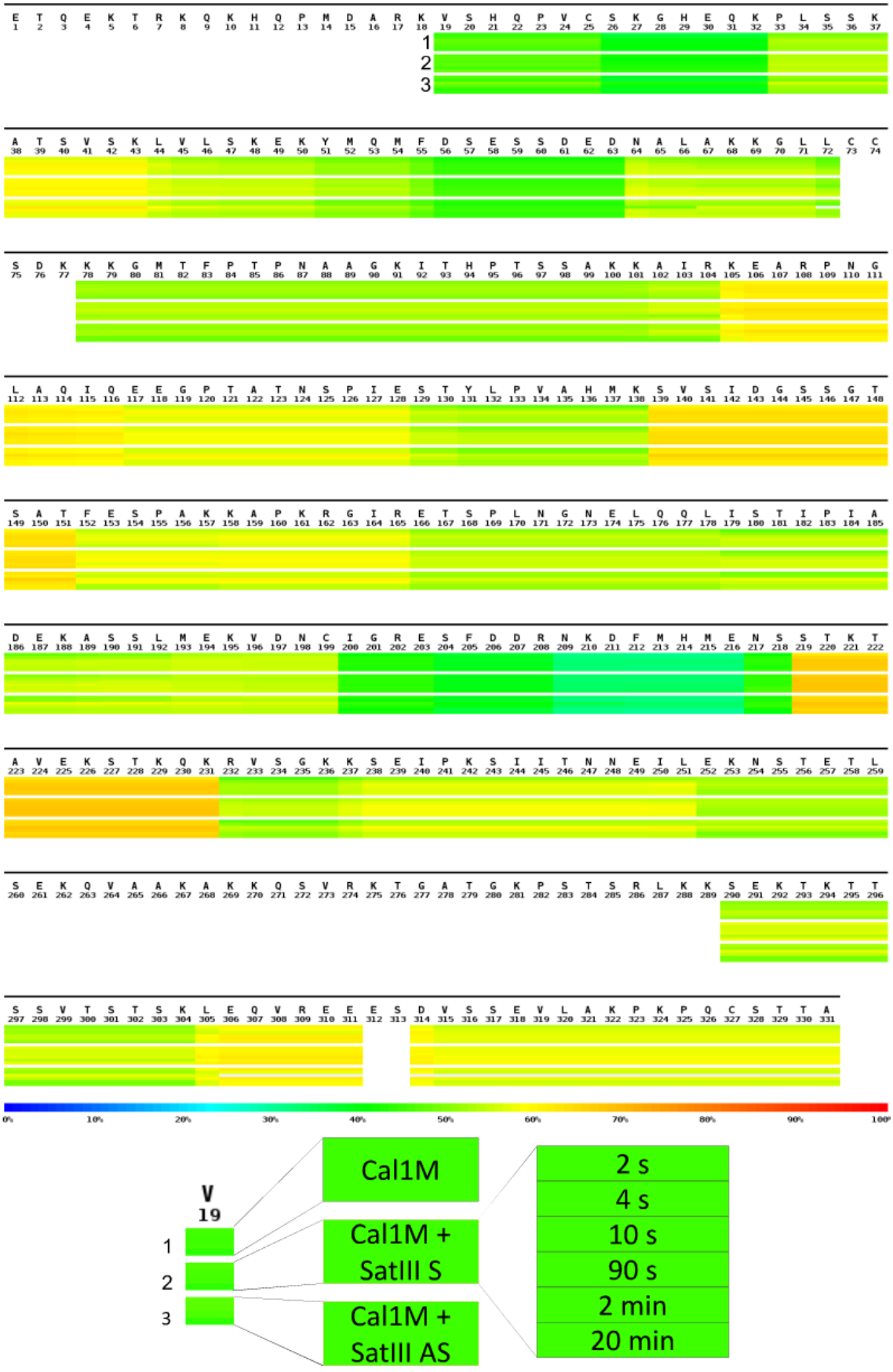

E T Q E K T R K Q K H Q P M D A R K V S H Q P V C S K G H E Q K P L S S S K
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
1
2
3
A T S V S K L V L S K E K Y M Q M F D S E S S D E D N A L A K K G L L C C
38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74
S D K K K G M T F P T P N A A G K I T H P T S S A K K A I R K E A R P N G
75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111
L A Q I Q E E G P T A T N S P I E S T Y L P V A H M K S V S I D G S S G T
112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148
S A T F E S P A K K A P K R G I R E T S P L N G N E L Q Q L I S T I P I A
149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185
D E K A S S L M E K V D N C I G R E S F D D R N K D F M H M E N S S T K T
186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222
A V E K S T K Q K R V S G K K S E I P K S I I T N N E I L E K N S T E T L
223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259
S E K Q V A A K A K K Q S V R K T G A T G K P S T S R L K K S E K T K T T
260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296
S S V T S T S K L E Q V R E E E S D V S S E V L A K P K P Q C S T T A
297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
V
19
1
2
3
Cal1M
Cal1M +
SatIII S
Cal1M +
SatIII AS
2 s
4 s
10 s
90 s
2 min
20 min

Figure 25. **Presence of SatIII RNA does not protect Cal1M surface from $^{1}H/^{2}H$ exchange even in shorter timescales**

Heat maps under the sequence of Cal1M illustrate the different HDX experiments performed at 4 °C and shorter reaction times. The scale for the heat maps is shown at the bottom of the figure. Each lane is further divided into smaller lanes showing different experiments with different exchange times, as schematically described in the diagram at the bottom of the figure.
The first lane shows HDX in the Cal1M sample in the absence of RNA. There was no significant difference between the different reaction times.
The second lane describes HDX in the Cal1M sample in the presence of a SatIII sense RNA fragment. The SatIII S RNA did not protect any of the Cal1M segments from exchange, resulting in a spectrum strikingly similar to the sample without RNA.
The third lane describes HDX in the Cal1M sample in the presence of a SatIII antisense RNA fragment. The SatIII AS RNA did not protect any of the peptides Cal1M segments from exchange, resulting in a spectrum strikingly similar to the sample without RNA. The results are not corrected for back-exchange, resulting in an apparent total exchange of 50%.

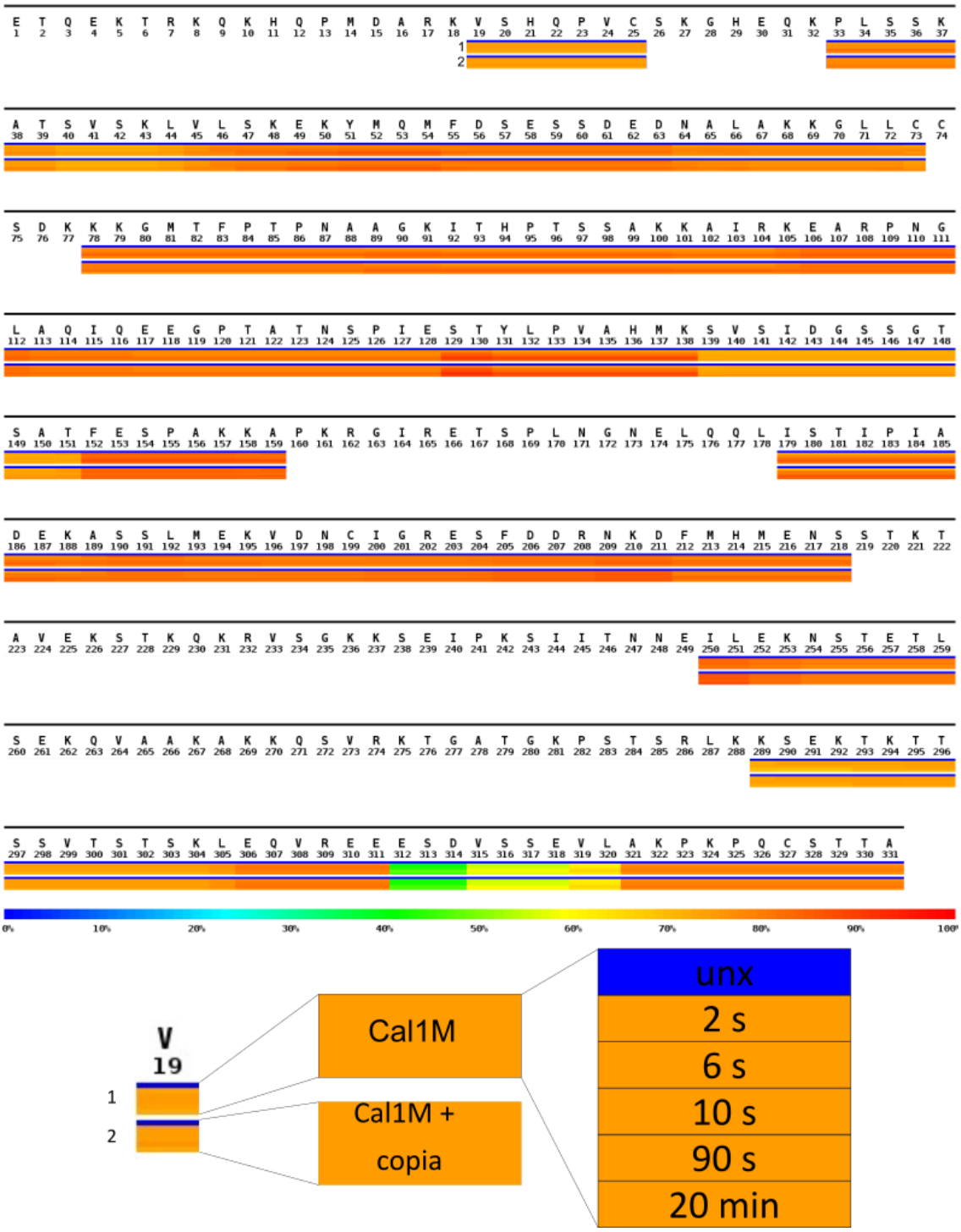

E T Q E K T R K Q K H Q P M D A R K V S H Q P V C S K G H E Q K P L S S K
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
1
2
A T S V S K L V L S K E K Y M Q M F D S E S S D E D N A L A K K G L L C C
38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74
S D K K K G M T F P T P N A A G K I T H P T S S A K K A I R K E A R P N G
75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111
L A Q I Q E E G P T A T N S P I E S T Y L P V A H M K S V S I D G S S G T
112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148
S A T F E S P A K K A P K R G I R E T S P L N G N E L Q Q L I S T I P I A
149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185
D E K A S S L M E K V D N C I G R E S F D D R N K D F M H M E N S S T K T
186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222
A V E K S T K Q K R V S G K K S E I P K S I I T N N E I L E K N S T E T L
223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259
S E K Q V A A K A K K Q S V R K T G A T G K P S T S R L K K S E K T K T T
260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296
S S V T S T S K L E Q V R E E E S D V S S E V L A K P K P Q C S T T A
297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
V
19
1
2
Cal1M
Cal1M +
copia
unx
2 s
6 s
10 s
90 s
20 min

Figure 26. **Presence of copia RNA does not protect Cal1M surface from $^{1}H/^{2}H$ exchange**

Under the sequence of Cal1M heat maps illustrate the different HDX experiments performed at 4 °C and with copia RNA. The white gaps in the spectrum are caused by the missing peptides, as addition of the copia RNA decreases the solubility of the protein and resulting quality of the spectrum.
The first lane shows HDX in the Cal1M sample in the absence of RNA. There was no significant difference between the different reaction times. The scale for the heat maps is shown at the bottom of the figure. Each lane is further divided into smaller lanes showing different experiments with different exchange times, as schematically described in the diagram at the bottom of the figure. In this spectrum, unexchanged control was also added to each lane, forming a clear blue line throughout the whole sequence.
The second lane describes HDX in the Cal1M sample in the presence of a copia RNA fragment. The copia RNA did not protect any of the peptides provide significant protection to any of the peptides, resulting in a spectrum strikingly similar to the sample without RNA.
The results are corrected for back-exchange, resulting in an apparent total exchange of nearly 100%. Some peptides are absent from the spectrum due to low signal quality, indicated by empty spaces.

3. Discussion

Centromeric RNA is reported to bind the inner kinetochore region of *Drosophila* (Rošić et al., 2014). This is also reported in other species like humans (Fukagawa & Earnshaw, 2014; Zhang et al., 2022) or frogs (Grenfell et al., 2016). Unlike the other listed species, Drosophila inner kinetochore complex is very simple, only three proteins – Cid, Cal1 and Cenp-C. This fact makes it a perfect model organism for detailed structural studies of protein-RNA interactions in centromeric chromatin. RNA dependence of inner kinetochore fits well to the whole picture of epigenetic maintenance of centromeric regions (Allshire & Karpen, 2008). In this work I provided new insights into the inner centromere protein-RNA interactions in *Drosophila melanogaster* and identified Cal1 as an RNA-binder.

Cal1 is RNA binding protein

Centromeric chromatin is not transcriptionaly silent even though it carries H3K9 methylation, does not contain any protein coding genes and is hetechromatinised. Its active transcription seems to be vital for maintaining centromere structure and proper function (Wong et al., 2007; Saffery et al., 2012). Transcripts of satellite DNA has been previously shown to colocalise with the *Drosophila* inner kinetochore and their absence leads to chromosome segregation problems. This is due to the failure of recruiting essential kinetochore components and subsequent failed connection to the mitotic spindle microtubules. It was previously shown that these transcripts interact with Cenp-C and are important for facilitating the interaction with Cid. The opposite was also true, without Cenp-C, SatIII was no longer associating with centromeres in imunoprecipitation experiments (Rošić et al., 2014). The interaction of satellite transcripts with Cenp-C was also observed in maize (Du et al., 2010b). I could not test this interaction *in vitro*, because of the instability of Cenp-C in solution and following difficulties with the purification. Instead, my experiments with *Drosophila* inner kinetochore proteins led me to the investigation of Cal1 RNA-binding properties. Human ortholog of Cal1, HJURP, is a known RNA-binder (Quénet & Dalal, 2014). As shown in ***Figure 18***, Cal1 is directly interacting with SatIII RNA as well. Without centromeric α-Satellite transcripts, HJURP fails to load CENP-A in humans, even if all the other known regulators, such as CDK1 kinase, are present (J. Wang et al., 2014). This is similar to the Cid loading failure in

Drosophila in absence of SatIII RNA (Rošić et al., 2014). Cal1 is a functional ortholog of HJURP with no significant sequence or structural similarity (C.-C. Chen et al., 2014). However, the fact that it seems to bind RNA suggests coevolution of centromeric and pericentromeric transcripts and proteins. Centromeric sequences are mutating faster than the coding regions, as there are no genes that could be damaged and the control mechanisms are less strict. This leads to comparatively faster divergence of centromeric sequences. Centromeric proteins therefore have to evolve too, to still be able to perform their function (Malik & Henikoff, 2002). RNA binding seems to be a vital part of this function. RNA plays an important role in maintaining centromere integrity even though the centromeric proteins can vary in different organisms and the centromeric transcripts do not have any conserved sequence (Corless et al., 2020). RNA was found interacting with centromeric proteins and maintaining centromeric function in, among others, humans, (Quénet & Dalal, 2014), fruit flies (Rošić et al., 2014), tamar wallabies (Carone et al., 2013), mice (Bouzinba-Segard et al., 2006) and yeast (Choi et al., 2011). Centromeric transcripts in all these organisms are divergent, but originate predominantly in satellite regions of the pericentromere. They have various lengths, usually few hundreds of nucleotides, and are AT-rich compared to the average genome composition (Talbert & Henikoff, 2020).

Closer investigation of Cal1 interaction with cenRNA was complicated due to the instability of the purified protein samples and RNA contamination. Full length Cal1 from insect cells [***Figure 10***] was not stable after freezing. Only viable experiment without freezing was EMSA, where Cal1 showed affinity towards SatIII RNA [***Figure 19***]. Cal1M fragment purified from bacteria was stable for all the described experiments but with notable amount of RNA impurities. Stability and secondary structure motifs of Cal1M were confirmed by CD spectroscopy (Ranjbar & Gill, 2009). By comparing the measured spectrum with standard curves of proteins with a high percentage of a particular structural motif, more specifically α-helix, β-sheet or random coil, I deducted the secondary structure of Cal1M. CD confirmed the predictions from AlphaFold. Cal1M has few structured motifs, even at the secondary structure level, with most of the sequence being random coil with a minor α-helix compound [***Figure 15***]. I was also able to confirm that the protein can be frozen and stored at -80 °C for several weeks without losing its native structure by comparing the spectra of both frozen and unfrozen samples.

Another experiment I conducted involved using CD to investigate the thermal stability of Cal1M fragment. If the protein has a native structure in the solution that is not rigid, but instead rapidly fluctuates between several possible conformations, it may not be detectable by standard CD. I measured the CD spectrum over time but did not observe any changes. I only measured over the timescales necessary for further experiments, and the protein remained stable during that time. To observe the melting curve, I included the temperature gradient in the measurement [***Figure 16***]. If the protein was as unstructured as previously predicted, there would be no melting curve, only a straight line (J. Miles et al., 2021). However, this was not the case, as the protein exhibited a clear melting curve. This result indicates that there is some native structure, no matter how insignificant, that can be disrupted by heating the protein. The inflection point of the measured curve represents the protein's melting point. For Cal1M, the temperature was measured at 53.3 °C. This temperature falls within the normal range for protein stability. (Ericsson et al., 2006) Although the result was initially confusing, it was reproducible. Despite being predicted to lack stability and structure, Cal1M exhibited a surprisingly high melting point. Cal1M therefore has a stable structure in solution. It suggests that more experimental approaches based on the thermal stability of proteins, such as DSF, may be possible. Due to the high absorption of nucleic acid I wasn't able to measure its effect on Cal1 using CD spectroscopy. This is not a problem for DSF, which I later used to probe the stability of Cal1M in the presence of RNA.

For further exploration of Cal1M-RNA interaction *in vitro* I needed a stable protein sample. The concentration of the protein had to be kept low, otherwise precipitation ensued. This makes measuring weak interactions difficult, as they require higher concentrations of one binding partner (Zagrovic et al., 2018). The results from EMSA and fluorescence anisotropy [***Figure 18-20***] show that there is a direct interaction of SatIII (and other) RNA with Cal1M, but the concentration was not sufficient for full thermodynamic description of the interaction. Apart from precipitation of Cal1M alone, addition of RNA further destabilised it. This is probably due to the neutralisation of the net positive charge of the protein by the net negative charge of the nucleic acid. Neutral molecules aggregate easier, as there is less electrical repulsion to keep them apart (W. Wang et al., 2010). This effect had been described before in cellular context (Ung et al., 2001) and in nanostructures, both natural and artificial (Bornholdt & Prasad, 2008; Weizmann et al., 2008). I observed precipitation in several experiments when the

concentration of RNA was increased above certain level [***Figure 20***]. This level was dependent on the size of the RNA molecule. This is probably mostly due to two effects. Firstly, Cal1 concentration in the cell is very low, only few molecules per cell are necessary (Schittenhelm et al., 2010). Secondly, I worked only with a fragment. Both N- and C-terminal parts were missing from Cal1M. They may be regulating the interaction either directly or through posttranslational modifications, similar to HJURP (Barnhart et al., 2011). In fact, Cal1M was binding bacterial RNA as well, more specifically rRNA [***Figure 17***]. This was despite the fact that there were nonspecific nucleases present during the protein purification. rRNA is very abundant, nevertheless it was not easily detectable in the purified protein sample. Only after phenol-chloroform extraction of RNA from the purified protein sample I was able to analyse it. This suggests that it is bound to Cal1 and this interaction protects it from nucleases. This was further supported by the fact that increasing the concentration of nucleases did not decrease the yield of co-purified RNA. Only longer incubation of bacterial lysate at room temperature with increased amounts of nucleases led to significant decrease of RNA in the solution. Unfortunately, it led to severely decreased yield of protein after purification as well. Adding RNAses to the purified protein helps as well, but once they are present, they cannot be disposed of again. Since all my following experiments depend on the stability of RNA in the sample, presence of additional RNAses is detrimental. Adding RNAses and later inhibiting them with specific inhibitors does not have the desired effect as the added RNA is still degraded too fast for any meaningful measurement.

I used several different RNAs, mostly derived from pericentromeric regions of *Drosophila*. For controls I also used mRNA and rRNA, as well as plasmid DNA. I was initially focusing at SatIII RNA, because that was the one identified in the inner kinetochore before (Rošić et al., 2014). I used both sense and antisense transcripts of SatIII and their affinity to Cal1M was comparable [***Figure 18-20***]. Other RNAs, such as Hsrω or copia, have shown similar affinity [***Figure 18, Figure 20***]. In vivo, Cal1 seems to bind more species of RNA as well (Valent, 2022). The specificity may not be driven by the sequence of the nucleic acid but rather by its availability in centromeric environment. It was observed before that centromeres in interphase nucleus colocalise with the nucleolus (Ochs & Press, 1992; Guttenbach et al., 1996). RNA-binding capacity of centromeric proteins may be the cause of this.

When I repeated the experiments to confirmed the interaction I also attempted to measure the thermodynamic parameters of it. For this I used three methods – fluorescence anisotropy, differential scanning fluorimetry and mass spectrometry.

It was necessary to measure the complete binding curve. The size of the curve is dependent on the strength of the interaction, which is represented by the binding constant K_D, or the numerical value of the inflection point of the curve. The weaker the interaction, the larger the K_D and the larger the curve (Jarmoskaite et al., 2020). To measure such weak interactions, it is necessary to increase the concentration of the reactants accordingly. In this case, the concentration of RNA in the solution was increased, but this resulted in protein precipitation. I improved the stability by changing the buffer composition, but it was still not sufficient to measure the full binding curve. I was only able to measure the left part of the curve [***Figure 22***], which is not sufficient to calculate the K_D. However, I was able to confirm that an interaction is occurring, since I observed a partial curve rather than a straight line (Ericsson et al., 2006). The interaction nevertheless appears to be weak, since I cannot measure the complete curve in the concentration range that is available. To get higher concentration range the solubility of Cal1M would have to be somehow significantly increased.

Next, I used DSF to investigate whether there were significant differences in protein structural stability in the presence of different RNAs. Measuring the fluorescence while gradually increasing the temperature of the sample can provide insight into the protein's stability in solution. The hydrophobic core of the protein is typically concealed within the structure and is not available to the solvent. Thermal dissolution of the protein's structure reveals the hydrophobic parts to the solvent and allows binding of the SYPRO Orange, thus increasing the fluorescence signal. A slower increase in fluorescent signal indicates greater protein stability and the slower unfolding rate in the increasing temperature (Ericsson et al., 2006). In addition to testing protein stability in various buffers, DSF can also determine the melting point of its tertiary structure. Binding partners can affect the stability of proteins in solutions. The strength of the bond between the binding partner and the protein can impact its thermal stability, generally improving it. This can be observed as a shift of the melting point in the thermal spectrum. To determine the thermodynamic parameters of the interaction, a comparison of the spectra of pure protein and protein with an interaction partner, such as RNA in this case, is necessary (Niesen et al., 2007). The shift in the stability can be used to calculate the aforementioned

thermodynamic parameters. I did several repeats of this experiment, but they were not comparable to each other. Differences in biological replicates could be explained by the slightly different composition of each sample [***Appendix 7***]. However, differences in technical replicates were too large to be explained by pipetting errors. My assumption was that the protein sample sometimes started precipitating when mixed with RNA. Despite my attempts to change the situation, I observed no major improvements and decided to not continue the experiment.

