# Gene regulation via nucleosome stability modulation and RNA:DNA triplex formation


## DISSERTATION

DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES
DER NATURWISSENSCHAFTEN (DR. RER. NAT.) DER FAKULTÄT FÜR BIOLOGIE
UND VORKLINISCHE MEDIZIN DER UNIVERSITÄT REGENSBURG

vorgelegt von

**Sara Wernig-Zorc (geb. Ambrozic)**

aus

**Ljubljana, Slowenien**

im Jahr

**2022**

Das Promotionsgesuch wurde eingereicht am:

12.05.2022

Die Arbeit wurde angeleitet von:

Prof. Dr. Gernot Längst

Unterschrift

(Sara Wernig-Zorc)

# Table of Contents

*ABBREVIATIONS* ..... *1*

*1 SUMMARY* ..... *3*

*2 INTRODUCTION* ..... *5*

**2.1 Regulation of transcription in Eukaryotes** ..... **5**

**2.2 Nucleosomes** ..... **5**

2.1.1 Nucleosomal properties ..... 6

2.2.2 Non-canonical structures ..... 8

2.2.2 Differential sensitivity to MNase ..... 9

2.2.3 The fragile nucleosome ..... 10

**2.3 RNA:DNA triplexes** ..... **11**

2.3.1 Biological significance ..... 12

2.3.2 Non-coding RNAs ..... 13

2.3.3 Triplex stability ..... 14

2.3.4 *In vivo* triplex assembly ..... 15

2.3.5 Computational predictions ..... 16

*3 OBJECTIVES* ..... *18*

**3.1 Nucleosome stability** ..... **18**

**3.2 The triplex code** ..... **18**

*4 NUCLEOSOME STABILITY* ..... *20*

**4.1 Results** ..... **20**

4.1.1 The MACC score ..... 22

4.1.2 The nucMACC score ..... 24

4.1.3 Nucleosome accessibility score ..... 29

4.1.4 Nucleosome stability score ..... 30

4.1.5 Gene regulation via modulation of nucleosome stability ..... 32

4.1.6 Pipeline robustness ..... 35

4.1.7 The nucMACC pipeline ..... 38

**4.2 Discussion** ..... **41**

*5 THE TRIPLEX CODE* ..... *43*

**5.1 Results** ..... **43**

5.1.1 Comparing triplex motifs ..... 43
5.1.2 Binding affinity cut-off ..... 45
5.1.3 Pyrimidine motif ..... 45
5.1.4 Mixed motif ..... 49
5.1.5 Purine motif ..... 53
5.1.6 Published triplexes ..... 56
5.1.7 Mismatches ..... 58
5.1.8 High throughput binding affinity studies ..... 59

**5.2 Discussion ..... 66**

***6 MATERIALS AND METHODS ..... 69***

**6.1 Nucleosome stability ..... 69**

6.1.1 Annotation and publicly available data ..... 69
6.1.2 Software ..... 69
6.1.3 Data and code availability ..... 70
6.1.4 Data analysis ..... 70

**6.2 The triplex code ..... 73**

6.2.1 Technical devices ..... 73
6.2.2 Consumables ..... 73
6.2.3 Solutions and buffers ..... 74
6.2.4 Experimental procedures ..... 75
6.2.2 Software ..... 79
6.2.3 Data and code availability ..... 79
6.2.4 Data analysis ..... 79

***7 REFERENCES ..... 81***

***8 ACKNOWLEDGEMENTS ..... 92***

***APPENDIX 1: Nucleosome stability - supplemental figures ..... 93***

***APPENDIX 2: Triplex oligonucleotides ..... 97***

***APPENDIX 3: Published triplex pairs ..... 101***

***APPENDIX 4: The triplex code – supplemental figures ..... 102***

# ABBREVIATIONS

| | |
|---|---|
| A | Adenine |
| ac | Acetylation |
| ATP | Adenosine triphosphate |
| BSA | Bovine serum albumin |
| bp | Base pair |
| C | Cytosine |
| chRNA | Chromatin-associated RNA |
| ds | Double stranded |
| DNA | Deoxyribonucleic acid |
| eRNA | Enhancer-associated RNA |
| EMSA | Electrophoretic Mobility Shift Assay |
| ESC | Embryonic stem cells |
| G | Guanine |
| h | Hour |
| H1 | Histone 1 |
| H2A | Histone 2A |
| H2B | Histone 2B |
| H3 | Histone 3 |
| H4 | Histone 4 |
| K | Guanine or Thymine or Uracil |
| $K_d$ | Dissociation constant |
| kDa | Kilo Dalton |
| LOESS | Locally weighted scatterplot smoothing |
| lncRNA | Long non-coding RNA |
| M | Mixed base (Guanine or Thymine or Uracil) |
| me | Methylation |
| mRNA | Messenger RNA |
| miRNA | Micro RNA |
| MST | MicroScale Thermophoresis |
| NGS | Next Generation Sequencing |
| MNase | Micrococcal Nuclease |
| MACC | MNase accessibility score |

| | |
|---|---|
| nucMACC | Nucleosome MNase accessibility score |
| nM | Nano molar |
| nt | Nucleotides |
| NDR | Nucleosome depleted region |
| NL | Non-labelled |
| PAA | Polyacrylamide |
| PTM | Post-transcriptional modifications |
| R | Purine base (Adenine or Guanine) |
| rRNA | Ribosomal RNA |
| RNA | Ribonucleic Acid |
| RNA pol-II | RNA polymerase II |
| ss | Single stranded |
| SD | Standard deviation |
| sf | Self-forming dsDNA |
| seq | High throughput sequencing |
| snoRNA | Small nucleolar RNA |
| snRNA | Small nuclear RNA |
| T | Thymine |
| TO | Thiazole orange |
| TAD | Topologically associated domain |
| tRNA | Transfer RNAs |
| TSE | Triplex-SELEX-EMSA |
| TSS | Transcription start site |
| TES | Transcription termination site |
| TRIS | Tris (hydroxymethyl) aminomethane |
| TrTS | Triplex Targeting Site |
| TF | Transcription factor |
| TFO | Triplex forming oligo |
| U | Uracil |
| V | Volt |
| Y | Pyrimidine base (Cytosine or Thymine or Uracil) |
| µM | Micro molar |
| °C | Degrees Celsius |

# 1 SUMMARY

Gene regulation is a tightly controlled process in Eukaryotes. Coding and non-coding regions work together to ensure proper spatiotemporal gene expression through 3D chromatin organization, nucleosome positioning, histone tail post-translational modifications, epigenetic DNA modifications, and non-canonical nucleic acid structures such as RNA:DNA triplexes. In this thesis, I investigate two of these mechanisms, nucleosomes, and RNA:DNA triplexes.

In the first chapter, I focus on characterizing nucleosomal properties, such as stability and accessibility, and explore how these properties change to regulate gene expression. I show that Eukaryotic organisms can modulate nucleosome stability and change the RNA polymerase pausing rate, which in turn regulates gene expression. Intriguingly, I find a distinct group of un-stable nucleosomes enriched at the TSS of promoters marked with motif one, an M1BP TF-specific motif, in *D. melanogaster*. Modulation of stability may ensure a fast response to environmental cues and proper spatiotemporal expression of developmental genes. Furthermore, I developed a bioinformatic pipeline called nucMACC, with which scientists can study nucleosomal properties and positioning in their projects. I show the nucMACC pipeline is consistent and robust and provide recommendations for minimum sequencing depth, MNase titrations, and spike-in use. In summary, the nucMACC pipeline provides high-resolution nucleosome positions and an automated way of calling non-canonical nucleosomes, un-stable nucleosomes, hyper-accessible nucleosomes, hypo-accessible nucleosomes, and stable canonical nucleosomes.

In the second chapter, I study the triplex binding code and explore how the code changes based on the triplex motif (Purine, pyrimidine, and mixed), sequence Guanine content, length, and the nucleic acid (RNA or DNA). Triplexes are non-canonical DNA/RNA structures consisting of three nucleotide strands, most commonly a double-stranded DNA molecule and a single-stranded RNA molecule in its major groove. I show that major differences exist between the triplex motifs and between RNA:DNA and DNA:DNA triplexes. Interestingly, I discovered that the mixed RNA:DNA triplex motif permits triplex formation only at very narrow Guanine contents, while DNA:DNA mixed motif triplexes are able to form at almost any Guanine content. Moreover, I confirm the newly defined triplex code by testing published triplex pairs, of which half are unable to form a triplex under physiological conditions. Furthermore, I developed a high throughput method to investigate the triplex binding code in the context of molecular crowding and competition. I find that certain mismatches in the motif stabilize triplex formation and are preferentially selected for triplex formation over the complementary Hoogsteen base-pairing motif. Moreover, I investigate how the location of a mismatch modifies triplex stability and show that the middle section of the triplex is more sensitive to mismatches than flanking

regions. In summary, I add to the existing triplex code by delineating the differences in the binding code between triplex motifs and RNA:DNA and DNA:DNA triplexes. I show how the binding code changes in the context of molecular crowding and competition and moreover, show the differential effect of different mismatch locations.

# 2 INTRODUCTION

## 2.1 Regulation of transcription in Eukaryotes

Eukaryotic genomes contain coding and non-coding genes, i.e., genes encoding and genes not encoding the blueprints for proteins, respectively. The non-coding regions represent more than 98% of the human genome, while coding regions represent only 2%. Non-coding regions were initially thought of as junk DNA (Ohno, 1972); however, high-throughput sequencing studies revealed that non-coding regions are essential for proper gene expression (Dunham *et al.*, 2012).

Coding and non-coding regions cooperate in regulating DNA transcription. On the one hand, regulatory proteins, such as transcription factors (TF), bind to specific DNA sequences and cooperatively interact amongst them, creating local molecular environments that bring regulatory elements together. DNA transcription is initiated when a TF binds to a gene promoter or enhancer and initiates a series of interactions between multiple proteins at the promoter region. Transcription starts when the transcription complex is formed at the promoter region, which assists RNA polymerase II (RNA pol-II) binding. RNA pol-II is a multiprotein complex that transcribes DNA into precursors of messenger RNA (mRNA), which are afterward translated into proteins. TF binding is often hindered by a nucleosome positioned at its target. Thus, nucleosome positioning represents a way of regulating gene expression. On the other hand, non-coding RNAs assist in the formation of the 3D chromatin structure, organizing it in chromatin domains, which is essential for correct spatiotemporal gene expression (Schubert *et al.*, 2012; Li and Fu, 2019).

Furthermore, gene transcription is regulated by histone tail post-translational modifications (PTM), epigenetic modifications such as DNA methylation, and non-canonical nucleic acid structures (RNA:DNA triplexes, G-quadruplexes, R-loops, etc.).

## 2.2 Nucleosomes

The total amount of DNA in a human cell joined together corresponds to an approximately 2 m long fiber that must be compacted in the cell nucleus with a diameter of 10 µm. To allow precise regulation, this requires compaction of DNA into a higher-order organization and spatially organized compaction of chromatin. Local chromatin compaction leads to differential accessibility of transcriptional machinery to genes in euchromatin and heterochromatin regions.

Nucleosomes form the building blocks of chromatin, consisting of 147 bp of DNA tightly wrapped around a histone octamer, composed of one H3/H4 tetramer and two H2A/H2B histone dimers (Figure

1). DNA is bound to histones by over 360 direct and indirect hydrogen bonds, rendering the nucleosome a stable particle (Davey *et al.*, 2002). Neighboring nucleosomes are connected by linker DNA, which varies in length from 15 to 95 bp depending on the species and cell type. Often linker DNA is bound by the H1 histone (Van Holde, 1985).

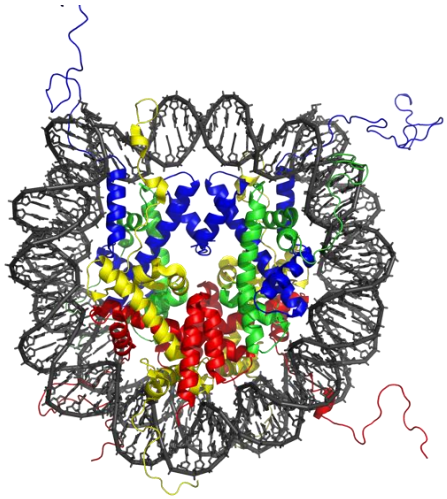


**Figure 1:** Diagram of a nucleosome core particle. DNA (black) is wrapped around a histone octamer, consisting of an H3 histone (blue) and H4 histone (green) tetramer and two H2A (yellow) - H2B (red) dimers (Davey *et al.*, 2002).

Nucleosomes are dynamic structures and continually cycle between wrapped and unwrapped states at the entry/exit site. This occurs several times per second and is termed nucleosome site exposure or nucleosome breathing (Polach and Widom, 1995; Li and Widom, 2004; Li *et al.*, 2005; Tims *et al.*, 2011). Site exposure provides access to DNA binding sites, allowing TFs and DNA repair complexes to bind on DNA otherwise concealed by nucleosomes. Histone PTM and DNA sequence composition influence the equilibrium and timescale of site exposure (North *et al.*, 2012). DNA unwrapping and site exposure are highly affected by PTM, especially on H3 and H2A histone tails, as these are in direct contact with DNA fragment ends (Figure 1) (Bowman and Poirier, 2015).

### 2.1.1 Nucleosomal properties

Nucleosomes vary in their properties, such as accessibility, structure, and stability. Nucleosome stability, i.e., the number and strength of histone-histone and histone–DNA interactions, is affected by histone variants, altered nucleosomal structure, underlying DNA sequence (Culkin *et al.*, 2017), histone post-translational modifications, and other factors associated with chromatin (Anderson, Lowary and Widom, 2001; North *et al.*, 2012).

Histone variants are highly conserved between species and have evolutionary evolved to perform specialized functions. Canonical histones are replaced with non-canonical histone variants in a replication-independent manner, which changes the composition of chromatin. For instance, histone variant H3.3 is deposited at actively transcribed genes which trigger changes in chromatin properties and enable further epigenetic modifications. Another H3 histone variant is the CENP-A histone, located at centromeres and represents the foundation of kinetochore assembly. Several H2A histone variants can also regulate chromatin, for example, H2A.X, H2A.Z, macroH2A, and H2A.Bbd. They have very diverse functions, from DNA double-stranded break repair (H2A.X), establishing transcriptional competence (H2A.Z), counteracting heterochromatic silencing (H2A.Z), hindering transcription

(macroH2A), and facilitating transcription (H2A.Bbd). The location of histone variants on chromatin correlates with their function, for example, H2A.Bbd is depleted at the inactivated X chromosome, while macroH2A is enriched (Henikoff and Smith, 2015). Intriguingly, histone variants, such as CENP-A and H2A.Bbd, cause weaker DNA-histone interactions at the entry-exit site of the octamer, in turn affecting DNA fragment size upon MNase digestion (Bao *et al.*, 2004; Gautier *et al.*, 2004; Tachiwana *et al.*, 2011). While macroH2A and H2A.Z more stably wrap DNA (Chakravarthy, Patel and Bowman, 2012) and are able to compact chromatin fibers more readily than the canonical H2A (Fan *et al.*, 2002), respectively.

Histone post-translational modifications change the properties of nucleosomes and contribute to gene regulation. They permit the binding of regulatory factors, which have specialized domains that recognize the modification (such as bromodomains recognizing histone acetylation) (Bowman and Poirier, 2015). Furthermore, PTMs at the histone tail region help stabilize chromatin higher-order structure (Pepenella, Murphy and Hayes, 2014). A single PTM can substantially reduce the free energy of nucleosome formation, which increases the probability that a nucleosome will fluctuate into an altered state, and in turn, modulate DNA accessibility (Bowman and Poirier, 2015). The most common PTMs are acetylation, phosphorylation, and methylation, with ubiquitination and ribosylation as the least common. Intriguingly several PTMs cause nucleosome destabilizations, such as H3K64me3/ac, H3K115ac, H3T118ph, K122ac, or disruption of histone-histone interactions, such as H4K91ac and H4R92me (Bowman and Poirier, 2015). Furthermore, PTM also serves as targets for chromatin remodelers, among which are H3K4me3, H3K9ac and H3K36me2/3 modifications, which recruit CHD1, SWI/SNF and ISWIb chromatin remodelers, respectively (Bowman and Poirier, 2015). Chromatin remodelers are complexes that provide a mechanism of modifying chromatin by sliding histone octamers along the DNA (nucleosome sliding), changing the histone composition of a nucleosome (histone exchange), and disrupting (nucleosome eviction) or *de novo* assembling nucleosomes (Längst and Manelyte, 2015).

DNA methylation also contributes to nucleosomal properties, such as nucleosome stability and accessibility. It induces tighter DNA-histone interactions and contributes to the formation of repressive chromatin (Lee and Lee, 2012; Jimenez-Useche *et al.*, 2013; Li *et al.*, 2022). Nucleosome accessibility can also be modulated by nucleosome reshaping. For instance, reshaping the octamer core by Swi6 chromatin remodeler of *Schizosaccharomyces pombe* increases nucleosome accessibility and, in turn, increases multivalent interactions between nucleosomes. Reshaping of the octamer core thus facilitates the formation of heterochromatin (Sanulli *et al.*, 2019).

In summary, nucleosomal properties are modulated by histone variants, PTMs, chromatin remodelers, and DNA methylation. These together influence nucleosomal stability, accessibility, and structure.

## 2.2.2 Non-canonical structures

In addition to the aforementioned changes to the nucleosomal structure by PTMs, histone variants, and DNA methylation, several non-canonical nucleosome structures exist. Over the last decades, many of these have been studied *in vitro*, while their existence and function *in vivo* remain to be elusive.

The first non-canonical structure is the lexosome, also referred to as the split-nucleosome, which is a nucleosome particle containing all eight canonical histones split into two heterotetrameric complexes. Evidence for the lexosome first emerged in 1976 when Tsanev and Petrov observed a split-nucleosome structure under the electron microscope (Tsanev, R., and Petrov, 1976). Evidence was further substantiated by uncovering unusual accessibility of the typically buried H3 cysteine 110 (Lee and Garrard, 1991). Lexosomes have been associated with poised transcription and hyper-acetylation and are believed to facilitate transcription and enable easier polymerase access (Johnson, Sterner and Allfrey, 1987).

Second, the hemisome, also called the half-nucleosome, is a particle containing only one heterotypic tetramer and is formed by a deposition of a single H3/H4 and H2B/H2A dimers. They were shown to exist *in vivo* in *D. melanogaster* (Dalal *et al.*, 2007) and humans (Dimitriadis *et al.*, 2010) and were associated with the CenH3 histone variant (CENP-A in humans). The hemisomes containing cenH3 histones are un-stable due to their partially unwrapped left-handed structure and have been proposed to prevent the formation of neocentromeres on chromosome arms (Henikoff and Furuyama, 2012).

The third non-canonical nucleosome particle is the hexasome which differs from the canonical nucleosome by lacking one histone H2A/H2B dimer and was proposed to appear during transcription as an intermediate of nucleosome assembly (Arimura *et al.*, 2012). Moreover, several studies have provided evidence that hexasomes can form independent of transcription. For instance, hexasomes were shown to be located near TSSs, using ChIP-Exo (Rhee *et al.*, 2014). In *D. melanogaster,* hexasomes were present at the +1 nucleosome position of expressed genes and are proposed to promote RNA Pol-II-dependent gene transcription (Ramachandran, Ahmad and Henikoff, 2017). This can, in part, be explained by RNA pol-II dissociation of H2A-H2B dimer during transcription through a nucleosome (González and Palacián, 1989; Kireeva *et al.*, 2002). Further, proof of the existence of hexasomes in vivo is the differential remodeling of CHD1, where the remodeler shifts hexasomes unidirectionally, opposite to canonical nucleosomes that can be moved bidirectionally. This mechanism contributes to the packing of nucleosome arrays observed *in vivo* (Levendosky *et al.*, 2016).

In summary, several non-canonical nucleosome structures exist, which may affect DNA accessibility, gene regulation, and affect nucleosomal properties, such as accessibility and stability. Nonetheless, their existence *in vivo* has been challenged on many notions and has not been conclusively proven.

### 2.2.2 Differential sensitivity to MNase

The advancement of sequencing technology allowed for a genome-wide investigation of *in vivo* nucleosome positioning and nucleosomal properties. Nucleosomes have been studied in high throughput sequencing experiments with MNase-seq, ChIP-seq, ATAC-seq, DNase-seq, NOME-seq, MPE-seq, and chemical mapping (Barski *et al.*, 2007; Schones *et al.*, 2008; Kelly *et al.*, 2012; Buenrostro *et al.*, 2013; Ishii, Kadonaga and Ren, 2015; Voong *et al.*, 2016; Zhong *et al.*, 2016).

Micrococcal nuclease (MNase) digestion followed by high-throughput sequencing (MNase-seq) is the most widely used method for determining genome-wide nucleosome positions and structure. MNase is an endo-exonuclease that preferentially cuts nucleosome-free regions and has the ability to induce double-stranded breaks. Past studies have reported differential MNase sensitivity to the core nucleosomes (Längst *et al.*, 1997) by using multiple MNase titrations per sample to release all nucleosomes from the genome and provide information on global chromatin structure and accessibility (Mieczkowski *et al.*, 2016; Chereji, Bryson and Henikoff, 2019). Nucleosomes are differentially released from chromatin based on their properties, such as structure, stability, and accessibility.

Two computational pipelines have been published for investigating nucleosome accessibility and global positioning using MNase-seq data. The first pipeline is the MACC pipeline. It uses linear regression to define an MNase ACCessibility (MACC) score by fitting a slope over 300bp bins, using fragments counts of four MNase titrations (Mieczkowski *et al.*, 2016). The MACC score provides information on chromatin accessibility in a defined window. Conversely, a newer pipeline, qMNase-seq, uses six MNase titrations and additionally requires spike-ins (Chereji, Bryson and Henikoff, 2019). Multiple MNase titrations per sample represent a labor-intensive and expensive experiment, limiting its use.

In summary, several methods have been developed to investigate nucleosomal properties, such as PTMs, histone variants, global positioning, and accessibility. Nevertheless, the most versatile and widely used is the MNase-seq method. MNase digestion kinetics have been extensively studied *in vitro* and *in vivo*, enabling the development of bioinformatics pipelines, where nucleosomal properties can be studied on a global scale.

### 2.2.3 The fragile nucleosome

In recent years, several studies have shown that nucleosomes are far more dynamic than initially proposed and can take several transient states to either promote transcription (Gallego *et al.*, 2020), heterochromatin formation (Sanulli et al., 2019), or form phase-separated condensates (Keenen *et al.*, 2021). Particular interest among the non-canonical nucleosomes has been the fragile nucleosome, which remains to be conclusively proven *in vivo*, as the current techniques do not allow their direct detection.

The term fragile nucleosome was first coined in 1976 and was proposed to take a lexosome form as an intermediate during RNA pol-II transcription (Tsanev, R., and Petrov, 1976). The term has been loosely defined in the literature and may exemplify a hemisome (Rhee *et al.*, 2014), hexasome (Brahma and Henikoff, 2019), lexosome, or an octamer with non-canonical histone variants. The commonality between these structures is a sub-nucleosomal-sized particle in an un-stable state.

Several publications have shown these unusually short DNA fragments following MNase digestion by ChIP-exo in *S. cerevisiae* (Rhee *et al.*, 2014), MPE-ChIP-seq in mouse embryonic stem cells (Ishii, Kadonaga and Ren, 2015), chemical mapping in mouse ESCs (Voong *et al.*, 2016), CUT&RUN-ChIP in *S. cerevisiae* (Brahma and Henikoff, 2019), and MNase-seq titrations in *C. elegans* embryos (Jeffers and Lieb, 2017), *D. melanogaster* (Chereji *et al.*, 2016), human HeLa cell line (Schwartz *et al.*, 2019) and *S. cerevisiae* (Xi *et al.*, 2011). In addition to a loosely defined definition, there is also no standardized way of studying fragile nucleosomes in sequencing experiments, which leads to confusion and contradicting results. For instance, in *C. elegans*, the fragile nucleosome has been associated with inducible genes poised for the transcriptional response to developmental and environmental cues (Jeffers and Lieb, 2017). In contrast, they were associated with nucleosome eviction in budding yeast (Brahma and Henikoff, 2019).

In summary, despite the recognized importance of nucleosomes as regulators of gene expression and building blocks of chromatin, a clear definition of non-canonical nucleosomes and a defined way of analyzing them in high-throughput sequencing data is lacking.

## 2.3 RNA:DNA triplexes

Triple helices or triplexes are oligonucleotide molecules made of three nucleotide strands and were first discovered in 1957 (Felsenfeld and Rich, 1957). They form in a 5' to 3' directional nucleation-zipping model with reference to the polypurine strand, possibly due to the right-handed structure of the dsDNA (Alberti *et al.*, 2002). Although most studied triplexes consist of a dsDNA and an ssRNA molecule in its major groove, triplexes can also be formed as DNA:DNA-DNA and RNA:RNA-RNA structures, where ":" specifies Hoogsteen or reverse Hoogsteen bonds and "-" refers to Watson-Crick bonds. RNA:DNA-DNA triplexes are the most stable structures, with DNA:DNA-DNA having a slightly reduced stability and RNA:RNA-RNA the least stable structures (Figure 2) (Kunkler *et al.*, 2019).

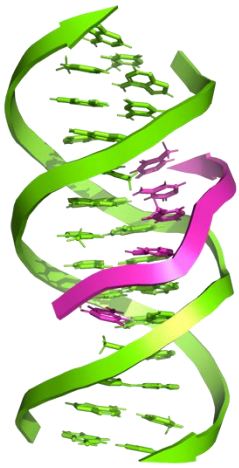


**Figure 2:** Triple helix structure. ssRNA/ssDNA molecule (pink) binds to the major groove of dsDNA (green) and wraps around it (source: Wikipedia).

The third strand interacts only with the purine strand of the duplex, and based on the sequence composition of the third strand, we differentiate three triplex motifs (Figure 3):

- Purine motif (R) consists of Adenine and Guanine bases, forming A:A-T and G:G-C base triplets.
- Pyrimidine motif (Y) consists of Cytosine and Thymine bases, forming C:G-C and T:A-T triplets.
- Mixed motif (M) consists of Guanine and Thymine bases (G:G-C and T:A-T).

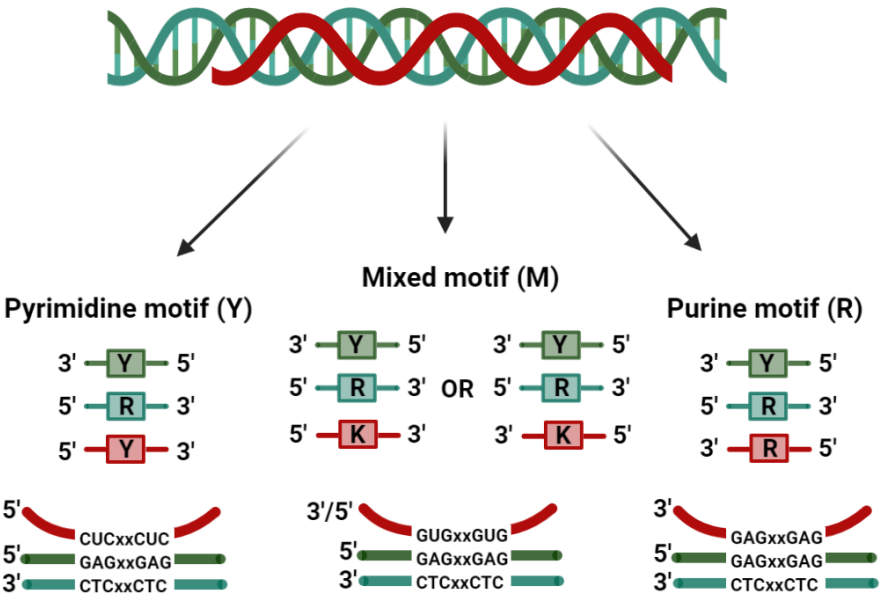



**Figure 3:** RNA:DNA triplexes can form three different motifs. (A) Pyrimidine motifs are formed in a parallel orientation, forming Hoogsteen bonds with the purine strand of the duplex. (B) Mixed motifs can form in both parallel and anti-parallel orientations. (C) Purine motifs are formed in an anti-parallel orientation, forming reverse Hoogsteen bonds. Y = C/U/T; R = G/A; K = G/U/T (Figure made with BioRender).

### 2.3.1 Biological significance

The formation of triplex structures *in vitro* poses the question of their potential functional role *in vivo*. Triplexes have been long used in biotechnology to modulate gene transcription *in vivo* via TFO-targeting of desired regions (Thuong and Hélène, 1993). Moreover, computational predictions indicate prospective TrTSs are enriched at promoter and enhancer regions of Eukaryotic genomes (Goñi et al., 2004; Goñi et al., 2006; Wu et al., 2007), while chromatin-associated RNAs (chRNA) contain a significantly higher amount of potential purine (GA) and mixed motif (GU) triplex-forming sequences, than other RNA molecules (Buske *et al.*, 2012). Together the data suggest a potential role in the modulation of transcription.

In general, triplexes can work in *cis* or *trans*, alone or as anchors for proteins, and can activate or repress gene transcription (Bacolla, Wang and Vasquez, 2015). Specifically, several mechanisms of triplex-mediated gene regulation are conceivable: They can form directly at the TF promoter DNA motif and, with that, prevent TF binding and, in turn, transcription initiation. Conversely, promoter occlusion can also lead to transcription initiation, not only repression, by preventing transcriptional repressors from binding. Triplexes may also have a role in protein recruitment by directly binding to proteins involved in RNA syntheses, such as TFs, or proteins modulating their activity. For instance, a lncRNA may bind to a specific protein on one end and form a sequence-specific triplex on the target site on the other, bringing factors to their target site. This mechanism allows for both transcription initiation and repression, depending on the recruited factor. Alternatively, triplexes may form on RNA transcripts, leading to RNase H-induced transcript degradation (van Dyke MW, 2005; Buske, Mattick and Bailey, 2011).