Finally, I used $^{1}H/^{2}H$ mass spectrometry to probe the interaction. $^{1}H/^{2}H$ MS is a suitable approach to continue this work with the full-length centromeric proteins (Karch et al., 2018) To analyse the $^{1}H/^{2}H$ data, the protein spectrum must be measured multiple times under different conditions. Firstly, the spectrum of protein itself must be measured. Secondly, the fully deuterated protein spectrum must be measured. This is achieved through several rounds of lyophilisation and rehydration with fully deuterated buffer. Finally, the spectrum of the experimental sample with the binding partner must be measured. By comparing the three spectra, it should be possible to identify which protons are exchanged for deuterons under different conditions. However, Cal1M lacks a stable structure, as shown by the near total $^{1}H/^{2}H$ exchange in all tested conditions [***Figure 24-26***]. This suggests that the protein is always either completely unstructured and fully accessible to the solvent, or that it has a dynamic structure that very rapidly transitions between open and closed states (Resetca & Wilson, 2013). The quickest experiment we were able to reproducibly pipette was 2 seconds long, indicating that the opening and closing of the protein must occur faster than that. However, this result presents a problem as it renders the binding data impossible to analyse. In theory, the binding partner should protect a portion of the protein's surface, preventing any $^{1}H/^{2}H$ exchange (Ozohanics & Ambrus, 2020). This is not what I observed. If the protein changes conformation too quickly and the binding RNA is not interacting strong enough, it may be displaced from the interaction site each time. Protein precipitation upon RNA interaction may be the cause. If the protein precipitates, it is not digested properly and is washed away instead of loaded into the mass spectrometer. In this case, the interaction with RNA remains invisible, because the spectrum shows only unbound protein fraction, which remains soluble. Shortening the length of the RNA used for MS experiments helped to keep Cal1M more soluble, but did not change the overall result and the possible interaction remained invisible in the spectrum.

My research has shown that Cal1M is an RNA-binding region of Cal1 protein, however the interaction between Cal1M and RNA is weak and further work is necessary to unravel its details.

Future work

To successfully measure the interaction of Cal1 or other centromeric protein, it is necessary to purify the full-length proteins. This has proven to be difficult in bacteria and insect cells, as the first mentioned lack the ability to deal with large unstructured proteins, and the later suffer from the native biological function of the expressed proteins and die. It might be possible to use another expression system, such as yeast or mammalian cells (Fletcher et al., 2016; McKenzie & Abbott, 2018), or even use cell-less expression with purified protein expression machinery (Bernhard & Tozawa, 2013).

HDX is theoretically capable of solving the problem and so are other methods described here. I do not think that major change of methodology is necessary. However, some additional methods may be complementary to those used here and may work better under some circumstances. One of these possible methods is XRNAX (Trendel et al., 2019), successfully used in our laboratory by Valent, 2022. Direct crosslinking and pulldown help with analysis of such low expressed proteins, such as Cal1 and Cenp-C. Covalently crosslinked protein-RNA complexes could be also further analysed by mass spectrometry, to elucidate which amino acids and nucleotides are directly responsible for the interaction.

Concluding remarks

The work presented in this thesis demonstrates that Cal1 is an RNA binding protein, however the interaction has eluded all attempts for closer investigation. A major remaining question concerns the regulation of this interaction. My results show that the Cal1M fragment binds RNA, but it does so in a very promiscuous manner. There must be a control mechanism involved, as other experiments show at least a preference for SatIII RNA, if not specificity. The control may be facilitated by the remaining parts of Cal1 that were not available during my research. It is necessary to work with the full-length protein in the future.

Another important regulator of the interaction may be RNA-modifications. It was previously described that noncoding RNAs can be modified, regulating their stability, and that this effect plays a role in cell maintenance and cancer (Cusenza et al., 2023). At least two have been described in lncRNAs: 4-acetylcytosine (Ac4C, Yu et al., 2023) and 6-methyladenine (m6A, He & Lan, 2021), but it is not known how they affect the interaction with kinetochore proteins.

Although I eventually focused on Cal1, it has been suggested that both Cid and Cenp-C are RNA binders, which also requires further investigation. This can be covered using the same approaches described here for Cal1. The only major obstacle is the difficulty in expressing and purifying these proteins.

In this thesis I showed that Cal1 is an RNA-binding protein and improved our understanding of the role of noncoding RNAs in the centromeric chromatin of *Drosophila melanogaster*.

4. Materials

4.1. Chemicals

acetic acid	Roth
acrylamide/bisacrylamide (37,5:1)	AppliChem
agar	AppliChem
agarose	Sigma-Aldrich
ammonium persulfate (APS)	AppliChem
ammonium sulfate	Invitrogen
ampicilin	AppliChem
aprotinin	AppliChem
β-mercapto ethanol	AppliChem
boric acid	Roth
bromphenol blue	AppliChem
bovine serum albumin (BSA)	Th. Geyer
chloramphenicol	AppliChem
chloroform	Sigma-Aldrich
4′,6-diamidino-2-phenylindole (DAPI)	Sigma-Aldrich
diethylpyrocarbonate (DEPC)	AppliChem
dimethyl sulfoxide (DMSO)	AppliChem
dithiothreitol (DTT)	Sigma-Aldrich
ethanol, spectroscopic grade	Sigma-Aldrich
ethanol, denatured	Sigma-Aldrich
ethidium bromide	AppliChem
ethylenediaminetetraacetic acid (EDTA)	Roth
formaldehyde	Sigma-Aldrich
glycerol	Roth
glycine	Sigma-Aldrich
guanidine hydrochloride	Sigma-Aldrich
4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES)	Sigma-Aldrich
isopropanol	Sigma-Aldrich
kanamycin	ApliChem
leupeptin	AppliChem
lithium chloride	Sigma-Aldrich
magnesium chloride	AppliChem
methanol	Sigma-Aldrich
N,N,N’,N’–tetramethylethylen-1,2-diamine (TEMED)	AppliChem
nickel sulphate	AppliChem
Nonidet P40 (NP40)	AppliChem
pepstatin	AppliChem
phenol/chloroform/isoamyl alcohol (25:24:1)	AppliChem
phenylmethylsulfonylfluoride (PMSF)	Sigma-Aldrich
Ponceau S	AppliChem
potassium chloride	AppliChem
potassium hydrogen phosphate	AppliChem
potassium dihydrogen phosphate	AppliChem
RNAse ZAP	Sigma-Aldrich

skim milk powder	Sigma-Aldrich
sodium acetate	Sigma-Aldrich
sodium chloride	Sigma-Aldrich
sodium citrate	Sigma-Aldrich
sodium dodecyl sulphate (SDS)	Sigma-Aldrich
sodium hydroxide	AppliChem
sodium hydrogen phosphate	AppliChem
sodium dihydrogen phosphate	Sigma-Aldrich
spectinomycin	Schiebel lab, ZMBH
streptomycin	Schiebel lab, ZMBH
tetracycline	Schiebel lab, ZMBH
tris(2-carboxyethyl)phosphine (TCEP)	Roth
tris(hydroxymethyl)aminomethane (tris base)	AppliChem
tris(hydroxymethyl)aminomethane-HCl (tris-HCl)	AppliChem
Triton X-100	Merck
trypan blue	Sigma-Aldrich
Tween-20	AppliChem
urea	Sigma-Aldrich

4.2. Reagents, kits, consumables, labware

1 kb Plus DNA ladder	NEB
100 bp DNA ladder	NEB
ÄKTA columns	GE Healthcare/Cytiva
blotting paper	Roth
BlueStar plus prestained protein marker	Nippon Genetics
Bradford kit	Sigma-Aldrich
cell culture plates	Greiner Bio-One
cell culture flasks	TPP
coverslips	Neolab
DNAse I	NEB
dNTPs	NEB
Dulbecco’s phosphate buffered saline (PBS)	Capricorn Scientific
foetal bovine serum (FBS)	Capricorn Scientific
gel loading dye, purple	NEB
Gibson assembly mastermix	NEB
GlycoBlue	Invitrogen
Megascript T7 in vitro transcription kit	Invitrogen
microscopy slides	Thermo Fisher Scientific
molecular cloning supplies	NEB
mounting medium	Polysciences
nitrocellulose membranes	Amersham Biosciences
Nucleospin gel and PCR clean up kit	Macherey-Nagel
Nucleospin plasmid kit	Macherey-Nagel
penicillin/streptomycin (penstrep)	Capricorn Scientific
pipette tips	Sarstedt, TipOne, Avant Guard
primers	Sigma-Aldrich
Protease inhibitor cocktail tablet	Roche

proteinase K	Thermo Fisher Scientific
Q5 DNA polymerase	NEB
QuantiTect reverse transcription kit	Qiagen
Quickstain Protein Ark	Serva
RedTaq DNA polymerase mastermix	Jena Bioscience
restriction enzymes	NEB
RNA clean-up and concentrator kit	Zymo Research
RNA ladder	NEB
RNA loading dye	NEB
RNAse A	Sigma-Aldrich
RNAsin Plus ribonuclease inhibitor	Promega
Schneider's Drosophila medium	Gibco
Script cDNA kit	Jena Bioscience
ssRNA ladder	NEB
SuperSignal™WestPico PLUS	Thermo Fisher Scientific
SuperSignal™WestFemto PLUS	Thermo Fisher Scientific
SYPRO orange	Mayer Lab, ZMBH
TRIsure	Bioline
TRIzol	Invitrogen
tubes, vials	Sarstedt, Eppendorf
TURBO DNAse	Invitrogen
Whatman paper	Sigma-Aldrich

4.3. Lab equipment

Agilent Infinity II UPLC	Agilent
Airyscan LSM900 microscope	Carl Zeiss AG
agarose electrophoresis chambers, gel trays and combs	ZMBH workshop
ÄKTA Pure liquid chromatography system	GE Healthcare/Cytiva
cell counter, LUNA	Logos Biosystems
cell culture CO_2 incubator	Thermo Fisher Scientific
centrifuges	Eppendorf
CLARIO Star plate reader	BMG Labtech
Cytospin cytocentrifuge	Thermo Fisher Scientific
Deltavision microscope	Olympus/GE Healthcare
Emulsiflex C3 cell homogeniser	Avestin
Gel Doc Go imaging system	Biorad
J-715 spectropolarimeter	Jasco
LightCycler 480	Roche
magnetic stirrers	Drehzahl, Heidolph
Nanodrop One	Thermo Fischer Scientific
PAL DHR autosampler	CTC Analytics AG
PAGE chambers, glasses and combs	Biorad
pH meter	Mettler Toledo
rollers	NeoLab, Phoenix Instruments
thermocyclers	BioRad, Nippon Genetics
timsTOF Pro mass spectrometer	Bruker Daltonics
WB chambers, gel trays	Biorad

4.4. Buffers, solvents, and mixtures

10× Phosphate buffered saline (PBS)	
NaCl	1.37M
KCl	27mM
Na_2HPO_4	100mM
NaH_2PO_4	18mM

1× SDS running buffer	
tris base	25mM
glycine	192mM
SDS	0.1% (w/v)

4× Lämmli sample buffer	
tris-HCl	50mM, pH=6.8
glycerol	10% (w/v)
SDS	2% (w/v)
β-mercaptoethanol	0.5% (v/v)
bromphenolblue	0.02% (w/v)

50× Tris-acetate-EDTA (TAE)	
tris-HCl	2M
EDTA	50mM
acetic acid	1M (to pH = 7.7)

Protease inhibitor solution A (1:1000)	
aprotinin	1 mg/ml
leupeptin hemisulphate	1 mg/ml
water	

Protease inhibitor solution B	
pepstatin	0.5 mg/ml
ethanol	

RIPA buffer	
NaCl	150mM
tris-HCl	10mM
EDTA	1mM
Triton X-100	1% (v/v)

separation gel buffer	
tris base	1.5M; pH=8.8

stacking gel buffer	
tris base	0.5M; pH=6.8

TBST	
tris base	20mM
NaCl	150mM
Tween 20	0.1% (w/v)

transfer buffer	
tris base	3.03 g
glycine	14.40g
CH_3OH	200 ml
H_2O	fill to 800 ml

Lysis buffer, Ni^{2+} affinity chromatography		
NaCl	500mM	
$Na_2HPO_4 + NaH_2PO_4$	50mM, pH = 8.0	
Imidazole	5mM	
Glycerol	10 % (w/v)	
Protease inhibitors mix A	1 ml/l	
Protease inhibitors mix B	0.5 ml/l	
PMSF	1 ml/l	
Lysozyme	1 ml/l	added right before using
DTT	1 ml/l	
Basemuncher	1 µl/l	

Elution buffer, Ni^{2+} affinity chromatography	
NaCl	500mM
$Na_2HPO_4 + NaH_2PO_4$	50mM, pH = 8.0
Imidazole	300mM
Glycerol	10 % (w/v)
DTT	1 ml/l, added right before using

Lysis buffer, MBP affinity chromatography		
NaCl	500mM	
$Na_2HPO_4 + NaH_2PO_4$	50mM, pH = 8.0	
Glycerol	10 % (w/v)	
Protease inhibitors mix A	1 ml/l	
Protease inhibitors mix B	0.5 ml/l	
PMSF	1 ml/l	
Lysozyme	1 ml/l	added right before using
DTT	1 ml/l	
Basemuncher	1 µl/l	

Elution buffer, MBP affinity chromatography	
NaCl	500mM
$Na_2HPO_4 + NaH_2PO_4$	50mM, pH = 8.0
maltose	10mM
Glycerol	10 % (w/v)
DTT	1 ml/l, added right before using

Lysis buffer, GST affinity chromatography		
NaCl	500mM	
Na_2HPO_4 + NaH_2PO_4	5mM, pH = 8.0	
Glycerol	10 % (w/v)	
Protease inhibitors mix A	1 ml/l	
Protease inhibitors mix B	0.5 ml/l	
PMSF	1 ml/l	
Lysozyme	1 ml/l	added right before using
DTT	1 ml/l	
Basemuncher	1 µl/l	

Elution buffer, GST affinity chromatography	
NaCl	500mM
Na_2HPO_4 + NaH_2PO_4	50mM, pH = 8.0
GST	10mM
Glycerol	10 % (w/v)
DTT	1 ml/l, added right before using

Size exclusion chromatography buffer	
NaCl	500mM
Na_2HPO_4 + NaH_2PO_4	50mM, pH = 8.0
Glycerol	10 % (w/v)
DTT	1 ml/l, added right before using

Quench buffers for MS - testing	
NaCl	500mM
Na_2HPO_4 + NaH_2PO_4	50mM, pH = 2; 2.5; 3; 3.5; 4
DTT	1mM

Quench buffers for MS – MS2	
Glycine-HCl	1M, pH = 2.3
Guanidine-HCl	6M
TCEP	500mM

Quench buffers for MS - $^1H/^2H$	
Glycine-HCl	1M, pH = 2.3
Thiourea	2M
TCEP	400mM

4.5. Antibodies

primary:

actin, mouse, 1:10000 WB	Milipore (MAB1501)
Cal1, rabbit, 1:3000 WB, 1:1000 IF	Erhardt Lab, (Bade et al., 2014)
Cenp-C, guinea pig, 1:5000 IF	Covance
Cenp-C, rabbit, 1:1000 WB	MSB
Cid, rabbit, 1:2000 WB	Active motif (39713)
Cid, chicken, 1:700 IF	Heun Lab, University of Edinburg
tubulin, mouse, 1:5000 WB	Sigma (T9026)

secondary:

AlexaFluor488 α-rabbit, goat, 1:500 IF	Thermo Fischer Scientific
AlexaFluor488 α-chicken, goat, 1:500 IF	Thermo Fischer Scientific
AlexaFluor546 α-guinea pig, goat, 1:500 IF	Thermo Fischer Scientific
AlexaFluor647 α-chicken, goat, 1:500 IF	Thermo Fischer Scientific
IgG HRP α-mouse, rabbit, 1:10000 WB	Sigma (A9044)
IgG HRP α-rabbit, goat, 1:5000 WB	Sigma (A0545)

4.6. Primers

Cal1, pGEX6p1, fwd - ccaggggccccctgggatccccggaattcatggcgaatgcggtggtg
Cal1, pGEX6p1, rew - tcgtcagtcagtcacgatgcggccgcttacttgtcaccggaattattctcgag
Cal1, pETM44, fwd - gaagttctgttccaggggcccatggaaatggcgaatgcggtggtg
Cal1, pETM44, rev - aagcttgtcgacggagctcgaattcttacttgtcaccggaattattctcgag
Cal1, pCA528, fwd - ccaccatcgggcgcgatgggtcatcaccatcatc
Cal1, pCA528, rev - gtaggcctttgaattttacttgtcaccggaattattctc
Cal1, pFastBac1, fwd - tcgacgagctcactagtcgcggccgcatgggtcatcaccatcatc
Cal1, pFastBac1, rew - gacaagcttggtaccgcatgcttacttgtcaccggaattattc
Cal1N, pCA528, fwd - aggctcacagagaacagattggtgggatggcgaatgcggtg
Cal1N, pCA528, rev - gatccggtctcccaccttaggcatccatcggttggt
Cal1M, pCA528, fwd - aggctcacagagaacagattggtggggagacacaggaaaagactagg
Cal1M, pCA528, rev - gatccggtctcccaccttatgccgtagttgaacattgtg
Cal1C, pCA528, fwd - aggctcacagagaacagattggtggggtcagggaagaggagtcc
Cal1C, pCA528, rev - gatccggtctcccaccttacttgtcaccggaattattctcga
Cenp-C, pCA528, fwd - ccaccatcgggcgcgatgggtcatcaccatcatc
Cenp-C, pCA528, rev – gtaggcctttgaattttaacccctgtttgcca
Cenp-C, pFastBac1, fwd - tcgacgagctcactagtcgcggccgcatgggtcatcaccatcatc
Cenp-C, pFastBac1, rew - gacaagcttggtaccgcatgcctaactgcgtatacacatc
Cenp-C1, pCA528, fwd - aggctcacagagaacagattggtgggatgtcgaagccccagaa
Cenp-C1, pCA528, rev - gatccggtctcccaccttacattagatttctacgtagcagctcc
Cenp-C2, pCA528, fwd - aggctcacagagaacagattggtggg aagcttatcctagaagaagagattcag
Cenp-C2, pCA528, rev - gatccggtctcccaccttatatacttgcttgcccatggtc
Cenp-C3, pCA528, fwd - aggctcacagagaacagattggtgggatactgaggagaagtggaaaaaaattg
Cenp-C3, pCA528, rev - gatccggtctcccaccttaggaagctgaggaactaaatactacc
Cenp-C4, pCA528, fwd - aggctcacagagaacagattggtgggaccggtattcggagatcaaa
Cenp-C4, pCA528, rev – gatccggtctcccaccttactaactgcgtatacacatcagcac

4.7. Plasmids

pCA528, N-terminal SUMO tag, bacterial expression
pFastBac1, insect cell expression, bacmid construction plasmid
pGEX6p1, N-terminal GST tag, bacterial expression
pET24a, C-terminal His tag, bacterial expression
pET19b, N-terminal His tag, bacterial expression
pCA535, Ulp1 SUMO protease bacterial expression plasmid
pETM, N-terminal MBP tag, bacterial expression
pMT, insect expression vector

4.8. Cell lines

E. coli:	
BL21DE3	Erhardt lab, ZMBH, Heidelberg
BL21DE3 codon plus	Schiebel lab, ZMBH, Heidelberg
DH5α	Erhardt lab, ZMBH, Heidelberg
DH10bac	Schiebel lab, ZMBH, Heidelberg
pLysS	Schiebel lab, ZMBH, Heidelberg
RIL	Schiebel lab, ZMBH, Heidelberg
Rosetta	Schiebel lab, ZMBH, Heidelberg
TOP10	Thermo Fischer Scientific
insect cells:	
Schneider S2	Erhardt lab, ZMBH, Heidelberg
High Five	Melchior lab, ZMBH, Heidelberg
Sf21	Melchior lab, ZMBH, Heidelberg

5. Methods

5.1 Molecular Biology

5.1.1. Construct design

The genes I used in this study were obtained from either our lab, friendly labs, Addgene (www.addgene.com), or amplified from cDNA. I acquired gene sequences from the plasmid maps of our lab, Addgene or from UniProt (www.uniprot.org). I designed primers using the NEBuilder online tool (https://nebuilder.neb.com/) to contain restriction sites for restriction the chosen endonucleases. I only chose unique cleavage sites, which I checked using the NEBCutter online tool (https://nc3.neb.com/NEBcutter/). I processed assembled sequences, plasmid maps and other sequence relevant information using the SnapGene software. I ordered primers from SigmaAldrich.

5.1.2. mRNA extraction and cDNA preparation

I extracted mRNA from S2 fly cells and from HEK293T human cells. The cells were detached from the bottom of the flask by repeated pipetting, collected in a 15 ml tube, and centrifuged at 300×g for 7 min. I resuspended the resulting pellet in TRIzol (Invitrogen) and stored it at -80 °C prior to use. Subsequently I purified the RNA from the mixture by phenol-chloroform extraction. I centrifuged the cell lysate at 20 000×g at 4 °C for 30 min and transferred the aqueous phase to another tube. I added an equal volume of pure $CHCl_3$ and then centrifuged it again at 20 000×g at 4 °C for 15 min. I repeated this step once more to improve the purity of the final sample. I mixed the final aqueous phase with an equal volume of isopropanol and kept for 30 min at -80 °C to precipitate the RNA. I pelleted the precipitate by centrifugation at 20 000×g at 4 °C for 30 min. I carefully decanted the supernatant and the washed the pellet with 500 µl of 70% EtOH and centrifuged it again at 20 000×g at 4 °C for 5 min. I repeated the same process with 200 µl of 100% EtOH and air dried the final pellet and dissolved it in water. I measured the concentration of the purified RNA using nanodrop and stored it at -80 °C until further use.

The cDNA was prepared using the Script cDNA kit (Jena bioscience).

5.1.3. Molecular cloning

DNA amplification

I amplified genes of interest by PCR from either plasmids or cDNA. The PCR reaction used Q5 DNA polymearase (NEB). The PCR reaction was performed using the following protocol:

5 µl	Q5 reaction buffer	98 °C	60 s	
5 µl	Q5 GC enhancer	98 °C	30 s	
2 µl	dNTPs	X °C	30 s, depending on primers	30×
2×1 µl	primers	72 °C	30 s per kbp	
0,25 µl	Q5 DNA polymerase	72 °C	5 min	
25 µg	DNA template	4 °C	∞	
to 25 µl	H_2O			

I analysed the product of the PCR reaction by agarose gel electrophoresis. Once I confirmed the correct size of the DNA fragment, I purified the DNA from the PCR reaction using the NucleoSpin PCR&gel clean-up kit (Macherey-Nagel) and measured its concentration by nanodrop.

T4 ligation

The plasmid of interest was digested using the same set of restriction endonucleases as the gene of interest to produce compatible DNA strand ends. I performed the cleavage reaction according to the NEB protocol specific for each restriction endonuclease used. I separated cleaved plasmid fragments by agarose gel electrophoresis. I excised the correct fragment from the agarose gel with a scalpel and dissolved it in NTI buffer (Macherey-Nagel). I purified both cleaved fragments, the linearised plasmid and cleaved insert, using the NucleoSpin PCR&gel clean-up kit (Macherey-Nagel) and measured their concentration by nanodrop. I used the purified fragments for the T4 ligase reaction according to the manufacturer's protocol (NEB, T4 ligation). I further used the resulting mixture for transformation of bacteria.

Gibson and HiFi assembly

Some of the constructs I used I generated using other ligation methods – Gibson and HiFi assembly and CPEC. (Gibson et al., 2019; Quan & Tian, 2009)

The primers I designed in NEBuilder allow for the use of both of these methods, as well as T4 ligation, without the need for a new set. I linearised the plasmid the same way as in the T4 ligation protocol, but the insert was not cleaved by restriction endonucleases. Instead, I purified it after PCR using the PCR&gel clean-up kit (Macherey-Nagel). I subsequently mixed the fragments and incubated them according to the manufacturer's protocol (NEB, HiFi assembly). I further used the resulting mixture for transformation of bacteria.

Transformation

I used ligated plasmids to transform the bacterial strains used for DNA amplification. Our laboratory uses either home-made DH5α or commercial TOP10 (Thermo Fisher Scientific) *E. coli* strains. I performed the transformation according to the following protocol. I took bacteria out from the -80 °C cold storage and thawed them on ice. I pipetted plasmid DNA into each tube (50-150 ng per tube) and incubated the mixture on ice for 10-30 min. Following incubation, I exposed the bacteria to a heat shock at 42 °C for 1 min and then cooled them on ice for 2 min. I allowed them to recover by adding 300 µl of LB or SOC media per tube and incubated them at 37 °C for 1 h. After that I seeded them onto agar plates containing appropriate antibiotic and incubated at 37 °C. After overnight incubation, I further analysed the colonies.

Construct analysis

I picked grown colonies with a pipette tip and transferred them to both LB with appropriate antibiotic and to the colony PCR reaction mixture. I used RedTaq (Jena Bioscience) mastermix for the colony PCR reaction along with plasmid backbone specific primers (T7 or plasmid specific tag sequence). I picked up each colony with a pipette tip and briefly immersed it in the prepared PCR mastermix as well as 50 µl LB medium. I performed colony PCR according to the following protocol:

3 µl	RedTaq mastermix	98 °C	2 min	
2×1 µl	primers	98 °C	30 s	30×
10 µl	H_2O	X °C	30 s, depending on primers	
tip dip	bacterial colony	72 °C	30 s per kbp	
		72 °C	5 min	
		4 °C	∞	

I visualised the PCR product by agarose gel electrophoresis. I then expanded the previously prepared LB cultures of positive clones to 3-5 ml liquid culture and incubated them at 37 °C for at least 8 h or overnight.

I subsequently harvested these minipreps by centrifugation and extracted the plasmid DNA using the NucleoSpin plasmid purification kit (Macherey-Nagel). The purified plasmid was sequenced by Sanger sequencing reaction (Microsynth, Eurofins). In the end I compared the resulting sequence to the plasmid map generated by NEBuilder.

5.1.4. Optimisation of protein expression and purification

The previous purification protocol used in the lab was suboptimal for several reasons. The elution buffer contained 2M NaCl which improves the stability of centromeric proteins in solution, but severely limits the possibilities to investigate any electrostatic interactions or using affinity and ion exchange chromatography for purification. This amount of NaCl would also have a significant impact on the subsequently used methods, such as electrophoretic mobility shift assay and mass spectrometry.

Based on the paper by Klare et al., (2015) and former lab protocols, I eventually chose a 50mM phosphate buffer with a pH of 8.0, containing 500mM NaCl, 5mM imidazole, 1mM DTT, and 10% w/v glycerol. This buffer was subsequently used for most of the following work. The proteins I selected for this work remained stable throughout the required experimental period and could be flash-frozen in liquid nitrogen and stored at -80 °C without significant degradation.

Another improvement of the original protocol was in the affinity chromatography purification step. The original protocol used beads coated with a Ni-chelating matrix. These beads were then incubated with the cell lysate to bind the proteins of interest, washed, and eluted using buffer with a higher imidazole concentration. I started using the liquid chromatography system ÄKTA Pure (GE Healthcare, later Cytiva) early on, and used it for most of the work described in this thesis. High pressure liquid chromatography (HPLC), or in biomedical fields often referred to as fast protein liquid chromatography (FPLC), offers better resolution, purer samples and various methods of purification depending on the columns used. It also provides a high degree of automation, which

accelerates the entire protocol and minimises any sudden changes in the environment that could harm the delicate protein sample.

5.1.5. Expression tests

I used plasmids containing the correct sequences to transform the bacterial expression strains. The transformation protocol was the same as described in the previous chapter.

I selected single colonies and used them to inoculate 5 ml cultures in LB medium with the appropriate antibiotic. I incubated these cultures overnight at 37 °C, 180 RPM and then used them to seed larger volumes of LB medium for the expression tests. ***Table 2*** lists all of the conditions that I tested.

I collected 1 ml samples at selected timepoints. The samples were then centrifuged at 11000×g for 1 min and the pellet was resuspended in Lämmli buffer (2x concentrated, 1:1 volume ratio). I loaded the mixture onto a polyacrylamide gel and subjected it to electrophoresis (0,04 A, 1 h). I used different percentage of the gel depending on the size of the protein of interest. 15% SDS-PAGE for proteins smaller than 50 kDa, 12% SDS-PAGE for proteins between 50 and 100 kDa, and 10% SDS-PAGE for proteins larger than 100 kDa. After electrophoresis, I stained half of the gels using quickstain solution to check the expression levels, while the other half I used for Western blot analysis.

5.1.6. Western blot

I placed the SDS-PAGE gels in the blotting chamber with the nitrocellulose membrane facing them in a sandwich. I subsequently covered the gel-membrane sandwich with Whatman paper and a sponge from both sides and closed the blotting chamber. The blotting chamber was submerged in the transfer buffer and I performed Western blotting at 4 °C, 100 V, and 1-2 h depending on the protein size. After the transfer was done, I briefly washed the nitrocellulose membrane, which now contained proteins, with TBST buffer and then blocked it using a blocking solution (5% skim milk in TBST). I added the primary antibody after blocking, using the recommended concentration, and then I incubated the submerged membrane at 4 °C overnight on the shaker. In the morning, I washed the membrane 3 times with TBST (5-10 ml). Then, I changed the blocking buffer for a fresh one and added the secondary antibody was added using recommended

concentration. I used horseradish peroxidase coupled antibodies. I incubated the secondary antibody at room temperature for 1 h on the shaking incubator. After the incubation, I briefly washed the membrane was 3 times with 5-10 ml TBST and once with water.

I developed the membrane using SuperSignal™ WestPico (or Femto) PLUS chemiluminescent substrate (Thermo Scientific) and visualised it with the Gel Doc Go (Biorad). I used the conditions and strains that showed the best yields for subsequent protein expression and purification.

5.1.7. Bacterial expression

The proteins used in this study were primarily expressed and purified using *E. coli* expression strains. Transformed cells were stored as glycerol stocks at -80°C. I initiated starter cultures by scraping a frozen glycerol stock with a pipette tip and transferring it to 5 ml of growth media with the appropriate antibiotics. These starter cultures were then incubated overnight at 37°C and 180 RPM. I used starter cultures to inoculate 1 L of growth media in 5 L Erlenmeyer flasks in the morning. The cultures were incubated at 37°C with constant shaking for several hours. Optical density at 600 nm (OD600) was measured using a nanodrop or spectrophotometer every hour. The culture was grown until the end of the exponential growth phase, depending on the growth media used. For LB, this is OD600 = 0.6-0.8, and for TB, it is OD600 = 1.0-1.2. Upon reaching the desired OD600, the culture was cooled to the specific expression temperature. Following cooling, I added 1mM IPTG to induce expression. Once expression was complete, I collected the culture and centrifuged it at 4°C and 4000×g for 15 minutes. I discarded the supernatant, and the pellet was washed with PBS. After washing, the cells were pelleted again using the same centrifugation step. The supernatant was once again discarded, and I stored the pellet at -80°C until purification.

5.1.8. Insect cells construct preparation and expression

I used the commercially available Invitrogen Bac-to-Bac system for the preparation of constructs for insect cell expression. Firstly, the gene of interest was inserted into the pFastBac1 plasmid using molecular cloning, as described in **5.1.3**. This plasmid carries sites for translocation flanking the inserted gene of interest. The resulting construct was

then used for the transformation of DH10Bac packaging cells. Translocase is used by these cells to transfer the sequence from the plasmid into the bacmid, which is a larger plasmid containing the baculoviral genome. I picked positive clones using the Blue/White selection process and were then used to inoculate 10 ml miniprep cultures. I confirmed the sequence quality by PCR and sequencing.

I purified the Bacmid DNA using the protocol from the Schiebel lab with a NucleoSpin plasmid purification kit (Macherey-Nagel). The pelleted bacteria were resuspended in 300 µl of A1 solution. Next, 300 µl of A2 solution was added, and the mixture was incubated at room temperature for 5 minutes. Following this, 350 µl of A3 solution was added, and the mixture was centrifuged for 10 minutes at 16,000×g. The supernatant was then collected in a new tube, and 640 µl of cold isopropanol was added. The resulting mixture was centrifuged again, this time for 30 minutes at 16,000×g, and the supernatant was collected. The supernatant was removed, and the pellet was washed with 1 ml of 70% EtOH. The mixture was then centrifuged again for 5 minutes at 16,000×g. The EtOH was pipetted out and the pellet was left to dry before being dissolved in water. At this point I performed another control PCR to ensure the quality of the sequence.

I used the bacmid DNA to transfect the expression strain of insect cells, either High Five or Sf21. The confluent cells (7×10^5 ml^{-1}) were seeded on a 6-well plate and allowed to settle. Meanwhile, 10 µg of bacmid DNA was dissolved in 200 µl of Sf900 media, and 14 µl of cellfectin was dissolved in another 200 µl. Both mixtures were combined and incubated at room temperature for 20 minutes to form a transfection mixture. The settled cells were undisturbed as 1 ml of the transfection mixture was pipetted on top, replacing the old medium. The cells were then incubated at 27 °C for 5 hours. After this time, the medium was discarded and replaced with fresh Sf900. The cells were then incubated for another 72 hours at 27 °C to produce the P0 generation of baculoviruses.

The culture was grown and collected, then pelleted by centrifugation. I collected the supernatant, which now contains viral particles, and filtered it through a 0.22 µm syringe filter. The filtered virus stock was stabilised by adding 5% FBS, flash frozen with N2(l), and stored at -80°C for a few weeks before usage. The virus stock was expanded by using this mixture for subsequent infection of insect cells and harvesting of the newly formed viral particles in the same way as described earlier.

I prepared the liquid insect cell culture for protein production by baculoviral infection. Cells were cultured to a density of 0.6-0.8×10^6 per ml in 100 ml of medium and infected with 1 ml of pre-prepared viral stock. The cells were incubated for 2-3 days, depending on their condition, and observed daily under a microscope to assess the progression of the infection. Once the majority of cells ceased dividing and became enlarged and deformed, they were allowed to produce proteins for one day before being harvested by centrifugation at 300×g. The cells were frozen rapidly in liquid nitrogen and stored at -80 °C for a few days before purification.

5.1.9. Protein purification

I resuspended harvested cell pellets in cold lysis buffer by shaking and vortexing. The mixture was kept on ice throughout the protocol. After resuspension, I lysed cells using Emulsiflex C3 (Avestine). The lysate was collected and cleared by centrifugation at 40,000×g for 40 minutes. I filtered the cleared lysate through a 0.45 µm filter and used it for affinity chromatography.

I used the ÄKTA Pure chromatography system to load the lysate onto an affinity column. The work utilized HisTrap, MBPTrap, and GSTrap columns (Cytiva), depending on the used protein affinity tag. The lysate was pumped through the column using the manufacturer's recommended pressure limit of 0.5 MPa. Once all of the lysate had been pumped through the column, it was washed with 5 CV of lysis buffer until the flowthrough had a constant UV absorbance and conductivity. The protein was eluted with an elution buffer gradient ranging from 0 to 100% over a period of 10 minutes. The eluate was collected using an automated fraction collector and stored at 4°C or on ice.

To remove excess imidazole from the previous elution, I loaded the eluate from nickel chelation chromatography onto a buffer exchange column using the ÄKTA Pure chromatography system. The column was washed with 5 CV of final buffer prior to use. The protein eluate was collected using an automated fraction collector and kept on ice. The SUMO-His tag was then cleaved by Ulp1 SUMO protease at 4 °C overnight. The resulting mixture was loaded onto the nickel chelating column again, and the flowthrough was collected.

I used NiNTA beads for some of the purifications. The beads were added to the bacterial lysate and incubated at 4 °C on the roller for 1 hour. After this time, the lysate was poured into a syringe with a glass frit at the bottom. The liquid part was discarded, and the beads

were collected. They were washed three times with lysis buffer to remove non-specifically bound proteins. The protein was eluted using the elution buffer with 300 mM imidazole and kept on ice.

For some of the experiments described in this work, the protein sample's purity was insufficient after affinity chromatography. In these cases, I added a round of size exclusion chromatography at the end to obtain purer protein samples. The eluate was loaded onto a Superdex 200 Increase size exclusion column using the ÄKTA Pure chromatography system. The eluate was passed through the column using the manufacturer's recommended pressure limit of 2 MPa, and the flowthrough was collected using an automated fraction collector. The eluate was kept on ice.

Insoluble protein purification protocol

The proteins described in the work are large and unstructured, which leads to their poor solubility. After cell lysis and clearing of the lysate, the pellet was usually discarded. Here I describe purification in denaturing conditions to determine whether the protein is in fact being expressed but is insoluble.

In this case, the pellet was resuspended in lysis buffer containing 6M guanidium hydrochloride and 2% Triton X-100. The mixture was shaken at room temperature for 2 hours and then centrifuged at 20,000×g at 4 °C for 20 minutes. I filtered the resulting supernatant through a 0.22 um syringe filter. I poured the protein sample into NiNTA beads contained in a syringe with a frit glass bottom. Excess liquid was washed through with a lysis buffer containing a urea gradient from 8M to 0. After the urea wash, the protein was eluted using an elution buffer containing 300mM imidazole and collected. The eluate was kept on ice for protein analysis and quality control.

Protein analysis and quality control

Following each chromatography step, 10 µl of every eluted fraction was collected and used for SDS-PAGE. The samples were mixed with SDS loading buffer, loaded onto a polyacrylamide gel, and subjected to electrophoresis (0.04 A, 1 h). The gels were then stained using a quickstain solution.

Finally, after the last purification step, the protein concentration was measured using a nanodrop with elution buffer as a blank solution. The protein samples were found to have

significant nucleic acid contamination, leading to incorrect results when analysed using UV-VIS spectroscopy. To address this issue, the Bradford method (kit) was employed to determine the concentration of the samples. Following this, the samples were flash frozen in 50 µl aliquots using liquid nitrogen and stored at -80 °C.

After every chromatography step 10 µl of each eluted fraction was collected and used for SDS-PAGE. Samples were mixed with SDS loading buffer, loaded to polyacrylamide gel and subjected to electrophoresis (0.04 A, 1 h). After electrophoresis the gels were stained using quickstain solution.

After the last purification step the protein concentration was measured using nanodrop with elution buffer as a blank solution. Some proteins had large contamination of nucleic acids and UV-VIS spectroscopy of the pure sample was not yielding correct results. In that case Bradford method (kit, Thermo Fisher) was used. When the concentration of the protein samples was determined they were flash frozen in 50 µl aliquots in liquid nitrogen and stored in -80 °C.

5.1.10. In vitro transcription

RNA was prepared using in vitro transcription reaction. Megascript T7 kit (Invitrogen) was used. Either cleaved plasmid or synthetised oligos were used as a template, both using T7 promoter. Manufacturers protocol was used:

2 µl	T7 buffer	37 °C overnight
4×2 µl	NTPs	
2 µl	T7 RNA polymerase	
0.2-1 µg	DNA template	
To 20 µl	RNAse free H_2O	

After overnight incubation, 2 µl of TURBO DNAse was added and the mixture was incubated at 37°C for 1 hour.

I purified the RNA from the reaction mixture using phenol-chloroform extraction. Following DNAse digestion, 115 µl of H2O and 15 µl of 5M NH4OAc were added. The reaction mixture was mixed with an equal volume of PhOH/CHCl3/isoamylol (25:24:1) and thoroughly mixed. It was then centrifuged at 20,000×g at 4 °C for 10 minutes, and

the aqueous phase was transferred to a new tube. Another extraction step was performed using pure $CHCl_3$. The resulting mixture was centrifuged under the same conditions as the previous step. The aqueous phase was collected in a new tube and mixed with an equal volume of cold isopropanol. The mixture was then kept at -80 °C for 30 minutes to precipitate the RNA. Afterward, it was centrifuged at 20,000×g at 4 °C for 30 minutes. The supernatant was carefully decanted, and the pellet was washed with 500 µl of 70% EtOH and 200 µl of 100% EtOH. After each washing step, the sample was centrifuged at 20,000×g at 4 °C for 5 minutes. The resulting pellet was allowed to dry and, once all traces of ethanol had evaporated, it was dissolved in water. The concentration of the purified RNA was measured using a nanodrop and it was stored at -80 °C until further use.

5.2. Biochemistry and biophysics

5.2.1. Immunofluorescence

I split Schneider S2 insect cells to a concentration of 1×10^5 cells per ml and allowed to recover overnight under normal cultivation conditions (room temperature, Schneider S2 medium) in a 12-well plate. The following morning, 500 µl of colcemid was added to each well to arrest the cell cycle at metaphase for 1 hour. The cells were then harvested, pelleted by centrifugation at 800×g for 5 minutes, resuspended in 500 µl of hypotonic solution (0.5% sodium citrate), and allowed to swell for 10 minutes. The cells were transferred to a microscopy slide using a cytospin centrifuge (Thermo Fisher Scientific, 900×g, 10 min). After attachment to the glass slide, RNAse A (10 µg per slide) or PBS in the control samples was applied for 1 h. The cells were then gently washed with PBS and fixed with 4% formaldehyde for 10 min. After fixation, the cells were gently washed with PBS and blocked for 1 h with 4% BSA. The cells that were blocked were washed with PBS and then incubated with the primary antibody. After incubation overnight in a humid chamber at 4°C, they were washed with PBS and incubated with the secondary antibody for 1 hour at room temperature. Finally, DNA was stained with DAPI (1:1000 dilution) for 5 minutes at room temperature. The cells were then washed with PBS, carefully dried, and covered with a cover slip. I performed the visualisation using an Airyscan LSM900 microscope.

5.2.2. Electrophoretic mobility shift assay

I used EMSA to investigate RNA-protein interactions in vitro. RNAse-free agarose and DEPC-treated TAE buffer were used for all EMSAs. RNAse Zap was used to clean all surfaces when preparing the gel and assembling the electrophoretic cell. Samples were mixed in PCR tubes and RNAsin was added to each tube. They were incubated at 4°C for 30 minutes, mixed with 2x RNA loading dye, and loaded onto a 1% agarose gel. The electrophoresis ran at 120 V for about an hour until the dye reached the end of the gel. To maintain a constant temperature, the electrophoretic chamber was placed in a larger container filled with ice. After the electrophoresis, the gel was stained for 10 minutes using a TAE buffer bath containing 0.1 ‰ ethidium bromide. The gels were visualised using Gel Doc Go (Biorad).

5.2.3. Fluorescence anisotrophy

I employed fluorescence anisotropy to measure the binding affinity between SatIII RNA and Cal1M protein. The RNA was prepared through in vitro transcription, with the addition of a 25% of Alexa-labelled UTPs. A dilution series of labelled satellite RNA was pipetted into a 96-well plate and measured using CLARIO Star plate reader to establish the optimal experimental settings. The optimal concentration of labelled RNA was selected, and the reaction mixture was consistently maintained at this concentration. For the experiment, a dilution series of Cal1M was used with a stable amount of labelled satellite RNA.

5.2.4. Circular dichroism

I used circular dichroism to assess the quality and stability of the purified protein sample. The sample was stored at -80°C for an extended period, and CD was used to ensure its stability under changing conditions before preparing it for mass spectrometry.

The protein sample was thawed on ice and transferred to a quartz cuvette, and its CD spectrum was measured using a Jasco CD spectropolarimeter. The spectrum was compared with standard curves using GraphPad Prism. Another spectrum was measured using temperature perturbation from 10-85 °C. A wavelength of 222 nm was selected as it corresponds to the α-helix absorption maximum. The cuvette chamber was heated

evenly from room temperature to 90 °C while spectrophotometric measurements were taken. The data were processed using Origin.

5.2.5. Differential scanning fluorimetry

I measured thermodynamic parameters of the Cal1M-RNA interaction using differential scanning fluorimetry. The protein amount was optimised through preliminary experiments, and a concentration of 8 μM was used for all experiments. The concentration of all RNAs used was kept equimolar to that of Cal1M. Both proteins and RNAs were thawed on ice, centrifuged at 20,000×g for 20 minutes to separate any potential precipitate, and mixed to a final concentration in a tube. SYPRO orange was used as a fluorescent dye in a 160× diluted concentration. The mixture was then transferred to a 384 well plate and shortly centrifuged at 1000×g to remove bubbles before being inserted into the CLARIO Star. I carried out data analysis using Excel and Origin.

5.3. Mass spectrometry

5.3.1. MS2 sequence coverage

For mass spectrometry, a different set of buffers was required compared to other methods I used previously. The solubility and stability of Cal1M and various RNAs in the quench buffer necessary for MS-$^{1}H/^{2}H$ had to be probed. The quench buffer is designed to rapidly decrease pH to stop the hydrogen/deuterium exchange and freeze the reaction in place, allowing it to be probed by the mass spectrometer.

The thawed Purified Cal1M was centrifuged for 10 minutes at 4 °C at 20,000×g and then diluted 1:3 in the quench buffer. To determine the best conditions, several quench buffers were tested, and two were ultimately used for the experiments: one with 6M guanidine for the protein samples and another with 2M thiourea for the samples containing RNA.

After quenching, the protein sample is injected into a liquid chromatography system that pumps it through a column with immobilised protease to digest the protein into defined peptides. Several different protease columns were used. The first experiments were done with pepsin, but later nepentesin was used with better results. The data were analysed using Bruker DataAnalysis software.

5.3.2. Hydrogen/deuterium exchange

In total 100 pmol of protein at a concentration of 20μM was used per sample and mixed with RNA at an equimolar ratio. Protein samples were 10× diluted by D_2O-based buffer. $^1H/^2H$ reaction followed for 20 s, 2 m, 20 m and 2 h at 20 °C or later for 2 s, 4 s, 10 s 90 s, 20 min and 2 h at 4 °C. The exchange reaction was quenched by the addition of ice-cold quenching buffer containing 1 M glycine-HCl, pH 2.3, 6M guanidine and 0.5M TCEP at 1:1 (v/v). The $^1H/^2H$ exchange reactions were prepared using a PAL DHR autosampler (CTC Analytics AG) controlled by Chronos software (AxelSemrau) or by manual pipetting for the shortest time scales. The injected sample was delivered onto a custom-made nepenthesin-2 protease column (bed volume 66 μL) and subsequently onto the trap column (SecurityGuard™ ULTRA Cartridge UHPLC Fully Porous Polar C18, 2.1mm ID, Phenomenex) under the flow of 0.4% formic acid in water driven by the 1260 Infinity II Quaternary pump under the flow rate of 200 μL min^{-1}. After 3 minutes, desalted peptides were eluted and separated using an analytical column (Luna Omega Polar C18, 1.6 μm, 100 Å, 1.0x100 mm, Phenomenex) under a water-ACN gradient (10 %–45 % in 6 min; solvent A: 0.1% formic acid in water, solvent B: 0.1% formic acid, 2% water in ACN). The water-ACN gradient was delivered by the 1290 Infinity II LC pump under the flow rate of 40 μL min^{-1}. After protein digestion and desalting step, protease column was washed by injection of 100 μL of 150 mM triethylammonium acetate, pH 7.5, 8 M urea. After water-ACN gradient, the wash step of the trap and analytical column proceeded for 5 min under the flow of 150 mM triethylammonium acetate, pH 7.5 in 80% ACN in water at a flow rate of 60 μL min^{-1} delivered by the 1290 Infinity II LC pump. To minimize the deuterium back-exchange, the LC system was refrigerated to 0 °C. The LC system involved the temperature-controlled box and Agilent Infinity II UPLC (Agilent Technologies) directly connected to an ESI source of timsTOF Pro (Bruker Daltonics).