Importantly, triplexes can form within the same nucleic acid as DNA:DNA-DNA triplexes or RNA:RNA-RNA triplexes, termed intramolecular triplexes, or with three separate nucleic acids, termed intermolecular triplexes. Additionally, to the above-mentioned mechanisms of gene regulation, intramolecular triplexes may also modulate transcription by formation proximal to the TSS or in the gene body. Intramolecular triplexes proximal to the TSS may be the result of increased negative superhelical tensions due to nearby transcription. Such triplexes may then inhibit subsequent transcription by protein displacement or recruitment of repressive proteins. Downstream intramolecular triplexes may arise from locally denatured DNA templates caused by transcription or replication. Here triplex formation would hinder transcription elongation (van Dyke MW, 2005; Buske, Mattick and Bailey, 2011).

Helicases of the RecQ family have been shown to specifically bind triplexes and actively unwind them in a 3' to 5' direction (Maine and Kodadek, 1994). Other helicases, such as DHX9, have also been shown

to unwind triplex structures; however, in contrast to RecQ helicases, they require a 3'single-stranded overhang of the third strand (Jain *et al.*, 2010).

In summary, substantiating albeit indirect evidence suggests triplexes may regulate gene expression *in vivo*. However, direct proof of their *in vivo* assembly is still lacking.

### 2.3.2 Non-coding RNAs

Non-coding RNAs consist of several RNA classes: transfer RNAs (tRNA), ribosomal RNA (rRNA), small nucleolar RNA (snoRNA), small nuclear RNA (snRNA), microRNA (miRNA), enhancer-associated RNA (eRNA) and long non-coding RNA (lncRNA) (Li and Fu, 2019). Non-coding RNAs can be associated directly (R-loops and RNA:DNA triplexes) or indirectly (via protein-binding partners) with chromatin, termed chromatin-associated RNAs. The largest class of chRNAs are the newly transcribed RNAs (nascent RNAs). However, other RNAs have also been shown to be associated with chromatin, such as snoRNAs, miRNAs and lncRNAs (Schubert *et al.*, 2012). ChRNAs were shown to have a vital role in higher-order chromatin structure formation, and their removal with RNaseA resulted in chromatin compaction, demonstrating their role in chromatin accessibility modulation (Schubert *et al.*, 2012).

Long non-coding RNAs (lncRNA) play an important role in regulating cell type-specific gene expression. They exert their function by direct (formation of RNA:DNA triplexes and R-loops) and indirect (binding to proteins) interactions with chromatin and recruitment of regulatory proteins. Among the direct chromatin interaction, several lncRNAs have been predicted to form RNA:DNA triplexes, whereas only a handful have been verified experimentally (Li, Syed and Sugiyama, 2016; Greifenstein, Jo and Bierhoff, 2021). For instance, lncRNA KHPS1 has been shown to form a triplex on a poised enhancer, where it recruits E2F1 and p300 and subsequently activates transcription of a proto-oncogene SPHK1, from which it is transcribed (Postepska-Igielska *et al.*, 2015).

Computationally predicted TrTSs were shown to be enriched in the boundaries of topologically associated domains (TAD) (Soibam and Zhamangaraeva, 2021), telomeres, poised promoters, and polycomb repressed regions (Farabella *et al.*, 2021). A clear distinction can be seen for TrTSs of miRNAs and lncRNAs, where lncRNA-TrTS are enriched at poised chromatin regions, whereas miRNA-TrTS are preferentially located at active chromatin regions (Paugh et al., 2016).

In summary, evidence suggests triplex formation may be a mode for lncRNAs to recruit chromatin modifiers (Mondal *et al.*, 2015), induce transcriptional repression (Kalwa et al., 2016), and aid in chromatin organization (Farabella *et al.*, 2021). Nonetheless, only a handful of lncRNAs have been shown to form triplexes *in vivo*.

### 2.3.3 Triplex stability

The triplex formation is a slow process (Rougée *et al.*, 1992), but when formed, triplexes are very stable. They experience a half-life of a few hours to several days (James, Brown and Fox, 2003). Triplex stability is affected by several different factors, such as sequence composition (mainly Guanine content), mismatches, length, pH, and salt concentration (Duca *et al.*, 2008; Buske, Mattick and Bailey, 2011). *In vivo*, stability is additionally affected by their proximity to nucleosomes, where the interaction with the H3 N-terminal tail increases their stability (Maldonado *et al.*, 2019). Other proteins have also been shown to stabilize triplex structures (Buske, Mattick and Bailey, 2011), such as argonaut proteins (Toscano-Garibay and Aquino-Jarquin, 2014). Furthermore, DNA methylation increases the thermal stability of intramolecular DNA:DNA triplexes (Carr *et al.*, 2018).

The combination of three negatively charged strands creates a charge repulsion, which must be mitigated by positively charged proteins. Interestingly, triplexes can also be stabilized by polyamines, present in sufficient amounts in cells to mitigate charge repulsion effects under physiological conditions (Kim *et al.*, 2002).

*In vitro* studies have shown that the pyrimidine motif is additionally constrained by Cytosine protonation at the N3 position, which favorably occurs in low pH /acidic conditions (Morgan and Ills, 1968; de los Santos, Rosen and Patel, 1989). In spite of this, it was demonstrated for i-motifs, where similar to the C:G-C$^+$triplex, a hemi protonated cytosine is required to form a C:G-C$^+$ triplet, that such a reaction also occurs in physiological conditions (Zeraati *et al.*, 2018). This notion was also challenged by the aforementioned KHPS1 lncRNA, which contains a CU motif and has been extensively shown to form a triplex *in vitro* with 10 mM $Mg^{2+}$ buffer conditions and pH 7.5 (Postepska-Igielska *et al.*, 2015). Divalent cations such as spermine and $Mg^{2+}$ and molecular crowding stabilize triplex formation and increase their melting temperature (Sugimoto *et al.*, 2001; Wu *et al.*, 2002; Chen and Chen, 2011; Li, Syed and Sugiyama, 2016).

*In vitro,* triplexes are also stabilized by Thiazole Orange (TO), a cyanine dye. TO binding to a triplex structure induces a stock shift in the absorption spectrum, resulting in a >1000-fold increase in the fluorescence signal, enabling their detection (Lubitz, Zikich and Kotlyar, 2010) and isolation (Maldonado *et al.*, 2019).

Most studies on triplex stability have been performed on DNA:DNA triplexes due to the high costs and complexity of RNA oligo synthesis. Besides, it was recently shown that significant differences exist between DNA:DNA, RNA:DNA and RNA:RNA triplex stability (Maldonado *et al.*, 2017; Kunkler *et al.*, 2019), demonstrating additional studies on RNA:DNA triplexes are required. In addition to the already

mentioned triplexes and R-loops, several other non-canonical DNA structures exist, such as G-quadruplexes, i-motifs, hairpins, and cruciforms. It is possible that different DNA structures compete with each other and are favored in certain conditions. In summary, several factors influence triplex stability, among which are Guanine content, mismatches, length, pH, and salt concentration. Additionally, triplex structures are stabilized *in vivo* via nucleosomes, proteins, and polyamines and via TO *in vitro*.

### 2.3.4 *In vivo* triplex assembly

Several antibodies were tested for specificity to triplexes, but none of them had a high specificity and thus a wide use was not adopted (Agazie, Lee and Burkholder, 1994). Instead, TO has been used for triplex stabilization and detection (Gorab and Pearson, 2018). Two methods for detecting triplexes *in vivo* have been published in recent years (Maldonado *et al.*, 2019; Sentürk Cetin *et al.*, 2019). Both methods take advantage of differential RNase H digestion of R-loops and triplexes, with R-loops being digested and triplexes being protected from digestion (Figure 4). Additionally, triplexes protect chromatin from DNase I digestion, enabling sequencing of protected regions, the premises for one of the methods, TrIP-seq (Maldonado *et al.*, 2019). Here TO was used to stabilize triplexes to ensure that the washing steps do not disrupt them. The drawbacks of this method are that only TrTS sites are detected and that the resolution of those sites is restricted by the DNaseI cutting sites.

The second method detects both TFOs (TriplexRNA-seq) and TrTSs (TriplexDNA-seq) but not in pairs. The method uses a pulldown approach with either an anti-DNA antibody to detect TFOs or streptavidin to detect TrTSs by ligation of a biotin linker oligos to RNA. No stabilization agent is used, possibly destabilizing some of the less stable triplexes. Not surprisingly, the methods show vastly different results. TrIP-seq mainly uncovered TrTS enriched with putative miRNA-TrTs, whereas the TriplexDNA-seq method mainly pulled down enhancer regions as putative TrTSs (Maldonado *et al.*, 2019; Sentürk Cetin *et al.*, 2019). Conversely, the TriplexRNA-seq method revealed mRNAs and lncRNAs as part of the triplex complexes, indicating trans-acting roles for enhancer RNAs (Sentürk Cetin *et al.*, 2019). Comparing R-loop data and TriplexDNA-seq, the authors found a 20% overlap. They interpret this as the triplex structures being more widespread than initially thought. Another interpretation could also be that due to incomplete RNaseH digestion and lack of triplex stabilization, indeed, a mix of R-loops and strong triplexes were sequenced. One may speculate miRNAs form less stable triplexes and are thus only revealed with the TrIP-seq method.

In summary, two methods for detecting triplex structures *in vivo* have been developed. However, neither of them detects triplex pairs, only TrTs or TFO.

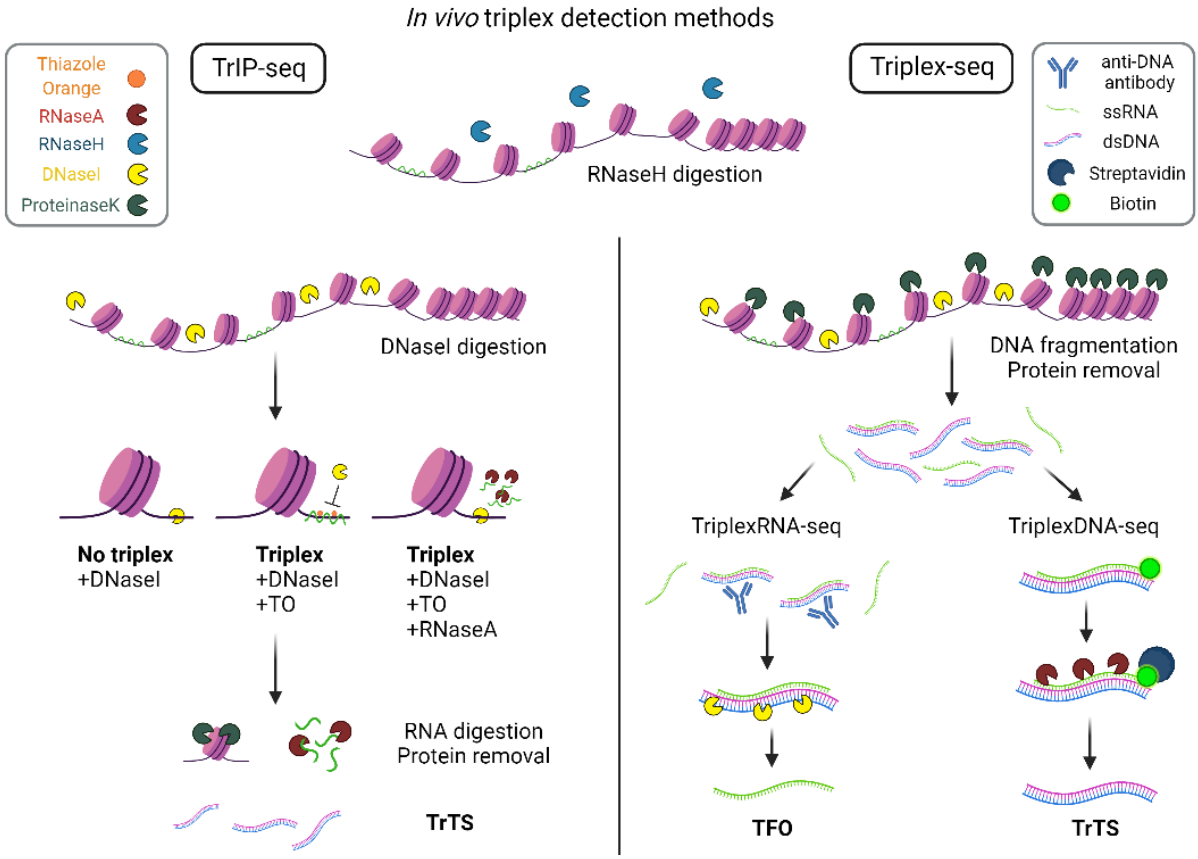



**Figure 4:** Comparison between the two triplex detection methods, TrIP-seq and TriplexDNA/RNA-seq. The commonality between the methods is RNaseH digestion to remove R-loops. TrIP-seq uses Thiazole Orange to stabilize triplex structures and differential DNaseI digestion to detect Triplex targeting sites on the DNA (TrTs). Whereas Triplex-seq takes a different approach. Here no stabilization agent is used. Chromatin is firstly fragmented, followed by protein/nucleosome removal. From here, samples are separated into two parts. To isolate TrTSs, RNA molecules are ligated to biotin-oligos and pulled down with streptavidin. RNA in the triplex structure is removed with RNaseA digestion. To isolate triplex-forming oligos (TFO), triplex structures are first removed from ssRNA molecules with anti-DNA antibody pulldown, followed by DNaseI digestion to remove the DNA from triplex structures (Figure made with BioRender).

### 2.3.5 Computational predictions

Several computational tools have been developed to predict and better understand triplex formation; Triplexator, triplexes, LongTarget, trident, TDF, and TTS mapping, with Triplexator being the most widely used (Jenjaroenpun and Kuznetsov, 2009; Buske *et al.*, 2012, 2013; Hon *et al.*, 2013; He *et al.*, 2015; Paugh, Coss, Bao, Laudermilk, Grace, Ferreira, Waddell, Ridout, Naeve, Leuze, LoCascio, Panetta, Wilkinson, C. H. Pui, *et al.*, 2016; C.-C. Kuo *et al.*, 2019). Even so, all prediction tools take only Hoogsteen base-pairing rules into consideration because a sufficient amount of biochemical data is not yet available. The users must set the parameters themselves, such as length, mismatch percentage and Guanine percentage. Which parameters are sensible is unclear as the triplex binding rules have not been fully defined. Several attempts were made to define these rules from available sequencing data (Zhang, Long and Kwoh, 2020); however, the machine learning algorithms were trained on

inadequate data, such as ChOP-seq, which detects all lncRNA interactions with chromatin, direct (R-loops, triplexes) or indirect (via protein-binding partners) (Mondal *et al.*, 2015; C. C. Kuo *et al.*, 2019). Additionally, the method uses crosslinking to stabilize lncRNAs binding to chromatin. Conversely, crosslinking does not stabilize RNA–DNA interactions and may even disrupt triplex formation. Consequent washing steps may also disrupt triplex formation unless they are stabilized by proteins (Postepska-Igielska et al., 2015). Therefore, it is unclear how such tools improve triplex binding prediction.

Computational predictions revealed enrichment of GA/CU/GU motifs in chromatin-associated small RNAs, with GU (mixed) motif as the most prevalent. On the contrary, one study proposed only DNA-DNA and not RNA:DNA triplexes with GU motifs can form in physiological conditions (Semerad and James maher, 1994). Additionally, mixed DNA:DNA triplexes were suggested to form only in anti-parallel orientation by another study (Beal and Dervan, 1991). In turn, not many studies focused on the mixed triplex motif, clearly underlying the need for further investigation. It remains unclear whether mixed triplex motifs differ in stability and if their formation is feasible under physiological conditions. Since mixed motif triplexes can form in both parallel and anti-parallel orientation, it also remains to be tested whether the binding rules differ between the orientations and if they differ between DNA:DNA and RNA:DNA triplexes. In summary, there is a clear lack of understanding of the triplex binding code *in vitro* and *in vivo*, underlying the need for additional studies to define it.

# 3 OBJECTIVES

Eukaryotic cells are regulated at the epigenetic, transcriptional, post-transcriptional, translational, and post-translational levels. Understanding the complex system of gene regulation is fundamental for disease treatment, where gene regulation is misregulated.

## 3.1 Nucleosome stability

Nucleosomes are essential for DNA compaction, regulating accessibility to regulatory machinery, replication, and recombination in a spatiotemporal manner. Base-pair resolution of nucleosome positioning is thus necessary to investigate gene regulation. Furthermore, non-canonical nucleosome structures create another layer of gene expression fine-tuning. Nevertheless, studying non-canonical structures at a genome-wide level and in the context of gene regulation has been neglected. The MNase-seq method has been wide-spread adopted as the gold-standard method for investigating nucleosome positions, yet current MNase-seq protocols and bioinformatics pipelines are hard to use, not reproducible, and require high labor and costs.

Additionally, up to now, no pipeline exists to study non-canonical nucleosome groups and nucleosome properties such as nucleosome stability and accessibility in an automated and genome-wide manner. This study aims to develop an automated and reproducible pipeline with an automated approach to detecting special nucleosome groups, such as un-stable nucleosomes and non-canonical nucleosome structures. I aimed at characterizing nucleosomes with un-typical properties and finding their function. Additionally, the pipeline should provide nucleosome positions with a base-pair resolution and accessibility and stability scores for each nucleosome. *D. melanogaster* is the ideal model for studying nucleosome stability and developing a pipeline due to its small genome size and availability of high-quality sequencing data with high sequencing depth. Furthermore, histones are highly conserved in eukaryotes, and as such, nucleosome properties are universal. Finally, I aim to provide recommendations for the minimum sequencing depth and the number of MNase titrations to reduce labor and costs.

## 3.2 The triplex code

DNA can form several non-canonical structures to regulate itself, such as R-loops, G-quadruplexes, and RNA:DNA triplexes. The latter has received more attention in recent years, yet the triplex binding code is not fully understood. Triple helix structures form via sequence-specific base-pairing and form Hoogsteen or reverse Hoogsteen hydrogen bonds. Several bioinformatic tools have been developed,

predicting triplex formation solely based on Hoogsteen base-paring rules. Yet such predictions are not capturing the complexity of triplex formation, with many predicted triplex pairs not forming a triplex under physiological conditions.

To further add to the complexity, most studies on triplex formation have been performed on DNA:DNA triplexes; however, we now know the triplex code is different for DNA:DNA and RNA:DNA triplexes, underlying the need to investigate the triplex binding code in a thorough manner. Both RNA:DNA and DNA:DNA triplexes have been suggested to form *in vivo*, with the latter forming intramolecular triplexes and RNA:DNA triplexes forming intermolecular triplexes. This study aims to complement the current knowledge on triplex formation by examining each motif (purine, pyrimidine, and mixed) separately and evaluating the differences between RNA:DNA and DNA:DNA triplexes.

# 4 NUCLEOSOME STABILITY

## 4.1 Results

The definitions for un-typical nucleosomes have been loosely defined, hindering progress in the field. To distinguish between different nucleosome populations, I use the following definitions (Figure 5):

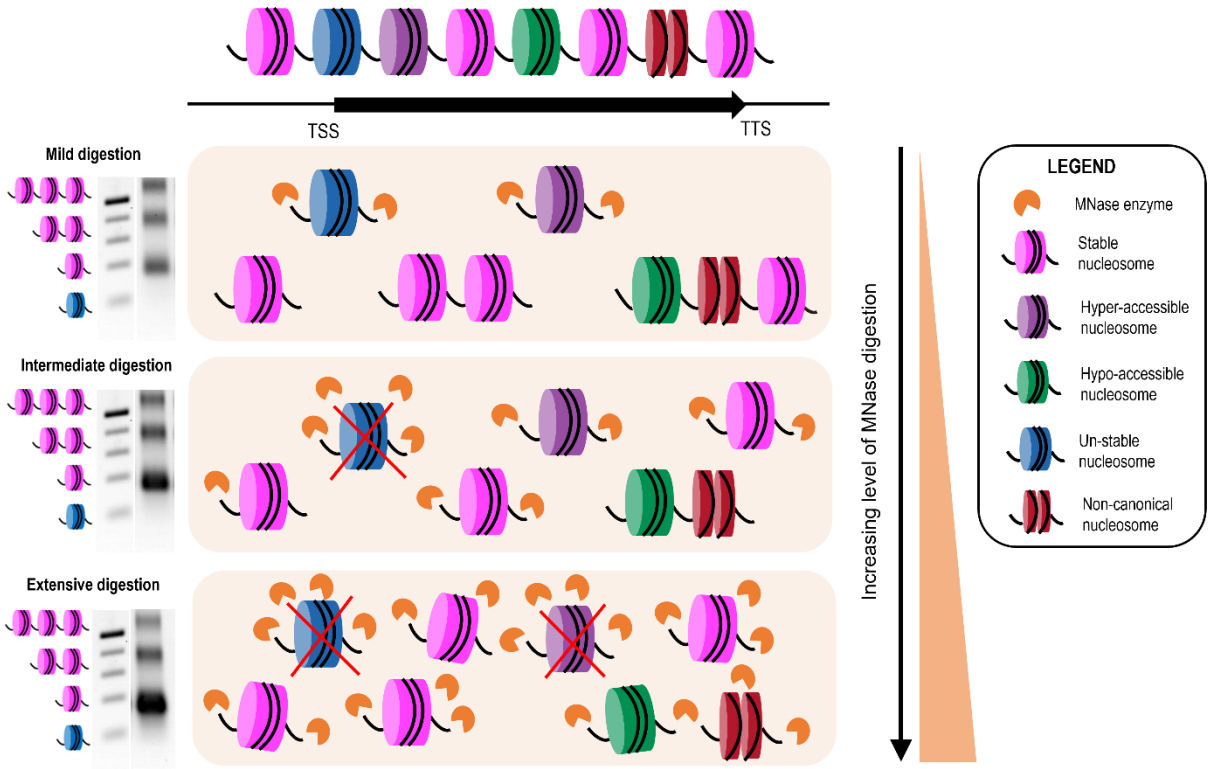



**Figure 5:** Release of nucleosomes from chromatin is dependent on the used MNase concentrations. The differential accessibility of cellular DNA to MNase, a 16.9 kDa endonuclease, is due to chromatin compaction or crowding, DNA sequence composition, differential nucleosome structure, and DNA linker accessibility. Under mild MNase conditions, hyper-accessible nucleosomes or nucleosomes with reduced stability are preferentially released. Under intermediate digestion, stable nucleosomes will be released, while un-stable nucleosomes will already get over-digested. Under extensive digestion, hypo-accessible or non-canonical nucleosomes will be released, while hyper-accessible nucleosomes will be over-digested. MNase is a processive enzyme; therefore, the digestion kinetics will result in the progressive trimming of nucleosomal particles, and even stable nucleosomes will eventually get over-digested at prolonged, extensive MNase digestion.

- **Stable nucleosomes:** Canonical nucleosomes with average stability, released at intermediate digestion levels. Present in the mono-nucleosomal fraction in the low and intermediate digestions and might be over-digested (present in the sub-nucleosomal fraction) at higher MNase concentrations (Figure 6).

**- Fragile/Un-stable nucleosomes:** Nucleosomes with reduced stability. MNase digestion at low MNase concentration leads to cleavage within the realm of the nucleosomal DNA and the release of un-stable nucleosomes. They are fully hydrolyzed at high MNase digestion conditions, resulting in a decrease of occupancy with increasing MNase concentration. They correspond to DNA fragments in the sub-nucleosomal fraction and at low MNase digestion (Figure 6).

**- Non-canonical nucleosomes:** This is a diverse group of nucleosomes with either a non-canonical nucleosome structure (hemisome – a nucleosome particle containing one H3/H4 dimer and one H2A/H2B dimer (Lavelle and Prunell, 2007); lexosome – a nucleosome particle containing one H3/H4 tetramer and two H2A/H2B dimer, split in half (Tsanev, R., and Petrov, 1976); hexasome - a nucleosome particle containing one H3/H4 tetramer and one H2A/H2B dimer (Hutcheon, Dixon and Levy-Wilson, 1980)) or histone variants. This group represents structurally altered nucleosomes with additional MNase cleavage sites within the realm of the nucleosome, generating a sub-nucleosomal fraction that also remains stable at higher MNase concentrations (Figure 6). Over-digested nucleosomes cannot be distinguished from nucleosomes with non-canonical structures or histone variants without additional experimental validation (For example, using ChIP-seq or MNase-ChIP-seq data on histone variants).

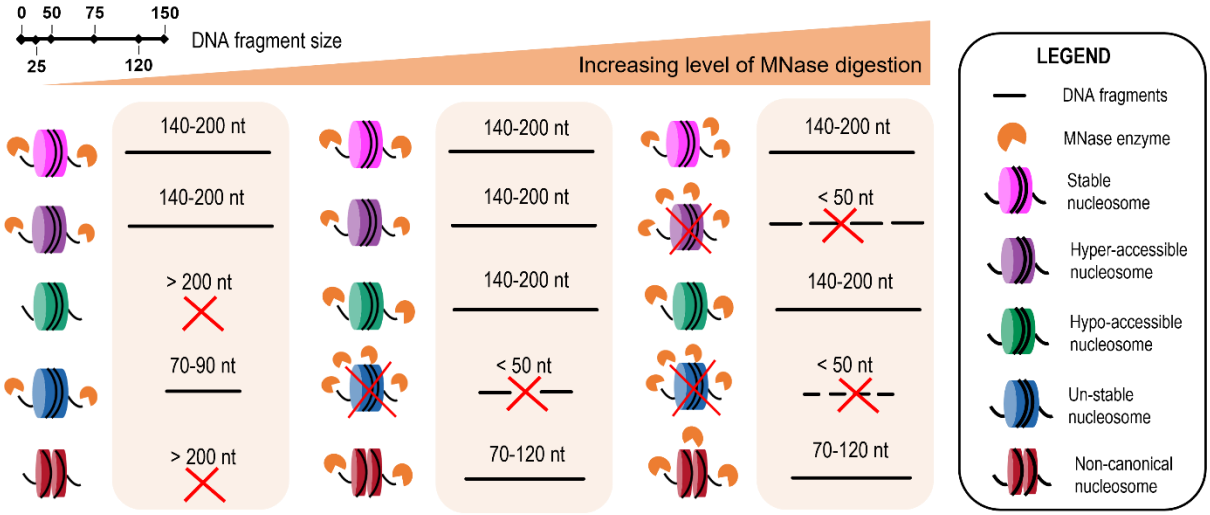



**Figure 6:** MNase kinetics reflect in different DNA fragment sizes of nucleosome groups and are dependent on MNase concentration.

**- Hyper-accessible nucleosomes:** Nucleosomes with increased accessibility to regulatory factors, such as TFs and transcriptional machinery. They are released at lower MNase concentrations as full-sized nucleosomal DNA fragments. They exhibit a regular and stable nucleosomal structure but reside in regions of reduced chromatin compaction, revealing the high accessibility of their DNA linkers for the endonuclease. These nucleosomes are present in the mono-nucleosomal fraction and have a decreased occupancy with higher digestion levels (Figure 6).

**- Hypo-accessible nucleosomes:** Nucleosomes with reduced accessibility to regulatory factors and transcriptional machinery. Nucleosomes residing in compacted chromatin regions correlate with reduced DNA linker accessibility. These nucleosomes are released at higher MNase concentrations in the mono-nucleosomal fraction and exhibit a higher occupancy with increasing MNase concentrations (Figure 6).

### 4.1.1 The MACC score

To establish a bioinformatics pipeline, I used *D. melanogaster* as a model system. Several publicly available datasets were used in this study, chosen based on the following criteria: At least two MNase titrations per sample, S2 cell line, sufficient sequencing depth, histone immunoprecipitation, and availability of at least two replicates per sample. The S2 cell line is derived from a primary culture of late-stage (20-24 hours) *Drosophila melanogaster* embryos originating from a macrophage-like cell type (Schneider, 1972), and it is the most widely studied cell line of the fruit fly.

The primary datasets used in this study come from the Tolstorukov and Kingston labs with H3 or H4 immunoprecipitation after MNase digestion, followed by high throughput sequencing. Both datasets contain four MNase concentration titrations (1.5U, 6.25U, 25U, and 100U) per sample and two replicates each (Mieczkowski *et al.*, 2016; Mueller *et al.*, 2017). In the original paper, the authors define an MNase accessibility (MACC) score and study genome-wide accessibility in 300bp non-overlapping bins. MACC score quantifies nucleosome accessibility by taking advantage of differential MNase nucleosome digestion (Mieczkowski *et al.*, 2016). They compare MNase-ChIP-seq and MNase-seq data and define two distinct groups. The first group consists of hyper-accessible bins (high MACC score) enriched in DNase hyper-accessible regions. In comparison, the second group consists of hypo-accessible bins (low MACC score) depleted in DNase hyper-accessible regions.

I started by using the MACC R package and replicated the main findings of the Tolstorukov and Kingston labs. Authors filter fragments below 50 nt and above 500 nt, ending up with sub-, mono- and di- nucleosome-sized fragments. I wondered if using only the mono-nucleosomal-sized fragments would significantly change the MACC score. I found that the MACC scores of fragment size 50 - 500 nt and 140 – 200 nt had a high correlation of 0.8, which indicated the main drivers of the MACC score might be the canonical mono-nucleosomes (Figure 7, A). Next, I compared the MACC score of mono- and sub-nucleosome-sized fragments. Interestingly I observed a peculiar group of sub-nucleosomes with a high MACC score (Figure 7, B – marked in yellow).

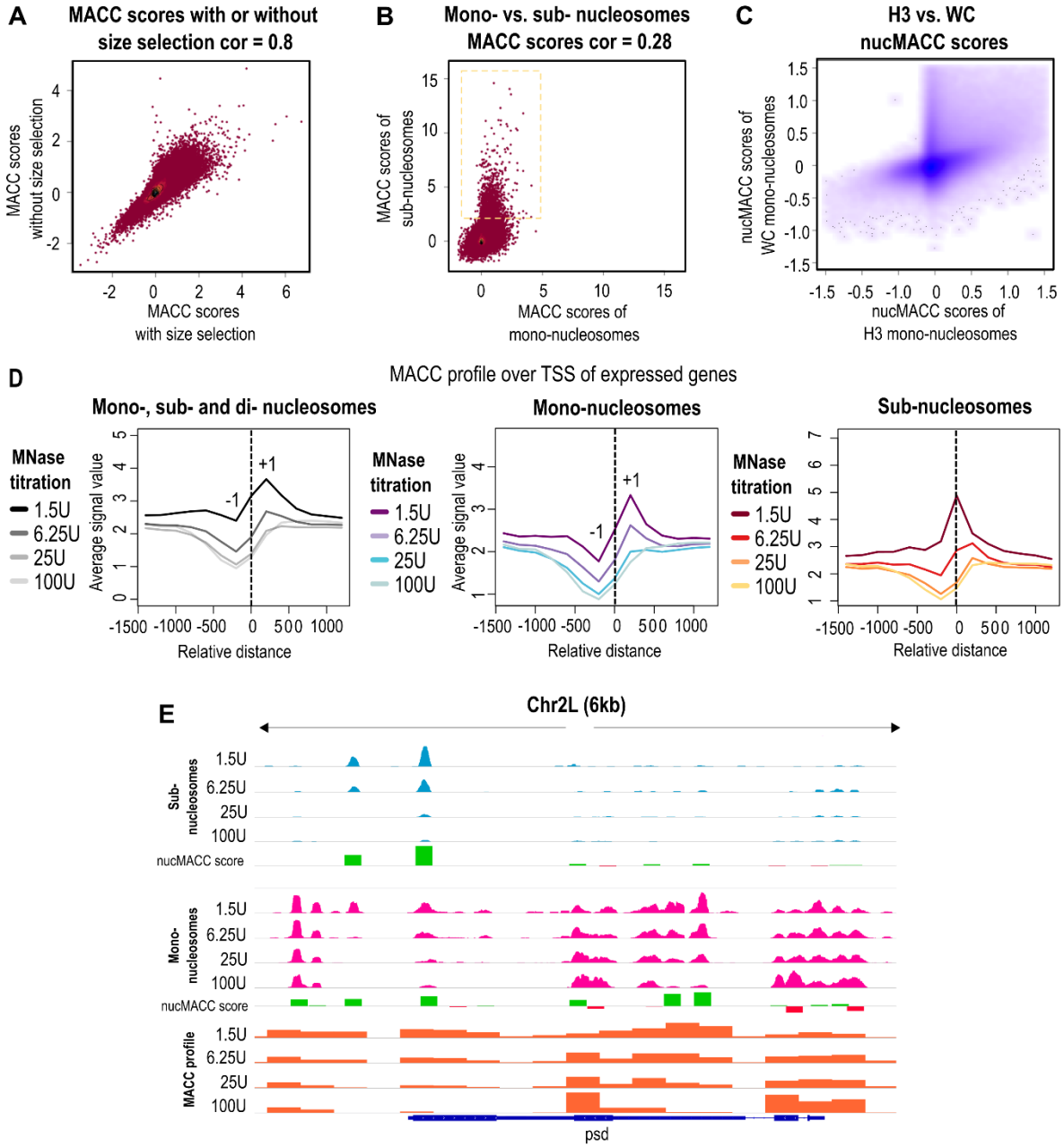



**Figure 7:** Comparing MACC and nucMACC scores. (A) Comparing MACC scores with (length 140-200 nt) or without size selection. (B) Comparing MACC scores of mono- and sub- nucleosomal-sized fragments. (C) Replicating the main finding from Mieczkowski et el., with nucMACC scores. (D) MACC profile over TSS of expressed genes, without size selection or with mono- and sub- nucleosomal size selection. (E) A selected region shows the difference between the MACC and nucMACC profiles.

Ultimately, I wanted to increase the resolution of nucleosome positions. I added a nucleosome position calling step before calculating the MACC score to obtain a nucleosome accessibility score, herein referred to as the nucMACC score. In order to confirm that the nucMACC score indeed gives an accessibility score, I replicated the main finding of the Mieczkowski *et al.* paper, where a clear distinction between the whole chromatin MNase-seq and anti-H3 ChIP-MNase seq can be observed (Figure 7, C). Looking at the MACC profile at the TSS of expressed genes, I observed a clear increase in

accessibility at the +1 nucleosomes and a decrease in the -1 nucleosomes (Figure 7, D – left). Intriguingly, the MACC profile for mono- and sub- nucleosomes revealed two distinct populations. On the one hand, the MACC profile of mono-nucleosomes matched the overall MACC profile, albeit with a slightly higher resolution (Figure 7, D – middle). On the other hand, the sub-nucleosomal MACC profile displayed an increase in accessibility at the 0 positions, where usually no nucleosomes reside (Figure 7, D – right). Lastly, I compared the MACC and nucMACC profiles and could visually confirm a higher resolution of the accessibility score with two distinct populations, mono- and sub- nucleosomes (Figure 7, E).

### 4.1.2 The nucMACC score

Having defined the nucMACC score, I went on to optimize the score calling and characterize the mono- and sub- nucleosomal fractions. Looking at the fragment sizes of MNase-seq data, I observed the disappearance of di-nucleosomes and increasing levels of mono- and sub- nucleosomes with increasing MNase digestion (Figure 8, B). Based on the shape of the curve of the fragment size distribution, I selected the size ranges corresponding to mono-nucleosomal DNA (140 – 200 base pairs, Figure 8, B – purple fraction and Figure 8, A) and sub-nucleosomal DNA (50 – 139 base pairs, Figure 8, B – pink fraction and Figure 8, A).

When plotting the two fractions relative to the transcription start site (TSS), I observed striking differences in the location of the mono-nucleosomal and sub-nucleosomal fractions on transcribed genes (Figure 8, C-D). Sub-nucleosomes show the highest coverage at the nucleosome depleted region (NDR) upstream of the TSS at very low MNase concentrations. The occupancy is increasingly lost with higher MNase concentrations (Figure 8, C). Since the data result from anti-histone ChIP-seq after MNase fragmentation, it is evident that very un-stable histone-DNA complexes cover the promoter regions of active genes. This is evident in the histone H4 ChIP (Figure 8, C) and the H3 ChIP experiments (Figure 9, C) and could already be observed in the MACC profile (Figure 7, D), albeit at a very low resolution.

The mono-nucleosomal DNA fraction, in contrast, has a distinct MNase concentration-dependent TSS profile over expressed genes. A precisely positioned mono nucleosome is located at the so-called +1 site, exhibiting the highest occupancy at low MNase concentrations (Figure 8, B). Not surprisingly, I find the +1 nucleosomes of expressed genes in a hyper-accessible configuration, meaning the fragment count decreases progressively with increasing MNase concentrations. In contrast, the fragment count does not change significantly for the nucleosomes +2, +3, +4, and subsequent nucleosomes (Figure 8, B and Figure 9, B). The binding of the initiation complex and ongoing transcription renders the +1 nucleosomes into an open conformation that allows preferential DNA

linker cleavage by MNase. I can observe this feature on a per gene basis (Figure 8, D), where a hyper-accessible +1 nucleosome and a hypo-accessible +2, +3, +4 nucleosome can be observed.

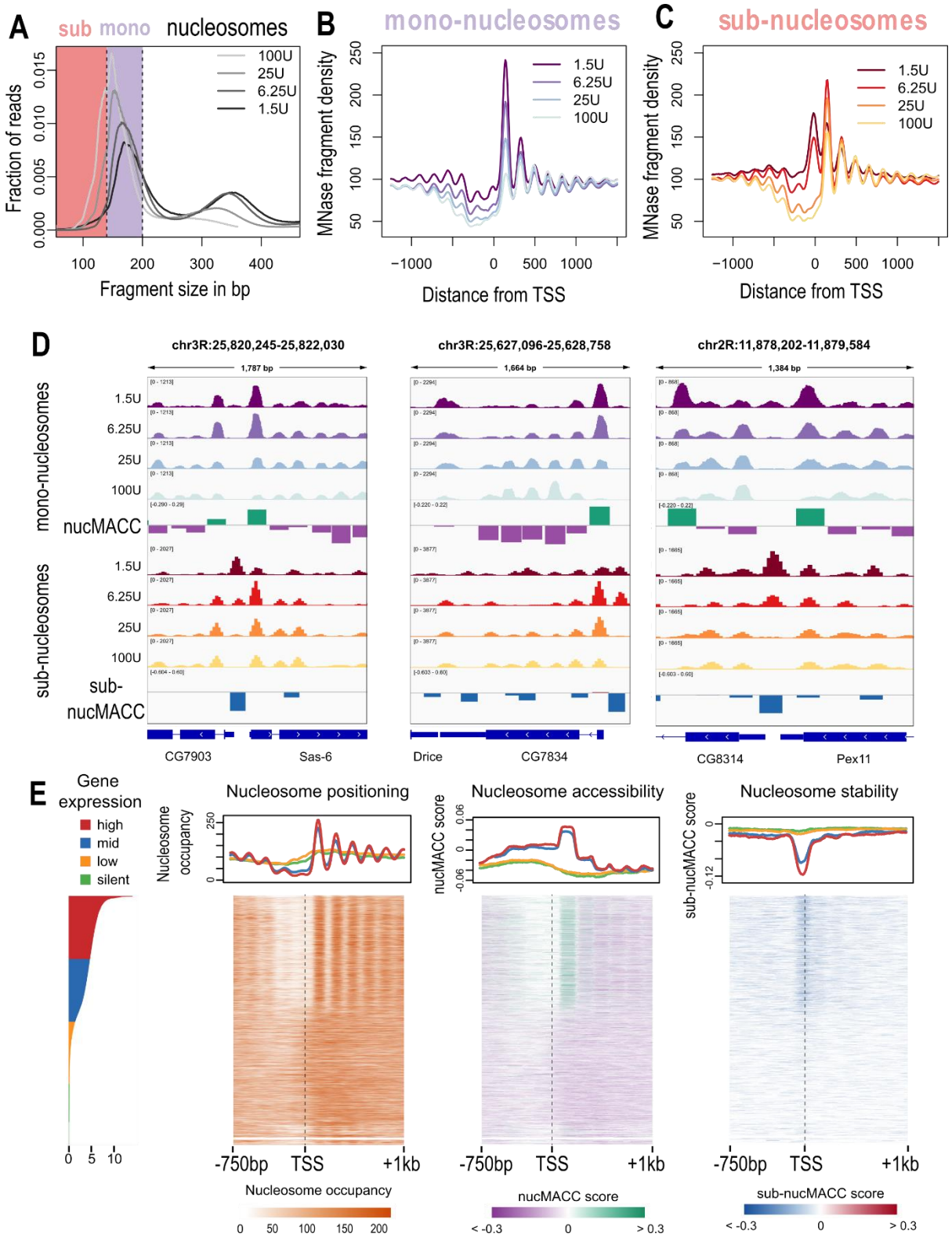



**Figure 8:** Characterization of the nucMACC score for ChIP-H4-MNase-seq data. (A) *In silico* fragment size selection based on the insert size distribution plot. In the blue square, the sub-nucleosomal fraction is indicated. In the pink square, the mono-nucleosomal fraction is indicated. (B-C) Average MNase signal over the TSS of expressed genes for mono- (B) and sub- (C) nucleosomes. (D) Examples

of the output of the nucMACC pipeline. (E) Nucleosome positions (left), nucleosome accessibility (middle), and stability (right) scores over TSSs, sorted by gene expression. Figure generated in collaboration with Dr. Uwe Schwartz.

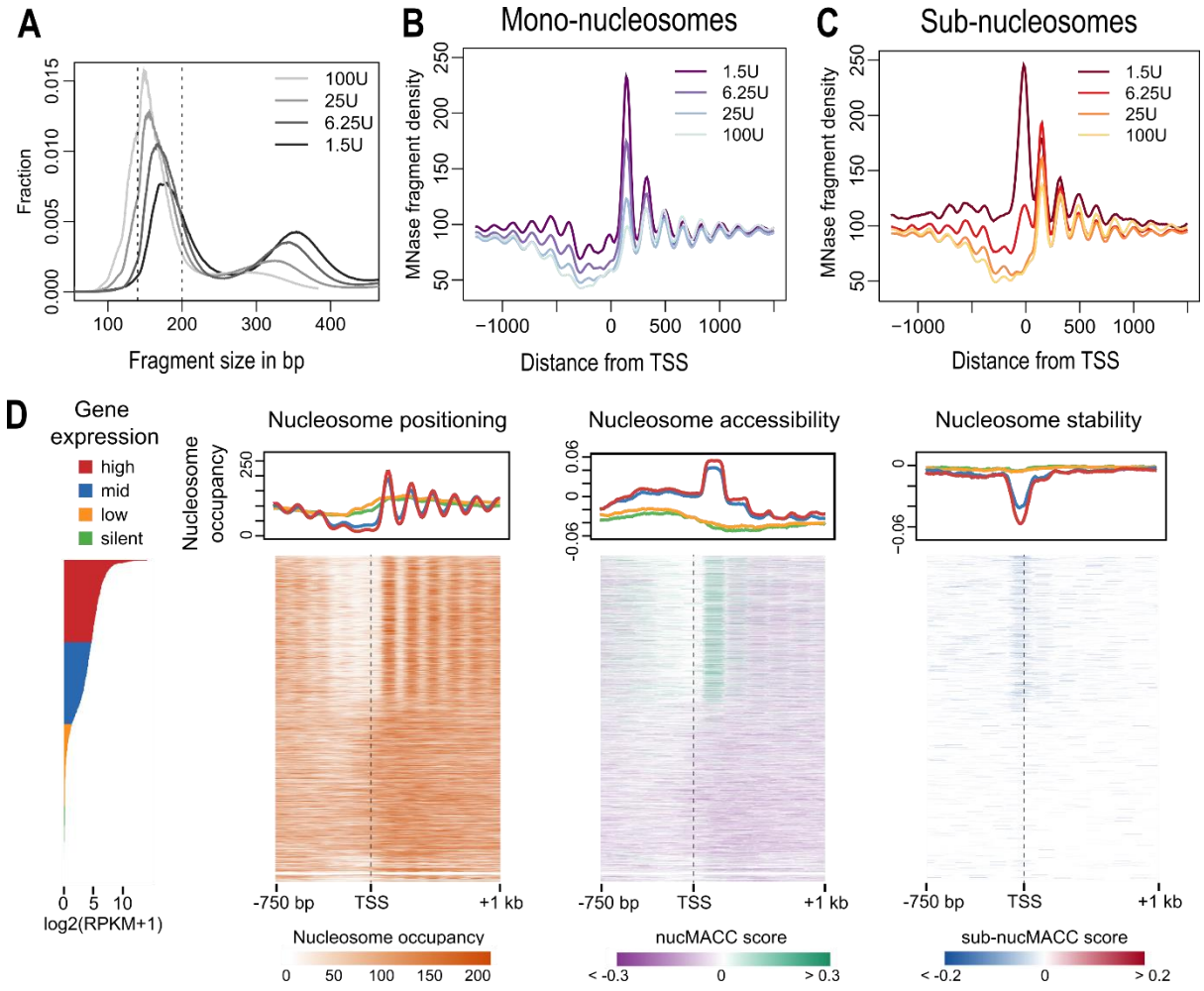



**Figure 9:** Characterization of the nucMACC score for ChIP-H3-MNase-seq data. (A) *In silico* fragment size selection based on the insert size distribution plot. (B-C) Average MNase signal over the TSS of expressed genes for mono- (B) and sub- (C) nucleosomes. (D) Nucleosome positions (left), nucleosome accessibility (middle), and stability (right) score over TSSs, sorted by gene expression. Figure generated in collaboration with Dr. Uwe Schwartz.

I define the nucleosome accessibility and stability scores by comparing the differential MNase digestion conditions, considering the normalized fragment count of the mono- and sub- nucleosomal fractions, respectively. I separate nucleosomes based on gene expression into high, mid, low expression, and silent genes, confirming previously published observation that non-expressed and lowly expressed genes exhibit greater overall fuzziness in nucleosome positioning (Figure 8, E and Figure 9, D). Similarly, I also observed a difference in nucleosome accessibility and stability scores around the TSS, based on gene expression levels, with most high and mid-expressed genes displaying a hyper-accessible +1 nucleosome and a sub-group of those with an un-stable nucleosome at the TSS (Figure 8, E and Figure 9, D).

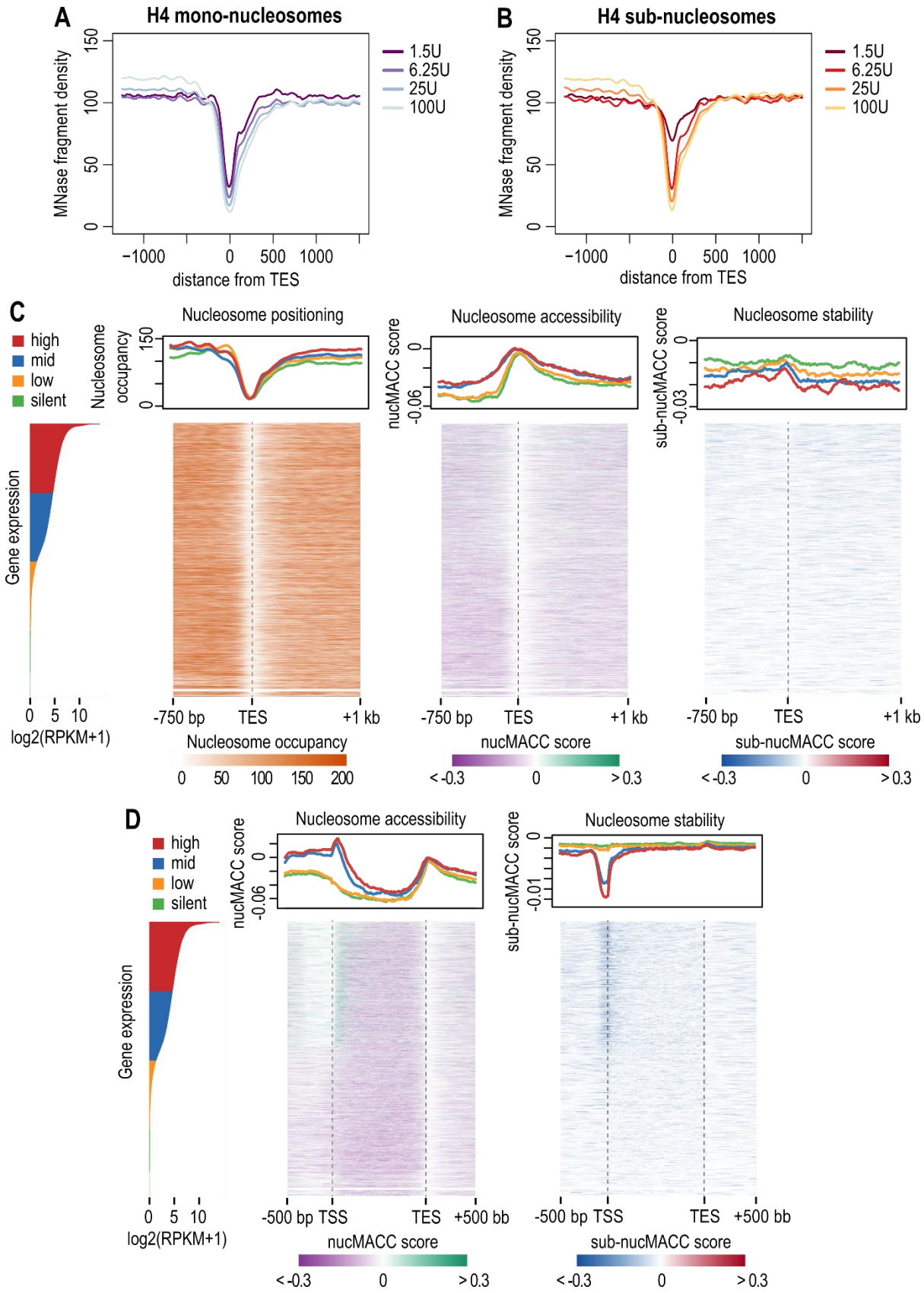



**Figure 10:** Characterization of the nucMACC score for ChIP-H4-MNase-seq data. (A-B) Average MNase signal over the transcription termination site (TES) of expressed genes for mono- (A) and sub- (B) nucleosomes. (C) Nucleosome positions (left), nucleosome accessibility (middle), and stability (right) score over TESs, sorted by gene expression. (D) Nucleosome positions (left), nucleosome accessibility

(middle), and stability (right) score over gene body, sorted by gene expression. Figure generated in collaboration with Dr. Uwe Schwartz.

Looking at the transcription termination site (TES) of expressed genes, I observed the typical dip in the MNase signal for mono-nucleosomes (Figure 10, A), whereas the sub-nucleosomes display a differential signal depending on the used MNase concentration (Figure 10, B). Not surprisingly, when the nucleosome stability scores were normalized to the underlying GC content, the MNase signal at low MNase concentration disappeared, revealing the TES is indeed nucleosome-free. As such, TES does not contain a non-canonical nucleosome (Figure 10, C and D).

A key to identifying differentially accessible regions in chromatin and defining functionally distinct nucleosomes is an unbiased statistical analysis. MNase has a known sequence bias, with a 30-x faster cleavage rate of DNA upstream of A or T, than G or C (Hörz and Altenburger, 1981). To exclude this bias from the nucMACC scores, I normalized them to the underlying GC content using LOESS regression (Figure 11, A and B).

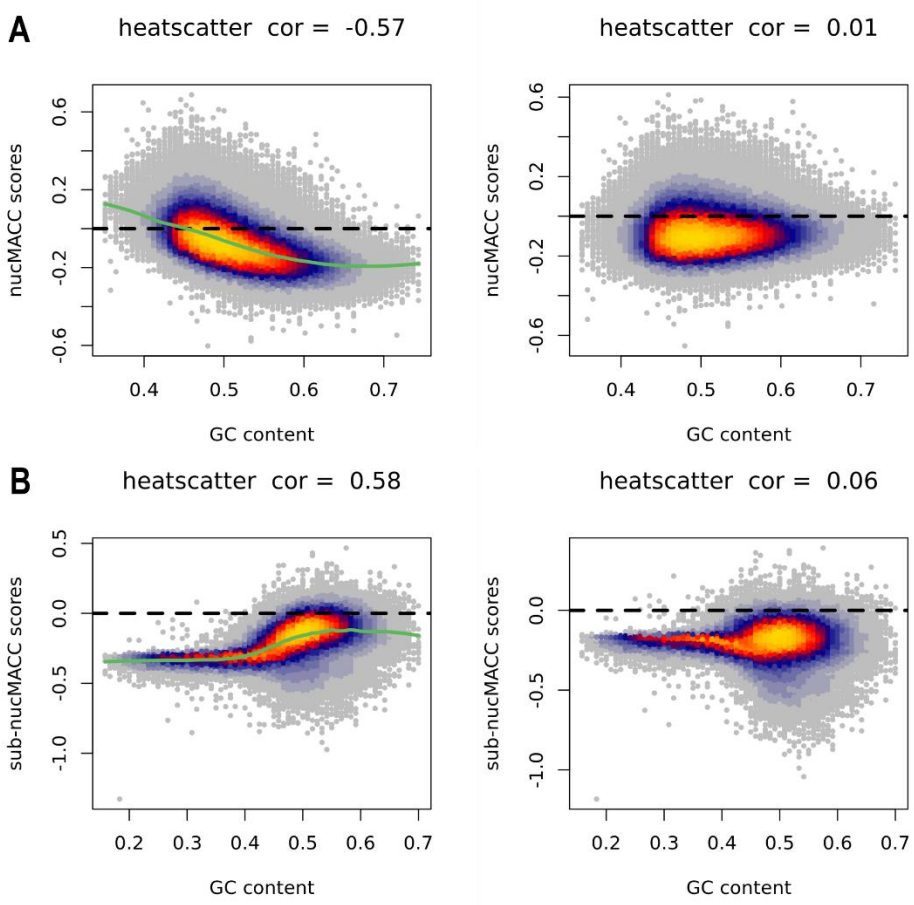



**Figure 11:** NucMACC scores before (left) and after (right) LOESS GC normalization for mono-nucleosomes (A) and sub-nucleosomes (B).

### 4.1.3 Nucleosome accessibility score

I set out to find a universal nucMACC score cut-off to detect nucleosomes exhibiting extraordinary accessibilities in an automated and data-independent manner.

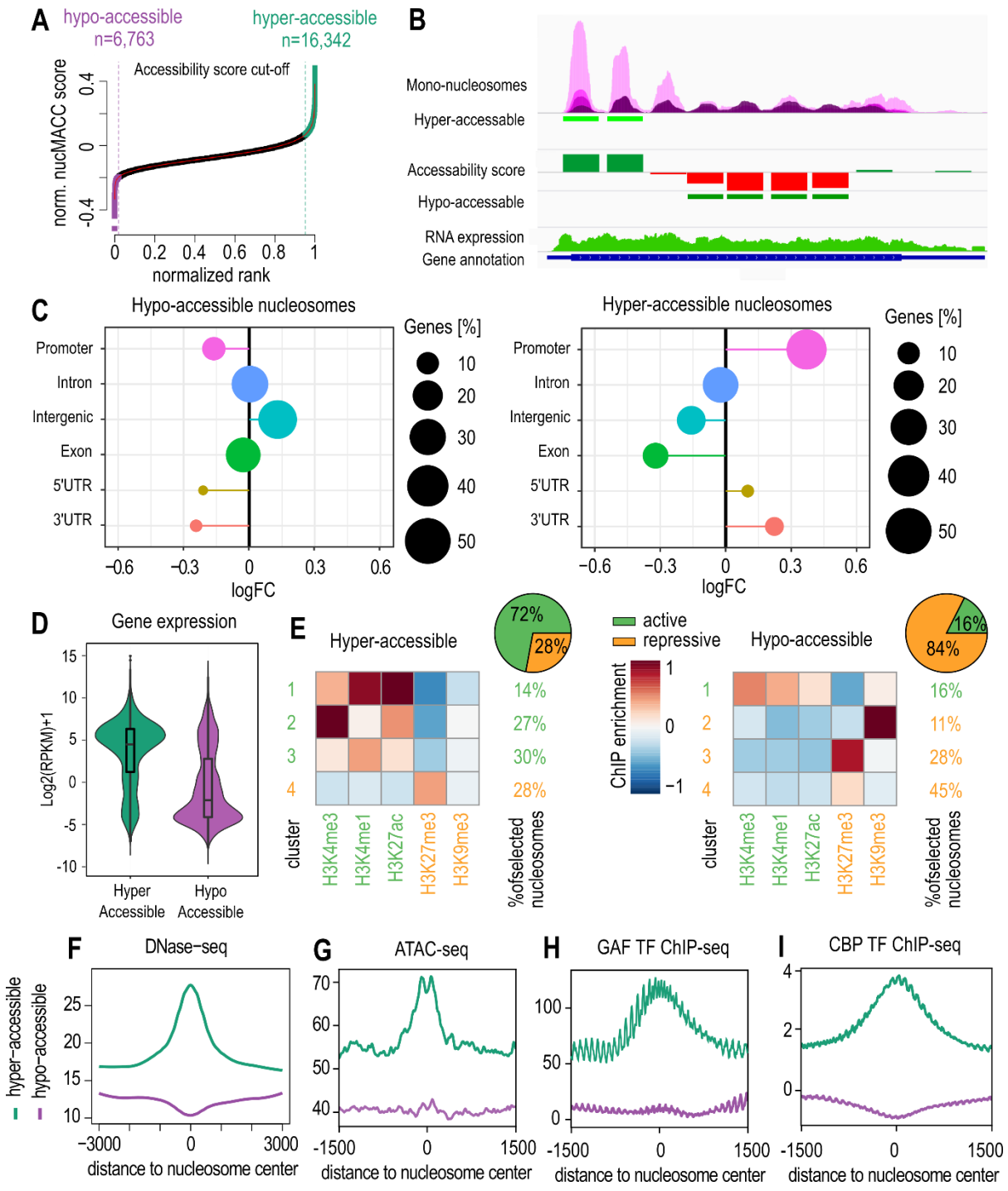



**Figure 12:** Characterization of the nucleosome accessibility score. (A) Nucleosome accessibility score cut-off. Hyper-accessible nucleosomes are indicated in green, whereas hypo-accessible nucleosomes are indicated in purple. (B) Example genomic region indicating a hyper-accessible +1 nucleosome and a hypo-accessible gene body. (C) Feature distribution of hypo- (left) and hyper- (right) accessible nucleosomes vs. all mono-nucleosomes. (D) Gene expression distribution for hyper- and hypo-accessible nucleosomes. (E) Enrichment of histone marks (H3K4me1, H3K27ac, H3K4me3, H3K9me3,

and H3K27me3) in hyper- and hypo-accessible nucleosomes. (F) Correlation between DNase-seq data and hyper-, hypo- accessible nucleosomes. (G) Correlation between ATAC-seq data and hyper-, hypo-accessible nucleosomes. (H) Enrichment of GAF ChIP-seq data over hyper-, hypo-accessible nucleosomes. (I) Enrichment of CBP ChIP-seq data over hyper-, hypo- accessible nucleosomes. Figure E was generated by Dr. Uwe Schwartz.