The mass spectrometer operated in the MS mode with a 1Hz data acquisition rate. Acquired LC-MS data were peak-picked and exported in DataAnalysis (v. 5.3, Bruker Daltonics) and further processed by the DeutEx software. (Trcka et al., 2019) Data visualization was performed using MSTools (http://peterslab.org/MSTools/index.php, (Kavan & Man, 2011)). For peptide identification, the same LC-MS system was used but the mass spectrometer was operated in data-dependent MS/MS mode using PASEF. The LC-MS/MS data were searched using MASCOT (v. 2.7, Matrix Science) against a customized database combining sequences of Cal1 and used proteases. Search parameters

were set as follows: no-enzyme, no modifications allowed, precursor tolerance 10 ppm, fragment ion tolerance 0.05 Da, decoy search enabled, FDR < 1%, IonScore > 20 and peptide length > 5.

Bibliography

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

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

Bade, D., Pauleau, A.-L., Wendler, A., & Erhardt, S. (2014). The E3 Ligase CUL3/RDX Controls Centromere Maintenance by Ubiquitylating and Stabilizing CENP-A in a CAL1-Dependent Manner. *Developmental Cell*, *28*(5), 508–519. https://doi.org/10.1016/j.devcel.2014.01.031

Barnhart, M. C., Kuich, P. H. J. L., Stellfox, M. E., Ward, J. A., Bassett, E. A., Black, B. E., & Foltz, D. R. (2011). HJURP is a CENP-A chromatin assembly factor sufficient to form a functional de novo kinetochore. *The Journal of Cell Biology*, *194*(2), 229–243. https://doi.org/10.1083/jcb.201012017

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), 1848–1857. https://doi.org/10.1261/rna.2167605

Bernhard, F., & Tozawa, Y. (2013). Cell-free expression—Making a mark. *Current Opinion in Structural Biology*, *23*(3), 374–380. https://doi.org/10.1016/j.sbi.2013.03.012

Bersani, F., Lee, E., Kharchenko, P. V., Xu, A. W., Liu, M., Xega, K., MacKenzie, O. C., Brannigan, B. W., Wittner, B. S., Jung, H., Ramaswamy, S., Park, P. J., Maheswaran, S., Ting, D. T., & Haber, D. A. (2015). Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. *Proceedings of the National Academy of Sciences*, *112*(49), 15148–15153. https://doi.org/10.1073/pnas.1518008112

Black, B. E., & Cleveland, D. W. (2011). Epigenetic centromere propagation and the nature of CENP-a nucleosomes. *Cell*, *144*(4), 471–479. https://doi.org/10.1016/j.cell.2011.02.002

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

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

Boopathi, R., Danev, R., Khoshouei, M., Kale, S., Nahata, S., Ramos, L., Angelov, D., Dimitrov, S., Hamiche, A., Petosa, C., & Bednar, J. (2020). Phase-plate cryo-EM structure of the Widom 601 CENP-A nucleosome core particle reveals differential flexibility of the DNA ends. *Nucleic Acids Research*, *48*(10), 5735–5748. https://doi.org/10.1093/nar/gkaa246

Bornholdt, Z. A., & Prasad, B. V. V. (2008). X-ray structure of NS1 from a highly pathogenic H5N1 influenza virus. *Nature*, *456*(7224), 985–988. https://doi.org/10.1038/nature07444

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 of the United States of America*, *103*(23), 8709–8714. https://doi.org/10.1073/pnas.0508006103

Brown, M. T. (1995). Sequence similarities between the yeast chromosome segregation protein Mif2 and the mammalian centromere protein CENP-C. *Gene*, *160*(1), 111–116. https://doi.org/10.1016/0378-1119(95)00163-z

Carone, D. M., Zhang, C., Hall, L. E., Obergfell, C., Carone, B. R., O'Neill, M. J., & O'Neill, R. J. (2013). Hypermorphic expression of centromeric retroelement-encoded small RNAs impairs CENP-A loading. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *21*(1), 49–62. https://doi.org/10.1007/s10577-013-9337-0

Carroll, C. W., Milks, K. J., & Straight, A. F. (2010). Dual recognition of CENP-A nucleosomes is required for centromere assembly. *Journal of Cell Biology*, *189*(7), 1143–1155. https://doi.org/10.1083/jcb.201001013

Carroll, C. W., & Straight, A. F. (2006). Centromere formation: From epigenetics to self-assembly. *Trends in Cell Biology*, *16*(2), 70–78. https://doi.org/10.1016/j.tcb.2005.12.008

Chan, G. K., Liu, S.-T., & Yen, T. J. (2005). Kinetochore structure and function. *Trends in Cell Biology*, *15*(11), 589–598. https://doi.org/10.1016/j.tcb.2005.09.010

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. (2019a). Islands of retroelements are major components of Drosophila centromeres. *PLOS Biology*, *17*(5), e3000241. https://doi.org/10.1371/journal.pbio.3000241

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. (2019b). *Islands of retroelements are the major components of* Drosophila *centromeres* [Preprint]. Genomics. https://doi.org/10.1101/537357

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

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, P., Zhao, J., Wang, Y., Wang, M., Long, H., Liang, D., Huang, L., Wen, Z., Li, W., Li, X., Feng, H., Zhao, H., Zhu, P., Li, M., Wang, Q., & Li, G. (2013). H3.3 actively marks enhancers and primes gene transcription via opening higher-ordered chromatin. *Genes & Development*, *27*(19), 2109–2124. https://doi.org/10.1101/gad.222174.113

Chik, J. K., Moiseeva, V., Goel, P. K., Meinen, B. A., Koldewey, P., An, S., Mellone, B. G., Subramanian, L., & Cho, U.-S. (2019). Structures of CENP-C cupin domains at regional centromeres reveal unique patterns of dimerization and recruitment functions for the inner pocket. *Journal of Biological Chemistry*, *294*(38), 14119–14134. https://doi.org/10.1074/jbc.RA119.008464

Chmátal, L., Gabriel, S. I., Mitsainas, G. P., Martínez-Vargas, J., Ventura, J., Searle, J. B., Schultz, R. M., & Lampson, M. A. (2014). Centromere Strength Provides the Cell Biological Basis for Meiotic Drive and Karyotype Evolution in Mice. *Current Biology*, *24*(19), 2295–2300. https://doi.org/10.1016/j.cub.2014.08.017

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

Cooper, J. L., & Henikoff, S. (2004). Adaptive Evolution of the Histone Fold Domain in Centromeric Histones. *Molecular Biology and Evolution*, *21*(9), 1712–1718. https://doi.org/10.1093/molbev/msh179

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

Cusenza, V. Y., Tameni, A., Neri, A., & Frazzi, R. (2023). The lncRNA epigenetics: The significance of m6A and m5C lncRNA modifications in cancer. *Frontiers in Oncology*, *13*. https://doi.org/10.3389/fonc.2023.1063636

Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W., & Richmond, T. J. (2002). Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. *Journal of Molecular Biology*, *319*(5), 1097–1113. https://doi.org/10.1016/S0022-2836(02)00386-8

Dawe, R. K., Reed, L. M., Yu, H. G., Muszynski, M. G., & Hiatt, E. N. (1999). A maize homolog of mammalian CENPC is a constitutive component of the inner kinetochore. *The Plant Cell*, *11*(7), 1227–1238. https://doi.org/10.1105/tpc.11.7.1227

Domon, B., & Aebersold, R. (2006). Mass Spectrometry and Protein Analysis. *Science*, *312*(5771), 212–217. https://doi.org/10.1126/science.1124619

Du, Y., Topp, C. N., & Dawe, R. K. (2010a). 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

Du, Y., Topp, C. N., & Dawe, R. K. (2010b). DNA Binding of Centromere Protein C (CENPC) Is Stabilized by Single-Stranded RNA. *PLOS Genetics*, *6*(2), Article 2. https://doi.org/10.1371/journal.pgen.1000835

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 G_1 phase. *Nucleus (Austin, Tex.)*, *2*(2), 146–157. https://doi.org/10.4161/nucl.2.2.15211

Earnshaw, W. C., & Rothfield, N. (1985). Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. *Chromosoma*, *91*(3–4), 313–321. https://doi.org/10.1007/BF00328227

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. *The Journal of Cell Biology*, *183*(5), 805–818. https://doi.org/10.1083/jcb.200806038

Ericsson, U. B., Hallberg, B. M., DeTitta, G. T., Dekker, N., & Nordlund, P. (2006). Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Analytical Biochemistry*, *357*(2), 289–298. https://doi.org/10.1016/j.ab.2006.07.027

Felsenfeld, G. (1978). Chromatin. *Nature*, *271*(5641), Article 5641. https://doi.org/10.1038/271115a0

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

Fletcher, E., Krivoruchko, A., & Nielsen, J. (2016). Industrial systems biology and its impact on synthetic biology of yeast cell factories. *Biotechnology and Bioengineering*, *113*(6), 1164–1170. https://doi.org/10.1002/bit.25870

Fukagawa, T., & Earnshaw, W. C. (2014). The Centromere: Chromatin Foundation for the Kinetochore Machinery. *Developmental Cell*, *30*(5), 496–508. https://doi.org/10.1016/j.devcel.2014.08.016

Gibson, B. A., Doolittle, L. K., Schneider, M. W. G., Jensen, L. E., Gamarra, N., Henry, L., Gerlich, D. W., Redding, S., & Rosen, M. K. (2019). Organization of Chromatin by Intrinsic and Regulated Phase Separation. *Cell*, *179*(2), 470-484.e21. https://doi.org/10.1016/j.cell.2019.08.037

Goshima, G., Wollman, R., Goodwin, S. S., Zhang, N., Scholey, J. M., Vale, R. D., & Stuurman, N. (2007). Genes Required for Mitotic Spindle Assembly in Drosophila S2 Cells. *Science*, *316*(5823), 417–421. https://doi.org/10.1126/science.1141314

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

Guttenbach, M., Martínez-Expósito, M. J., Engel, W., & Schmid, M. (1996). Interphase Chromosome Arrangement in Sertoli Cells of Adult Mice1. *Biology of Reproduction*, *54*(5), 980–986. https://doi.org/10.1095/biolreprod54.5.980

Hake, S. B., & Allis, C. D. (2006). Histone H3 variants and their potential role in indexing mammalian genomes: The 'H3 barcode hypothesis'. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(17), 6428–6435. https://doi.org/10.1073/pnas.0600803103

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

Hayashi, T., Fujita, Y., Iwasaki, O., Adachi, Y., Takahashi, K., & Yanagida, M. (2004). Mis16 and Mis18 Are Required for CENP-A Loading and Histone Deacetylation at Centromeres. *Cell*, *118*(6), 715–729. https://doi.org/10.1016/j.cell.2004.09.002

He, C., & Lan, F. (2021). RNA m6A meets transposable elements and chromatin. *Protein & Cell*, *12*(12), 906–910. https://doi.org/10.1007/s13238-021-00859-2

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), Article 1. https://doi.org/10.1038/srep42520

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

Henikoff, S., Ahmad, K., & Malik, H. S. (2001). The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA. *Science*, *293*(5532), 1098–1102. https://doi.org/10.1126/science.1062939

Henikoff, S., Ahmad, K., Platero, J. S., & van Steensel, B. (2000). Heterochromatic deposition of centromeric histone H3-like proteins. *Proceedings of the National Academy of Sciences*, *97*(2), 716–721. https://doi.org/10.1073/pnas.97.2.716

Henikoff, S., & Dalal, Y. (2005). Centromeric chromatin: What makes it unique? *Current Opinion in Genetics & Development*, *15*(2), 177–184. https://doi.org/10.1016/j.gde.2005.01.004

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

Holmes, D. S., Mayfield, J. E., Sander, G., & Bonner, J. (1972). Chromosomal RNA: Its Properties. *Science*, *177*(4043), 72–74. https://doi.org/10.1126/science.177.4043.72

Huang, R. C., & Bonner, J. (1965). Histone-bound RNA, a component of native nucleohistone. *Proceedings of the National Academy of Sciences*, *54*(3), 960–967. https://doi.org/10.1073/pnas.54.3.960

Ideue, T., & Tani, T. (2020). Centromeric Non-Coding RNAs: Conservation and Diversity in Function. *Non-Coding RNA*, *6*(1), Article 1. https://doi.org/10.3390/ncrna6010004

Ishii, K., Ogiyama, Y., Chikashige, Y., Soejima, S., Masuda, F., Kakuma, T., Hiraoka, Y., & Takahashi, K. (2008). Heterochromatin Integrity Affects Chromosome Reorganization After Centromere Dysfunction. *Science*, *321*(5892), 1088–1091. https://doi.org/10.1126/science.1158699

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

Jarmoskaite, I., AlSadhan, I., Vaidyanathan, P. P., & Herschlag, D. (2020). How to measure and evaluate binding affinities. *eLife*, *9*, e57264. https://doi.org/10.7554/eLife.57264

Jenuwein, T., & Allis, C. D. (2001). Translating the Histone Code. *Science*, *293*(5532), 1074–1080. https://doi.org/10.1126/science.1063127

J. Miles, A., W. Janes, R., & A. Wallace, B. (2021). Tools and methods for circular dichroism spectroscopy of proteins: A tutorial review. *Chemical Society Reviews*, *50*(15), 8400–8413. https://doi.org/10.1039/D0CS00558D

Johnson, W. L., & Straight, A. F. (2017). RNA-mediated regulation of heterochromatin. *Current Opinion in Cell Biology*, *46*, 102–109. https://doi.org/10.1016/j.ceb.2017.05.004

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), Article 7873. https://doi.org/10.1038/s41586-021-03819-2

Karch, K. R., Coradin, M., Zandarashvili, L., Kan, Z.-Y., Gerace, M., Englander, S. W., Black, B. E., & Garcia, B. A. (2018). Hydrogen-Deuterium Exchange Coupled to Top- and Middle-Down Mass Spectrometry Reveals Histone Tail Dynamics before and after Nucleosome Assembly. *Structure*, *26*(12), 1651-1663.e3. https://doi.org/10.1016/j.str.2018.08.006

Kavan, D., & Man, P. (2011). MSTools—Web based application for visualization and presentation of HXMS data. *International Journal of Mass Spectrometry*, *302*(1), 53–58. https://doi.org/10.1016/j.ijms.2010.07.030

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, *10*(6), 845–858. https://doi.org/10.1038/nprot.2015.053

Ketel, C., Wang, H. S. W., McClellan, M., Bouchonville, K., Selmecki, A., Lahav, T., Gerami-Nejad, M., & Berman, J. (2009). Neocentromeres Form Efficiently at Multiple Possible Loci in Candida albicans. *PLOS Genetics*, *5*(3), e1000400. https://doi.org/10.1371/journal.pgen.1000400

Kilchert, C., Sträßer, K., Kunetsky, V., & Änkö, M.-L. (2020). From parts lists to functional significance—RNA–protein interactions in gene regulation. *WIREs RNA*, *11*(3), e1582. https://doi.org/10.1002/wrna.1582

Klare, K., Weir, J. R., Basilico, F., Zimniak, T., Massimiliano, L., Ludwigs, N., Herzog, F., & Musacchio, A. (2015). CENP-C is a blueprint for constitutive centromere–associated network assembly within human kinetochores. *The Journal of Cell Biology*, *210*(1), 923–934. https://doi.org/10.1083/jcb.201412028

Klug et al., T., F. ,. Koller, T. (1979). Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. *The Journal of Cell Biology*, *83*(2), 403–427.

Konermann, L., Pan, J., & Liu, Y.-H. (2011). Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chemical Society Reviews*, *40*(3), 1224–1234. https://doi.org/10.1039/C0CS00113A

Koonin, E. V. (2014). The origins of cellular life. *Antonie Van Leeuwenhoek*, *106*(1), 27–41. https://doi.org/10.1007/s10482-014-0169-5

Kouzarides, T. (2007). Chromatin Modifications and Their Function. *Cell*, *128*(4), 693–705. https://doi.org/10.1016/j.cell.2007.02.005

Kufareva, I., & Abagyan, R. (2012). Methods of protein structure comparison. *Methods in Molecular Biology (Clifton, N.J.)*, *857*, 231–257. https://doi.org/10.1007/978-1-61779-588-6_10

Kursel, L. E., & Malik, H. S. (2018). The cellular mechanisms and consequences of centromere drive. *Current Opinion in Cell Biology*, *52*, 58–65. https://doi.org/10.1016/j.ceb.2018.01.011

Kyriacou, E., & Heun, P. (2023). Centromere structure and function: Lessons from Drosophila. *Genetics*, iyad170. https://doi.org/10.1093/genetics/iyad170

Leger, A., Amaral, P. P., Pandolfini, L., Capitanchik, C., Capraro, F., Miano, V., Migliori, V., Toolan-Kerr, P., Sideri, T., Enright, A. J., Tzelepis, K., van Werven, F. J., Luscombe, N. M., Barbieri, I., Ule, J., Fitzgerald, T., Birney, E., Leonardi, T., & Kouzarides, T. (2021). RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-27393-3

Liang, G., Lin, J. C. Y., Wei, V., Yoo, C., Cheng, J. C., Nguyen, C. T., Weisenberger, D. J., Egger, G., Takai, D., Gonzales, F. A., & Jones, P. A. (2004). Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(19), 7357–7362. https://doi.org/10.1073/pnas.0401866101

LiCata, V. J., & Wowor, A. J. (2008). Applications of Fluorescence Anisotropy to the Study of Protein–DNA Interactions. In *Methods in Cell Biology* (Vol. 84, pp. 243–262). Academic Press. https://doi.org/10.1016/S0091-679X(07)84009-X

Ling, Y. H., & Yuen, K. W. Y. (2019). Centromeric non-coding RNA as a hidden epigenetic factor of the point centromere. *Current Genetics*, *65*(5), 1165–1171. https://doi.org/10.1007/s00294-019-00988-6

Liu, J., Peng, Y., & Wei, W. (2022). Cell cycle on the crossroad of tumorigenesis and cancer therapy. *Trends in Cell Biology*, *32*(1), 30–44. https://doi.org/10.1016/j.tcb.2021.07.001

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

Maher, S., Jjunju, F. P. M., & Taylor, S. (2015). Colloquium: 100 years of mass spectrometry: Perspectives and future trends. *Reviews of Modern Physics*, *87*(1), 113–135. https://doi.org/10.1103/RevModPhys.87.113

Malik, H. S. (2009). The Centromere-Drive Hypothesis: A Simple Basis for Centromere Complexity. In D. Ugarkovic (Ed.), *Centromere: Structure and Evolution* (pp. 33–52). Springer. https://doi.org/10.1007/978-3-642-00182-6_2

Malik, H. S., & Henikoff, S. (2002). Conflict begets complexity: The evolution of centromeres. *Current Opinion in Genetics & Development*, *12*(6), 711–718. https://doi.org/10.1016/S0959-437X(02)00351-9

Malik, H. S., & Henikoff, S. (2003). Phylogenomics of the nucleosome. *Nature Structural & Molecular Biology*, *10*(11), 882–891. https://doi.org/10.1038/nsb996

Malik, H. S., & Henikoff, S. (2009). Major evolutionary transitions in centromere complexity. *Cell*, *138*(6), 1067–1082. https://doi.org/10.1016/j.cell.2009.08.036

McKenzie, E. A., & Abbott, W. M. (2018). Expression of recombinant proteins in insect and mammalian cells. *Methods*, *147*, 40–49. https://doi.org/10.1016/j.ymeth.2018.05.013

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), e103234. 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

Meluh, P. B., & Koshland, D. (1995). Evidence that the MIF2 gene of Saccharomyces cerevisiae encodes a centromere protein with homology to the mammalian centromere protein CENP-C. *Molecular Biology of the Cell*, *6*(7), 793–807.