Here I ranked the nucMACC scores, and only the nucleosomes where the score was exceptionally different from the main distribution were selected. The cut-off was defined as the point of the curve where the slope is greater than 1, a strategy that is also used to find super-enhancers (Figure 12, A) (Whyte *et al.*, 2013). In my analysis, I could identify a set of nucleosomes with an extraordinary low nucMACC score, distinguishing 2% (6763) of all nucleosomes, indicative of hypo-accessible nucleosomes. I also find hyper-accessible nucleosomes with high nucMACC scores, representing 4.9% (16342) of all nucleosomes (Figure 12, A). Characterization of hyper- and hypo-accessible nucleosomes revealed enrichment in promoter and intergenic regions, respectively, suggesting nucleosomes with distinct accessibility preferentially occupying regulatory regions (Figure 12, C).

Interestingly, hyper-accessible nucleosomes were present mainly on expressed genes, while hypo-accessible nucleosomes were mainly present on silent genes (Figure 12, D). This suggests an association with local chromatin states and is also reflected in histone modifications. Hyper-accessible nucleosomes show enrichment in active enhancers and promoters (H3K4me1, H3K27ac, and H3K4me3), and hypo-accessible nucleosomes in heterochromatin (H3K9me3 and H3K27me3) (Figure 12, E). Additionally, I confirmed and validated the accuracy of nucleosome accessibility scores with DNase-seq data from Dunham et al., 2012 (Figure 12, F) and ATAC-seq data from Ibrahim et al., 2018 (Figure 12, G). Moreover, hyper-accessible nucleosomes are enriched in GAF TF and CBP transcriptional co-activator occupancy. GAF and CBP have been shown to be involved in nucleosome displacement and transcriptional activation, respectively (Figure 12, H and I) (Fuda *et al.*, 2015; Philip *et al.*, 2015).

### 4.1.4 Nucleosome stability score

Next, I characterized nucleosomes with extraordinary stability or non-canonical structure. First, sub-nucleosomes were filtered if they overlapped with a mono-nucleosome position or had less than 4x higher signal than the mono-nucleosomal fraction. This resulted in 63014 positions. Second, in the same way as for mono-nucleosomes, a cut-off for sub-nucMACC scores was devised to define extreme nucleosome groups. Genes marked with a low sub-nucMACC score, i.e., un-stabile nucleosomes, showed significant enrichment in promoter regions (Figure 13, C) of expressed genes (Figure 13, D) and represented 6.7% (n = 4232) of all sub-nucleosomes (Figure 13, A).

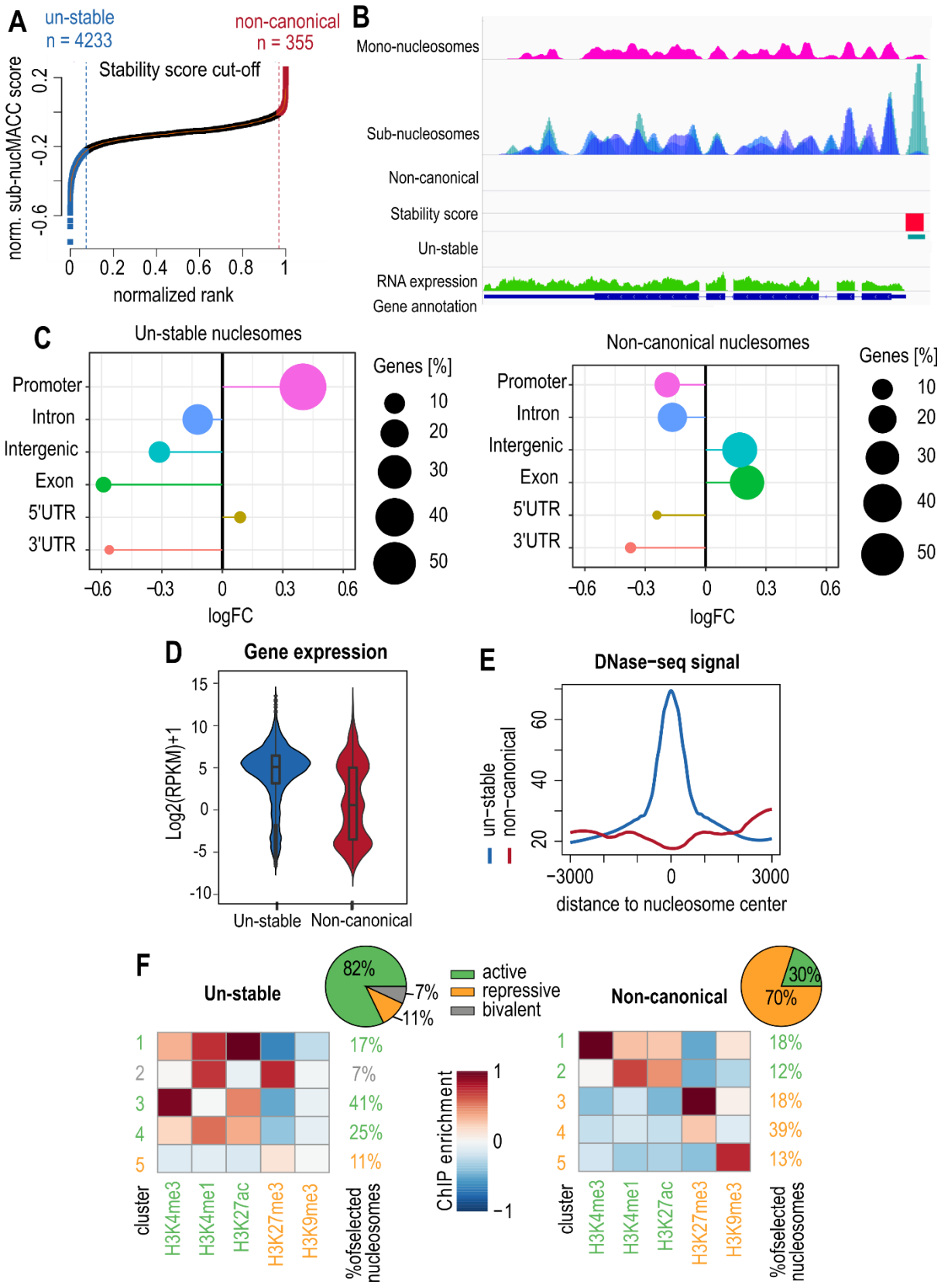



**Figure 13:** Characterization of the nucleosome stability score. (A) Nucleosome stability score cut-off. Un-stable nucleosomes are indicated in blue, whereas non-canonical nucleosomes are indicated in red. Stable canonical nucleosomes are indicated in black. (B) Example region indicating an un-stable nucleosome at TSS. (C) Feature distribution of un-stable and non-canonical nucleosomes over all sub-nucleosomes. (D) Gene expression distribution of un-stable and non-canonical nucleosomes. (E)

Correlation between DNase-seq data and un-stable, non-canonical nucleosomes. (F) Enrichment of histone modifications (H3K4me1, H3K27ac, H3K4me3, H3K9me3, and H3K27me3) in un-stable and non-canonical nucleosomes. Figure F generated by Dr. Uwe Schwartz.

Not surprisingly, un-stable nucleosomes showed enrichment in H3K27ac and H3K4me3 histone marks, representing active promoters (Figure 13, F). I could also detect sub-populations of un-stable nucleosomes with significant enrichment in H3K27me3 histone modification, marking repressive chromatin state and in H3K27ac and H3K4me1, marking poised enhancers (Figure 13, F).

Nucleosomes with a high sub-nucMACC score, i.e., nucleosomes with non-canonical structures, represent 0.5% of all sub-nucleosomes (n = 355) and favor silent genes (Figure 13, D). Non-canonical nucleosomes are enrichment in intergenic and exonic regions (Figure 13, C). They showed enrichment in H3K27me3 and H3K9me3 histone modifications, marking heterochromatin (Figure 13, F).

### 4.1.5 Gene regulation via modulation of nucleosome stability

I identified a specific group of un-stable nucleosomes, positioned at the TSS of expressed genes, referred to as TSS-un-stable nucleosomes (Figure 15, B). When comparing expressed genes with a TSS-un-stable-nucleosomes (n = 1418) with expressed genes without a TSS-un-stable-nucleosome (n = 6010), I do not see any difference in gene expression between the groups (Figure 15, A); however, I could observe a clear difference in nucleosome occupancy at the TSS (Figure 14, A-B and Figure 15, B). Strikingly, I find enrichment in the M1BP TF motif within the promoters of genes harboring the un-stable nucleosomes (Figure 15, C).

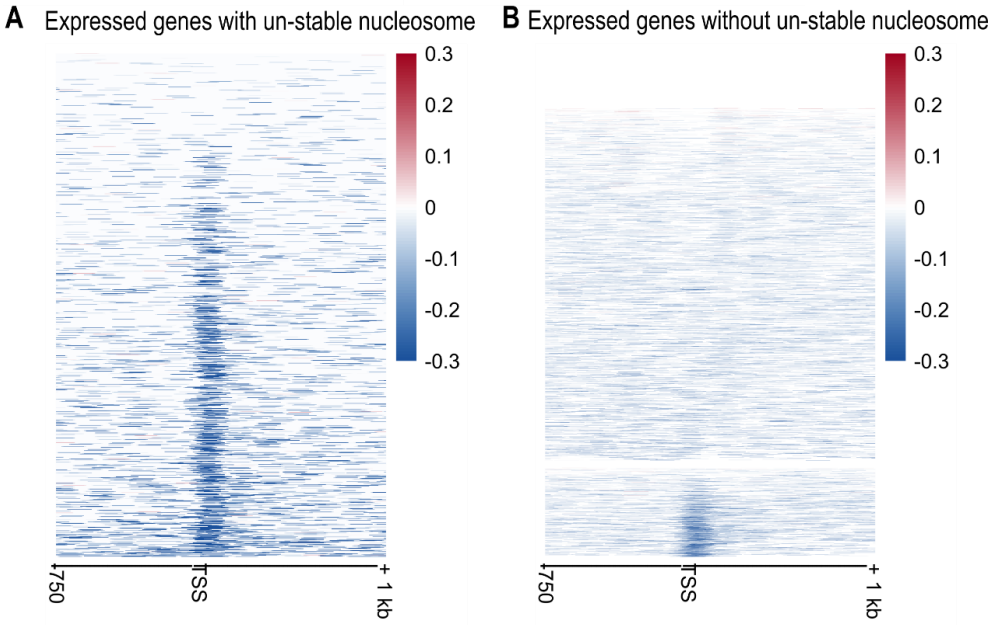



**Figure 14:** Nucleosome stability score distribution over TSS of genes with (A) and without (B) an un-stable nucleosome at TSS.

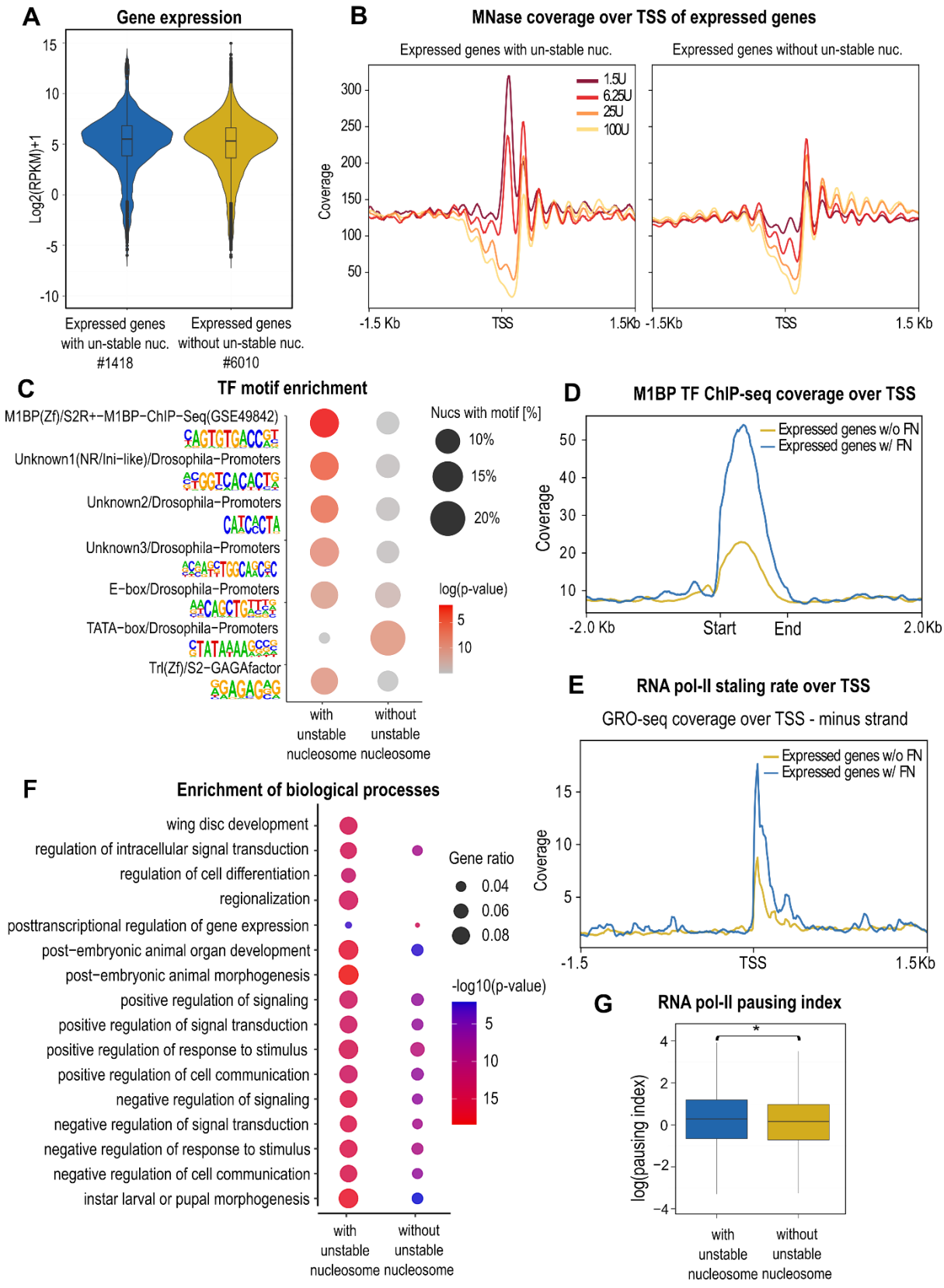



**Figure 15:** Characterization of un-stable nucleosomes. (A) Gene expression distribution of expressed genes with (blue) or without (gold) an un-stable nucleosome. (B) MNase coverage over TSS of expressed genes with (left) or without (right) an un-stable nucleosome. (C) TF motif enrichment of expressed genes with (left) or without (right) an un-stable nucleosome. (D) M1BP ChIP-seq seq coverage over TSS of genes with (blue) or without (gold) an un-stable nucleosome. (E) GRO-seq

coverage over genes with (blue) or without (gold) an un-stable nucleosome. (F) Enrichment of biological processes associated with genes with (left) or without (right) an un-stable nucleosome. (G) RNA pol-ii pausing index for genes with (blue) or without (gold) an un-stable nucleosome.

I next asked whether un-stable nucleosomes prevent TF recruitment and gene activation or do the un-stable nucleosome allow both to bind simultaneously. I overlapped the un-stable nucleosomes with ChIP-seq data for M1BP and found significant overlaps and enrichment over expressed genes without a TSS-un-stable nucleosome (Figure 15, D). M1BP TF has been shown to bind to specific promoters marked with “motif one", having a higher RNA pol-II transcription rate than other promoters in *D. melanogaster* (Li and Gilmour, 2013).

Therefore, I analyzed GRO-seq data (Core *et al.*, 2012) and looked for RNA pol-II staling rate for genes occupied with or without a TSS-un-stable nucleosome. I observed a clear difference between the two gene groups and their RNA pol-II occupancy and pausing index, with genes with the un-stable nucleosome having a higher pol-II occupancy at TSS and a higher pausing index (Figure 15, E-G).

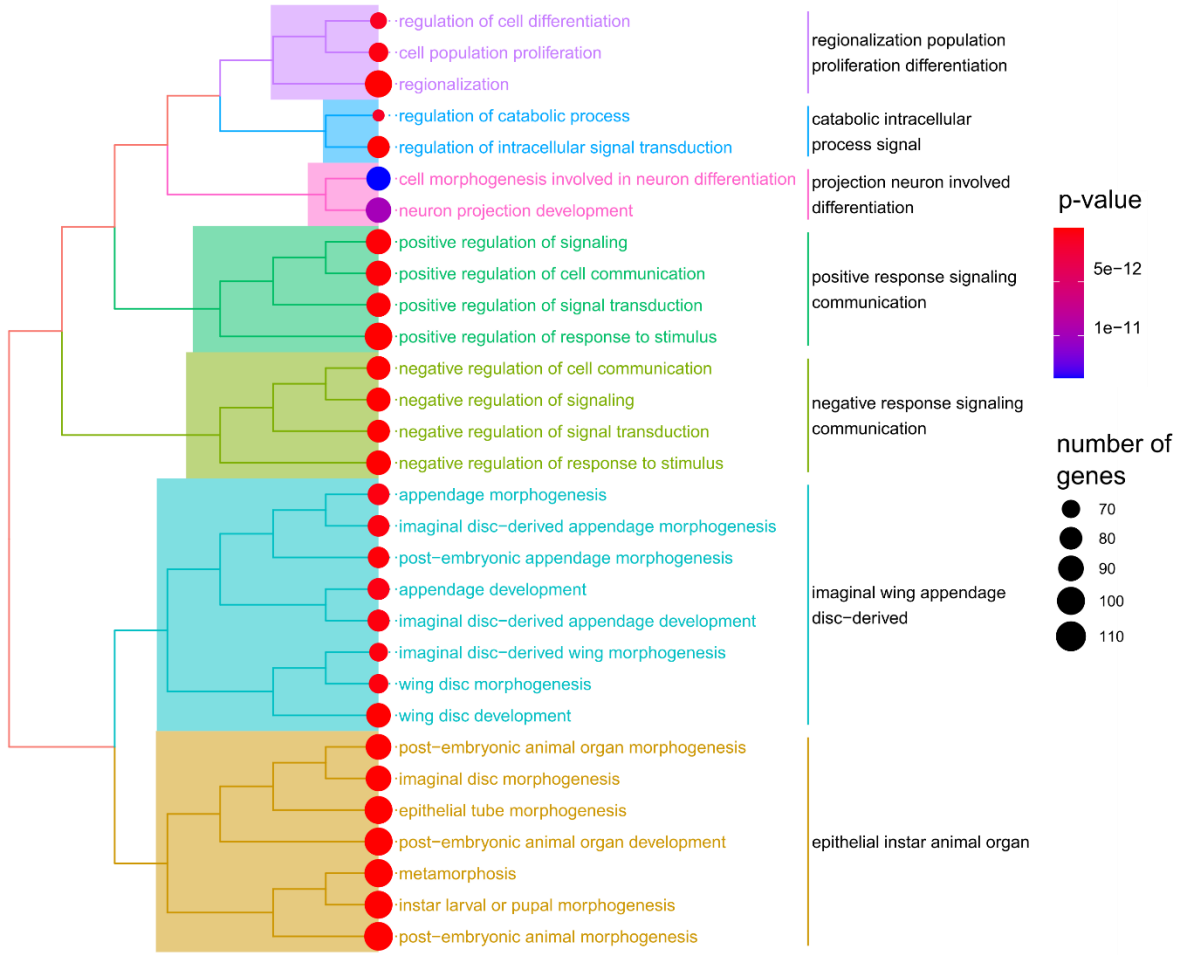



**Figure 16:** Molecular pathway enrichment for expressed genes with an un-stable nucleosome at TSS.

Next, I tested if un-stable nucleosomes occupy genes with a common biological function. Monitoring for GO terms, I observed enrichment in developmental, stimuli response, and morphological terms for TSS-un-stable nucleosomes associated genes (Figure 15, F and Figure 16). In comparison, expressed genes without TSS-un-stable nucleosomes are enriched in endosomal transport, mitochondrial translation and expression, mitotic division, DNA repair, and RNA splicing (Figure 15, F and Figure 17). Results clearly reflect the developmental stage of the S2 cell line and suggest genes with TSS-un-stable nucleosomes have a regulatory and predictive role in response to stimuli and spatiotemporal expression during the development of the fruit fly.

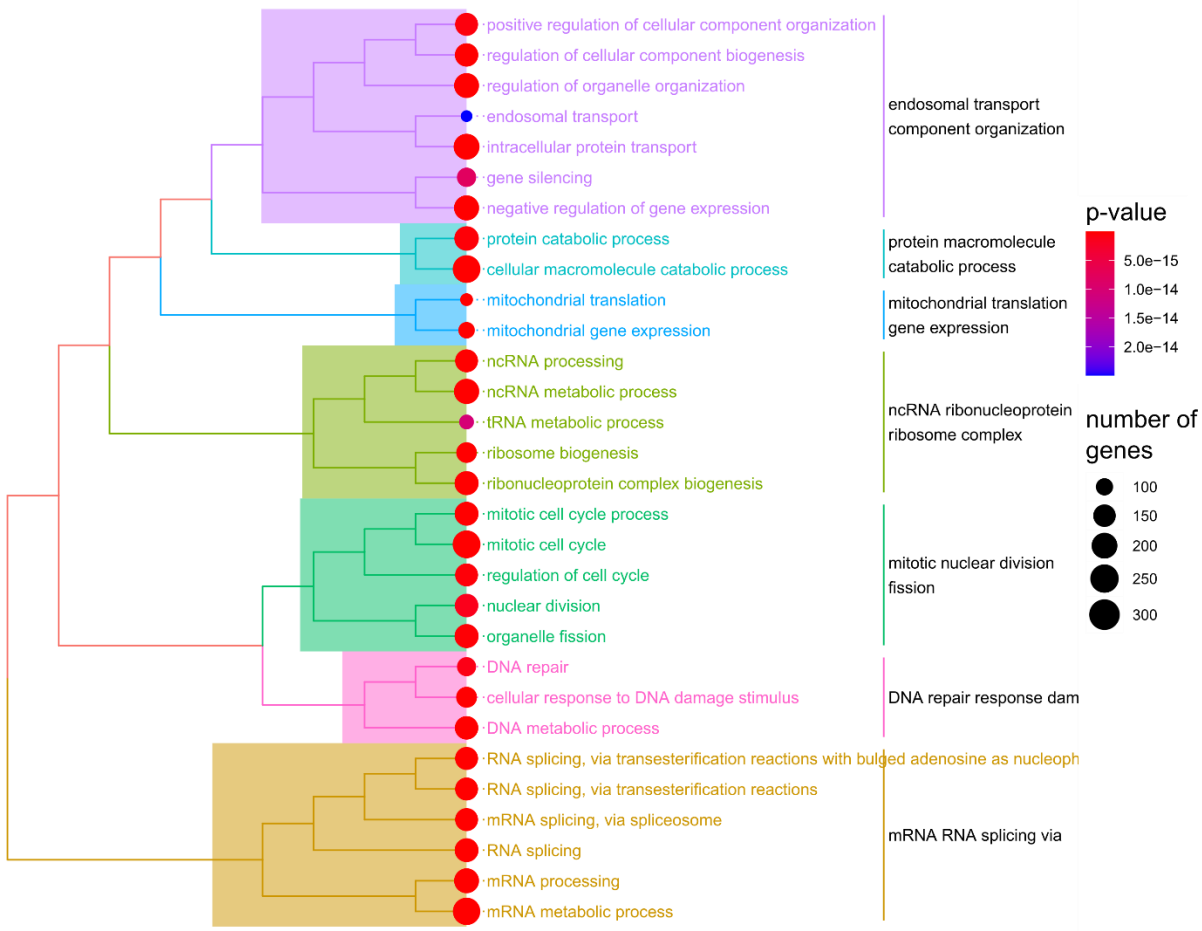



**Figure 17:** Molecular pathway enrichment for expressed genes without an un-stable nucleosome at TSS.

## 4.1.6 Pipeline robustness

**Comparing H3 and H4 data**

Next, I wanted to test the robustness of the pipeline by comparing MNase-H3-ChIP-seq and MNase-H4-ChIP-seq data. I find that 90% of mono-nucleosomes overlap between MNase-H3-ChIP-seq and MNase-H4-ChIP-seq data and exhibit a correlation of 0.68 between the nucMACC scores (Figure 18,

A). Sub-nucleosomal-sized fragments show a higher variation in sequencing depth than mono-nucleosomal-size fragments and thus, not surprisingly, show a lower overlap of 39.8% between experimental setups; nevertheless, the correlation between the nucMACC scores was high - corr = 0.83 (Figure 18, B).

**Minimum number of MNase titration**

Previously four or more MNase time points or concentrations have been used to determine DNA accessibility scores (Mieczkowski *et al.*, 2016; Mueller *et al.*, 2017; Chereji, Bryson and Henikoff, 2019). I sought to determine the minimum number of titration points per sample to obtain reliable and robust results with the pipeline. I compared the nucMACC scores of two or four titration points of the same experimental dataset. I found the combination of 1.5U + 25U, 1.5U + 100U titrations to have a high correlation for both mono- and sub-nucleosomes, with correlation scores of 0.8, 0.92 and 0.79, 0.79, respectively (Figure 18, C; Figure S2, A; Figure S3, A). The number of called mono-nucleosome positions also correlated with the nucMACC scores (Figure S2, B). In summary, the choice of one low and one high MNase concentration used to analyze mono-nucleosomes is not critical unless the difference in MNase concentration or digestion time is at least 15-fold. In contrast, to detect the sub-nucleosomal population, a very low MNase concentration is required due to their fragile nature.

**Estimation of required sequencing depth**

Next, I addressed the importance of sequencing depth on data quality, as MNase-seq is a rather expensive experiment. I devised a sequencing coverage matrix independent of the genome size (Equation 1), defined as the number of fragments per kilobase of genome size (FPKG).

**Equation 1:** Sequencing depth equation.

$$\text{FPKG (fragments per kilobase of genome size)} = \frac{\text{Number of fragments} \times 1\,\text{kb}}{\text{Genome size (dm3: 162367812)}}$$

For mono-nucleosomes, the pooled data had a coverage of 634 FPKG. I found that using half of the fragments, with coverage of 312 FPKG, preserves 73% of the mono-nucleosome positions, and the nucMACC score correlation is reasonably high, 0.88 (Figure 18, E and Figure S3, B). Whereas further reducing the sequencing depth to 156 FPKG results in a considerable loss of called mono-nucleosome positions (25%), albeit the nucMACC score correlation remains stable, with a correlation of 0.80. For sub-nucleosomes, I start with 183 FPKG, already at low coverage. Reducing the coverage to 135 FPKG (75% of sequencing fragments) amounts to a decrease in called sub-nucleosomes to 49% (Figure 18, F and Figure S3, C). Nonetheless, the correlation between nucMACC scores remains very high – 0.90

(Figure 18, F and Figure S3, C). Below this coverage, the sub-nucleosome analysis produces a very low amount of called nucleosomes (17%) with a high sub-nucMACC score correlation of 0.86 (Figure 18, F and Figure S3, C). According to the results, it is recommended that a sequencing depth of at least 180 FPKG or higher is used to discover nucleosomes with differential stability.

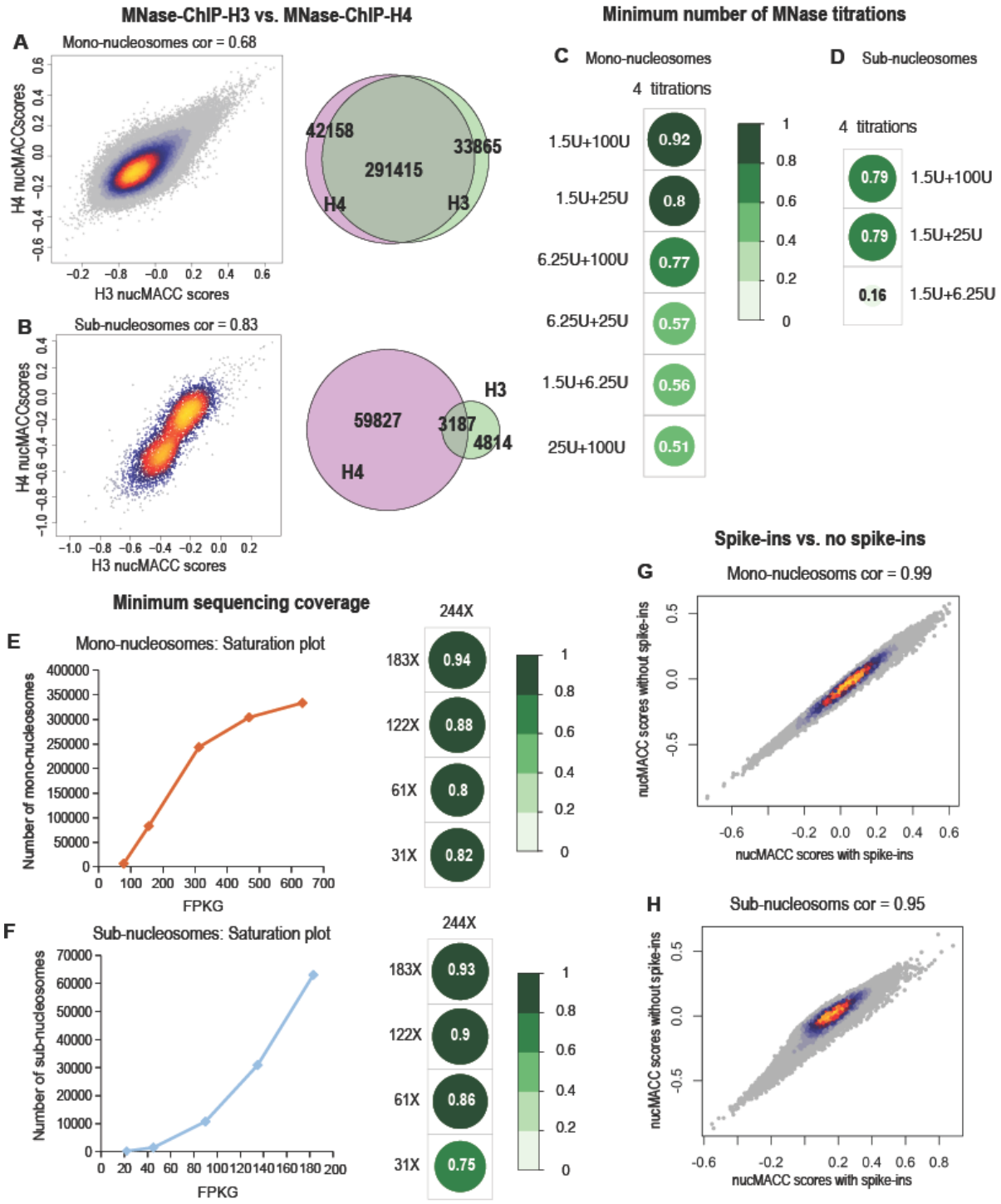



**Figure 18:** Evaluation of the nucMACC pipeline robustness. (A-B) Correlation between nucMACC scores and the overlap between called nucleosome between H3 and H4 data in mono- (A) and sub- (B) nucleosomes. (C-D) Correlation between nucMACC scores for samples with two or four MNase titrations in mono- (C) and sub- (D) nucleosomes. (E-F) The number of called nucleosomes based on

sequencing depth (FPKG – fragments per kilobase of the genome size) and the correlation between nucMACC scores of samples with different sequencing depths in mono- (E) and sub- (F) nucleosomes. (G-H) Correlation between nucMACC scores with or without spike-in information for mono- (G) and sub- (H) nucleosomes.

Ultimately, I tested the reliability and robustness of our accessibility and stability scores and determined if sequencing depth would bias the scores. This prompted me to use the pipeline on a published dataset with spike-ins and compare the results with or without spike-in information. Remarkably, I find almost no difference in nucMACC scores when comparing the pipeline with or without spike-in information for mono- (Figure 18, G and Figure S4) and sub- nucleosomes (Figure 18, H and Figure S4).

In conclusion, the nucMACC pipeline is robust, reproducible, doesn't require spike-ins, and provides clear data with only two MNase titrations. The pipeline is completely automated and publicly available on GitHub. I am the first to provide an automated way for calling special nucleosome groups in addition to nucleosome positions.

### 4.1.7 The nucMACC pipeline

The final nucMACC pipeline requires raw data in .fastq format and a minimum of two MNase titrations per experimental condition (Figure 19). The pipeline works on both MNase-seq and MNase-ChIP-seq data; however, I would strongly recommend using MNase + anti-histone immunoprecipitation data when analyzing sub-nucleosomal particles. The pipeline starts with read mapping to the reference genome and quality filtering. In the next step, fragments are divided into sub-nucleosomal- and mono-nucleosomal- sized-fragments which are processed separately.

**Mono-nucleosomes**

MNase titrations are pooled to obtain high sequencing depth and are used to call mono-nucleosome positions. A nucMACC score is calculated by counting fragments per each nucleosome position and each MNase titration. A linear regression slope is then fitted through MNase titrations for each nucleosome. The slope of the regression line represents the raw score, which is normalized to the underlying GC% content in the following step. The normalized score is referred to as the nucleosome MNase accessibility score (nucMACC). In the final step, I characterize nucleosomes, whose nucMACC score considerably deviates from the mean, into two categories, hypo-, and hyper- accessible nucleosomes.

**Sub-nucleosomes**

The lowest MNase titration is used to call sub-nucleosomal positions and is consequently filtered by previously called mono-nucleosomal positions to obtain proper sub-nucleosome positions. Similar to the mono-nucleosomes, a regression slope is fitted on the MNase titration fragment counts and is normalized to the underlying GC% content. The normalized score is referred to as the sub-nucleosome MNase accessibility score (sub-nucMACC). Here too, I select extreme nucleosome groups.

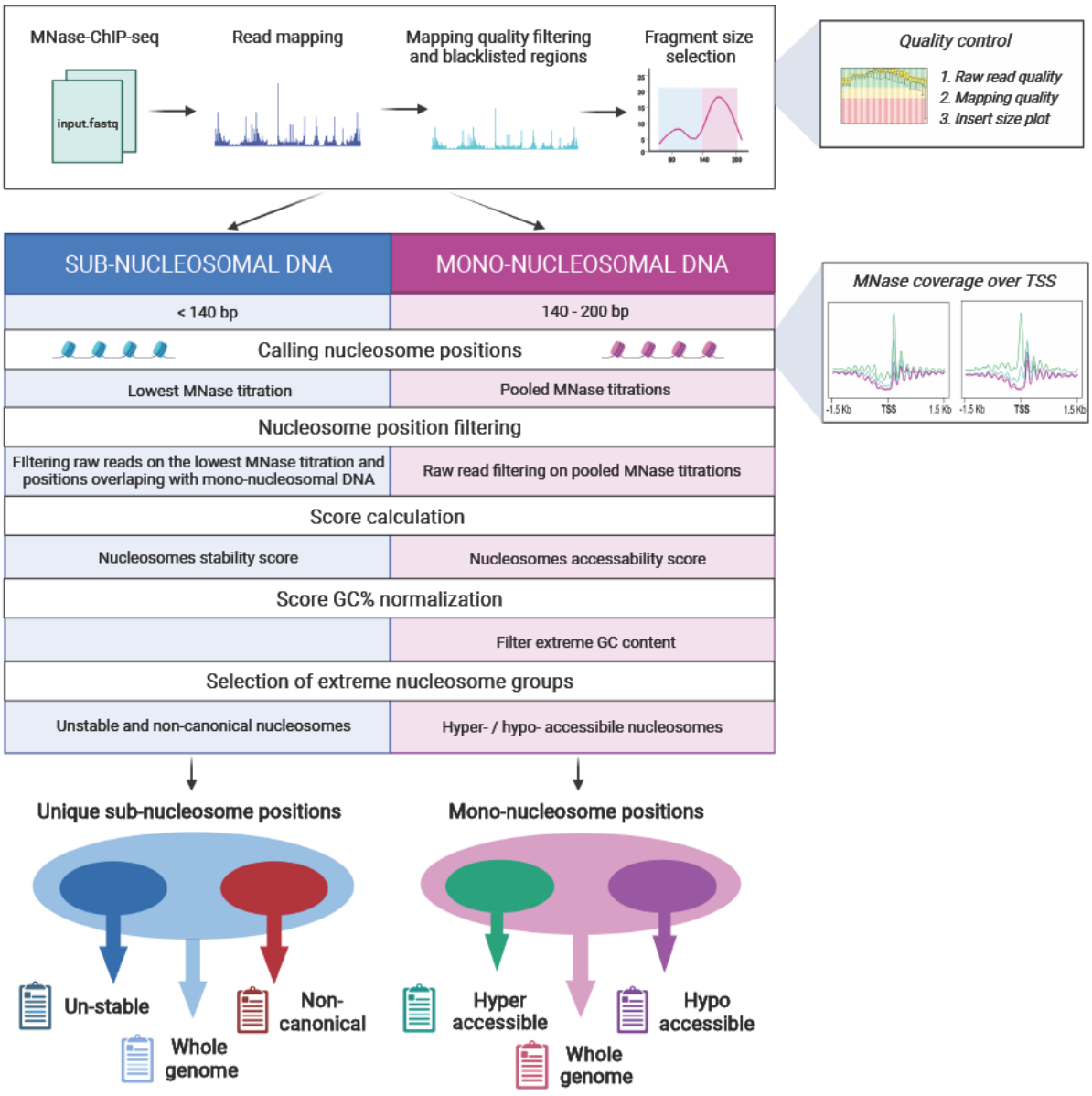



**Figure 19:** Workflow of the nucMACC pipeline (Figure made with BioRender).

**Output**

The final output of the pipeline is the quality matrix, unique sub-nucleosome positions with sub-nucMACC scores, mono-nucleosome positions with nucMACC scores, and positions of hyper-

accessible hypo-accessible nucleosomes, un-stable (fragile) nucleosomes, and non-canonical nucleosomes (Figure 20).

**Sub-nucleosomes output**

| Chromosome | Start | End | Strand | GC_cont | nucID | sub.nucMACC | slope | R2 | category | selection |
|---|---|---|---|---|---|---|---|---|---|---|
| chr2R | 3639101 | 3639221 | . | 0.525 | sub.nuc421778 | -0.299 | -0.240 | 0.318 | normal | not selected |
| chr3L | 526901 | 527021 | . | 0.475 | sub.nuc13605 | -0.299 | -0.291 | 0.705 | normal | enriched |
| chr2L | 490561 | 490681 | . | 0.433333 | sub.nuc117620 | -0.299 | -0.361 | 0.810 | normal | unique |
| chr3L | 621761 | 621881 | . | 0.633333 | sub.nuc14135 | 0.219 | 0.286 | 0.779 | non-canonical | enriched |
| chrX | 21804181 | 21804301 | . | 0.558333 | sub.nuc408396 | 0.237 | 0.310 | 0.815 | non-canonical | unique |
| chr2L | 8825321 | 8825441 | . | 0.55 | sub.nuc155270 | -0.899 | -0.817 | 0.936 | un-stable | enriched |
| chr3R | 15470121 | 15470241 | . | 0.533333 | sub.nuc279772 | -0.888 | -0.821 | 0.870 | un-stable | unique |

**Mono-nucleosomes output**

| Chromosome | Start | End | Strand | GC_cont | nucID | nucMACC | slope | R2 | category |
|---|---|---|---|---|---|---|---|---|---|
| chr3R | 18914486 | 18914634 | . | 0.432 | nuc333904 | -0.074 | 0.032 | 0.193 | normal |
| chr2L | 8428794 | 8428942 | . | 0.480 | nuc176795 | -0.653 | -0.602 | 0.800 | hypo-accessible |
| chr2R | 12845033 | 12845181 | . | 0.520 | nuc534680 | 0.488 | 0.482 | 0.985 | hyper-accessible |

**Figure 20:** Example output of the nucMACC pipeline. The mono-nucleosomes can be categorized as normal, hypo-, and hyper-accessible. In contrast, sub-nucleosomes can be categorized as normal, non-canonical, and un-stable. Here nucleosomes are additionally selected based on mono-nucleosome positions, as unique, i.e., not overlapping with a mono-nucleosome position, enriched, i.e., overlapping with a mono-nucleosome position, and have at least 4-times higher signal for the sub-nucleosome than mono-nucleosome and as not selected, i.e., neither unique not enriched. Slope = Slope of linear regression; R2 = R square value of the slope; The nucMACC = nucMACC score for mono-nucleosomes; sub-nucMACC = The nucMACC score for sub-nucleosomes; GC_cont = GC% of the underlying DNA sequence; Chromosome + Start + End = Coordinates of the nucleosomes.

## 4.2 Discussion

In the current study, I set out to optimize the MNase-seq pipeline and characterize nucleosomes based on their accessibility and stability. I observed local changes in nucleosome accessibility, which represent around 7% of all nucleosomes. The majority of these exhibit an increase in accessibility (4.9%), and 2% of nucleosomes show a decreased accessibility. Hyper-accessible nucleosomes were enriched in promoter regions, whereas hypo-accessible nucleosomes showed enrichment in intergenic regions (Figure 12, C) (Schwartz *et al.*, 2019).

I show by ChIP-MNase-seq that un-stable nucleosomes contain H4, H3, and H2B histones (Figure 8, D; Figure 9, C; Figure S1, A) and exhibit shorter DNA length. However, the nucMACC pipeline cannot discriminate between nucleosome structures. Interestingly I find a distinct group of un-stable nucleosomes enriched at TSS of promoters marked with motif one, which has been shown to harbor an M1BP TF-specific motif (Li and Gilmour, 2013). Undeniably, I see a high enrichment of M1BP ChIP-seq peaks over these un-stable nucleosomes. This made me speculate about their role in RNA pol-II pausing, which I confirmed with GRO-seq data (Figure 15, E). I could detect a clear difference in RNA pol-II pausing between genes marked with a TSS-un-stable nucleosome (Figure 15, E and G). I hypothesize that un-stable nucleosomes located directly on the TSS allow TF binding, despite nucleosomes normally occluding binding. This may enable a rapid response to stimuli and cell type-specific gene expression. The un-stable nucleosome represents a novel gene regulation machinery, adding to the complexity of eukaryotic gene regulation.

Additionally, I also find un-stable nucleosomes on local heterochromatin (Figure 13, F), which is consistent with a recent study by Sanulli et al., where they show nucleosomes on heterochromatic regions are hyper accessible to promote multivalent interactions with other nucleosomes and histone tails, enabling chromatin compaction into liquid condensates (Sanulli *et al.*, 2019). In this study, I separate nucleosome accessibility and stability and show that heterochromatic nucleosomes are un-stable.

I also provide a robust pipeline and guidelines for the experimental setup of the MNase-seq protocol by analyzing different datasets with two or four MNase titrations per sample, with or without spike-in information, different sequencing depths, and correlation between different anti-histone antibody pulldowns. In short, I show that only two MNase titrations are required, albeit at a high sequencing depth. I suggest a minimum sequencing depth of 312 FPKG per sample for mono- and 183 FPKG for sub- nucleosomes to attain high-quality results (Figure 18, E and F). This will, of course, depend on the experimental design. For example, if mono- and sub- nucleosomes are excised from the gel, the

sequencing depth required per sample will be lower. Mono-nucleosomes correlate well between H3 and H4 data (Figure 18, A and Figure S1, A). Likewise, the correlation between sub-nucMACC scores of H3 and H4 data is high (Figure 18, B and Figure S2, A). The differences observed between H3 and H4 data could be due to biological differences in histone composition, antibody efficiency, and the dynamic nature of sub-nucleosomal particles. For sub-nucleosomes, it is crucial to verify them by anti-histone ChIP after MNase treatment to differentiate them from TF bound to DNA at these sites. Ideally, an anti-TF ChIP-seq would be used for comparison to see whether it leads to structural changes in the nucleosome.

I also analyzed the nucMACC scores with or without spike-in information. Spike-ins are typically used to compare antibody pulldowns with different efficiencies or to make an absolute quantification of RNA expression or, in our case, nucleosome occupancy. I see no improvement or difference in nucMACC scores when using spike-in information (Figure 18, G-H and Figure S3, A-D). In conclusion, for determining absolute nucleosome occupancy, spike-ins are still compulsory and should be included in the experimental design, whereas using relative differences in the occupancy is sufficient to calculate the nucleosome accessibility and stability scores. Sequencing depth seems to have a more detrimental effect on the accuracy of nucMACC scores than the absence of spike-in information. In a recent publication, authors suggest spike-ins are necessary to accurately estimate the nucleosome digestion rate and subsequently accessibility scores (Chereji, Bryson and Henikoff, 2019). Unfortunately, I could not directly compare their scores and mine in the same dataset due to the lack of IP after MNase digestion. However, I have shown that the accessibility scores coincide with ATAC-seq and DNase-seq data (Figure 13, F and G; Figure ). In contrast to the publication mentioned above, it requires a less laborious and cheaper experimental design.

# 5 THE TRIPLEX CODE

## 5.1 Results

### 5.1.1 Comparing triplex motifs

Most biochemical studies on triplexes were performed on DNA:DNA triplexes. However, an indication that the triplex binding code is different between DNA:DNA and RNA:DNA triplexes (Maldonado *et al.*, 2017) and between the three triplex motifs (Kunkler *et al.*, 2019) has been published in recent papers.

To determine the triplex binding code, I first investigated the binding affinity of the three triplex motifs at the same Guanine percentage of the TrTS. Here I wanted to know if indeed significant differences exist between the motifs and between DNA:DNA and RNA:DNA triplexes. Using MST measurements, I have determined the binding affinity of the three motifs (Pyrimidine, purine, and mixed) with Guanine percentage of 76% in the TrTS and 29 nt length of the DNA or RNA TFO. Guanine percentage here refers to the sequence composition of the dsDNA since it was shown to have a more significant impact on triplex stability than the sequence composition of the TFO (Kunkler *et al.*, 2019). Each measurement was additionally verified with an electrophoretic Mobility Shift Assay (EMSA) to validate the triplex formation. The buffers used for triplex formation were designed to mimic the cell environment.

I have observed distinct differences in the triplex binding code for each of the three motifs (Figure 21, B and C). For instance, at 76% Guanines and RNA TFO, the mixed motif has the lowest $K_d$ value ($K_d$ = 70.3 nM +/- 68.7 nM), i.e., the highest binding affinity. In contrast, the pyrimidine motif has the lowest binding affinity (Kd = 1.5 µM +/- 272.2 nM), with no observable shift in the EMSA. Moreover, there are also key differences between RNA:DNA and DNA:DNA triplexes for the same motif (Figure 21, B and C).

Similar to the RNA:DNA triplexes, DNA:DNA follow the same order of binding, with the mixed motif having the highest binding affinity, followed by the purine motif and the pyrimidine motif, which do not form a triplex either in the MST or EMSA. Given that clear differences in the binding code of triplex motifs exist, I further evaluated each motif based on the Guanine sequence composition, TFO length and RNA or DNA TFO.

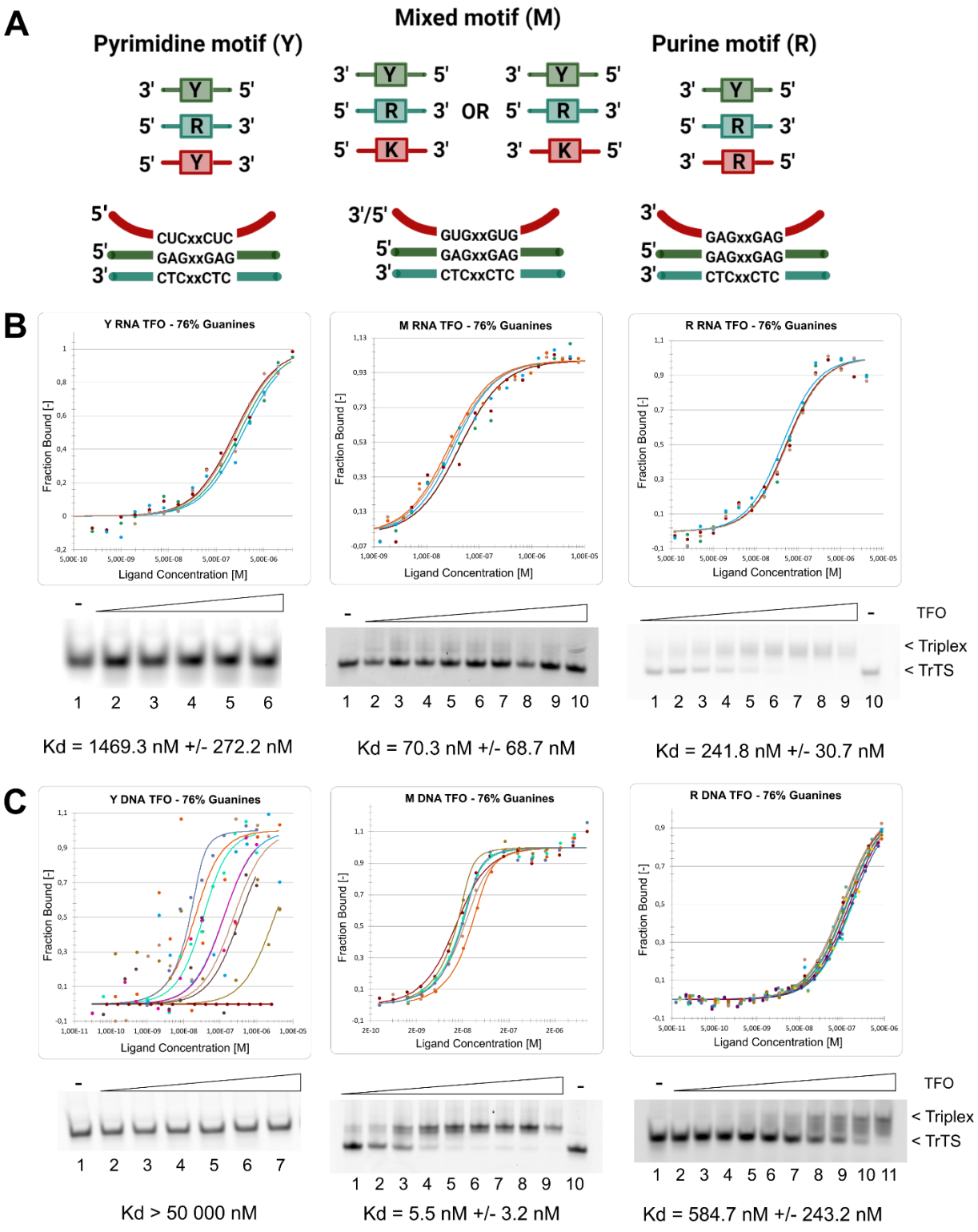



**Figure 21:** The triplex binding code. (A) Triplex motifs. (B-C) MST and EMSA measurements of RNA:DNA triplexes(B) and DNA:DNA triplexes (C). Y = C/G; R = A/U/T; K = G/U/T.

## 5.1.2 Binding affinity cut-off

Based on the number of RNA molecules per cell, the volume of the nucleus, and the *in vitro* binding affinities, I have chosen a loose cut-off for triplex formation (Equation 2). Since many lncRNAs are nucleus specific (Sun, Hao and Prasanth, 2018), it is reasonable to expect the RNA concentration in the nucleus, where triplex formation occurs, will be higher than in the whole cell.

**Equation 2:** Calculation for the number of RNA molecules required per binding affinity**.**

$$\text{Number of RNA molecules} = (\text{"Triplex binding affinity"}\left(\frac{\text{mol}}{\text{L}}\right) \text{ x "Volumen of the nucleus"}\left(\frac{\text{L}}{1}\right)) \text{ x } 6.022\text{E}23\ \left(\frac{1}{\text{mol}}\right)$$

The first cut-off is set at > 1µM, where more than 1440 RNA molecules are required for triplex formation, a requirement not met *in vivo* (Cabili *et al.*, 2015). Still, I consider this as a plausible binding *in vivo*, where stabilization factors, such as spermine concentration, salt concentration, proteins, crowding effect, and nucleosome stabilization help with triplex formation and stability. The second cut-off is set as > 50 µM, where 72024 RNA molecules are required for triplex binding. I cannot completely exclude *in vivo* triplex formation at Kd > 50 µM, but based on all the evidence currently available, it does not seem likely that a stable triplex will be formed. One can imagine a transient triplex formation, which could for a short time obscure binding of a protein or TF to a specific site, but not sufficient, for example, for recruitment of regulatory machinery.

## 5.1.3 Pyrimidine motif

### Guanine content of the TrTS

I investigated the influence of Guanine percentage in the TrTS on pyrimidine triplex stability. The pyrimidine motif shows a high tolerance for different Guanine contents. Interestingly the binding affinity of DNA:DNA triplexes drops with lower Guanine percentage and Guanine content between 70-80% (Figure 21, C - left; Figure 22, A; Table 1; Figure S4). In contrast, RNA:DNA triplexes form at any Guanine percentage, with higher Guanine content slightly reducing the binding affinity (Figure 21, B – left; Figure 22, B; Table 2; Figure S4). At 28% Guanine composition, both DNA:DNA and RNA:DNA triplexes are able to form (Figure 22, C and D), albeit a triplex-specific band shift can only be observed for DNA:DNA triplexes (Figure 22, E), whereas only a faint band can be seen for the RNA:DNA triplexes (Figure 22, F).

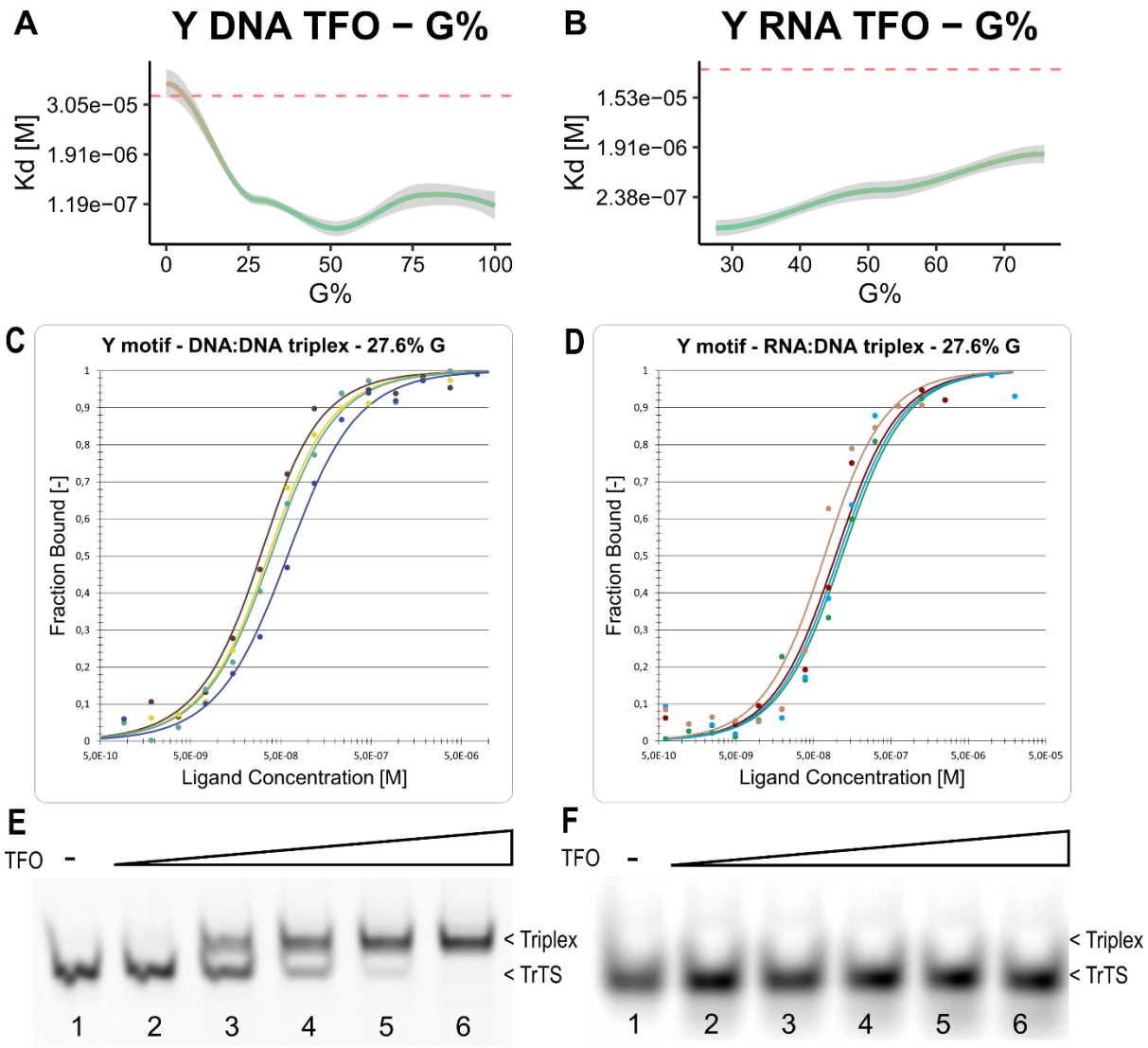



**Figure 22:** Summary of the binding affinities for the Pyrimidine motif across different triplex sequence Guanine compositions. (A-B) Binding affinity measurements of the pyrimidine motif with varying content of Guanines for (A) DNA:DNA triplex and (B) RNA:DNA triplex. (C-D) Binding affinity graph for DNA:DNA (C) and RNA:DNA (D) Pyrimidine motif triplexes at 27.6% Guanine content. (E-F) EMSA for DNA:DNA (E) and RNA:DNA (F) Pyrimidine motif triplexes at 27.6% Guanine content.

**Table 1**: $K_d$ values derived from microscale thermophoresis analysis of the Pyrimidine motif DNA:DNA triplexes with varying Guanine content. SD = Standard deviation. Replicates = Number of independent replicates. G% = Guanine percentage of the TrTS.

| G% | Kd [nM] | SD [nM] | Replicates |
|---|---|---|---|
| 0.0 | 100000.0 | 0.0 | 2 |
| 27.6 | 172.2 | 88.7 | 16 |
| 51.7 | 32.5 | 11.8 | 7 |
| 75.9 | 100000.0 | 0.0 | 3 |
| 100.0 | 114.7 | 45.4 | 2 |

**Table 2:** $K_d$ values derived from microscale thermophoresis analysis of the Pyrimidine motif RNA:DNA triplexes with varying Guanine content. SD = Standard deviation. Replicates = Number of independent replicates. G% = Guanine percentage of the TrTS.

| G% | Kd [nM] | SD [nM] | Replicates |
|---|---|---|---|
| 27.6 | 69.0 | 24.0 | 5 |
| 51.7 | 339.0 | 117.7 | 4 |
| 75.9 | 1469.3 | 272.2 | 4 |

**Length of the TFO**

Next, I looked at the influence of the TFO length on the triplex binding affinity.