Mitra, G. (2021). Emerging Role of Mass Spectrometry-Based Structural Proteomics in Elucidating Intrinsic Disorder in Proteins. *PROTEOMICS*, *21*(3–4), 2000011. https://doi.org/10.1002/pmic.202000011

Moore, L. L., & Roth, M. B. (2001). HCP-4, a CENP-C-like protein in Caenorhabditis elegans, is required for resolution of sister centromeres. *The Journal of Cell Biology*, *153*(6), 1199–1208. https://doi.org/10.1083/jcb.153.6.1199

Morrison, O., & Thakur, J. (2021). Molecular Complexes at Euchromatin, Heterochromatin and Centromeric Chromatin. *International Journal of Molecular Sciences*, *22*(13), Article 13. https://doi.org/10.3390/ijms22136922

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), e202007210. https://doi.org/10.1083/jcb.202007210

Niesen, F. H., Berglund, H., & Vedadi, M. (2007). The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature Protocols*, *2*(9), 2212–2221. https://doi.org/10.1038/nprot.2007.321

Ninomiya, K., Iwakiri, J., Aly, M. K., Sakaguchi, Y., Adachi, S., Natsume, T., Terai, G., Asai, K., Suzuki, T., & Hirose, T. (2021). m6A modification of HSATIII lncRNAs regulates temperature-dependent splicing. *The EMBO Journal*, *40*(15), e107976. https://doi.org/10.15252/embj.2021107976

Ninomiya, K., Yamazaki, T., & Hirose, T. (2023). Satellite RNAs: Emerging players in subnuclear architecture and gene regulation. *The EMBO Journal*, *42*(18), e114331. https://doi.org/10.15252/embj.2023114331

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

Ochs, R. L., & Press, R. I. (1992). Centromere autoantigens are associated with the nucleolus. *Experimental Cell Research*, *200*(2), 339–350. https://doi.org/10.1016/0014-4827(92)90181-7

Ohshiro, T., Konno, M., Asai, A., Komoto, Y., Yamagata, A., Doki, Y., Eguchi, H., Ofusa, K., Taniguchi, M., & Ishii, H. (2021). Single-molecule RNA sequencing for simultaneous detection of m6A and 5mC. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-98805-z

Orr, B., & Sunkel, C. E. (2011). Drosophila CENP-C is essential for centromere identity. *Chromosoma*, *120*(1), 83–96. https://doi.org/10.1007/s00412-010-0293-6

Ozohanics, O., & Ambrus, A. (2020). Hydrogen-Deuterium Exchange Mass Spectrometry: A Novel Structural Biology Approach to Structure, Dynamics and Interactions of Proteins and Their Complexes. *Life*, *10*(11), 286. https://doi.org/10.3390/life10110286

Pesenti, M. E., Raisch, T., Conti, D., Walstein, K., Hoffmann, I., Vogt, D., Prumbaum, D., Vetter, I. R., Raunser, S., & Musacchio, A. (2022). Structure of the human inner kinetochore CCAN complex and its significance for human centromere organization. *Molecular Cell*, *82*(11), 2113-2131.e8. https://doi.org/10.1016/j.molcel.2022.04.027

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

Popova, V. V., Kurshakova, M. M., & Kopytova, D. V. (2015). Methods to study the RNA-protein interactions. *Molecular Biology*, *49*(3), 418–426. https://doi.org/10.1134/S0026893315020107

Quan, J., & Tian, J. (2009). Circular Polymerase Extension Cloning of Complex Gene Libraries and Pathways. *PLoS ONE*, *4*(7), e6441. https://doi.org/10.1371/journal.pone.0006441

Quénet, D., & Dalal, Y. (2014). A long non-coding RNA is required for targeting centromeric protein A to the human centromere. *eLife*, *3*, e26016. https://doi.org/10.7554/eLife.03254

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

Ranjbar, B., & Gill, P. (2009). Circular Dichroism Techniques: Biomolecular and Nanostructural Analyses- A Review. *Chemical Biology & Drug Design*, *74*(2), 101–120. https://doi.org/10.1111/j.1747-0285.2009.00847.x

Resetca, D., & Wilson, D. J. (2013). Characterizing rapid, activity-linked conformational transitions in proteins via sub-second hydrogen deuterium exchange mass spectrometry. *The FEBS Journal*, *280*(22), 5616–5625. https://doi.org/10.1111/febs.12332

Richards, E. J., & Elgin, S. C. R. (2002). Epigenetic Codes for Heterochromatin Formation and Silencing: Rounding up the Usual Suspects. *Cell*, *108*(4), 489–500. https://doi.org/10.1016/S0092-8674(02)00644-X

Rieder, C. L. (1979). Ribonucleoprotein staining of centrioles and kinetochores in newt lung cell spindles. *Journal of Cell Biology*, *80*(1), 1–9. https://doi.org/10.1083/jcb.80.1.1

Righetti, P. G. (2005). Electrophoresis: The march of pennies, the march of dimes. *Journal of Chromatography A*, *1079*(1), 24–40. https://doi.org/10.1016/j.chroma.2005.01.018

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

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

Saitoh, H., Tomkiel, J., Cooke, C. A., Ratrie, H., Maurer, M., Rothfield, N. F., & Earnshaw, W. C. (1992). CENP-C, an autoantigen in scleroderma, is a component of the human inner kinetochore plate. *Cell*, *70*(1), 115–125. https://doi.org/10.1016/0092-8674(92)90538-n

Sawyer, I. A., & Dundr, M. (2017). Chromatin loops and causality loops: The influence of RNA upon spatial nuclear architecture. *Chromosoma*, *126*(5), 541–557. https://doi.org/10.1007/s00412-017-0632-y

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

Scholz, J., & Suppmann, S. (2017). A new single-step protocol for rapid baculovirus-driven protein production in insect cells. *BMC Biotechnology*, *17*(1), 83. https://doi.org/10.1186/s12896-017-0400-3

Schotanus, K., & Heitman, J. (2020). Centromere deletion in Cryptococcus deuterogattii leads to neocentromere formation and chromosome fusions. *eLife*, *9*, e56026. https://doi.org/10.7554/eLife.56026

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

Shivaraju, M., Camahort, R., Mattingly, M., & Gerton, J. L. (2011). Scm3 is a centromeric nucleosome assembly factor. *The Journal of Biological Chemistry*, *286*(14), 12016–12023. https://doi.org/10.1074/jbc.M110.183640

Simpson, R. J. (2010). Stabilization of Proteins for Storage. *Cold Spring Harbor Protocols*, *2010*(5), pdb.top79. https://doi.org/10.1101/pdb.top79

Steiner, F. A., & Henikoff, S. (2015). Diversity in the organization of centromeric chromatin. *Current Opinion in Genetics & Development*, *31*, 28–35. https://doi.org/10.1016/j.gde.2015.03.010

Stellfox, M. E., Bailey, A. O., & Foltz, D. R. (2013). Putting CENP-A in its place. *Cellular and Molecular Life Sciences*, *70*(3), 387–406. https://doi.org/10.1007/s00018-012-1048-8

Sullivan, & 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), Article 11. https://doi.org/10.1038/nsmb845

Sullivan, L. L., Boivin, C. D., Mravinac, B., Song, I. Y., & Sullivan, B. A. (2011). Genomic size of CENP-A domain is proportional to total alpha satellite array size at human centromeres and expands in cancer cells. *Chromosome Research*, *19*(4), 457–470. https://doi.org/10.1007/s10577-011-9208-5

Tachiwana, H., Kagawa, W., Shiga, T., Osakabe, A., Miya, Y., Saito, K., Hayashi-Takanaka, Y., Oda, T., Sato, M., Park, S.-Y., Kimura, H., & Kurumizaka, H. (2011). Crystal structure of the human centromeric nucleosome containing CENP-A. *Nature*, *476*(7359), 232–235. https://doi.org/10.1038/nature10258

Talbert, P. B., & Henikoff, S. (2020). What makes a centromere? *Experimental Cell Research*, *389*(2), 111895. https://doi.org/10.1016/j.yexcr.2020.111895

Talbert, P. B., Kasinathan, S., & Henikoff, S. (2018). Simple and Complex Centromeric Satellites in Drosophila Sibling Species. *Genetics*, *208*(3), 977–990. https://doi.org/10.1534/genetics.117.300620

Thakur, J., & Henikoff, S. (2020). Architectural RNA in chromatin organization. *Biochemical Society Transactions*, *48*(5), 1967. https://doi.org/10.1042/BST20191226

Thakur, J., & Sanyal, K. (2013). Efficient neocentromere formation is suppressed by gene conversion to maintain centromere function at native physical chromosomal loci in Candida albicans. *Genome Research*, *23*(4), 638–652. https://doi.org/10.1101/gr.141614.112

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

Trcka, F., Durech, M., Vankova, P., Chmelik, J., Martinkova, V., Hausner, J., Kadek, A., Marcoux, J., Klumpler, T., Vojtesek, B., Muller, P., & Man, P. (2019). Human Stress-inducible Hsp70 Has a High Propensity to Form ATP-dependent Antiparallel Dimers That Are Differentially Regulated by Cochaperone Binding. *Molecular & Cellular Proteomics: MCP*, *18*(2), 320–337. https://doi.org/10.1074/mcp.RA118.001044

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

Trivedi, R., & Nagarajaram, H. A. (2022). Intrinsically Disordered Proteins: An Overview. *International Journal of Molecular Sciences*, *23*(22), 14050. https://doi.org/10.3390/ijms232214050

Ung, T. L., Cao, C., Lu, J., Ozato, K., & Dever, T. E. (2001). Heterologous dimerization domains functionally substitute for the double-stranded RNA binding domains of the kinase PKR. *The EMBO Journal*, *20*(14), 3728–3737. https://doi.org/10.1093/emboj/20.14.3728

Unhavaithaya, Y., & Orr-Weaver, T. L. (2013). Centromere proteins CENP-C and CAL1 functionally interact in meiosis for centromere clustering, pairing, and chromosome segregation. *Proceedings of the National Academy of Sciences*, *110*(49), 19878–19883. https://doi.org/10.1073/pnas.1320074110

Vafa, O., & Sullivan, K. F. (1997). Chromatin containing CENP-A and alpha-satellite DNA is a major component of the inner kinetochore plate. *Current Biology: CB*, *7*(11), 897–900. https://doi.org/10.1016/s0960-9822(06)00381-2

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

Wang, J., Liu, X., Dou, Z., Chen, L., Jiang, H., Fu, C., Fu, G., Liu, D., Zhang, J., Zhu, T., Fang, J., Zang, J., Cheng, J., Teng, M., Ding, X., & Yao, X. (2014). Mitotic regulator Mis18β interacts with and specifies the centromeric assembly of molecular chaperone holliday

junction recognition protein (HJURP). *The Journal of Biological Chemistry*, *289*(12), 8326–8336. https://doi.org/10.1074/jbc.M113.529958

Wang, W., Nema, S., & Teagarden, D. (2010). Protein aggregation—Pathways and influencing factors. *International Journal of Pharmaceutics*, *390*(2), 89–99. https://doi.org/10.1016/j.ijpharm.2010.02.025

Waye, J. S., & Willard, H. F. (1985). Chromosome-specific alpha satellite DNA: Nucleotide sequence analysis of the 2.0 kilobasepair repeat from the human X chromosome. *Nucleic Acids Research*, *13*(8), 2731–2743. https://doi.org/10.1093/nar/13.8.2731

Weizmann, Y., Braunschweig, A. B., Wilner, O. I., Cheglakov, Z., & Willner, I. (2008). A polycatenated DNA scaffold for the one-step assembly of hierarchical nanostructures. *Proceedings of the National Academy of Sciences*, *105*(14), 5289–5294. https://doi.org/10.1073/pnas.0800723105

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

Wysocka, J., Swigut, T., Xiao, H., Milne, T. A., Kwon, S. Y., Landry, J., Kauer, M., Tackett, A. J., Chait, B. T., Badenhorst, P., Wu, C., & Allis, C. D. (2006). A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature*, *442*(7098), 86–90. https://doi.org/10.1038/nature04815

Yan, K., Yang, J., Zhang, Z., McLaughlin, S. H., Chang, L., Fasci, D., Ehrenhofer-Murray, A. E., Heck, A. J. R., & Barford, D. (2019). Structure of the inner kinetochore CCAN complex assembled onto a centromeric nucleosome. *Nature*, *574*(7777), Article 7777. https://doi.org/10.1038/s41586-019-1609-1

Yu, X.-M., Li, S.-J., Yao, Z.-T., Xu, J.-J., Zheng, C.-C., Liu, Z.-C., Ding, P.-B., Jiang, Z.-L., Wei, X., Zhao, L.-P., Shi, X.-Y., Li, Z.-G., Xu, W. W., & Li, B. (2023). N4-acetylcytidine modification of lncRNA CTC-490G23.2 promotes cancer metastasis through interacting with PTBP1 to increase CD44 alternative splicing. *Oncogene*, *42*(14), 1101–1116. https://doi.org/10.1038/s41388-023-02628-3

Zagrovic, B., Bartonek, L., & Polyansky, A. A. (2018). RNA-protein interactions in an unstructured context. *FEBS Letters*, *592*(17), 2901–2916. https://doi.org/10.1002/1873-3468.13116

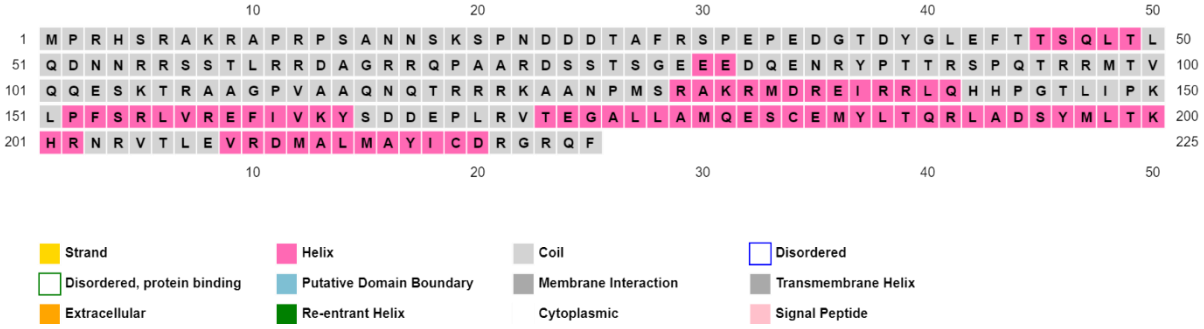
Zhang, C., Wang, D., Hao, Y., Wu, S., Luo, J., Xue, Y., Wang, D., Li, G., Liu, L., Shao, C., Li, H., Yuan, J., Zhu, M., Fu, X.-D., Yang, X., Chen, R., & Teng, Y. (2022). LncRNA CCTT-mediated RNA-DNA and RNA-protein interactions facilitate the recruitment of CENP-C to centromeric DNA during kinetochore assembly. *Molecular Cell*, *82*(21), 4018-4032.e9. https://doi.org/10.1016/j.molcel.2022.09.022

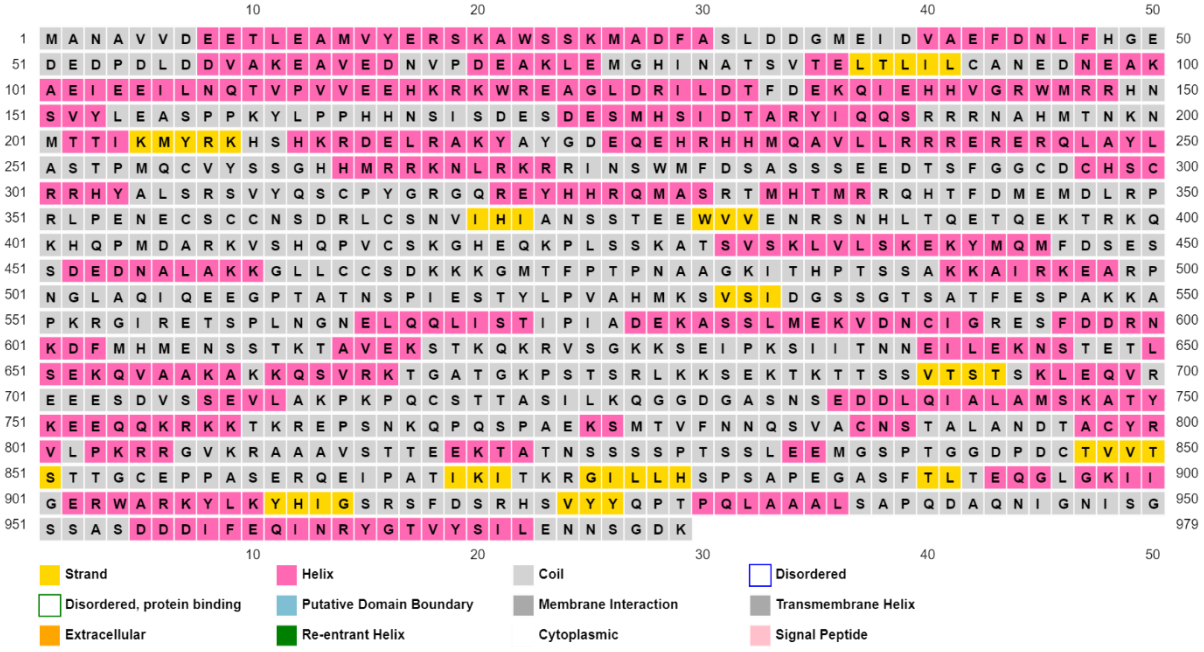
Zhou, B.-R., Yadav, K. N. S., Borgnia, M., Hong, J., Cao, B., Olins, A. L., Olins, D. E., Bai, Y., & Zhang, P. (2019). Atomic resolution cryo-EM structure of a native-like CENP-A nucleosome aided by an antibody fragment. *Nature Communications*, *10*(1), 2301. https://doi.org/10.1038/s41467-019-10247-4

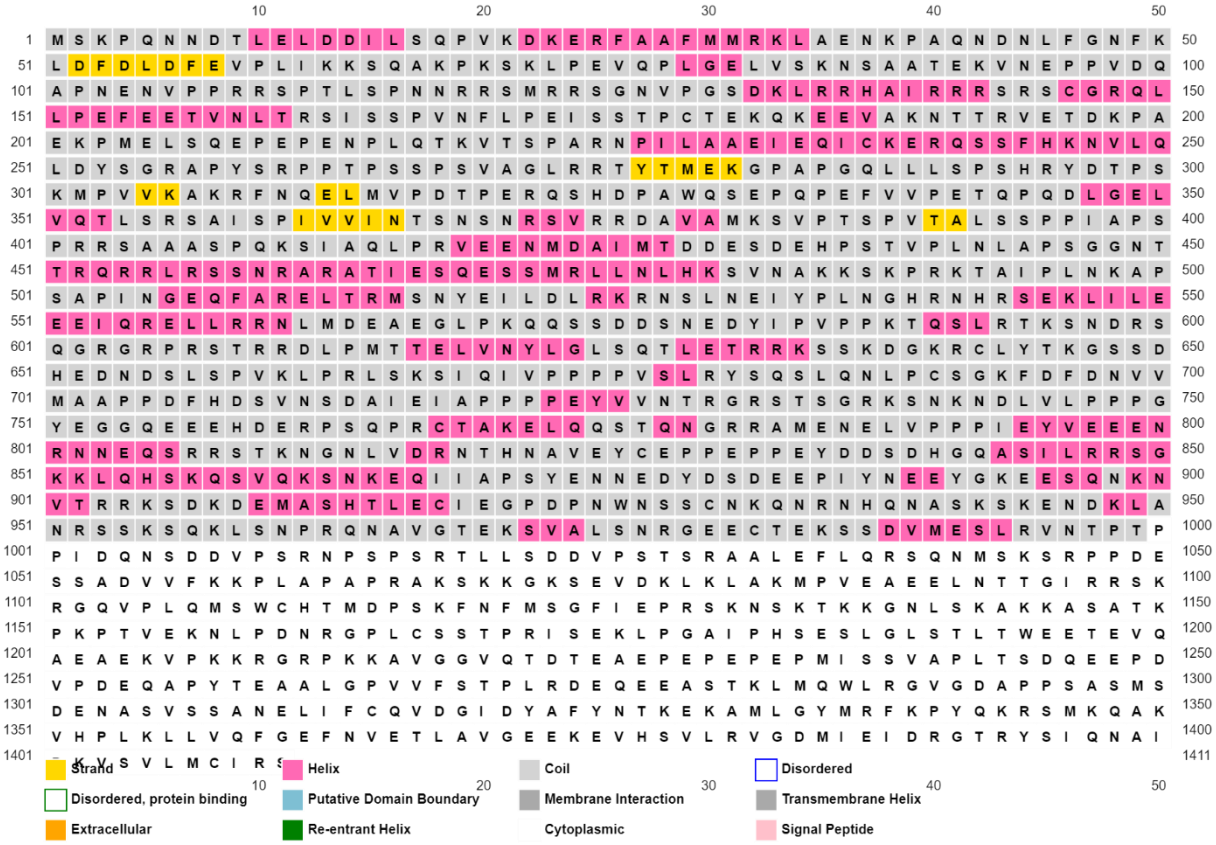
Other sources and used software

Addgene	https://www.addgene.org
AlphaFold	https://alphafold.ebi.ac.uk
BioRender	https://www.biorender.com
BLAST NCBI	https://blast.ncbi.nlm.nih.gov/
Chronos software	https://www.axelsemrau.de
DataAnalysis (v. 5.3)	https://www.bruker.com
DeutEx	(Trcka et al., 2019)
Expasy ProtParam tool	https://web.expasy.org/protparam/
FATCAT structure alignment tool	https://fatcat.godziklab.org
MASCOT (v. 2.7)	https://www.matrixscience.com
Mass spectrometry analysis	https://peterslab.org/MSTools
Microsoft Office	https://www.office.com
NEBuilder	https://nebuilder.neb.com/
NEBCutter	https://nc3.neb.com/NEBcutter/
OriginPro (2023, 10.0.0.154)	https://www.originlab.com
Phyre	http://www.sbg.bio.ic.ac.uk/phyre2
PsiPred	http://bioinf.cs.ucl.ac.uk/psipred/
RCSB PDB	https://www.rcsb.org/
SnapGene (v. 7.1)	https://www.snapgene.com
thesis of Iris Valent	10.11588/heidok.00032271
UniProt	https://www.uniprot.org

Appendix


Strand
Helix
Coil
Disordered
Disordered, protein binding
Putative Domain Boundary
Membrane Interaction
Transmembrane Helix
Extracellular
Re-entrant Helix
Cytoplasmic
Signal Peptide


Strand
Helix
Coil
Disordered
Disordered, protein binding
Putative Domain Boundary
Membrane Interaction
Transmembrane Helix
Extracellular
Re-entrant Helix
Cytoplasmic
Signal Peptide



Appendix 1. **Secondary structure predictions of Cid, Cal1 and Cenp-C generated by PsiPred**

Cid has α-helical C-terminal histone fold domain and random coil N-terminus. Cal1 has mostly random coil with large portion of α-helices, but a great number of those are too short to represent physical reality. Later predictions with AlphaFold show much smaller contribution of α-helices. Cenp-C follows the same trend and Cal1, with even less α-helices and more random coil. AlphaFold predicts less α-helices here as well. PsiPred could process only the first thousand, so the C-terminal β-sheet cupin domain is not visible here.

http://bioinf.cs.ucl.ac.uk/psipred/

>SatIII sense full length
ACAAAACUCGUCGAUUAAUGGUCGCGAUUUCUAGGGAUAGUAAAAAAUUUGUAAAUCCCUUUAG
UUAAAACCCAUUUAAAAGCGUAAAAAAGGUUCCCAAUAGUCCCAGUAGUUUUAACCCUCCUAUA
CCGGUUUUUUAAAUUAAAGGUAAAAACUUGUGUCAAACUAACCUUUAAAAUAAUGCUCGAGUUG
CUCCAUGCCGUAAGGUAUAAGUCUGUUAAUAAAAAUUUCAACGCCGCUUUUUUGACUAAUAAAU
UACUGGCUUUAAACCUUUUUGCCUAAAGCGGGUUUUCAACUAUAAAUGUUUGCCUUAAAAGCAG
UAUUCAACCGAUUUUUACCAGUGUAUCUACAUUCUUAUUGACAAAACUCUUUGAUUAAUGGUCG
CGAUUGUUAGGGAUAGUGAAAAACUUCUAAAUCCAUCUAGUUAAAACCCAGUUAAAAACGUAAA
AAACAUUCCCAGUAUUCCCAGUAGUUUUAAACCUUUUUUACGGGUUUUUUUAAAUUAAAAGUAA
AAACUUGUGUCAAACUAACCUUUAAAAUAAUGCUCGAGUUGAUCCAUACCGUAAGGUAUAAGUC
UGUUAAUAAAAAUUUCAACACCGUUUUCUUGACUAACAAACUACUGGCUUUGAACCUUUUUGCC
UAAAACCGUUUUUCAACUAUAAAUGUUUGCCUUAAAAACAGUAUUGAACCGAUUUUUACCAGUG
UAUCUACAUUCUUAUUCGGCUUAAGGUCGUGUGACCGCCGGCAAUGAUCACCUAGGCUCGAGCC
AUGGCGUCGAUCGACGGAGCUCCGGCCAGAGGGAUAUCACUCAGCAUAAU