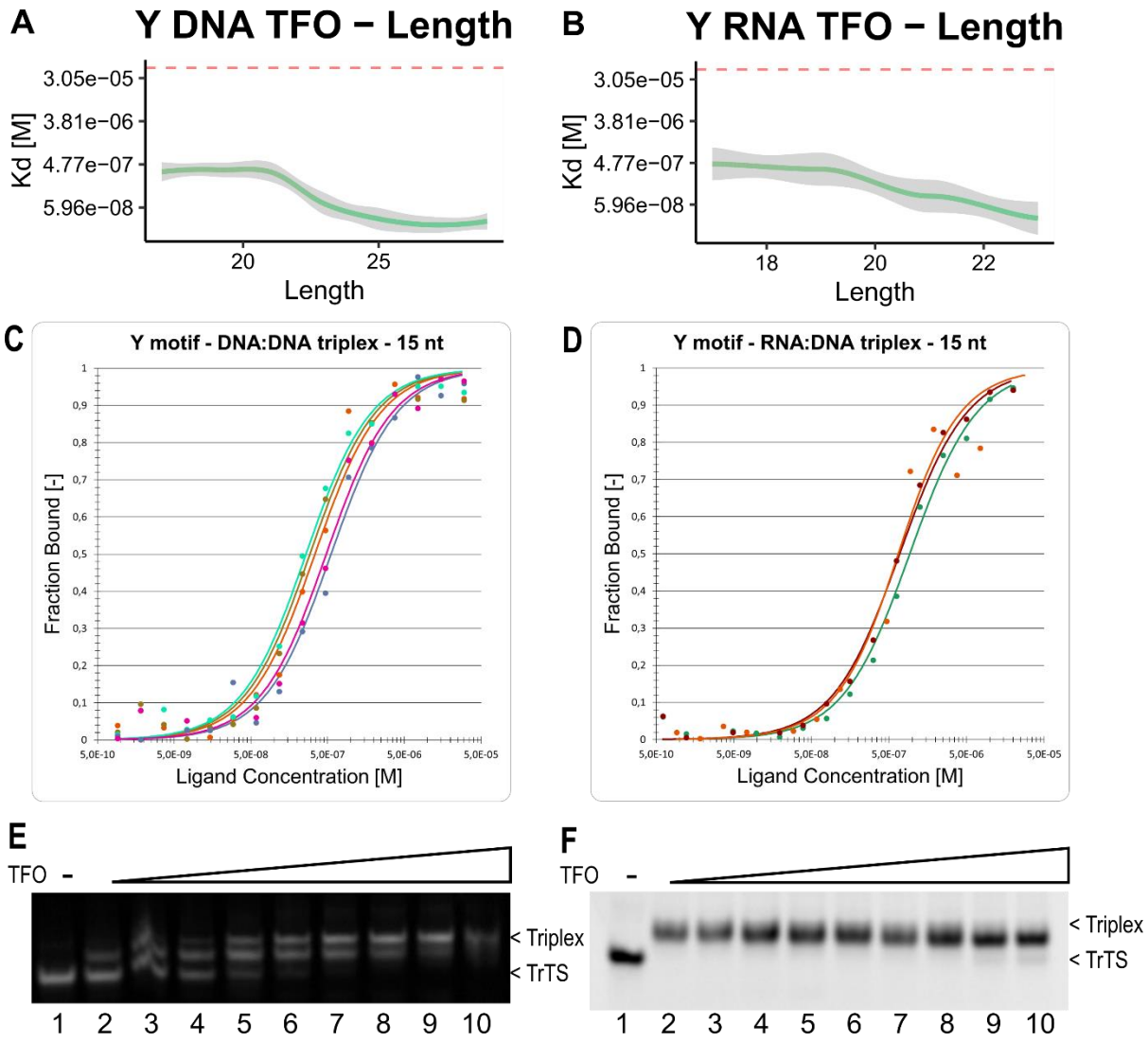



**Figure 23:** Summary of the binding affinities for the Pyrimidine motif across different TFO lengths. (A-B) Binding affinity measurements of the pyrimidine motif with varying lengths of TFO for (A) DNA:DNA triplex and (B) RNA:DNA triplex. (C-D) Binding affinity graph for DNA:DNA (C) and RNA:DNA (D)

Pyrimidine motif triplexes with 17 nt length. (E-F) EMSA for DNA:DNA (E) and RNA:DNA (F) Pyrimidine motif triplexes with 17 nt length.

For the pyrimidine motif, both RNA:DNA and DNA:DNA triplexes behave in a similar manner. All tested lengths, 17-29 nt, permitted triplex formation in the nM range, with minor differences in the binding affinities between DNA (Figure 23, A; Table 3; Figure S4) and RNA TFOs (Figure 23, B; Table 4; Figure S4). This is contrary to a recent publication asserting pyrimidine motif RNA:DNA triplexes cannot form when the length of the TFO is below 19 nt (Kunkler *et al.*, 2019). A good example is the TFO at 17 nt length, where both DNA:DNA (Figure 23, C) and RNA:DNA (Figure 23, D) triplexes form in the nM range (Table 3 and 4). Similarly, the EMSA experiments show a band shift, indicating triplex formation for both DNA:DNA (Figure 23, E) and RNA:DNA (Figure 23, F) triplexes. A third band can be observed in the EMSA experiments for DNA:DNA triplexes at 17nt length (Figure 23, E), which can be explained by the TrTS length (29 nt) and its mirror sequence (Table S2). Here multiple possible binding configurations are possible with at 17nt TFO. Importantly, only one band can be observed for the TrTS only well (Figure 23, E).

**Table 3:** $K_d$ values derived from microscale thermophoresis analysis of the Pyrimidine motif DNA:DNA triplexes with varying TFO length. SD = Standard deviation. Replicates = Number of independent replicates.

| Length | Kd [nM] | SD [nM] | Replicates |
|---|---|---|---|
| **17** | 374.2 | 125.8 | 5 |
| **19** | 301.5 | 78.4 | 6 |
| **21** | 337.1 | 60.6 | 5 |
| **23** | 117.5 | 112.9 | 4 |
| **25** | 35.5 | 10.9 | 4 |
| **27** | 27.3 | 12.6 | 4 |
| **29** | 32.5 | 11.8 | 7 |

**Table 4:** $K_d$ values derived from microscale thermophoresis analysis of the Pyrimidine motif RNA:DNA triplexes with varying TFO length. SD = Standard deviation. Replicates = Number of independent replicates.

| Length | Kd [nM] | SD [nM] | Replicates |
|---|---|---|---|
| **17** | 569.7 | 335.5 | 4 |
| **19** | 493.3 | 410.9 | 4 |
| **21** | 92.4 | 12.8 | 4 |
| **23** | 35.7 | 26.2 | 4 |

## 5.1.4 Mixed motif

**Guanine content of the TrTS**

The next motif I investigated is the mixed motif. The mixed motif has been largely neglected in biochemical studies, and as such, not much is known about the binding code nor the stability of mixed motif triplexes. First, I investigated the binding affinity of mixed motif triplexes with varying Guanine content. Compared to the pyrimidine motif, the mixed motif shows distinctly different behavior. The DNA:DNA triplex can form at almost all Guanine percentages, with the expectation of Guanine content below 35% (Figure 24, A; Table 5; Figure S5). The mixed motif has a higher G% cut-off than the pyrimidine motif, where binding can still be observed at 27% Guanines (Figure 22, A).

Conversely, the mixed motif does not form a triplex at 27% Guanines (Figure 22, C and D). Furthermore, the differences between the RNA and DNA TFOs are immense (Figure 24, A and B). RNA:DNA triplex shows a narrow tolerance for Guanine content (Figure 24, B). Specifically, binding is only permitted at 73-83% Guanines (Figure 24, B; Table 6; Figure S5).

An excellent example of this is at 52% Guanine percentage, where DNA:DNA triplexes form at a high binding affinity (Figure 24, C), whereas RNA:DNA triplexes have a very low binding affinity and quickly dissociate (Figure 24, D). This is corroborated with EMSA experiments, where a band shift can be observed for DNA:DNA triplexes (Figure E); however, RNA:DNA triplexes show only a faint band (Figure 24, F).

**Table 5:** $K_d$ values derived from microscale thermophoresis analysis of the Mixed motif DNA:DNA triplexes with varying Guanine content. SD = Standard deviation. Replicates = Number of independent replicates. G% = Guanine percentage of the TrTS.

| G% | Kd [nM] | SD [nM] | Replicates |
|---|---|---|---|
| 27.6 | 100000.0 | 0.0 | 3 |
| 51.7 | 14.2 | 5.8 | 6 |
| 75.9 | 5.5 | 3.2 | 6 |

**Table 6:** $K_d$ values derived from microscale thermophoresis analysis of the Mixed motif RNA:DNA triplexes with varying Guanine content. SD = Standard deviation. Replicates = Number of independent replicates. G% = Guanine percentage of the TrTS.

| G% | Kd [nM] | SD [nM] | Replicates |
|---|---|---|---|
| 27.6 | 100000.0 | 0.0 | 4 |
| 51.7 | 100000.0 | 0.0 | 4 |
| 69.0 | 428987.6 | 248802.2 | 3 |
| 71.4 | 1430900.0 | 1116978.6 | 3 |
| 75.9 | 70.3 | 68.7 | 7 |
| 86.2 | 428986.7 | 248798.4 | 3 |

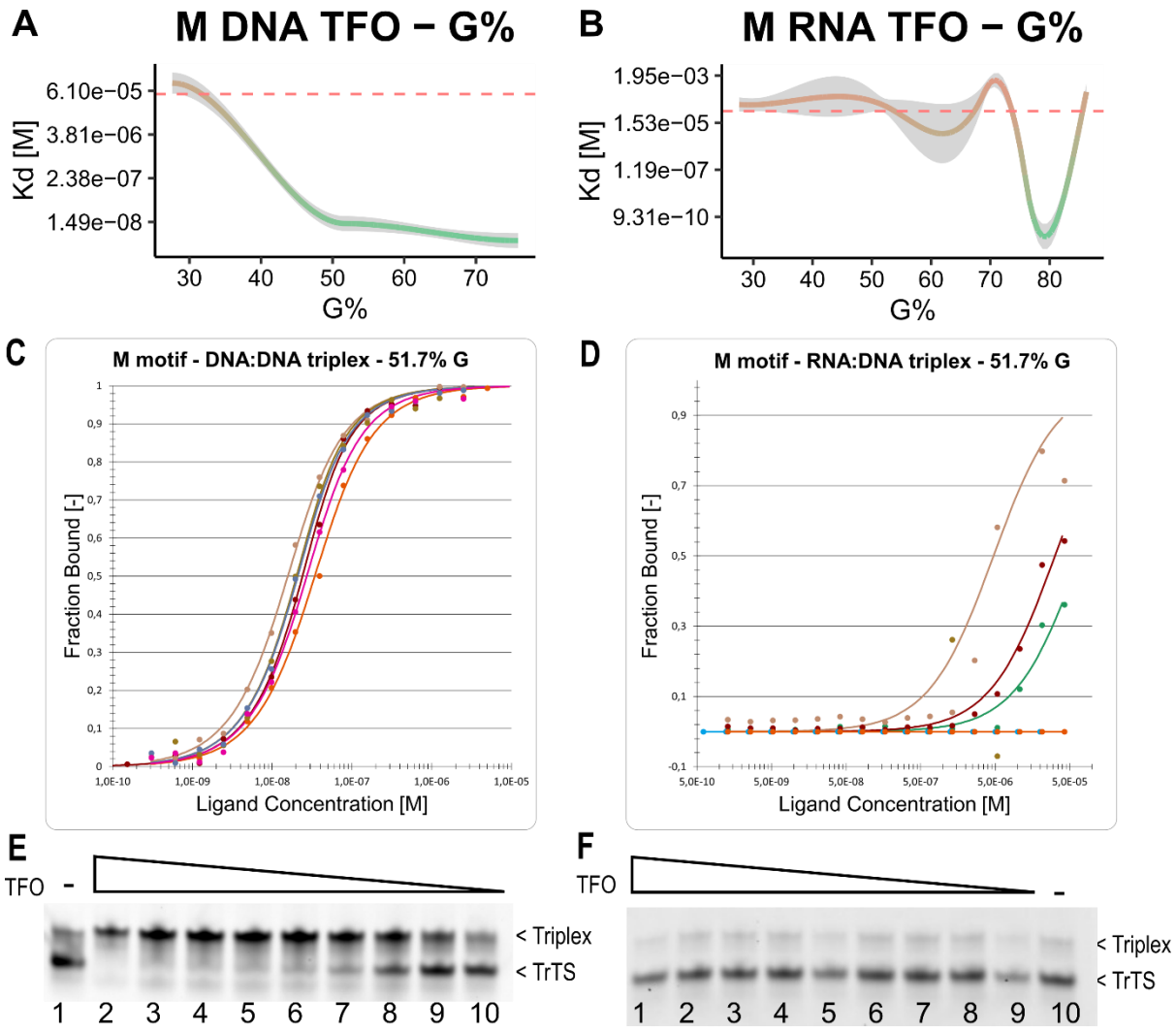



**Figure 24:** Summary of the binding affinities for the Mixed motif across different triplex sequence Guanine compositions. (A-B) Binding affinity measurements of the mixed motif with varying content of Guanines for (A) DNA:DNA triplex and (B) RNA:DNA triplex. (C-D) Binding affinity graph for DNA:DNA

(C) and RNA:DNA (D) Mixed motif triplexes at 51.7% Guanine content. (E-F) EMSA for DNA:DNA (E) and RNA:DNA (F) Mixed motif triplexes at 51.7% Guanine content.

**Length of the TFO**

Looking at the TFO length cut-off of the mixed motif, I observed RNA:DNA and DNA:DNA triplexes also differ in length constraints (Figure 25, A and B). For DNA:DNA, the length must be above 17 nt, with TFOs between 17 and 19 nt binding in the µM range (Figure 25, A; Table 7; Figure S5). On the contrary, RNA:DNA triplexes form above 21 nt, with a length between 21 and 23 nt already in the µM range (Figure 25, B and D; Table 8; Figure S5). This can also be seen with EMSA experiments, where DNA:DNA triplexes with a 21nt TFO show a typical triplex band (Figure 25, E), whereas no triplex band can be observed for the RNA:DNA triplex (Figure 25, F).

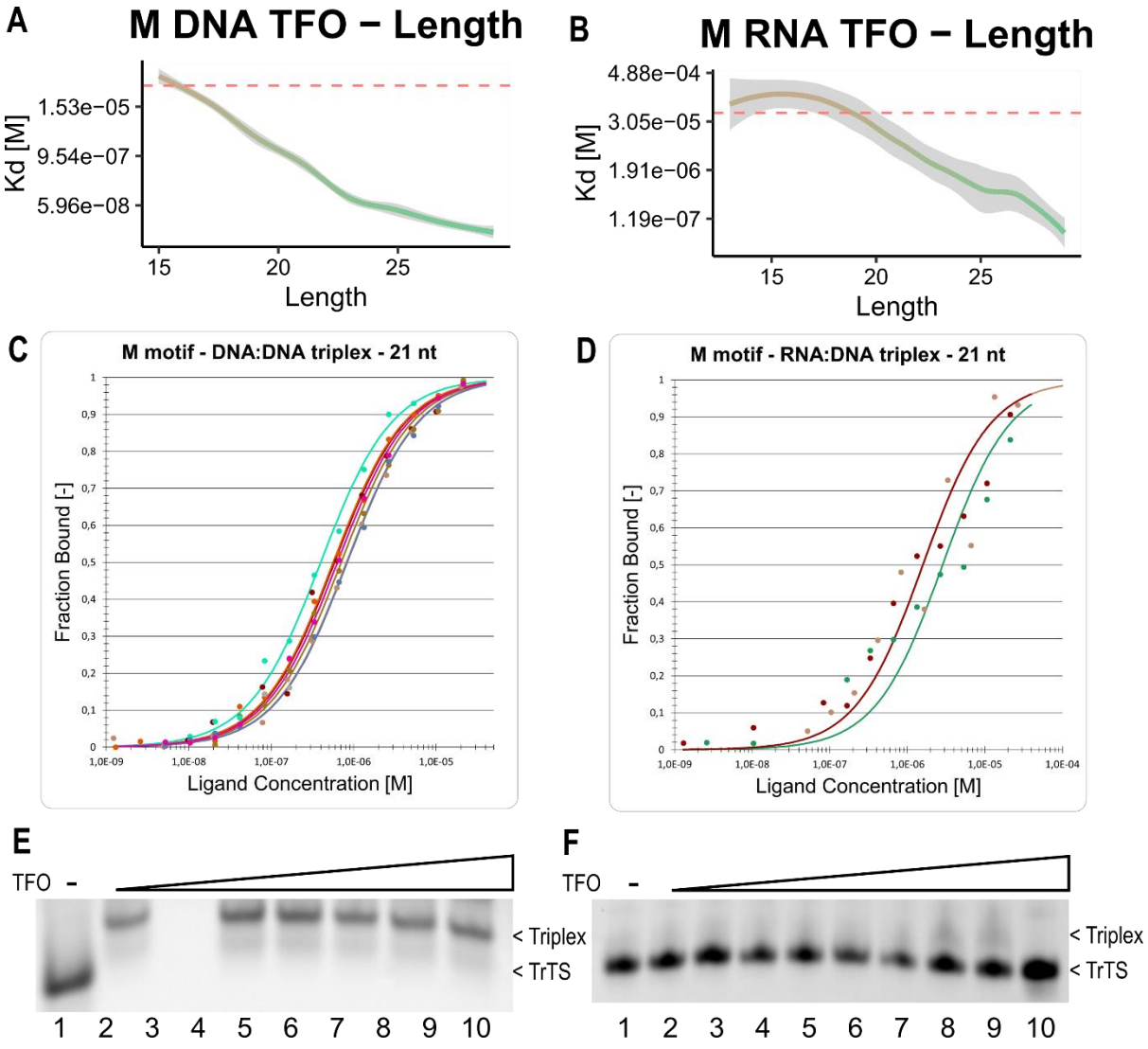



**Figure 25:** Summary of the binding affinities for the Mixed motif across different TFO lengths. (A-B) Binding affinity measurements of the mixed motif with varying lengths of TFO for (A) DNA:DNA triplex and (B) RNA:DNA triplex. (C-D) Binding affinity graph for DNA:DNA (C) and RNA:DNA (D) Mixed motif

triplexes with 21 nt length. (E-F) EMSA for DNA:DNA (E) and RNA:DNA (F) Mixed motif triplexes with 21 nt length.

**Table 7:** $K_d$ values derived from microscale thermophoresis analysis of the Mixed motif DNA:DNA triplexes with varying TFO length. SD = Standard deviation. Replicates = Number of independent replicates.

| Length | Kd [nM] | SD [nM] | Replicates |
|---|---|---|---|
| **15** | 80755.2 | 25008.1 | 4 |
| **17** | 28586.3 | 17366.2 | 7 |
| **19** | 3039.1 | 930.0 | 5 |
| **21** | 648.6 | 153.3 | 7 |
| **23** | 108.0 | 70.6 | 9 |
| **25** | 45.7 | 7.9 | 6 |
| **27** | 20.5 | 3.2 | 6 |
| **29** | 14.2 | 5.8 | 6 |

**Table 8:** $K_d$ values derived from microscale thermophoresis analysis of the Mixed motif RNA:DNA triplexes with varying TFO length. SD = Standard deviation. Replicates = Number of independent replicates.

| Length | Kd [nM] | SD [nM] | Replicates |
|---|---|---|---|
| **13** | 100000.0 | 0.0 | 2 |
| **15** | 100000.0 | 0.0 | 2 |
| **17** | 100000.0 | 0.0 | 4 |
| **19** | 318517.6 | 13307.0 | 2 |
| **21** | 2014.8 | 727.5 | 3 |
| **23** | 8339.2 | 385.3 | 3 |
| **25** | 661.4 | 45.3 | 3 |
| **27** | 903.5 | 95.2 | 3 |
| **29** | 70.3 | 68.7 | 7 |

## 5.1.5 Purine motif

### Guanine content of the TrTS

Lastly, I looked at the purine motif triplexes, starting with varying Guanine content in the TrTS. The purine motif permits vast differences in the Guanine content. Both DNA:DNA and RNA:DNA triplexes were able to form at all tested sequence compositions (Figure 26, A and B; Table 10; Figure S6). Nonetheless, some minor differences exist. For instance, RNA:DNA triplexes have a higher binding affinity at lower Guanine content (Figure 26, D), whereas DNA:DNA triplexes display a distinct preference for Guanine content around 50% and 100% (Figure 26, A and C). Surprisingly, no triplex-specific band shift could be observed for DNA:DNA (Figure 26, E) nor RNA:DNA (Figure 26, F) at lower Guanine percentages.

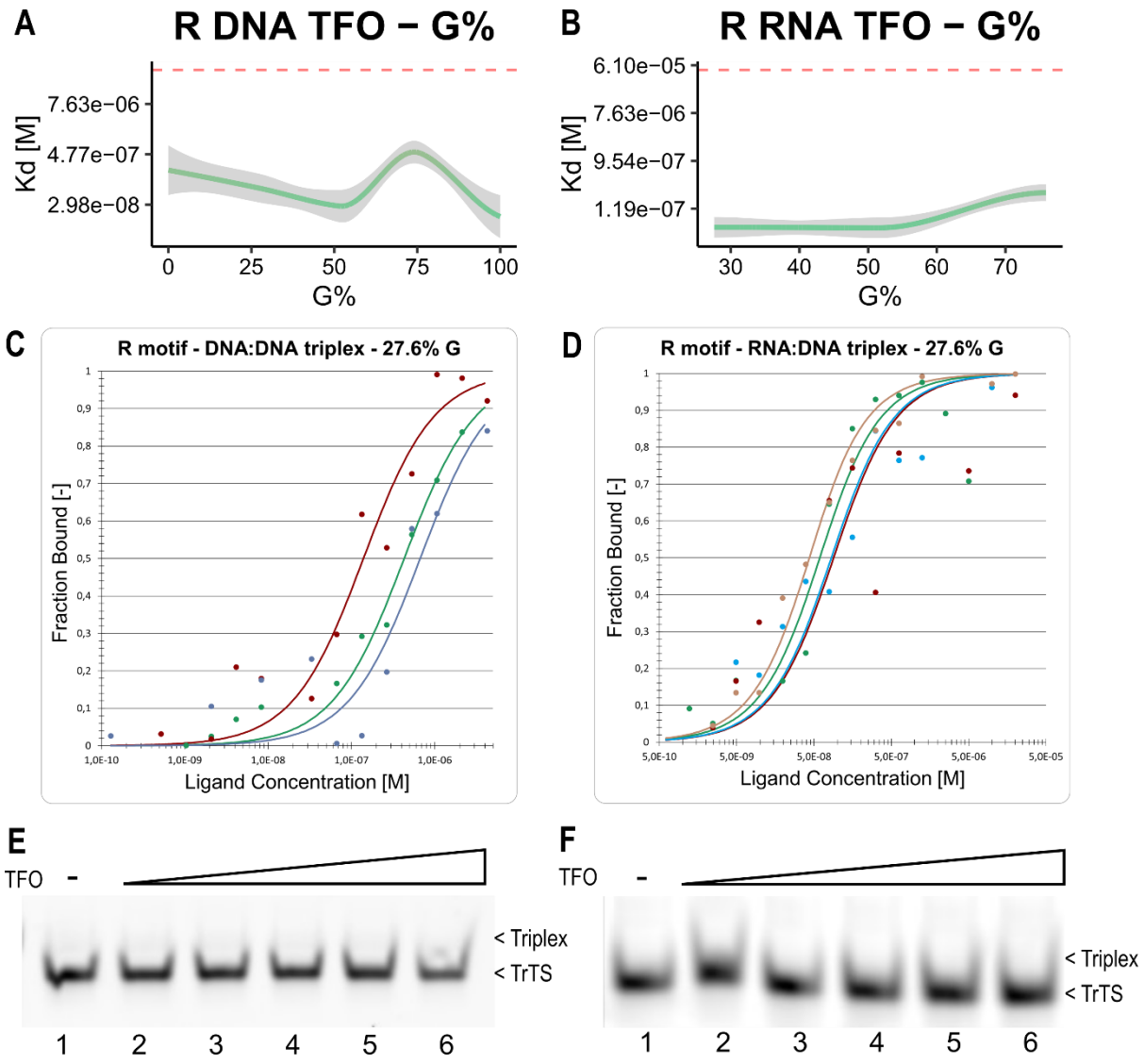



**Figure 26:** Summary of the binding affinities for the Purine motif across different triplex sequence Guanine compositions. (A-B) Binding affinity measurements of the purine motif with varying content

of Guanines for (A) DNA:DNA triplex and (B) RNA:DNA triplex. (C-D) Binding affinity graph for DNA:DNA (C) and RNA:DNA (D) Purine motif triplexes at 27.6% Guanine content. (E-F) EMSA for DNA:DNA (E) and RNA:DNA (F) Purine motif triplexes at 27.6% Guanine content.

**Table 9:** $K_d$ values derived from microscale thermophoresis analysis of the Purine motif DNA:DNA triplexes with varying Guanine content. SD = Standard deviation. Replicates = Number of independent replicates. G% = Guanine percentage of the TrTS.

| G% | Kd [nM] | SD [nM] | Replicates |
|---|---|---|---|
| 0.0 | 518.7 | 719.5 | 3 |
| 27.6 | 152.7 | 216.8 | 9 |
| 51.7 | 28.9 | 9.1 | 7 |
| 75.9 | 584.7 | 243.2 | 15 |
| 100.0 | 60.2 | 69.1 | 4 |

**Table 10:** $K_d$ values derived from microscale thermophoresis analysis of the Purine motif RNA:DNA triplexes with varying Guanine content. SD = Standard deviation. Replicates = Number of independent replicates. G% = Guanine percentage of the TrTS.

| G% | Kd [nM] | SD [nM] | Replicates |
|---|---|---|---|
| 27.6 | 56.6 | 21.4 | 4 |
| 51.7 | 59.1 | 29.2 | 4 |
| 75.9 | 241.8 | 30.7 | 6 |

### Length of the TFO

Similar to the Guanine content, the length dependence of the DNA:DNA and RNA:DNA triplexes of the purine motif is very similar. Both permit triplex formation above 11 nt, with triplexes at 13 nt having the binding affinity in the µM range and above 17 nt in the nM range (Figure 27, A and B). For DNA:DNA triplexes, the binding affinities are overall higher (Table 11; Figure S6) than for RNA:DNA triplexes (Table 12; Figure S6), which is in line with previously published data (Kunkler *et al.*, 2019). Here too, some discrepancies between the MST measurements and EMSA experiments can be observed (Figure 27 A and B; Figure S6). For instance, looking at 15nt TFO length, DNA:DNA triplexes have a binding affinity of 255 nM +/- 37 nM (Figure 27, C; Table 11), while RNA:DNA triplexes have a binding affinity of 7 nM +/- 1.6 nM (Figure 27, D; Table 12). Conversely, a triplex-specific band shift could only be observed for RNA:DNA triplexes (Figure 27, F) and not for DNA:DNA triplexes (Figure 27, E).

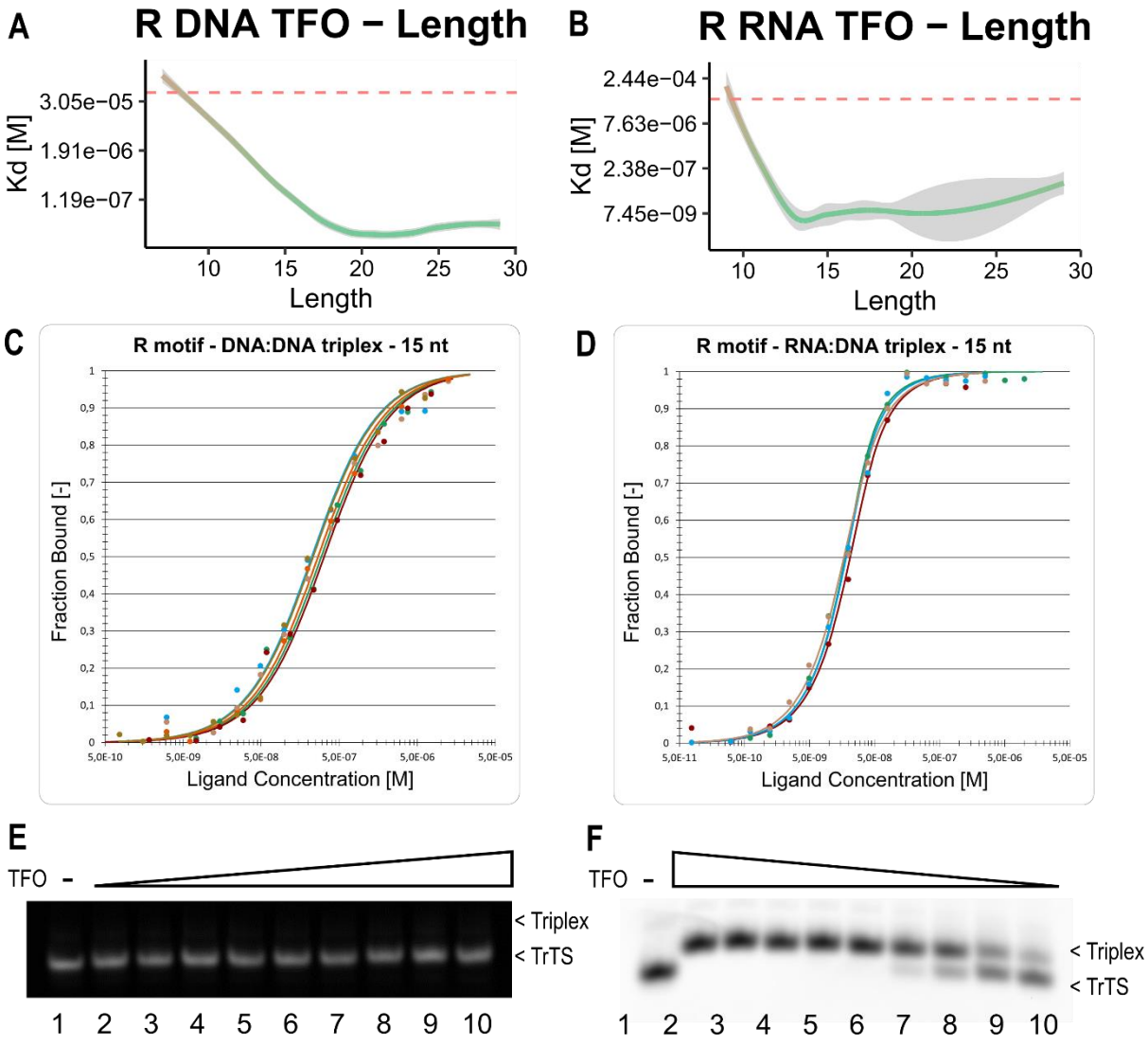



**Figure 27:** Summary of the binding affinities for the Purine motif across different TFO lengths. (A-B) Binding affinity measurements of the purine motif with varying lengths of TFO for (A) DNA:DNA triplex and (B) RNA:DNA triplex. (C-D) Binding affinity graph for DNA:DNA (C) and RNA:DNA (D) Purine motif triplexes with 15 nt length. (E-F) EMSA for DNA:DNA (E) and RNA:DNA (F) Purine motif triplexes with 15 nt length.

**Table 11:** $K_d$ values derived from microscale thermophoresis analysis for the Purine motif DNA:DNA triplexes with varying TFO length. SD = Standard deviation. Replicates = Number of independent replicates.

| Length | Kd [nM] | SD [nM] | Replicates |
|---|---|---|---|
| 7 | 100000.0 | 0.0 | 4 |
| 9 | 100000.0 | 0.0 | 2 |
| 11 | 4410.0 | 1530.1 | 6 |
| 15 | 254.9 | 37.0 | 6 |
| 17 | 42.2 | 6.4 | 6 |
| 19 | 22.4 | 9.7 | 6 |
| 21 | 16.5 | 4.5 | 6 |
| 23 | 19.4 | 5.3 | 6 |
| 25 | 23.2 | 8.7 | 6 |
| 27 | 39.8 | 14.4 | 6 |
| 29 | 28.9 | 9.1 | 7 |

**Table 12:** $K_d$ values derived from microscale thermophoresis analysis for the Purine motif RNA:DNA triplexes with varying TFO length. SD = Standard deviation. Replicates = Number of independent replicates.

| Length | Kd [nM] | SD [nM] | Replicates |
|---|---|---|---|
| 9 | 100000.0 | 0.0 | 2 |
| 11 | 1591.7 | 1746.4 | 3 |
| 13 | 8.2 | 8.0 | 4 |
| 15 | 7.0 | 1.6 | 4 |
| 17 | 9.4 | 0.8 | 5 |
| 19 | 8.5 | 1.3 | 5 |
| 29 | 115.6 | 115.4 | 4 |

### 5.1.6 Published triplexes

Next, I wanted to test published triplexes in a standardized manner (salt concentration, pH, etc.). In total, I tested fourteen triplex pairs from ten publications. Interestingly, these primarily represent the purine motifs with varying lengths (Table 13). For two of the triplex pairs, CCL2 + LNMAT and miR-24-

1, a triplex motif could not be determined as they do not resemble any of the three triplex motifs (Appendix 2). Not surprisingly, neither of the triplex pairs could form a triplex (Table 13 and Figure S7).

Half of the tested triplex pairs did not form a triplex in our hands, with four triplex pairs exhibiting a binding affinity in the µM range and four in the nM range. I cannot completely exclude the possibility of their formation in the cell, where various stabilization factors exist. Nonetheless, their formation seems unlikely based on the newly defined triplex binding code. One explanation for the discrepancy between the published data and my findings could be that these lncRNAs form R-loops or target chromatin via protein binding partners instead. For instance, lncRNA SARRAH, has been tested with an R-loop specific antibody, S9.6, which would explain the lack of triplex formation (Trembinski *et al.*, 2020). Another possibility is the buffer conditions used in these studies, where lower pH or non-physiological salt concentrations are often used to induce triplex formation.

**Table 13**: $K_d$ values derived from microscale thermophoresis analysis of the published triplex pairs. SD = Standard deviation. Replicates = Number of independent replicates. G% = Guanine percentage of the TrTS.

| triplex name | Kd [nM] | SD [nM] | Replicates | Motif | Length | G% |
|---|---|---|---|---|---|---|
| BCL9 + RP1184A10 | 312.4 | 103.3 | 6 | R | 39 | 51.3 |
| CCL2 + LNMAT | 100000.0 | 0.0 | 3 | ? | 16 | 50.0 |
| FG1 + AG30 | 100000.0 | 0.0 | 3 | M | 30 | 76.7 |
| KIAA1324 + RP1184A10 | 58.6 | 25.1 | 6 | R | 39 | 51.3 |
| miR-24-1 | 100000.0 | 0.0 | 3 | ? | 22 | 31.8 |
| miR-483-5p | 100000.0 | 0.0 | 3 | R | 22 | 50.0 |
| PCDH7 | 543.9 | 382.1 | 3 | R | 16 | 56.2 |
| PITX2 + FENDRR | 100000.0 | 0.0 | 3 | R | 41 | 65.9 |
| PROX1 + RP11-84A10 | 139.7 | 36.2 | 6 | R | 39 | 51.3 |
| SARRAH | 100000.0 | 0.0 | 4 | Y | 15 | 60.0 |
| SMAD2 + MEG3 | 1542.0 | 228.0 | 6 | R | 12 | 55.0 |
| TGFBR1 + MEG3 | 12087.7 | 4645.1 | 16 | R | 16 | 55.0 |
| TGFBR2 + MEG3 | 100000.0 | 0.0 | 10 | R | 19 | 55.0 |
| WWOX1 + PARTICLE_1 | 36007.1 | 31991.0 | 4 | M | 16 | 62.5 |
| WWOX2 + PARTICLE_2 | 2543.0 | 1162.3 | 5 | M | 16 | 50.0 |

### 5.1.7 Mismatches

Mismatches in the triplex motif and non-canonical base pairs significantly impact the triplex stability and formation (Maldonado *et al.*, 2017; Kunkler *et al.*, 2019). I sought to determine if the position of the mismatch had a differential effect on triplex stability.

I observed a higher sensitivity to mismatches in the middle region of the triplex (base triplet 12 and 19 from 29), with almost no difference in binding affinity from mismatches at the sides (base triplet 5 and 28 of 29) (Figure 28; Table 14; Figure S8). Only T12C mismatch completely abolished triplex formation, while G12C and A12C mismatches at the same position merely decreased the binding affinity (Table 14). Interestingly, at position 19, mismatches had a significant influence on the binding affinity; however, none of the mismatches abolished triplex formation.

**Table 14:** $K_d$ values derived from microscale thermophoresis analysis of pyrimidine DNA:DNA triplex with varying location and identity of mismatches. SD = Standard deviation. Replicates = Number of independent replicates.

| Mismatch | Kd [nM] | SD [nM] | Replicates |
|---|---|---|---|
| **5A** | 25.9 | 8.9 | 3 |
| **5G** | 19.6 | 3.9 | 3 |
| **5T** | 16.6 | 8.4 | 3 |
| **12A** | 264.7 | 41.9 | 3 |
| **12G** | 300.7 | 136.5 | 3 |
| **12T** | 72633.3 | 47400.5 | 3 |
| **19A** | 466.8 | 125.4 | 3 |
| **19G** | 665.4 | 259.7 | 3 |
| **19C** | 1009.5 | 252.8 | 3 |
| **28A** | 19.7 | 6.5 | 3 |
| **28G** | 8.8 | 2.8 | 3 |
| **28T** | 17.7 | 3.6 | 3 |

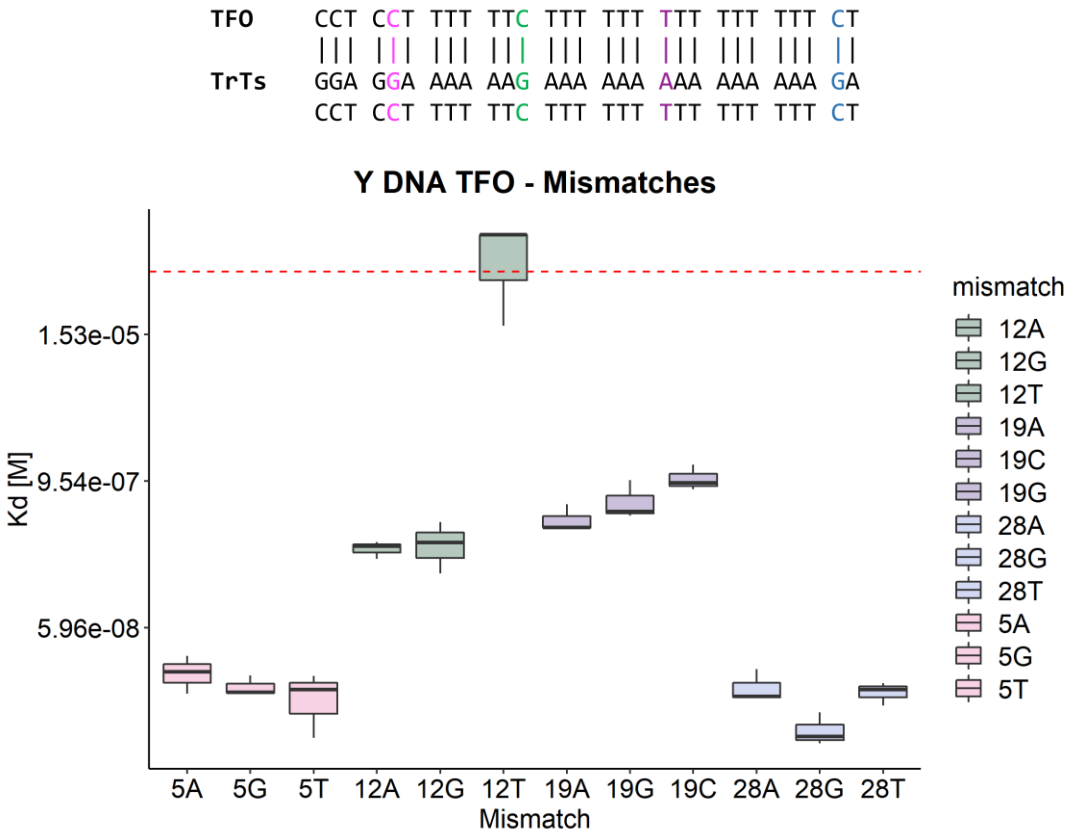



**Figure 28:** Summary of $K_d$ values for different mismatches in the triplex motif, with mismatch positions indicated above.

## 5.1.8 High throughput binding affinity studies

Ultimately, testing each possible triplex pair is labor-intensive. Thus, I have developed a high throughput biochemical method, termed Triplex-SELEX-EMSA-Sequencing (TSE-Seq) method, to study triplex binding affinities in a comprehensive manner. The TSE-seq method additionally allows for investigating molecular competition in triplex formation. One downside of the TSE-seq method is that triplex binding affinities are relative to each other and cannot be determined in absolute numbers. For this reason, the TSE-seq is used to complement the MST measurements and not replace them. Currently, no methods for genome-wide *in vivo* investigation of triplex pairs exist. However, two *in vivo* methods have been developed, giving information on either only TrTSs (Maldonado *et al.*, 2019) or TFOs and TrTSs (C.-C. Kuo *et al.*, 2019), but not the pairs. As such, these methods are not appropriate for investigating the triplex binding code.

## TSE-seq design

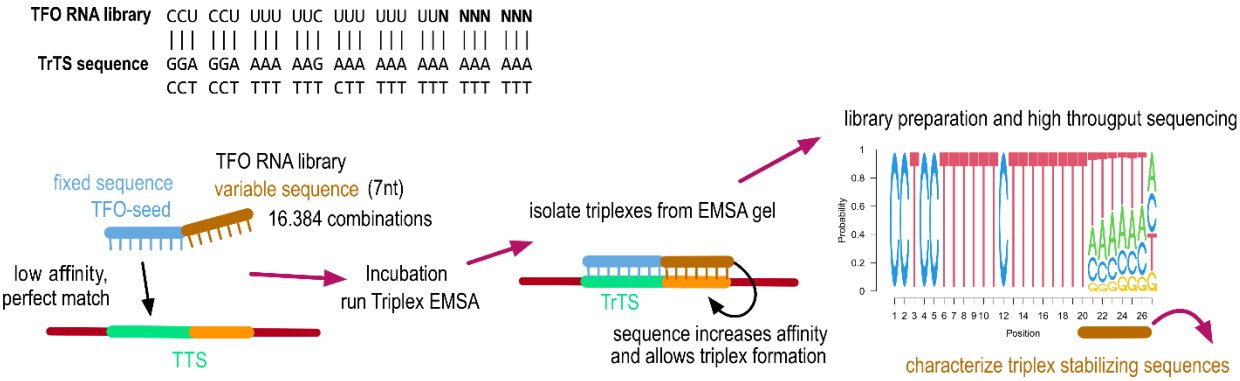



**Figure 29:** TSE-seq design. Thymine represents a Uracil in the TSE-seq experiment. Figure created by Dr. Gernot Längst and modified with permission.

I designed an RNA TFO library with 27nt long sequences with a 20 nt base (seed) region and a 7 nt randomized (variable) region (Figure 29). I used the EN3 motif, for which ample data exists, including my data on mismatches (Figure 28). The 20nt seed region formed a triplex with a very low binding affinity. Thus, I was able to study which sequences will stabilize the triplex formation and increase the binding affinity.

## TSE-seq results

Remarkably, the complemental TFO (CCUCCUUUUUUCUUUUUUUUUUUUUUU) was not present in the top 30 enriched motifs, even though it was 8-fold enriched in the triplex library over the input control (Number 13704) (Table 15). Still, the top 15 enriched TFOs tended to complete the pyrimidine motif with Uracil at positions 21, 22, and 23 (Figure 30, A). Starting at the fifth top enriched TFO, position 23 became permissible for a Guanine or Adenine, forming purine base triplets. This mismatch in the pyrimidine motif exhibited only a minor reduction in triplex stability (Figure 30, A). Conversely, base triplets at positions 24 – 27 did not significantly impact the triplex stability and exhibited only a minor tendency for Adenines at positions 24 and Cytosines at positions 27 (Figure 30, A). Top 30 enriched TFOs, showed a higher tendency to complete the EN3 motif until position 24. With a preference for Cytosines, Uracils, or Adenines at the last three bases (positions 24- 27) (Figure 30, B and Table 15).

Next, I looked at the motifs enriched at least 10-times over the input TFO library (n = 1102) (Figure 31, B) or at least 32-times (n = 97) (Figure 31, C) and compared those with all enriched TFOs (Figure 31, A). I used the Jensen-Shannon divergence, which is a metric of similarity between two probability distributions, to determine how similar or dis-similar the triplex and input libraries are. I found that for all enriched triplex motifs (Figure 31, A), positions 24, 25, and 26 did have a large impact on triplex

stability, albeit they significantly differed between the two libraries. This was consistent between replicates. Unsurprisingly, I found the triplex library had a much higher presence of Cytosines and Uracils at positions 21, 22 and 27. Additionally, position 23 exhibited a preference for Adenines and Uracils.

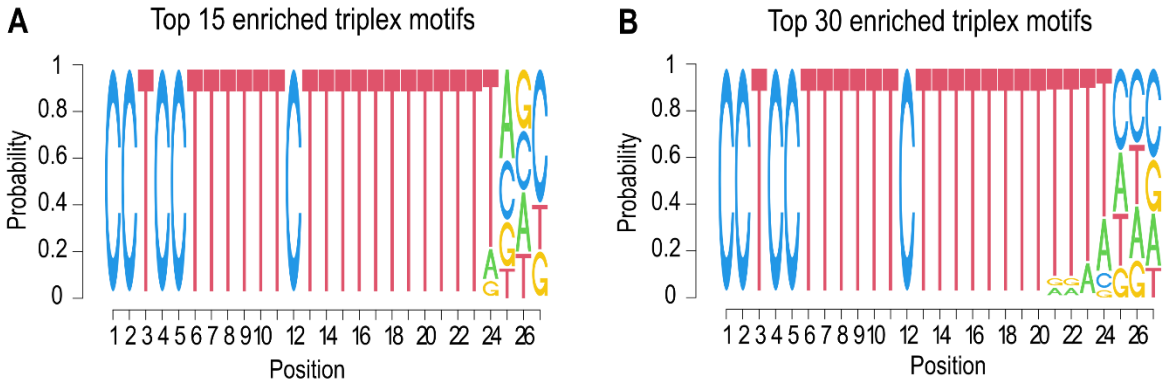



**Figure 30**: Motif of TSE-seq top enriched triplexes for top 15 (A) and top 30 (B) TFOs. Thymine (T) represents a Uracil (U) in the TSE experiment.

Contrarily, highly enriched triplex motifs showed a high tendency to complete the complementary pyrimidine motif and form canonical base triplets (Figure 31, B and C). Unexpectedly, I also found a high tendency for Adenines, forming a canonical purine motif, A:A-T (Figure 31, B and C). Triplex motifs with the highest enrichments showed no tolerance for changes in the triplex motif at positions 21 and 22 (Figure 31, C), as previously observed (Figure 30, A). Positions towards the end of the motif showed less weight on the triplex stability than positions 21 – 23 (Figure 31, C).

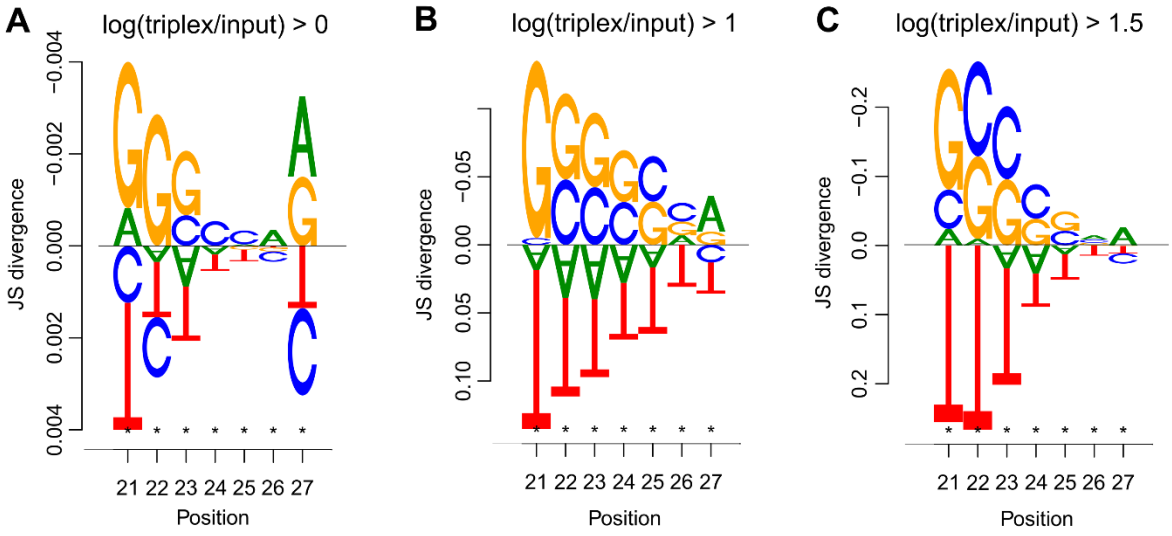



**Figure 31:** TSE-seq motifs. (A)Triplex motifs enriched over input TFO library. (B) Triplexes with more than 1.0 log10 fold enrichment over TFO input library. (C) Triplexes with more than 1.5 log10 fold enrichment over TFO input library. Thymine represents a Uracil in the TSE experiment. Positions marked with a star * have a statistically significant difference between the triplex and input libraries.

**Table 15:** Sequences of the top 30 TFOs. TPE = Number of TFOs per motif, normalized to the sequencing library size. Thymine (T) represents a Uracil (U) in the TSE experiment.

| # | Name | Seq | TPM (Input) | TPM (Triplex) | log(Triplex/Input) | Triplex/Input |
|---|---|---|---|---|---|---|
| 1 | Random_seq_22939 | CCTCCTTTTTTCTTTTTTTTTTTTACC | 1.42 | 251.95 | 2.25 | 176.95 |
| 2 | Random_seq_7616 | CCTCCTTTTTTCTTTTTTTTTTTTGTC | 1.40 | 191.48 | 2.13 | 136.43 |
| 3 | Random_seq_3308 | CCTCCTTTTTTCTTTTTTTTTTTTATG | 1.47 | 172.70 | 2.07 | 117.85 |
| 4 | Random_seq_26346 | CCTCCTTTTTTCTTTTTTTTTTTTACG | 4.03 | 431.72 | 2.03 | 107.04 |
| 5 | Random_seq_16746 | CCTCCTTTTTTCTTTTTTTTTTTGCTC | 1.51 | 159.96 | 2.03 | 106.27 |
| 6 | Random_seq_12915 | CCTCCTTTTTTCTTTTTTTTTTTTGAC | 1.98 | 193.66 | 1.99 | 97.78 |
| 7 | Random_seq_7971 | CCTCCTTTTTTCTTTTTTTTTTTATGC | 1.82 | 177.50 | 1.99 | 97.32 |
| 8 | Random_seq_1087 | CCTCCTTTTTTCTTTTTTTTTTTCAG | 3.51 | 321.38 | 1.96 | 91.67 |
| 9 | Random_seq_31901 | CCTCCTTTTTTCTTTTTTTTTTTTCC | 2.03 | 178.20 | 1.94 | 87.75 |
| 10 | Random_seq_42216 | CCTCCTTTTTTCTTTTTTTTTTTAGC | 3.98 | 340.57 | 1.93 | 85.55 |
| 11 | Random_seq_8736 | CCTCCTTTTTTCTTTTTTTTTTTCAC | 2.54 | 199.79 | 1.89 | 78.52 |
| 12 | Random_seq_35818 | CCTCCTTTTTTCTTTTTTTTTTAAAC | 1.53 | 120.28 | 1.89 | 78.38 |
| 13 | Random_seq_31237 | CCTCCTTTTTTCTTTTTTTTTTTGCT | 2.60 | 202.62 | 1.89 | 78.00 |
| 14 | Random_seq_13649 | CCTCCTTTTTTCTTTTTTTTTTTAGT | 1.00 | 77.29 | 1.89 | 77.04 |
| 15 | Random_seq_14795 | CCTCCTTTTTTCTTTTTTTTTTTCGT | 1.49 | 114.25 | 1.89 | 76.80 |
| 16 | Random_seq_12027 | CCTCCTTTTTTCTTTTTTTTTTTCCA | 2.37 | 180.74 | 1.88 | 76.28 |
| 17 | Random_seq_8210 | CCTCCTTTTTTCTTTTTTTTTTTAGTC | 1.39 | 101.56 | 1.86 | 73.22 |
| 18 | Random_seq_39484 | CCTCCTTTTTTCTTTTTTTTTTATCCA | 1.55 | 111.72 | 1.86 | 72.01 |
| 19 | Random_seq_37373 | CCTCCTTTTTTCTTTTTTTTTTTCTCG | 1.74 | 117.46 | 1.83 | 67.67 |
| 20 | Random_seq_2971 | CCTCCTTTTTTCTTTTTTTTAAAAAAT | 1.59 | 106.51 | 1.83 | 66.86 |
| 21 | Random_seq_21344 | CCTCCTTTTTTCTTTTTTTTTTTACGA | 2.83 | 177.82 | 1.80 | 62.86 |
| 22 | Random_seq_24006 | CCTCCTTTTTTCTTTTTTTTTTTTCTG | 3.00 | 187.88 | 1.80 | 62.64 |
| 23 | Random_seq_8384 | CCTCCTTTTTTCTTTTTTTTTTAATCA | 1.46 | 91.32 | 1.80 | 62.37 |
| 24 | Random_seq_17304 | CCTCCTTTTTTCTTTTTTTTTTTTCCG | 4.93 | 304.92 | 1.79 | 61.81 |
| 25 | Random_seq_19049 | CCTCCTTTTTTCTTTTTTTTTTTACTC | 1.49 | 88.38 | 1.77 | 59.30 |
| 26 | Random_seq_3195 | CCTCCTTTTTTCTTTTTTTTTTACTCA | 1.66 | 97.39 | 1.77 | 58.74 |
| 27 | Random_seq_3928 | CCTCCTTTTTTCTTTTTTTTTTTTAAC | 1.53 | 88.89 | 1.76 | 57.97 |
| 28 | Random_seq_11491 | CCTCCTTTTTTCTTTTTTTTGTTTTTG | 2.76 | 150.39 | 1.74 | 54.41 |
| 29 | Random_seq_3261 | CCTCCTTTTTTCTTTTTTTTTTTTCAA | 1.37 | 73.10 | 1.73 | 53.34 |
| 30 | Random_seq_5422 | CCTCCTTTTTTCTTTTTTTTTGTTTTA | 2.53 | 133.31 | 1.72 | 52.66 |

Negatively enriched triplex motifs, or enriched motifs in the input TFO library over the triplex samples, revealed which bases are detrimental to the triplex formation (Figure 32, A and B). For example, Guanines at position 27 seem to highly reduce triplex stability (Figure 32, A and B). Furthermore, Adenines at positions 21 and, to some extent, at position 22 also reduce triplex stability (Figure 32, A). This does not mean triplex formation is not possible with these sequences, as these were also present in the triplex library. But it does tell us which positions are permissible for mismatches and which positions are more sensitive to mismatches and which ones. A great example is the 3$^{rd}$ sequence in Table 15, with the variable region: AAAAAAT. This sequence was highly enriched in the triplex samples (67-fold enrichment), despite containing an Adenine at positions 21 and 22. In the same manner, the 3$^{rd}$, 4$^{th}$, 8$^{th}$, 19$^{th}$, 22$^{nd}$, 24$^{th,}$ and 28$^{th}$ sequences all contain Guanine at position 27 and are more than

50-fold enriched (Table 15). This suggests that the identity of the neighboring bases has a high impact on the triplex stability and is not independent of the mismatch. Looking at the top 25 negatively enriched sequences, we can see all variable regions contain a high percentage of Guanines, and almost all sequences have Guanine at position 27 (Table 16).