>SatIII antisense full length
AAUAAGAAUGUAGAUACACUGGUAAAAAUCGGUUCAAUACUGCUUUUAAAGCAAACAUUUAUAG
UUGUAAAAACGUCUCAGACAAAAAGGUUUAAAGCCAGUAGUUUAUUAGUAAAUAAAACGGUGUU
GUAUUUUUUAUUAACAGACUUAUACCUUACAGUAUGGAGUGACUCGAGCAUUAUUUUAAAGGUU
AGUUUGACACAAGUUUUUACCUUUAAUGUAAAAGACCGGUAUAAAACGUUUAAAACUACUGGGG
GUAAGAAUGUUUUUUACGCCUUCAAUUAGGUUUUUAAUUAAAGGGAUUUAGGAAGUUUUCAUUA
UCCCUAGCAAUCGUGACCAUUAAUCGACGAGUUUUGUCUAUAAGCAUGUAGAUACACUGGGAAA
AAUCGGUUCAAUAUUGCUUUUAAAGCAAACAUUUAUAGUUGUAAAAACGUCUCAGACAAAAAGG
UUUAAAGCCAGUAGUUUUUUAGUAAAUAAAACGGUGUUGUACUUUUUAUUAACAGACUUAUACC
UUACAGUAUGGAGUGACUCGAGCAUUAUUUUAAAGGUUAGUUUGACACAAGUUUUUACCCUUAA
UUUAAAAAAACCGGUAUAAAACGUUUAAAACUACUGGGGGAAUGGAGGAAUGCUUUUUACGCUUU
UAACUAGGUUUUUAAUUAAAGGGAUUUAGGAAGUUUUCCAUUAUCCCUAGCAAUCGUGACCAUU
AAUCGACGAGUUUUGUCAAUAAGAAUGUAGAUACACUGGUAAAAAUCGGUUCAAUACUGCUUUU
AAAGCAAACAUUUAUAGUUGUAAAAACGUCUCAGACAAAAAGGUUUAAAGCCAGUAGUUUAUUA
GUAAAUAAAACGGUGUUGUAUUUUUUAUUAACAGACUUAUACCUUACAGUAUGGAGUGACUCGA
GCAUUAUUUUAAAGGUUAGUUUGACACAAGUUUUUACCUUUAAUGUAAAAAACCGGUAUAAAAC
GUUUAAAACUACUGGGGGGAGGUAUGUCUUUUAUGCUUUUAACUAGGUUUUUAAUCAAAGGAUU
UAGGAAGUUUUUCAUUAUCCCUAGCAAUCGUGACCAUUAAUCGACGAGUUUUGUCAAUAAGAAU
GUAGAUACACUGGUAAAAAUCGGUUCAAUACUGCUUUUAAAGCAAACAUUUAUAGUUGUAAAAA
CGUCUCAGACAAAAAGGUUUAAAGCCAGUAGUUUAUUAGUAAAUAAAACGGUGUUGUAUUUUUU
AUAAACAGACUUAUACCUUACAGUAUGGAGUGACUCGAGCAUUAUUUUAAAGGUUAGUUUGACA
CAAGUUUUUACCUUUAAUUUAAAAAACCGGUAUAAAACGUUUAAAACUACUGGGGGGAGGUAUG
UCUUUUAUGCUUUUAACUAGGUUUUUAAUCAAAGGAUUUAGGAAGUUUUUCAUUAUCCCUAGCA
AUCGUGACCAUUAAUCGACGAGUUUUGUUCGGCUUAAGGUCGUGUGACCGCCGGCAAUGAUCAC
CUAGGCUCGAGCCAUGGACGCGAUCGACGGAGCUCCGGCCAGAGGGAUAUCACUCAGCAUAAU

>SatIII S1
GGGAGACCGGCCUCGAGCGGCCGCCAGUGUGAUGGAUAUCUGCAGAAUUCGGCUUGUUUUGAGC
AGCUAAUUACCAGCGCUA

>SatIII S3
UCAAAAUUGGGAGGAUAUGGCCAAAAAAUUUAAUUUCCAUUUUUGAACACAGUUUGAUUGGAAA
UUUUAUUACGAGCUCAAC

>SatIII AS1
GGGAGACCGGCCUCGAGCGGCCGCCAGUGUGAUGGAUAUCUGCAGAAUUCGGCUUUAUUCUUAC
AUCUAUGUGACCAUUUUU

>SatIII AS3
UCAUUUAUUUUGCCACAACAUAAAAAAUAAUUGUCUGAAUAUGGAAUGUCAUACCUCACUGAGC
UCGUAAUAAAAUUUCCAA

>Copia_I
GGTTATGGGCCCAGTCCATGCCTAATAAACAATTAAATTGTGAATTAAAGATTGTGAAAATAAA
TTGTGAAATAGCATTTTTTCACATTCTTGTGAAATAGCTTTTTTTTTCACATTCTTGTGAAATT
ATTTCCTTCTCAGAATTTGAGTGAAAAATGGACAAGGCTAAACGTAATATTAAGCCGTTTGATG
GCGAGAAGTACGCGATTTGGAAATTTAGAATTAGGGCTCTTTTAGCCGAGCAAGATGTGCTTAA
AGTAGTTGATGGTTTAATGCCTAACGAGGTAGATGACTCCTGGAAAAAGGCAGAGCGTTGTGCA
AAAAGTACAATAATAGAGTACCTAAGCGACTCGTTTTTAAATTTCGCAACAAGCGACATTACGG
CGCGTCAGATTCTTGAGAATTTGGACGCCGTTTATGAACGAAAAAGTTTGGCGTCGCAACTGGC
GCTGCGAAAACGTTTGCTTTCTCTGAAGCTATCGAGTGAGATGTCACTATTAAGCCATTTTCAT
ATTTTTGACGAACTTATAAGTGAATTGTTGGCAGCTGGTGCAAAAATAGAAGAGATGGATAAAA
TTTCTCATCTACTGATCACATTGCCTTCGTGTTACGATGGAATTATTACAGCGATAGAGACATT
ATCTGAAGAAAATTTGACATTGGCGTTTGTGAAAAATAGATTGCTGGATCAAGAAATTAAAATT
AAAAATGACCACAACGATACAAGCAAGAAAGTTATGAACGCGATCGTGCACAACAATAATAACA
CTTATAAAAATAATTTGTTTAAAAATCGGGTAACTAAACCAAAGAAAATATTCAAGGGAAATTC
AAAGTATAAAGTCAAGTGTCACCACTGTGGCAGAGAAGGCCACATTAAAAAAGATTGTTTCCAT
TATAAAAGAATATTAAATAATAAAAATAAAGAAAATGAAAAACAAGTTCAAACTGCAACATCAC
ACGGCATTGCGTTTATGGTAAAAGAAGTGAATAATACTTCAGTGATGGACAACTGCGGGTTTGT
CCTTGATTCTGGTGCTAGTGACCATCTTATAAATGATGAGTCGCTGTATACCGACAGTGTGGAG
GTTGTGCCTCCACTTAAGATTGCAGTGGCCAAGCAAGGCGAATTTATTTATGCCACTAAGCGTG
GTATTGTCCGACTACGGAATGACCATGAGATTACACTGGAGGATGTACTCTTTTGTAAGGAAGC
TGCTGGTAATTTGATGTCCGTAAAGCGTCTCCAAGAGGCAGGAATGTCGATCGAATTTGACAAA
AGCGGTGTAACCATTTCGAAAAATGGGTTAATGGTTGTCAAAAATTCAGGTATGTTAAACAATG
TACCTGTGATCAATTTTCAAGCATATTCTATAAATGCTAAGCATAAAAATAATTTTCGTTTATG
GCATGAGAGGTTTGGCCATATAAGCGATGGCAAATTATTAGAAATAAAACGAAAGAATATGTTT
AGTGATCAAAGTCTTCTAAACAACTTAGAGTTATCATGTGAAATTTGTGAACCCTGTTTAAATG
GTAAACAGGCAAGACTTCCTTTTAAACAATTGAAAGATAAGACCCATATTAAAAGACCACTTTT
TGTAGTACACTCAGATGTCTGTGGGCCTATTACTCCAGTTACTTTAGATGATAAAAATTATTTT
GTGATCTTTGTTGATCAGTTTACACATTATTGTGTAACTTATTTAATTAAATATAAATCTGATG
TGTTTAGCATGTTTCAAGATTTTGTAGCCAAGAGTGAAGCTCATTTTAATTTAAAGGTTGTGTA
CTTATACATTGACAATGGTAGAGAATACTTGTCAAATGAGATGAGACAATTTTGTGTTAAGAAA
GGAATTTCTTATCACTTAACAGTGCCACATACACCTCAGTTAAATGGTGTTTCTGAGAGAATGA
TAAGAACCATTACGGAAAAAGCTCGAACCATGGTTAGTGGTGCAAAGCTAGATAAAAGCTTTTG
GGGCGAAGCAGTATTAACTGCTACTTATTTAATCAACAGAATTCCTAGTAGAGCACTTGTTGAT
AGTTCAAAGACCCCATATGAGATGTGGCACAATAAGAAGCCATACTTAAAACATTTGAGAGTGT
TTGGTGCAACTGTTTATGTGCATATTAAAAACAAACAAGGAAAGTTTGATGATAAATCATTTAA
AAGTATTTTTGTGGGCTATGAACCCAATGGTTTTAAGTTGTGGGATGCTGTAAATGAAAAATTT
ATTGTCGCAAGAGATGTTGTTGTCGATGAAACCAATATGGTTAATTCTAGAGCTGTTAAATTTG
AAACAGTGTTCCTGAAAGATAGTAAGGAAAGTGAAAATAAAAATTTTCCGAATGACAGTAGGAA
AATAATACAAACAGAATTCCCGAATGAGAGTAAGGAATGCGACAACATACAATTCCTGAAAGAT
AGTAAGGAAAGTGAAAATAAAAATTTTCCGAATGACAGTAGGAAAATAATACAAACAGAATTCC
CGAATGAGAGTAAGGAATGCGACAACATACAATTCCTGAAAGATAGTAAGGAAAGTAATAAATA
TTTTCTGAATGAGAGTAAGAAAAGAAAGCGAGATGATCACCTGAATGAAAGTAAGGGATCAGGC
AACCCGAATGAGAGTAGGGAAAGTGAAACAGCAGAGCACTTAAAAGAAATTGGAATTGATAATC
CAACTAAAAATGATGGCATAGAAATTATTAATAGAAGAAGTGAGAGATTAAAGACTAAGCCTCA
GATATCCTATAATGAAGAGGATAATAGTCTAAATAAAGTTGTTCTAAATGCTCACACTATATTT
AACGATGTCCCAAATTCATTTGATGAAATTCAATATAGGGATGATAAATCTTCTTGGGAAGAAG
CCATCAATACAGAGTTAAATGCTCATAAAATTAATAATACTTGGACAATTACAAAAAGGCCTGA
AAACAAAAATATTGTAGATAGCAGATGGGTATTTTCTGTTAAATATAATGAACTTGGAAATCCA
ATTAGATACAAAGCTAGATTGGTTGCACGAGGATTCACTCAAAAATACCAAATAGACTATGAAG
AGACATTTGCTCCTGTAGCTAGAATTTCAAGTTTCCGATTTATATTGTCATTAGTAATACAGTA
TAACTTGAAAGTCCATCAAATGGATGTAAAAACAGCTTTCTTAAATGGCACGTTAAAAGAGGAA
ATTTATATGAGACTTCCTCAAGGTATATCGTGTAATAGTGACAATGTGTGTAAATTGAATAAGG
CAATTTACGGACTCAAGCAAGCGGCTAGATGCTGGTTTGAAGTATTTGAGCAAGCATTGAAAGA
GTGTGAGTTTGTAAACTCTTCAGTTGATCGCTGTATATATATTTTAGACAAAGGTAACATCAAT
GAAAACATATATGTATTATTATATGTAGATGATGTGGTTATAGCTACAGGAGATATGACAAGAA
TGAATAACTTCAAAAGGTATTTAATGGAAAAGTTTAGGATGACTGACCTAAATGAAATAAAACA

TTTTATTGGAATTAGGATAGAGATGCAGGAAGATAAAATCTATTTAAGCCAATCTGCATATGTT
AAAAAAATTTTAAGTAAATTTAACATGGAAAATTGTAATGCAGTTAGTACTCCTTTACCTAGTA
AAATAAATTATGAATTACTTAATTCAGATGAAGACTGCAATACCCCATGCCGTAGCCTCATAGG
ATGTTTAATGTACATAATGCTTTGTACACGCCCAGATTTAACTACTGCAGTAAATATCTTGAGC
AGATATAGTAGCAAAAATAACTCCGAATTATGGCAGAACTTAAAAAGAGTTCTTAGATATTTGA
AGGGCACTATCGATATGAAATTGATTTTTAAAAAGAACTTGGCATTTGAAAATAAAATTATTGG
TTATGTGGATTCTGATTGGGCTGGTAGTGAAATTGATAGAAAAAGTACAACAGGGTATTTATTC
AAAATGTTTGATTTTAATCTCATTTGTTGGAATACAAAGAGACAGAACTCAGTAGCAGCCTCAT
CAACTGAAGCTGAGTATATGGCCCTATTTGAAGCCGTGAGAGAAGCTCTATGGCTTAAATTTTT
ATTAACTAGTATTAACATTAAACTAGAAAACCCCATTAAAATTTACGAAGACAATCAAGGCTGT
ATTAGCATAGCAAACAACCCCTCATGTCATAAACGAGCTAAACATATTGATATTAAATATCATT
TTGCCAGAGAGCAAGTTCAGAATAATGTGATTTGTCTTGAGTATATTCCTACAGAGAATCAACT
GGCTGACATATTTACAAAACCGTTGCCTGCTGCGAGATTTGTGGAGTTACGAGACAAATTGGGT
TTGCTGCAAGACGACCAATCGAATGCTGAATGAAATTTTTATATATATTTTTCAAATTTAAATT
CCTGTAAACATATTTTGTTACAATGATCTGATCGGGTTTTTCTGGGTTTTCCCCGTATCCTCGC
AGCAAATGCTGGATCAGTTAACACTTCCCAGAATGCACACCACCCACATTTGATAGTTACTAAT
GAATATTATTGTTATGTTTTTAATTATAGACGTTATTTTTGAGGGGGCG

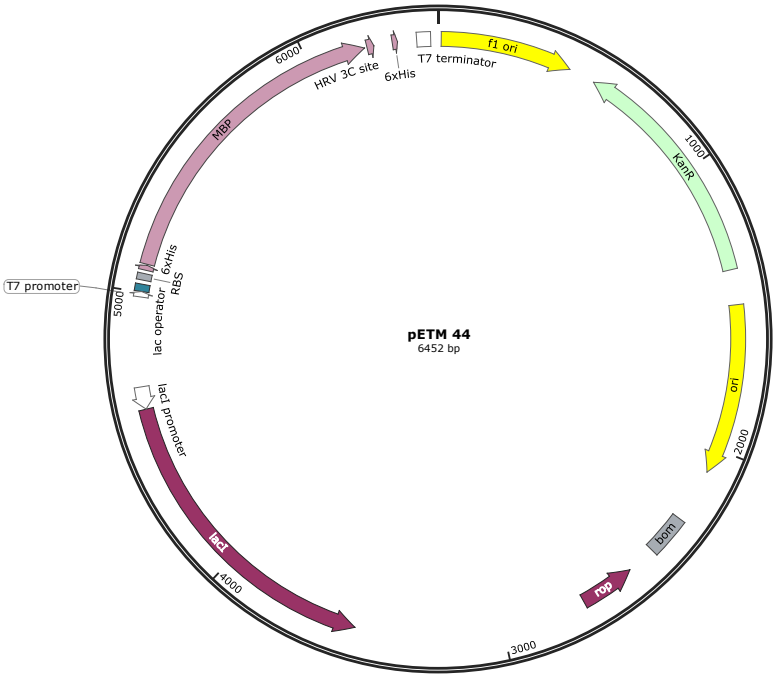
>FBgn0058469_CR40469
CACGTTCCTCACTAATTGTGGCTATTTGCGCCATCGTCTCATGCAATGTTATTTGAGAGATGGC
AAAATATATAGTATGTTTGTCTCCAATGTGTTGAGACTGAGAAGATATTGTACCCGTGAATTGA
TGAAAATTGATTGATTATATTGTAATGTTGATTTCATGAAAAACACGCTGTGTTGGAGGAACTC
AAACAAAACAAGCAAAAAATCC

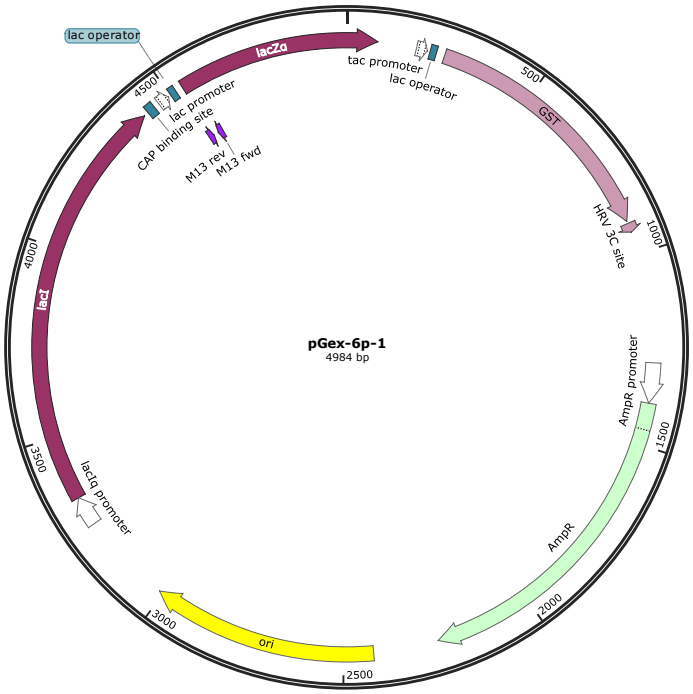
>353bp_SAT
ATGAAACTGTGTTCAACAATGGAAATTAAATTTCTTTGACATAGTGTGCAAATTTTGATGATGT
TACAAAATATGTGAAAATTTGCCGAAAAATTGATTTCCCTAAATCCTTCAAAAAGTAATGGAGA
TCGTTAGCACTGGTAATTAACTGCTGAAAACAGTTATTCTTGCATCTATATGACCCTTTTTAGC
CAAGTTATACCGAAAATTCCGTTTCTAAATATCAACTTTTTGGCAAAATCCGTTTTTCCAAGTT
TCGGTCATCAAATAATCAGTCTTTTCTGCCACAACTTTAAAAATAATTGTCTGAATATGGAATG
TCATACCTCGCTGAGCTCGTAATTAAATTTCCA

>260bp_SAT
TGGAAATTTAATTACGAGCTCAGCGAGGTATGACATTCCATATTCAGACAATTAATTTTAAAGT
TGTGGCAAAAAAACTGATTATTTAATGACCGAAATTTGGAAAAACGGATTTTGCCAATAAGTTA
ATATTTACTTTTTGAACGATTTAGGGAAGTTAATTTTTGGTTTAATTTTCGAATTTTTTTGAAA
GGGGGGTCATCAAAATTTGCATATATGCCGAAAAAATGTAATTTTCATTGTTGAACACAGTTTC
AT

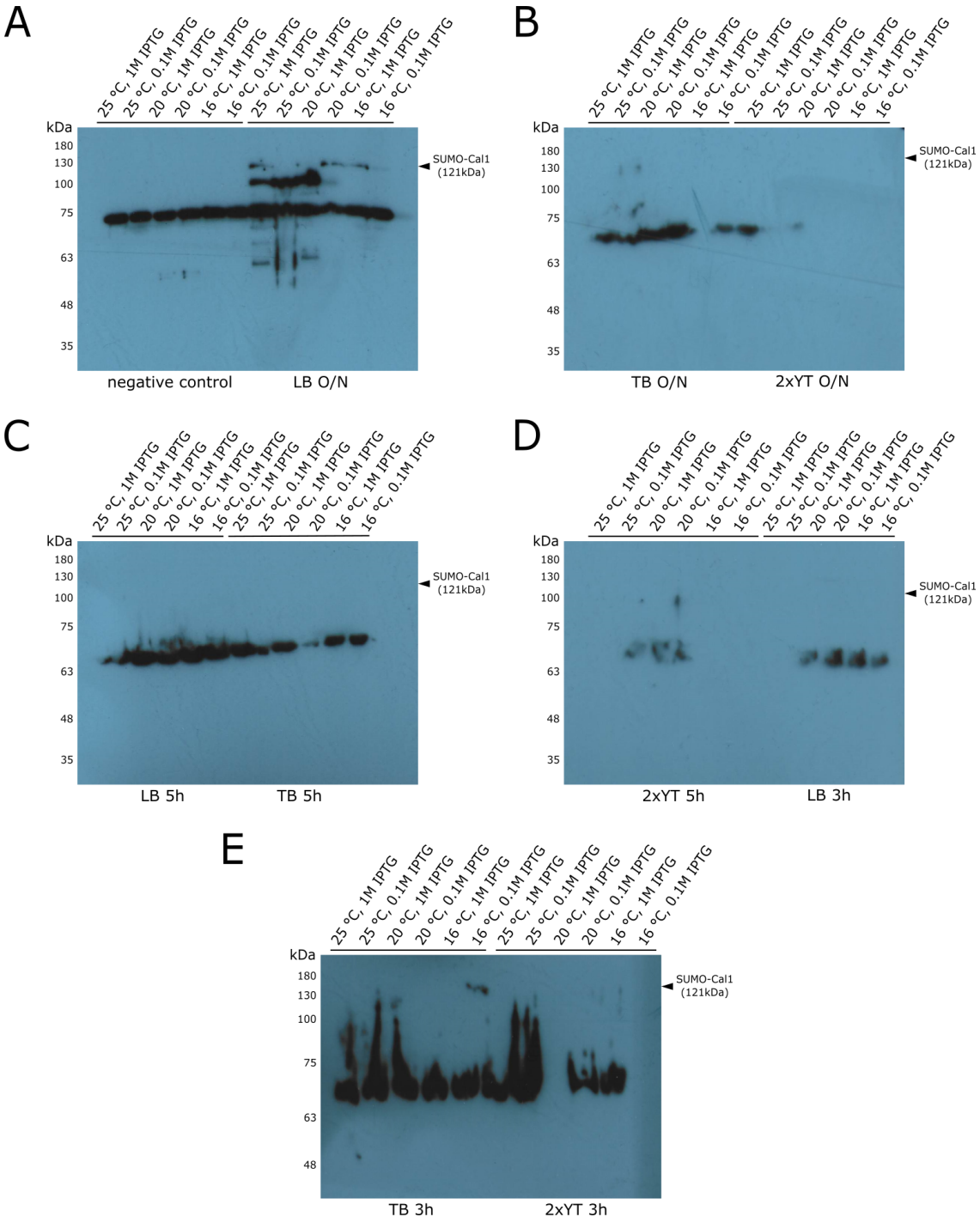
Appendix 2. **Sequences of used RNAs.**

Underlined sequences are fragment that I used for some experiments, where work with full length RNA was problematic.





Appendix 3. **Maps of used bacterial expression plasmids**

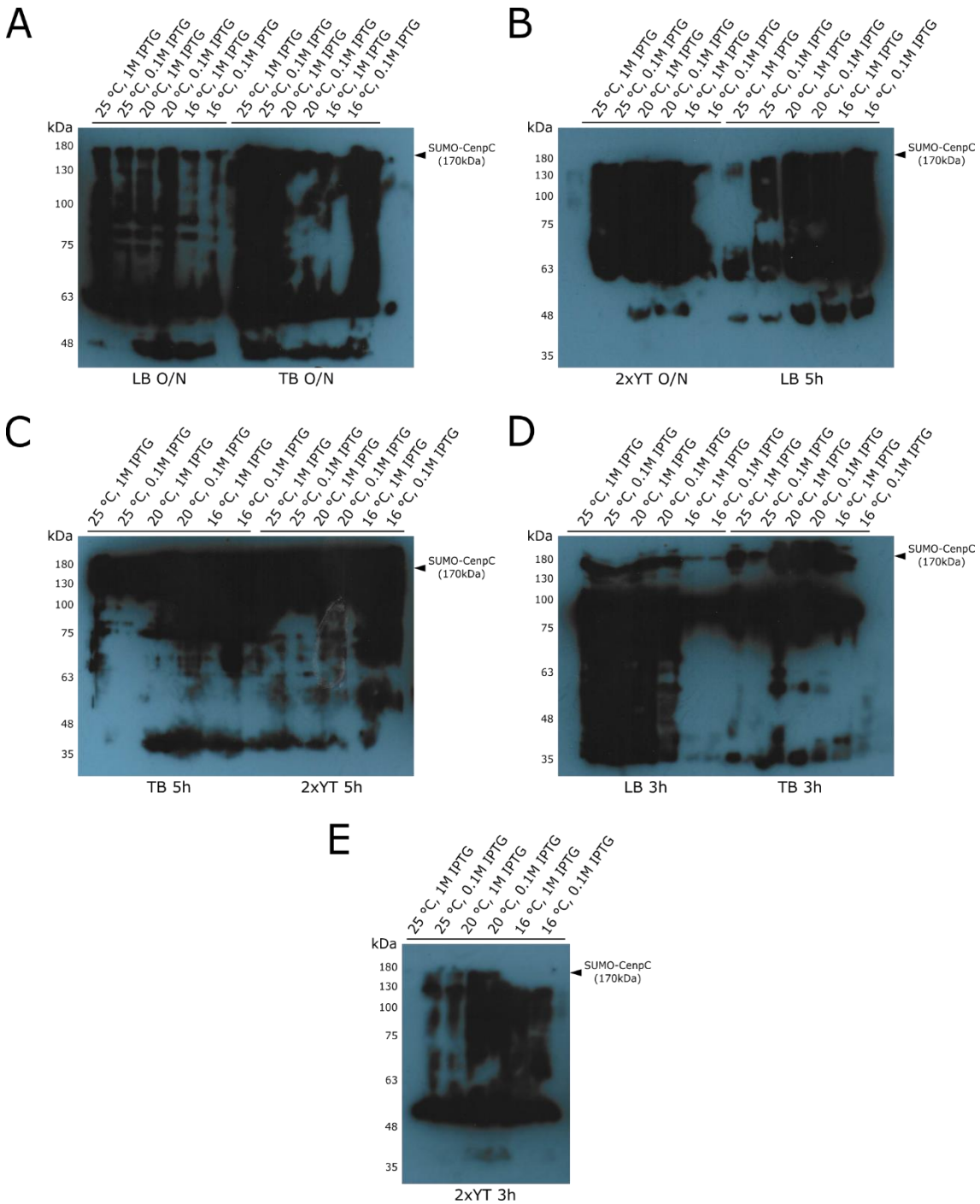

A
25 °C, 1M IPTG
25 °C, 0.1M IPTG
20 °C, 1M IPTG
20 °C, 0.1M IPTG
16 °C, 1M IPTG
16 °C, 0.1M IPTG
kDa
180
130
100
75
63
48
35
SUMO-Cal1
(121kDa)
negative control
LB O/N
B
TB O/N
2xYT O/N
C
LB 5h
TB 5h
D
2xYT 5h
LB 3h
E
TB 3h
2xYT 3h

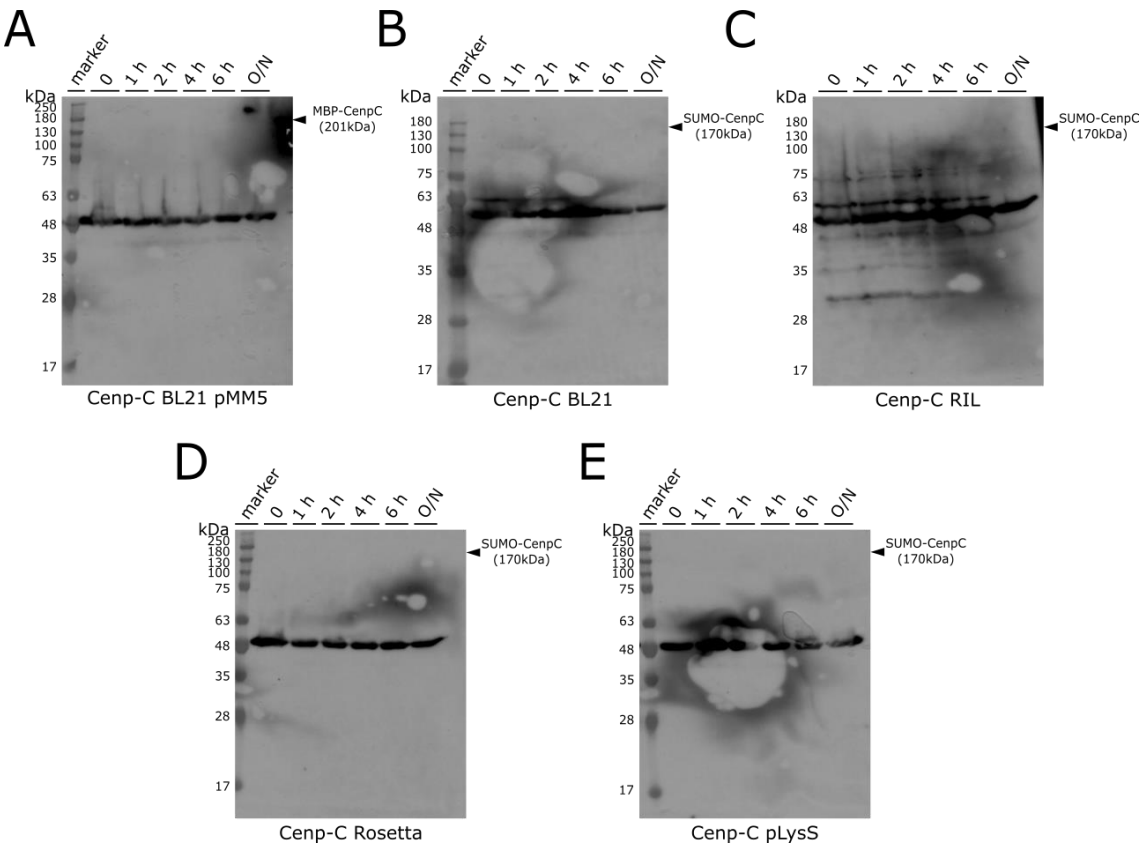
Appendix 4. **Western blots of expression tests of SUMO-Cal1 in different conditions**

Cal1 was expressed in various conditions in BL21DE3 strain. Samples of bacteria were taken and lysed in Lämmli buffer. Lysate was used for polyacrylamide gel electrophoresis and subsequent Western blot (WB).
(A) WB showing wild type cells (left half) and O/N expression of SUMO-Cal1 in LB medium (right half). Conditions of each sample are written above each lane.
(B) WB showing O/N expression of SUMO-Cal1 in TB medium (left half) and O/N expression of SUMO-Cal1 in 2xYT medium (right half). Conditions of each sample are written above each lane.
(C) WB showing 5 h expression of SUMO-Cal1 in LB medium (left half) and 5 h expression of SUMO-Cal1 in TB medium (right half). Conditions of each sample are written above each lane.
(D) WB showing 5 h expression of SUMO-Cal1 in 2xYT medium (left half) and 3 h expression of SUMO-Cal1 in LB medium (right half). Conditions of each sample are written above each lane.
(E) WB showing 3 h expression of SUMO-Cal1 in TB medium (left half) and 3 h expression of SUMO-Cal1 in 2xYT medium (right half). Conditions of each sample are written above each lane.
The expression of SUMO-Cal1 in BL21DE3 strain is poor. No matter the conditions the correct band are mostly not there and if, then barely visible. Unfortunately, the Cal1 antibody seems to bind some bacterial protein, because there is very strong unspecific band in each blot. This expression protocol was not the right one.

Appendix 5. **Western blots of expression tests of SUMO-Cenp-C in different conditions**

Cenp-C was expressed in various conditions in BL21DE3 strain. Samples of bacteria were taken and lysed in Lämmli buffer. Lysate was used for polyacrylamide gel electrophoresis and subsequent Western blot (WB).
(A) WB showing O/N expression of SUMO-Cenp-C in LB (left half) and O/N expression of SUMO-Cenp-C in TB medium (right half). Conditions of each sample are written above each lane.
(B) WB showing O/N expression of SUMO-Cenp-C in 2xYT (left half) and 5 h expression of SUMO-Cenp-C in LB medium (right half). Conditions of each sample are written above each lane.
(C) WB showing 5 h expression of SUMO-Cenp-C in TB medium (left half) and 5 h expression of SUMO-Cenp-C in 2xYT medium (right half). Conditions of each sample are written above each lane.
(D) WB showing 3 h expression of SUMO-Cenp-C in LB medium (left half) and 3 h expression of SUMO-Cenp-C in TB medium (right half). Conditions of each sample are written above each lane.
(E) WB showing 3 h expression of SUMO-Cenp-C in 2xYT medium. Conditions of each sample are written above each lane.
The expression of SUMO-Cenp-C in BL21DE3 strain is also poor. It seems to be marginally better than that of Cal1 because there is at least band of the correct size visible on the blots. Nevertheless, the expression is not only poor, but it also seems that the bacteria are rapidly degrading whatever is produced, since we can see the Cenp-C antibody binding to a large amount of protein debris of various sizes. This expression protocol was not the right one for Cenp-C either.


A
25 °C, 1M IPTG
25 °C, 0.1M IPTG
20 °C, 1M IPTG
20 °C, 0.1M IPTG
16 °C, 1M IPTG
16 °C, 0.1M IPTG
kDa
180
130
100
75
63
48
SUMO-CenpC
(170kDa)
LB O/N
TB O/N
B
2xYT O/N
LB 5h
C
TB 5h
2xYT 5h
D
LB 3h
TB 3h
E
2xYT 3h



Appendix 6. **Western blots of expression tests of Cenp-C in different bacterial strains**

Cenp-C was expressed in various bacterial strains in LB medium. Bacterial samples were taken and lysed in Lämmli buffer. The lysate was used for polyacrylamide gel electrophoresis and subsequent Western blot (WB).

(A) The WB shows the expression of MBP-Cenp-C in the BL21DE3 strain over time. However, the bands corresponding to the MBP-Cenp-C are not visible at all. This strain is not the correct one.

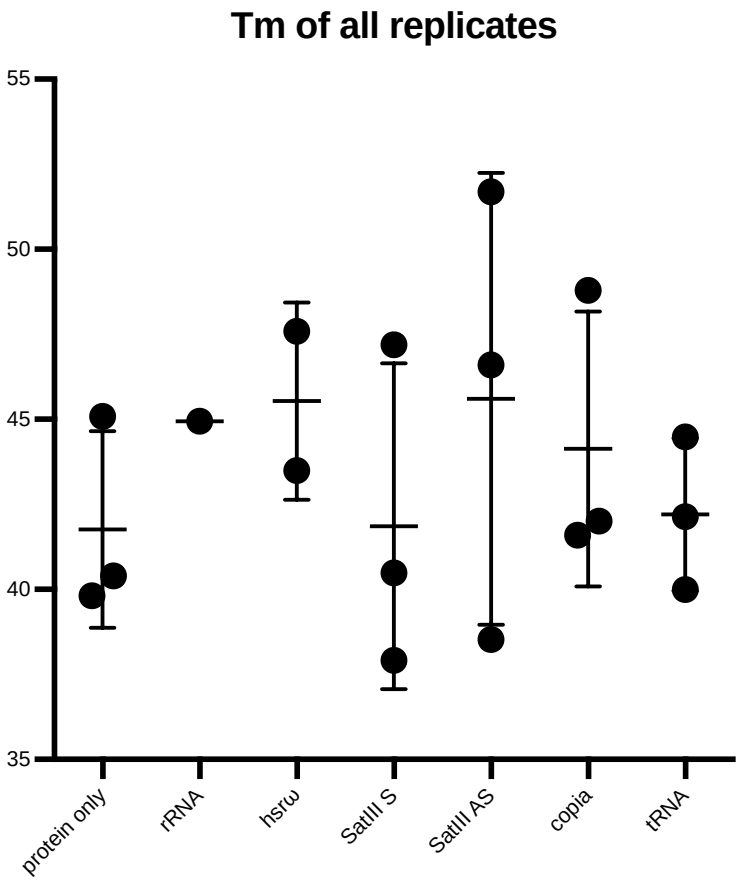
(B) The WB shows the expression of SUMO-Cenp-C in BL21DE3 codon plus strain over time. The bands corresponding to the SUMO-Cenp-C are not visible at all. This strain is not the correct one.

(C) The WB shows the expression of SUMO-Cenp-C in the RIL strain over time. The bands corresponding to the SUMO-Cenp-C are not visible at all. This strain is not the correct one.

(D) The WB shows the expression of SUMO-Cenp-C in the Rosetta strain over time. The bands corresponding to the SUMO-Cenp-C are not visible at all. This strain is not the correct one.

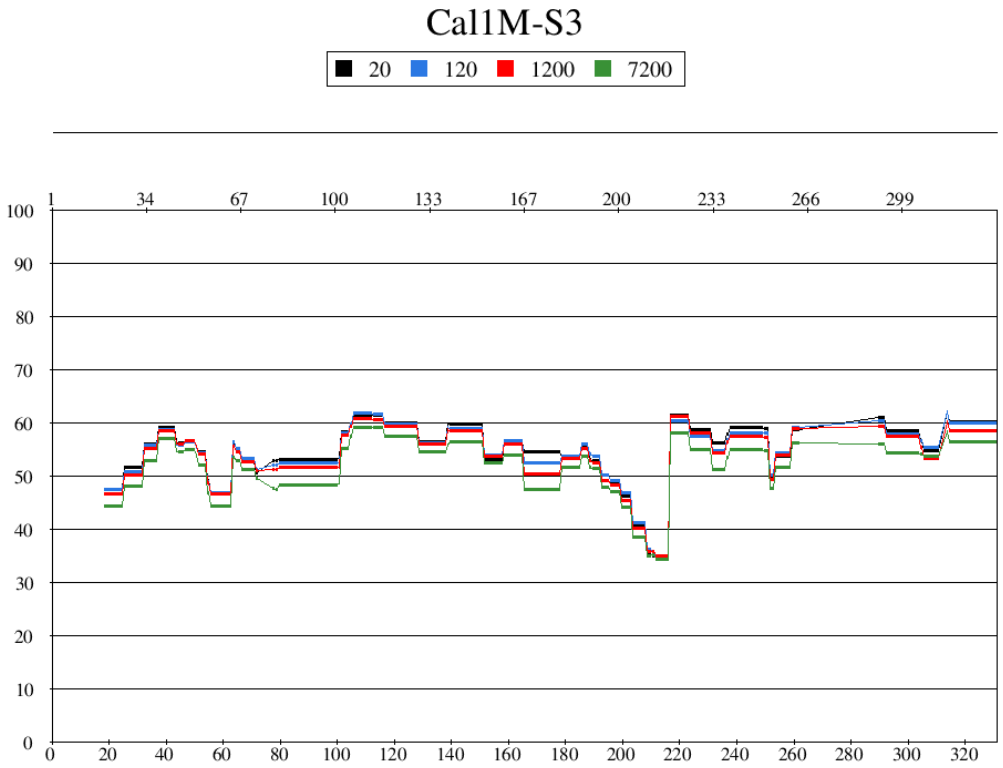
(E) The WB shows the expression of SUMO-Cenp-C in the pLysS strain over time. The bands corresponding to the SUMO-Cenp-C are visible but weak. This strain is not the correct one.

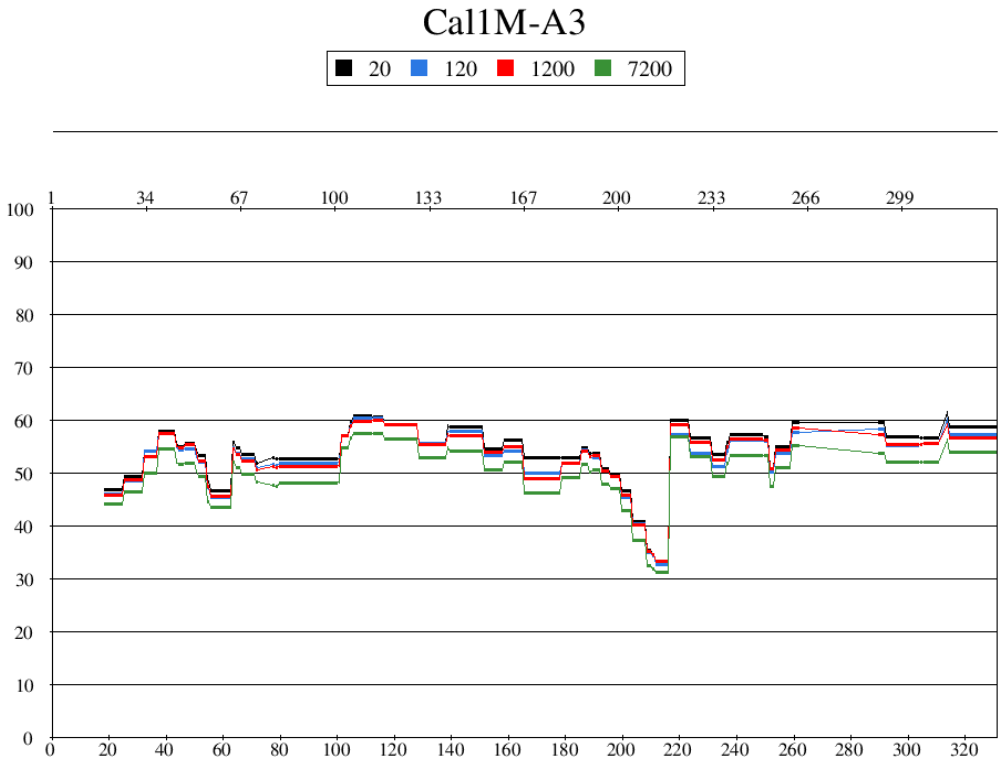
The results of the expression tests indicate that expressing Cenp-C in bacteria may not be feasible. Unlike Cal1, which is at least detectable by WB, Cenp-C appears to be completely absent. The observed bands are likely unspecific bacterial proteins.

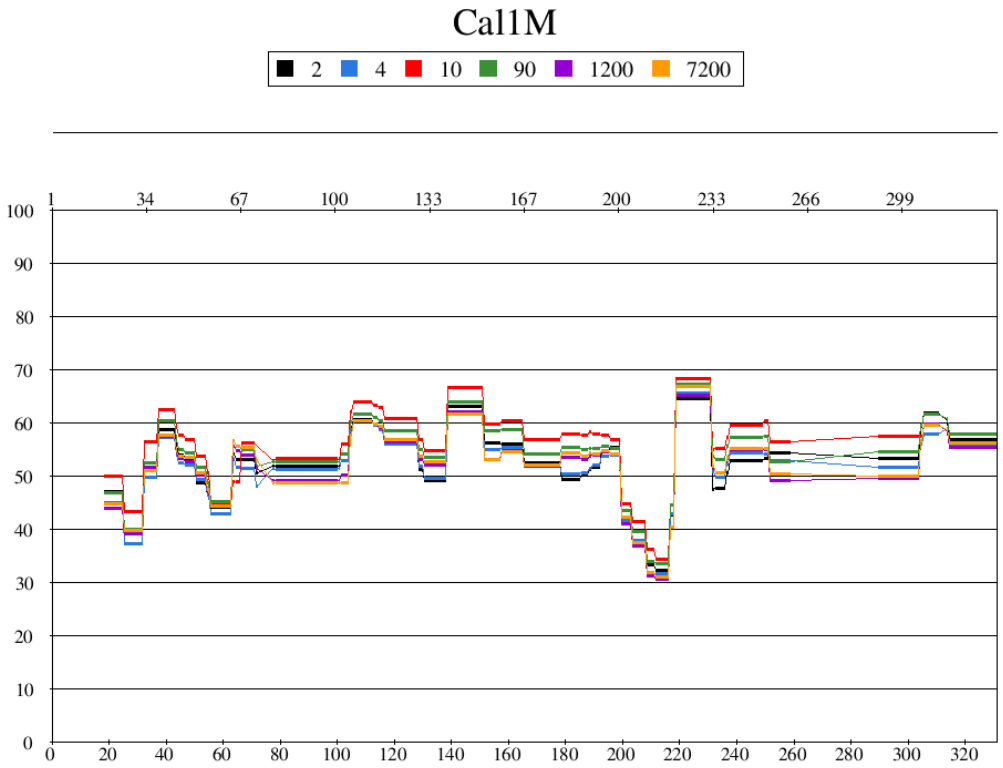


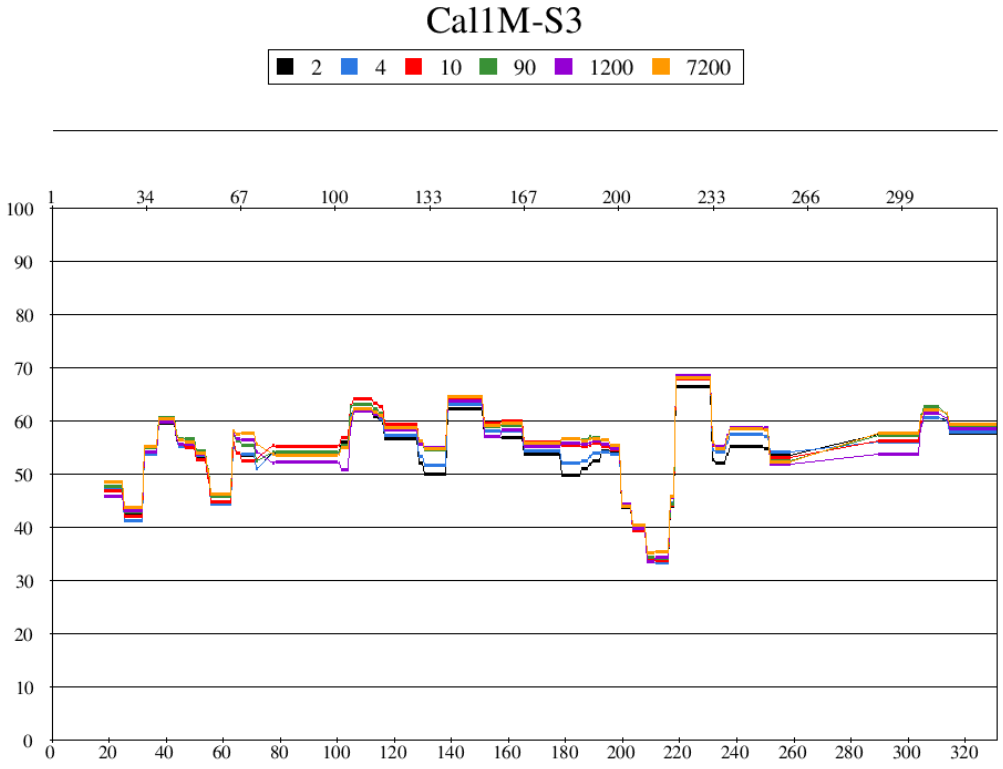
Appendix 7. **Differential scanning fluorimetry results were not replicable**

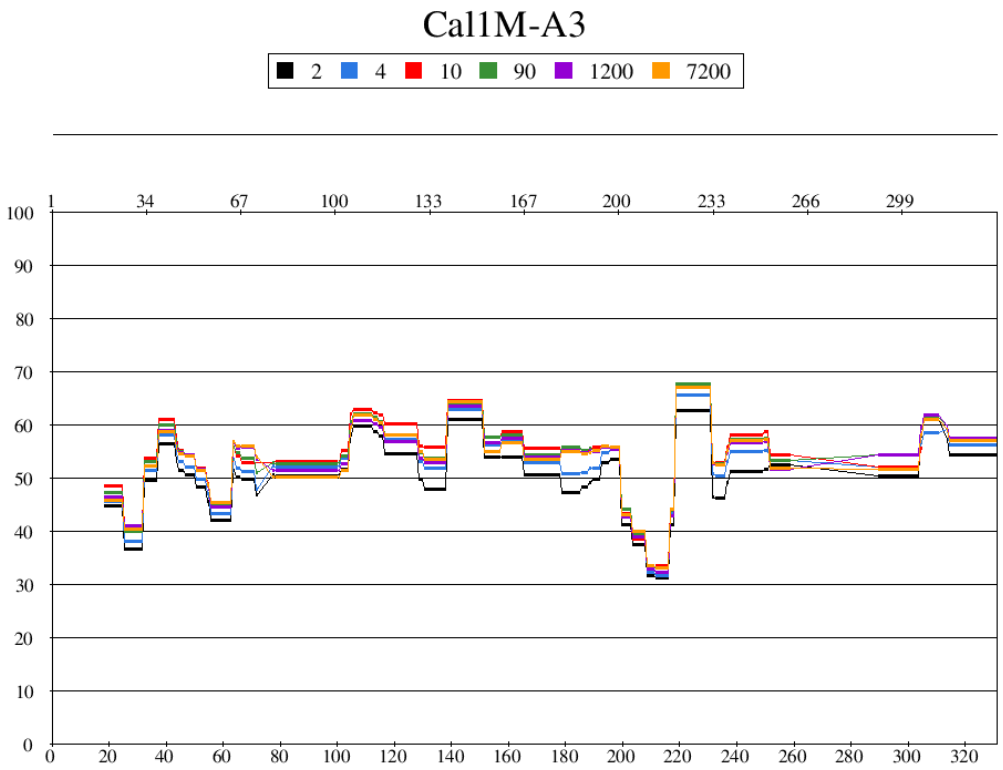
Technical replicates of various samples had shown great divergence, that could not be explained by pipetting errors. Graph shows Cal1M alone or in mixture with RNAs (X-axis) and its experimentally measured melting point (Y-axis, ° C). The observed differences could be explained by the fact that the protein is not stable enough during the experiments and precipitates. This leads to aggregation of the fluorescent dye and non-replicable results.


Cal1M-S3
20
120
1200
7200
1
34
67
100
133
167
200
233
266
299
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320


Cal1M-A3
20
120
1200
7200
1
34
67
100
133
167
200
233
266
299
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320

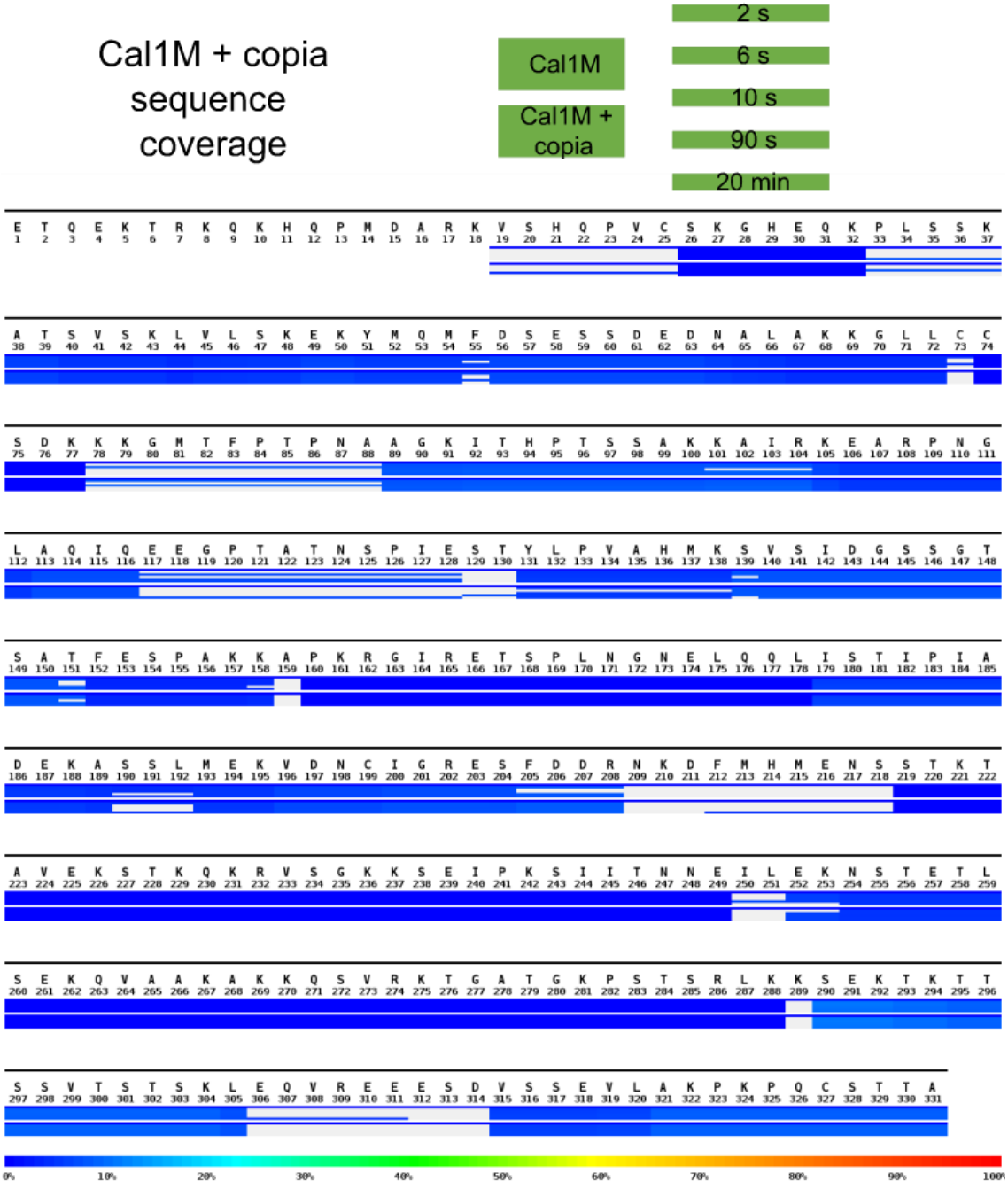

Cal1M
2
4
10
90
1200
7200
1
34
67
100
133
167
200
233
266
299
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320


Cal1M-S3
2
4
10
90
1200
7200
1
34
67
100
133
167
200
233
266
299
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320



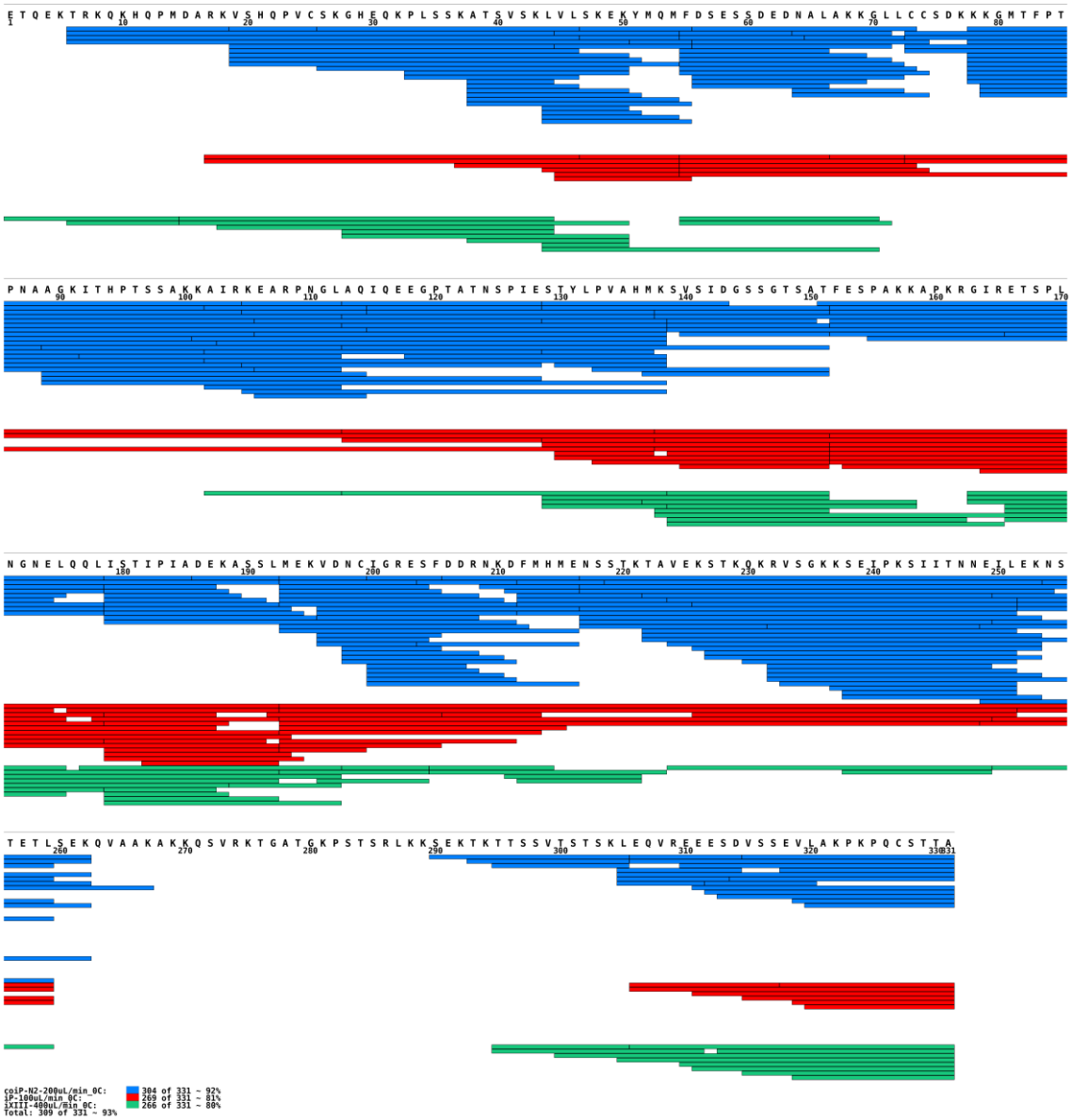
Appendix 8. **^{1}H/^{2}H of Cal1M in presence of SatIII RNA**

HDX of Cal1M peptides (X-axis) and the deuteration with calculated back-exchange (Y-axis). Different colours represent experimental times between starting the HDX and quenching it.

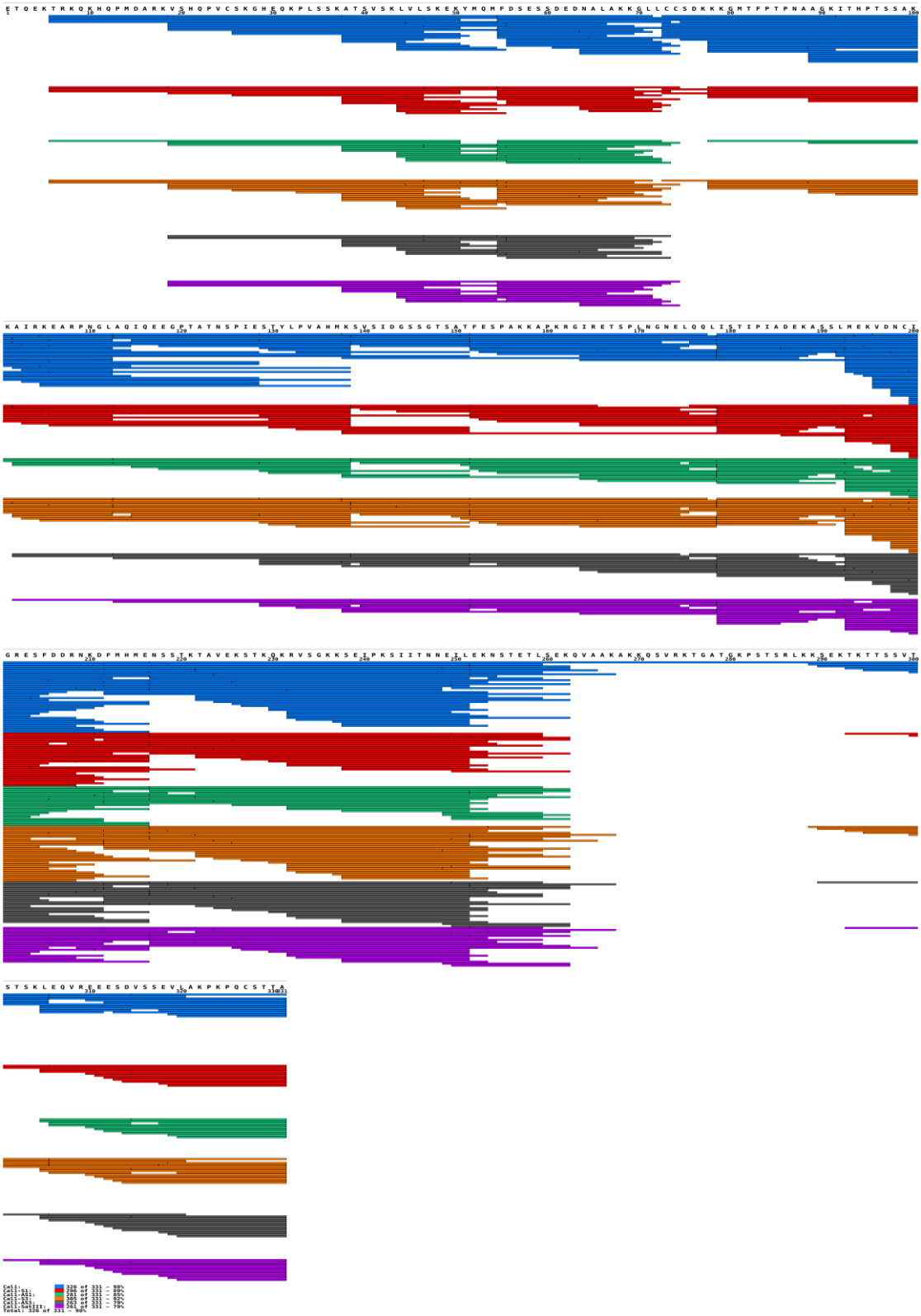


Appendix 9. **Mass spectrum of Cal1M in presence of copia RNA**

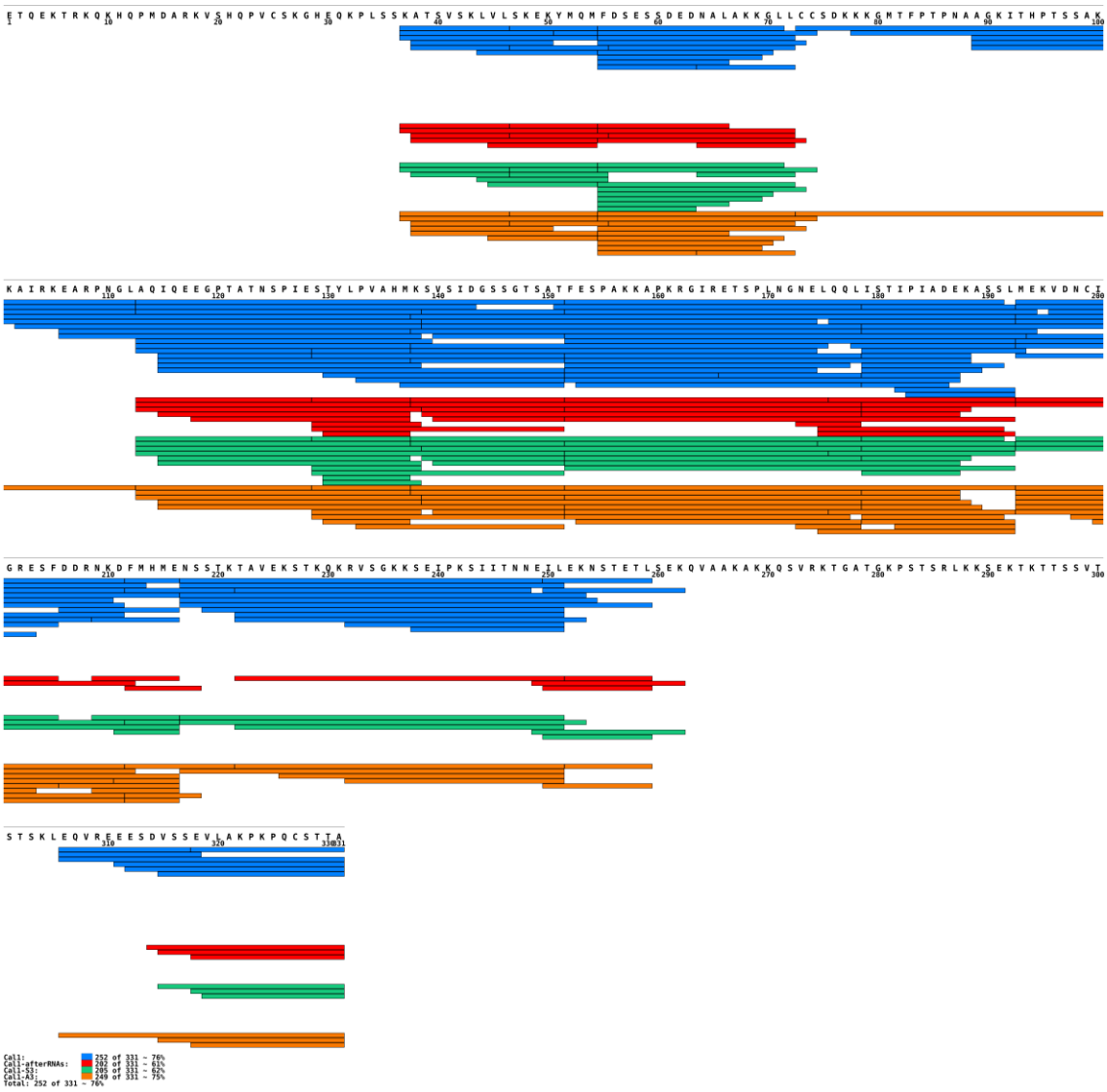
This initial experiment was designed to show, whether RNA presence in the sample leads to precipitation of some peptides. If it would, the precipitated peptides would not be visible in the spectrum in comparison to the sample without RNA. This was not confirmed, as both spectrums show comparable results.



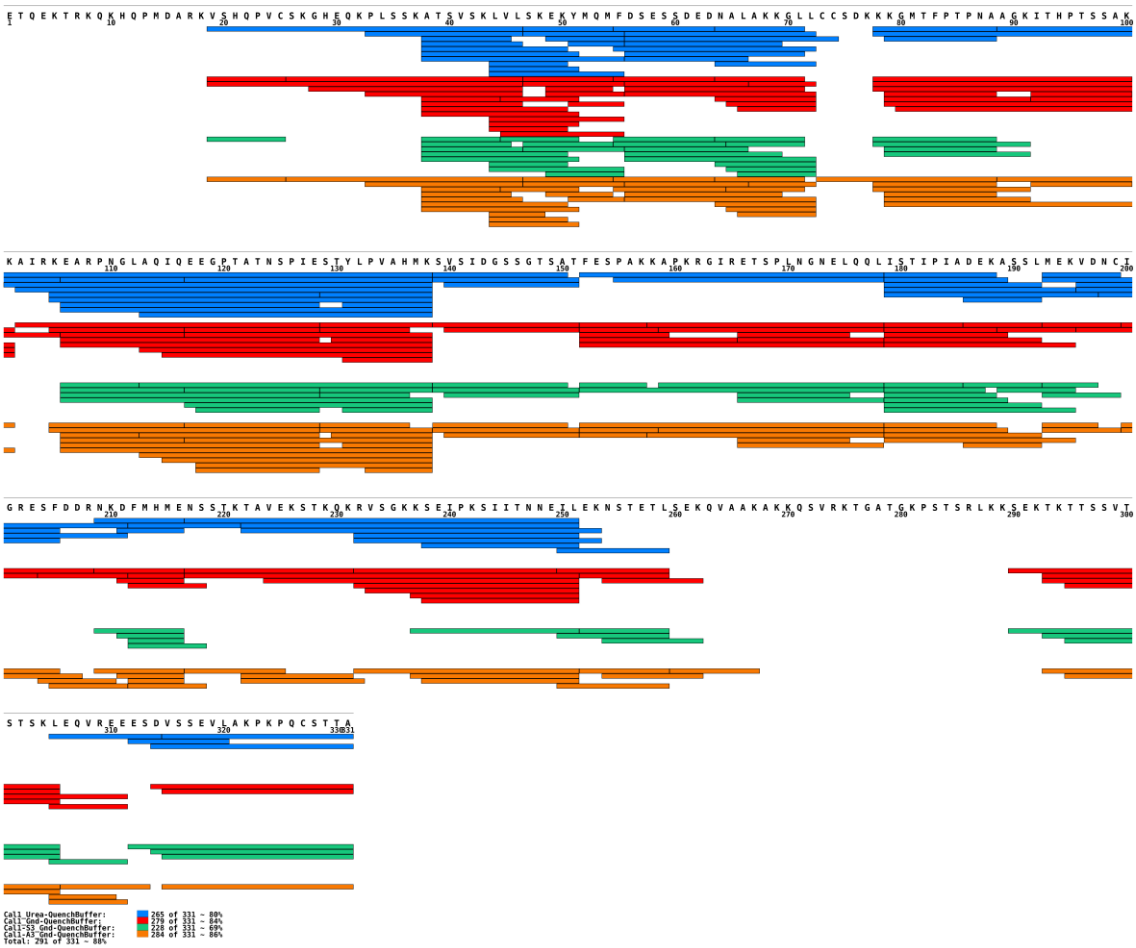
Appendix 10. **Different cleavage conditions of Cal1M**



Appendix 11. **Cleavage tests with different RNAs**



Appendix 12. **Guanidium-based quench tests**



Appendix 13. **Thiourea-based quench tests**

List of Abbreviations

ACN – acetonitrile
AF – anisotropy of fluorescence
Cal1 – chromosome alignment defect 1 in *Drosophila*
CD – circular dichroism spectrometry
Cenp-A – centromeric protein A
Cenp-C – centromeric protein C
cDNA – complementary DNA
Cid – centromere indentifier in *Drosophila*, aka Cenp-A or CenH3
CV – column volume
DSF – differential scanning fluorimetry
DTT – dithiothreitol
EMSA – electrophoretic mobility shift assay
ESI – electrospray ionisation
FPLC – fast protein liquid chromatography
GST – glutathione S-tranferase
HDX – hydrogen deuterium exchange
$^{1}H/^{2}H$ – hydrogen/deuterium
HPLC – high pressure liquid chromatography
IF - immunofluorescence
IPTG – isopropyl β-D-1-thiogalactosid, galactose mimetics, inducing agent
LB – Luria broth/Luria-Bertani medium/lysogenic broth
MBP – maltose binding protein
MS – mass spectrometry
OD_{600} – optical density at 600 nm
PAGE – polyacrylamide gel electrophoresis
PMSF – phenylmethylsulfonyl fluorid
SDS – sodium dodecyl sulphate
TB – terrific broth
TCEP – tris(2-carboxyethyl)phosphine

List of Figures

Acknowledgements

First acknowledgement goes to Sylvia, who took me to her lab, supervised me and supported me throughout my Ph. D. Thank you for everything!

I thank the whole community of Erhardt lab for creating a nice environment for work and afternoon beers and social activities. Most of all Aga, Saskia and Iris but all the others as well.

Then I would like to thank Matthias, who helped me with the more technical parts of the thesis, always offered a way out of any scientific problem and was overseeing my work throughout the years as my TAC member.

To Stefan for his support during all the TAC meetings and his encouragement.

To the people I cooperated with, from the mass spectrometry facilities in Heidelberg and in Vestec - Nicole, Pavla and Petr, and from Mayer lab – Szymon and Yang, for their technical expertise and help.

To prof. Klein for giving me space in his lab when my own institute was no longer viable workplace.

To Samuel, who, apart from being a great study and discussion partner, also convinced me to not give up so easily and try it abroad.

To all the people from ZMBH community, who made my life easier and more entertaining. Most of all Anna, Adam, Erik and Milan.

To Samantha, Kathy and Jasmin for lending me their language expertise and making my English and German better.

To all the people who helped me during the dark times of writing, who were not mentioned before – Goks, Tereza and Maria.

And last, but not the least, to my family, who supported me while I worked on something they do not understand without pushing me too much with constant annoying questions about timelines and life plans.