**Table 16:** Sequences of the top negative binders. TPE = Number of TFOs per motif, normalized to the sequencing library size. Thymine (T) represents a Uracil (U) in the TSE experiment.

| # | Name | Sequence | TPM(Input) | TPM(Triplex) | log(triplex/input) | triplex/input |
|---|---|---|---|---|---|---|
| 1 | Random_seq_55589 | CCTCCTTTTTTCTTTTTTTTAGTGGCG | 269.432452 | 15.99523 | -1.2264594142 | 0.0593663825 |
| 2 | Random_seq_2986 | CCTCCTTTTTTCTTTTTTTTAGGCCAG | 189.037476 | 11.901695 | -1.2009390934 | 0.0629594473 |
| 3 | Random_seq_5925 | CCTCCTTTTTTCTTTTTTTTGTTGGCG | 409.295185 | 25.791771 | -1.2005554715 | 0.0630150853 |
| 4 | Random_seq_25512 | CCTCCTTTTTTCTTTTTTTTGTGTCGG | 155.840787 | 10.205702 | -1.1838382497 | 0.0654880035 |
| 5 | Random_seq_1178 | CCTCCTTTTTTCTTTTTTTTGGCTCAG | 379.979727 | 26.381427 | -1.1584621431 | 0.0694285119 |
| 6 | Random_seq_24141 | CCTCCTTTTTTCTTTTTTTTAGCGGTG | 152.838381 | 11.555136 | -1.1214573673 | 0.0756036273 |
| 7 | Random_seq_42297 | CCTCCTTTTTTCTTTTTTTTAAGCGCG | 388.597988 | 31.142235 | -1.0961507701 | 0.08013998 |
| 8 | Random_seq_21199 | CCTCCTTTTTTCTTTTTTTTGCGACTG | 266.827781 | 21.669632 | -1.0903795085 | 0.0812120534 |
| 9 | Random_seq_3415 | CCTCCTTTTTTCTTTTTTTTGAGCCAG | 388.403599 | 31.742741 | -1.0876388199 | 0.0817261763 |
| 10 | Random_seq_2891 | CCTCCTTTTTTCTTTTTTTTGTGTCGA | 264.587744 | 22.101035 | -1.078157111 | 0.0835300784 |
| 11 | Random_seq_19323 | CCTCCTTTTTTCTTTTTTTTGCTGGTG | 264.547838 | 22.623259 | -1.0679490491 | 0.0855167034 |
| 12 | Random_seq_27031 | CCTCCTTTTTTCTTTTTTTTGTCGGTG | 140.537255 | 12.077921 | -1.0657992819 | 0.0859410624 |
| 13 | Random_seq_11100 | CCTCCTTTTTTCTTTTTTTTGACGCTG | 313.286882 | 26.991702 | -1.0647119399 | 0.0861565024 |
| 14 | Random_seq_10376 | CCTCCTTTTTTCTTTTTTTTGTGCCTG | 358.836449 | 31.816934 | -1.0522382232 | 0.0886669514 |
| 15 | Random_seq_1563 | CCTCCTTTTTTCTTTTTTTTATGGCGA | 176.597857 | 15.745804 | -1.049820588 | 0.08916192 |
| 16 | Random_seq_57072 | CCTCCTTTTTTCTTTTTTTTGTGGCTG | 441.264384 | 40.084124 | -1.0417264785 | 0.0908392462 |
| 17 | Random_seq_5439 | CCTCCTTTTTTCTTTTTTTTAGACGCG | 199.788814 | 18.28548 | -1.0384648036 | 0.091524043 |
| 18 | Random_seq_1203 | CCTCCTTTTTTCTTTTTTTTGATGGCG | 303.037615 | 27.778438 | -1.0377887179 | 0.0916666335 |
| 19 | Random_seq_15072 | CCTCCTTTTTTCTTTTTTTTGGACGCG | 165.043498 | 15.192237 | -1.0359766929 | 0.0920498971 |
| 20 | Random_seq_24445 | CCTCCTTTTTTCTTTTTTTTGGTGCG | 117.361585 | 11.036156 | -1.0267081354 | 0.0940355057 |
| 21 | Random_seq_26786 | CCTCCTTTTTTCTTTTTTTTGGCCTAG | 303.296434 | 28.600546 | -1.02549298 | 0.0942989854 |
| 22 | Random_seq_74232 | CCTCCTTTTTTCTTTTTTTTCGGTGAT | 32.470439 | 3.063945 | -1.0252071952 | 0.0943610587 |
| 23 | Random_seq_31966 | CCTCCTTTTTTCTTTTTTTTAGTGCTG | 152.70147 | 14.824594 | -1.0128604097 | 0.0970821957 |
| 24 | Random_seq_23540 | CCTCCTTTTTTCTTTTTTTTTAGGCG | 25.61264 | 2.492927 | -1.011744784 | 0.0973319033 |
| 25 | Random_seq_52810 | CCTCCTTTTTTCTTTTTTTTGGTCACG | 232.748576 | 22.946752 | -1.0061658106 | 0.0985903003 |

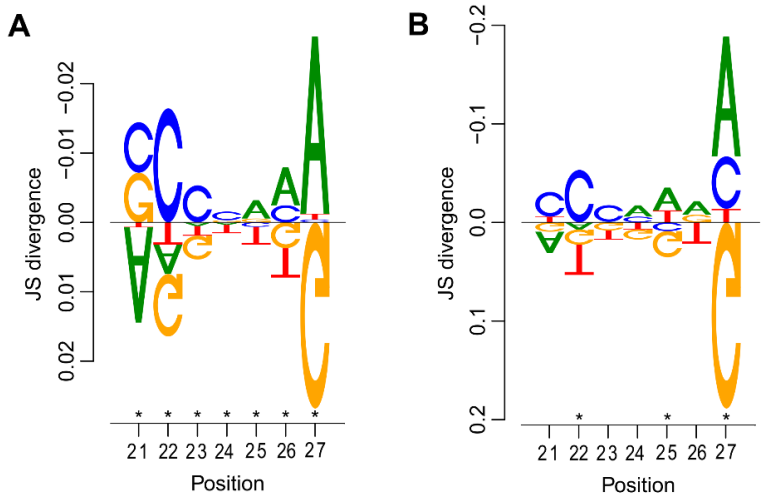


**Figure 32:** Negatively enriched triplex motifs. Motifs were negatively enriched in the triplex library, with less than -0.5 (A) and -1.0 (B) log10 fold enrichment over the input library, with 961 and 25 sequences.

### Verification of the top binders

The top five enriched TFO sequences (Table 15) were additionally verified with MST and EMSA (Figure 33). All of them were able to form a triplex in MST experiments (Figure 33, B - F and Table 17), confirming the validity of the TSE-seq experiments. Surprisingly the binding affinities of the top five motifs were much lower than the complementary sequence (Figure 33, A - F). The complementary sequence has a binding affinity of 22 nM +/- 11 nM (Figure 33, A), whereas the top five motifs have binding affinities between 4 µM and 10.8 µM (Figure 33, B – F).

As previously, MST measurements were complemented with EMSAs. I observed a faint triplex band for the top four triplex motifs and no triplex band for the fifth motif. This is in line with my previous EMSAs, where I noticed that triplexes with binding affinities in the µM range, according to MST experiments, often did not display a triplex band in the EMSA or displayed only a faint band. The TSE-seq experiments unravel an additional complexity in the triplex binding code, molecular competition.

**Table 17:** $K_d$ values derived from microscale thermophoresis analysis of the top five enriched triplex motifs from the TSE-seq experiments. SD = Standard deviation. Replicates = Number of independent replicates.

| triplex name | Kd [nM] | SD [nM] | Replicates | Sequence |
|---|---|---|---|---|
| **Nr1** | 7161.0 | 1485.8 | 4 | CCUCCUUUUUUCUUUUUUUUUUUUACC |
| **Nr2** | 4141.2 | 793.3 | 3 | CCUCCUUUUUUCUUUUUUUUUUUUGUC |
| **Nr3** | 1602.4 | 57.9 | 3 | CCUCCUUUUUUCUUUUUUUUUUUUAUG |
| **Nr4** | 1815.5 | 97.6 | 3 | CCUCCUUUUUUCUUUUUUUUUUUUACG |
| **Nr5** | 10829.9 | 409.9 | 3 | CCUCCUUUUUUCUUUUUUUUUUUUGCUC |

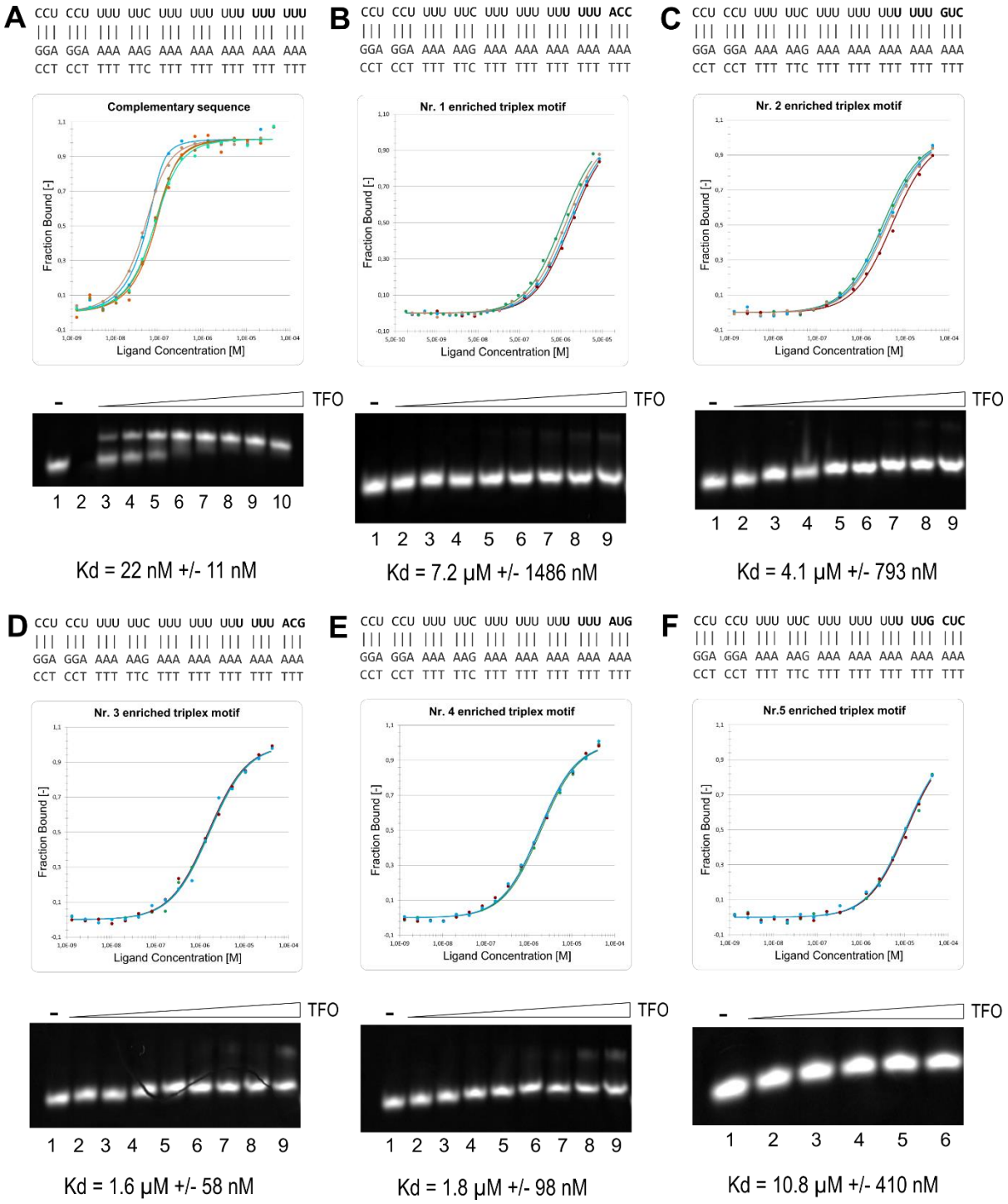



**Figure 33:** Validation of top binders from TSE-seq experiments with MST and EMSA. (A) Reference sequence. (B) Number 1 top enriched motif. (C) Number 2 top enriched motif. (D) Number 3 top enriched motif. (E) Number 4 top enriched motif. (F) Number 5 top enriched motif.

## 5.2 Discussion

In summary, triplex stability is dependent on the base affinity of the three motifs (Pyrimidine, Mixed, and Purine motifs). The base affinity is modulated by the TrTS Guanine content, TFO length, non-canonical base pairs, and mismatches in the respective order.

Comparing the triplex motifs and RNA:DNA triplexes, several distinct behaviors could be observed. For instance, the pyrimidine motif has a cut-off of 15% Guanine content for the DNA:DNA triplex (Figure 22, A), while the RNA:DNA triplex can form at all tested Guanine compositions (Figure 22, B). Conversely, the purine motif was able to form at all tested Guanine compositions for both DNA:DNA (Figure 24, A) and RNA:DNA (Figure 24, B) triplexes. Looking at the TFO length, the pyrimidine motif DNA:DNA and RNA:DNA triplexes were able to form at all tested lengths, including 17 nt, which is contrary to a recent publication where pyrimidine triplexes were shown to not form below 19nt (Kunkler *et al.*, 2019). One reason for this could be the difference in the Guanine compositions between the triplexes I tested (52% G) and the ones that were tested in the publication (11% G).

Further enforcing the newly defined triplex code by demonstrating that G% in the TrTS sequence has a more significant influence on the triplex formation than the TFO length. The Purine motif had the highest tolerance for different Guanine sequence compositions, with all tested sequences forming a stable triplex for both DNA:DNA (Figure 26, A) and RNA:DNA triplexes (Figure 26, B). The length constraint of the purine motif is 10 nt for both DNA:DNA (Figure 27, A) and RNA:DNA (Figure 27, B) triplexes.

It has been proposed that Mixed motif RNA:DNA triplexes cannot form under physiological conditions (Semerad and James maher, 1994). Conversely, I show that Mixed motif RNA:DNA can form triplexes under physiological conditions, albeit at a very narrow and specific Guanine content (Figure 25, B). Most likely, RNA molecules with higher Guanine content would preferentially form G-quadruplexes (Olivas and Maher, 1995). It is unclear why Guanine content below 72% does not permit triplex formation. Conversely, DNA:DNA mixed motif triplexes behave similarly to the other motifs (Purine and Pyrimidine) and form with high binding affinity with medium and high Guanine content (Figure 25, A). The length constraint of the mixed motif is much higher than for the other motifs, with DNA:DNA triplexes still forming at 17 nt, while more stable triplexes were observed with 21nt length and higher. On the contrary, RNA:DNA mixed motif triplexes form at 21 nt or higher, with more stable triplexes starting to form at 25 nt (Figure 25, B).

I have tested the newly defined triplex formation rules with published triplex sequences (Table 13). Interestingly, many of the published sequences did not form a triplex in physiological conditions or

had binding affinities that are unlikely for triplex formation *in vivo*. These experiments were performed *in vitro* and do not entirely encapsulate the complexity of *in vivo* conditions. We know that *in vivo*, nucleosomes and regulatory proteins stabilize triplex structures and as such, triplexes with low binding affinity may nonetheless form. Regardless, even with the stabilizing element, the absolute number of non-coding RNA molecules in the cell does not support triplex formation beyond $K_d$ of 50 µM. The binding affinities of the top five enriched motifs in the TSE-seq experiment also support the choice of the binding affinity cut-off at > 50 µM while still permitting triplex formation at $K_d$ > 1 µM (Figure 33, B – F and Equation 2).

Looking at different positions and identities of mismatches in the pyrimidine motif, I observed the middle sequences are more sensitive to mismatches while flanking regions display a high tolerance for mismatches (Figure 28). Specifically, only C12T mismatch disrupted triplex formation (Figure 28), which could partly be explained by a reduction in C:G-C triplets, similar to G% measurements for the pyrimidine motif (Figure 22, A). Nonetheless, G:G-C and A:G-C base triplets did not abolish triplex formation, although the binding affinity was slightly reduced (Table 14). All possible combinations of non-canonical base triplets for the pyrimidine motif were tested in a recent publication using EMSA experiments (Kunkler *et al.*, 2019). Here both G:G-C and A:G-C base triplets showed only a minor reduction in binding affinity, which is in line with my data.

On the contrary, the T:G-C base triplets did not disrupt triplex formation in their experiments, while my MST measurements (Figure 28) and EMSA experiments (Table S9) show an evident disruption in triplex formation. I hypothesize that the differences in sequence composition may explain this discrepancy. Specifically, the distance between C:G-C base triplets in the sequence may influence the sensitivity to mismatches.

Furthermore, I tested the triplex binding code in the context of molecular crowding. I designed a randomized RNA TFO library with a core 20 nt sequence and a variable 7 nt sequence and performed a triplex formation with the randomized library, following an EMSA experiment. At 20 nt length, triplex formation is possible but has a low binding affinity and fast dissociation rate. I was able to test which sequences will stabilize the triplex structure by analyzing the enriched TFOs that formed a triplex with the TrTs. Here I choose the pyrimidine EN3 motif, for which ample data exists (Figure 28) (Maldonado *et al.*, 2017). Surprisingly, the expected TFO sequence, a complementary sequence to the TrTS (Figure 33, A), was not enriched in the top 30 TFO sequences (Table 15). However, it was nonetheless considerably enriched compared to the input library. I found that although there was a tendency to complete the complementary pyrimidine motif, especially at positions 21 - 23 (Figure 31, B and C), some positions showed a considerable enrichment for Adenines (position 23 and 24) and Cytosines (position 27) (Figure 31, B and C).

On the contrary, Guanine at position 27 and Adenine at position 21 was incompatible with triplex formation (Figure 32, A and B). A clear difference can be observed for positively and negatively enriched triplexes, with the majority of enriched triplexes having a Uracil at positons 21 – 23 (Table 15), while negatively enriched triplexes, or rather sequences enriched in the input TFO library, have rather randomized sequence composition at all seven positions (Table 16). Moreover, I confirmed the validity of TSE-seq results by testing the top five enriched sequences with MST and EMSA experiments (Figure 33, B-F). All five sequences were able to form triplexes in MST experiments, while their formation in the EMSA was rather poor, and only faint bands or no triplex-specific bands were observed (Figure 33, B-F). The data suggest a high impact of molecular crowding on triplex stability, as has been observed before (Jiang *et al.*, 2015; Aviñó *et al.*, 2016). The results of TSE-seq experiments reveal the complexity of the triplex binding code, where a significant difference can be observed in the binding code between triplex formation in the context of molecular crowding and no molecular crowding.

The results of this study represent a valuable addition to the knowledge on triplex formation and a solid base for new triplex binding prediction software. Presently available prediction tools may also be used with carefully defined parameters. Nonetheless, I would suggest not using the triplex reactivity score or any other currently available scores, as these do not reflect the true binding affinity.

In summary, further studies on the orientation of mixed motifs, mismatches, and non-canonical base triplets for mixed and purine motifs are essential to complete the triplex binding code. Furthermore, a conclusive *in vivo* confirmation of triplex formation and characterization of different stabilization factors *in vivo* are the necessary next steps in the triplex field.

# 6 MATERIALS AND METHODS

## 6.1 Nucleosome stability

### 6.1.1 Annotation and publicly available data

**Annotation**

| Type | Source | Version |
|---|---|---|
| Reference genome | https://hgdownload.soe.ucsc.edu/goldenPath/dm3/bigZips/ | DM3 |
| Annotation | https://hgdownload.soe.ucsc.edu/goldenPath/dm3/bigZips/ | UCSC |

**Publicly available data**

| Dataset | Identifier | Reference |
|---|---|---|
| MNase-seq | GSE78984 | (Mieczkowski *et al.*, 2016) |
| MNase-ChIP-H3-seq | GSE78984 | (Mieczkowski *et al.*, 2016) |
| MNase-ChIP-H4-seq | GSE78984 | (Mieczkowski *et al.*, 2016) |
| RNA-seq | GSE78984 | (Mieczkowski *et al.*, 2016) |
| MNase-ChIP-H2B-seq | GSE69336 | (Chereji *et al.*, 2016) |
| MNase-ChIP-H3-seq | GSE69336 | (Chereji *et al.*, 2016) |
| M1BP-ChIP-seq | SRX3011239 | (Celniker *et al.*, 2009) |
| GRO-seq | GSE23544 | (Core *et al.*, 2012) |
| RNApol-II-seq | GSE23544 | (Core *et al.*, 2012) |
| MNase-seq with spike-ins | PRJNA528497 | (Chereji, Bryson and Henikoff, 2019) |
| H3K27ac ChIP-seq | modENCODE_296 | (Celniker *et al.*, 2009) |
| H3K27me3 ChIP-seq | modENCODE_298 | (Celniker *et al.*, 2009) |
| H3K4me1 ChIP-seq | modENCODE_304 | (Celniker *et al.*, 2009) |
| H3K9me3 ChIP-seq | modENCODE_313 | (Celniker *et al.*, 2009) |
| RNA pol-II ChIP-seq | modENCODE_329 | (Celniker *et al.*, 2009) |
| H3K36me3 ChIP-seq | modENCODE_3189 | (Celniker *et al.*, 2009) |
| H3K4me3 ChIP-seq | modENCODE_3761 | (Celniker *et al.*, 2009) |
| CBP TF ChIP-seq | modENCODE_858 | (Celniker *et al.*, 2009) |
| GAF TF-seq | modEncode_3238 | (Celniker *et al.*, 2009) |
| DNase-seq | GSM5258764 | (Dunham *et al.*, 2012) |
| ATAC-seq | GSM2756640 | (Ibrahim *et al.*, 2018) |

### 6.1.2 Software

| Software | Version | Reference |
|---|---|---|
| fastQC | 0.11.9 | (Andrews, 2010) |
| multiQC | 1.0 | (Ewels *et al.*, 2016) |
| Trimmomatic | 0.38 | (Bolger, Lohse and Usadel, 2014) |
| Bowtie2 | 2.3.5.1 | (Langmead and Salzberg, 2012) |

| | | |
|---|---|---|
| Qualimap | 2.2.1 | (Okonechnikov, Conesa and García-Alcalde, 2016) |
| DeepTools | 3.1.3 | (Ramírez *et al.*, 2016) |
| DANPOS | 2.2.2 | (Chen *et al.*, 2013) |
| BEDtools | 2.29.0 | (Quinlan and Hall, 2010) |
| featureCounts | 1.6.2 | (Liao, Smyth and Shi, 2014) |
| SAMtools | 1.9 | (Li *et al.*, 2009) |
| STAR | 2.7.2a | (Dobin *et al.*, 2013) |
| R | 4.1.0 | (R Core Team, 2020) |
| MACC R package | 0.2 | (Mieczkowski *et al.*, 2016) |
| edgeR R package | 3.34.1 | (Robinson, Mccarthy and Smyth, 2010) |
| limma R package | 3.48.3 | (Ritchie *et al.*, 2015) |
| LSD R package | 4.1.0 | (Schwalb *et al.*, 2020) |
| clusterProfiler R package | 4.0.5 | (Yu *et al.*, 2012) |
| ChIPpeakAnno R package | 3.26.4 | (Zhu *et al.*, 2010) |
| ChIPseeker R package | 1.28.3 | (Yu, Wang and He, 2015) |
| ggplot2 R package | 3.3.5 | (Hadley Wickham, 2016) |
| TxDb.Dmelanogaster.UCSC.dm3.ensGene R package | 3.2.2 | (Carlson and Maintainer, 2015) |

### 6.1.3 Data and code availability

All datasets used in this project are publicly available on the European Nucleotide Archive (ENA) and modENCODE databases with the above-stated identifiers. The code used for the analysis and figure generation is available on my GitHub page: https://github.com/sarawernig/TheNucMACCpipeline.

### 6.1.4 Data analysis

Computational analysis was performed on an Apple Mac Pro Late 2013, macOS 10.15.

**Processing of MNase-seq data**

Adapters were removed from raw sequencing reads with Trimmomatic and mapped to the UCSC Drosophila melanogaster dm3 genome, using the Bowtie2 with the following parameters --very-sensitive-local --no-discordant. Aligned reads were filtered for their mapping quality (MAPQ > 30), fragment size (length 140 – 200 nt for mono-nucleosomes and length 50 – 139 nt for sub-nucleosomes), and ambiguous chromosomes and blacklisted regions were removed using DeepTools2 alignmentSieve. Often there is a shift in the fragment sizes due to variation in MNase enzyme activity; thus, each dataset should be assessed individually. In case of no fragment size selection after MNase digestion, a dominant peak around 150 nt can be observed and a second smaller peak around 340 nt, representing mono- and di- nucleosomes, respectively. Larger di-nucleosomal fragments only reduce the quality of nucleosome positions and should be excluded from the analysis. In order to achieve

higher sequencing depth, mono-nucleosome positions were called on pooled MNase titrations, using DANPOS, with dpos -m 1 --extend 70 -c 162367812 -u 0 -z 1 -a 1 -e 1, parameters. Sub-nucleosome positions were called using only the mild digestion samples, with the following parameters: dpos -m 1 --extend 70 -z 70 -c 162367812 -u 0 -e 1. Here library size was normalized to effective dm3 genome size (162367812), paired-end reads centered to their midpoint and extended to 70bp to obtain nucleosome-sized positions. Sub-nucleosome positions were further filtered by two parameters; a non-overlapping position with a mono-nucleosome or an overlapping position with an enriched signal over a mono-nucleosome signal. Here a cut-off of > 4x higher signal in sub-nucleosomal data as opposed to mono-nucleosomal data was chosen.

**NucMACC score calculation**

Reads per sample per nucleosome position were counted using featureCounts, using a custom SAF file with nucleosome positions and the following parameters -SAF -p -B -C -largestOverlap. Information on the GC content of each nucleosome was added to the read count file using BEDtools nuc. Nucleosome positions were obtained by converting the DANPOS annotation file to SAF format using AWK. NucMACC scores were calculated using R and edgeR R package. Firstly, raw fragments were filtered for sequencing coverage, with cut-off > 30 fragments per mono-nucleosome and > 5 fragments for sub-nucleosomes. Then, fragments were normalized based on the library size (counts per million, CPM) or to the calculated library size based on normalization factors for spike-in analysis. Linear regression was calculated using normalized fragment counts and log2 MNase concentrations or time points to preserve the similar distance between samples. Finally, nucMACC scores were normalized for the inherent GC bias using the locally weighted scatterplot smoothing (LOESS) algorithm. Henceforth, the loess function implemented in R was applied to estimate the influence of GC content on MNase concentration/time titrations, using the slope of the linear regression line. The function predicts a local score based on the GC content and trained values. The predicted scores were used as a normalization factor for the final nucMACC scores. The nucMACC score can be positive or negative. For mono-nucleosomes, the nucMACC score was multiplied by -1 for a more intuitive interpretation. Here, a negative score represents hypo-accessible DNA, whereas a positive score represents hyper-accessible DNA. For sub-nucleosomes, a negative score represents un-stable nucleosomes, whereas a positive score characterizes nucleosomes with non-canonical structures.

**NucMACC score cut-off**

A universal cut-off of the nucMACC score was obtained by ranking nucleosomes based on the score and the first derivative and followed by a local regression, which predicts the slope based on the

scores. The cut-off for differentially accessible nucleosomes is when the slope = 1; this happens twice for positive and negative scores.

**Spike-in analysis**

Yeast reads were aligned to the sacCer3 Saccharomyces cerevisiae genome using the Bowtie2 with the following parameters --very-sensitive-local --no-discordant. Aligned reads were filtered for their mapping quality (MAPQ > 30) and fragment size (length 140 – 200 nt for mono-nucleosomes and length 50 – 139 nt for sub-nucleosomes) using DeepTools alignmentSieve. The number of reads per mono- and sub- nucleosome fraction was counted, and normalization factors were calculated by dividing the number of reads by 10e6.

**Nucleosome characterization**

Characterization of nucleosomes and figures used in the manuscript were done in R. Nucleosomes were assigned to genes using bitr R package and TxDb.Dmelanogaster.UCSC.dm3.ensGene annotation R package. Pathway enrichment of genes was performed with clusterProfiler R package. For expression correlation with nucleosome positioning, nucleosomes on promoter and gene body (exon and intron) regions were used. For heatmaps, LSD R package was used. For figures ChIPpeakAnno, ChIPseeker and ggplot2 R packages were used. For TF enrichment, HOMER findMotifsGenome was used, with a custom BED file containing the positions of TSS-un-stable nucleosomes and the *D. melanogaster* TF database.

**Pausing index**

Pausing index is defined as the ratio of the GRO-seq read density at TSS to gene body density. It was calculated as read density +/- 100 nt from TSS divided by read density +150 nt from TSS to the TES.

**Minimum sequencing coverage analysis**

Samples were sub-samples using SAMtools view -bs 42.X, where X represents the percentage to which the data was reduced to, and 42 as the seed number to obtain reproducible results.

**Processing of RNA-seq data**

Sequencing reads were aligned to the UCSC Drosophila melanogaster dm3 genome using STAR v2.7.2a with the following parameters --outFilterType BySJout --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 –

alignMatesGapMax. Aligned reads were filtered for their mapping quality (MAPQ > 30), using SAMtools v1.9. Reads per exon were counted using featureCounts v1.6.2 and parameters -C -p -B -t exon -g gene_id. Differential gene expression was done in R v.3.6.1, using the edgeR and limma Bioconductor packages.

## 6.2 The triplex code

### 6.2.1 Technical devices

| **Device name** | **Manufacturer** |
|---|---|
| Monolith NT.115 MST device | NanoTemper technologies |
| PCR machine | Peqlab |
| Qubit 2.0 Fluorometer | Invitrogen |
| NanoDrop 2.0 | Thermo Fischer Scientific |
| Fluorescence Image Reader FLA-3000 | Fujifilm |
| UV transilluminator | Thermo Fischer Scientific |
| Tabletop centrifuge | Eppendorf |
| Vortex | Carl Roth |
| Electrophoresis XCell SureLock ™ Mini Cell | Invitrogen |
| Electrophoresis power supply | Thermo Fischer Scientific |
| Tape station 4200 | Agilent |
| MiSeq | Illumina |

### 6.2.2 Consumables

| **Description** | **Manufacturer** |
|---|---|
| 0.2 mL PCR tubes | Sarstedt |
| 1.5 mL tubes | Sarstedt |
| Glass pipettes 5/10/25 mL | Sarstedt |
| Monolith N.T115 standard treated capillaries | NanoTemper technologies |
| Orange G | Sigma |
| NP40 | Sigma |
| Pipette tips | Sarstedt |
| TEMED | Carl Roth |
| TRIS | Carl Roth |
| Glycerin | Carl Roth |
| ROTIPHORESE®Gel 30 | Carl Roth |
| Magnesium acetate tetrahydrate | Carl Roth |
| Acetic acid | Carl Roth |
| Ammonium persulphate (APS) | Carl Roth |
| Bovine Serum Albumin (BSA) | Carl Roth |
| HEPES | Carl Roth |
| Sodium chloride | Carl Roth |
| Magnesium chloride | Carl Roth |
| Potassium chloride | Carl Roth |
| Sodium hydroxide | Carl Roth |

| | |
|---|---|
| Empty Mini Gel Cassettes | Thermo Fischer Scientific |
| RNasin Ribonuclease Inhibitor | Promega |
| DNA GeneRuler 50bp | Fermentas |
| SYBR Green I RNA gel stain | Sigma |
| ZR Small-RNA PAGE recovery kit | Zymo research |
| TURBO DNA-free™ Kit | Invitrogen |
| CATS Small RNA-seq kit | Diagenode |
| Tape station DNA ScreenTape | Agilent |
| MiSeq sequencing kit | Illumina |
| Triplex oligonucleotides | Sigma, IDT, Eurofins |

**Triplex oligonucleotides**

Oligonucleotide sequences are available in Appendix 2 and 3.

## 6.2.3 Solutions and buffers

1X OLIGO ANNEALING (OA) BUFFER
20 mM TRIS-HCl pH 7.4
2 mM $MgCl_2$
50 mM NaCl

1X TRIPLEX ANNEALING (TA) BUFFER
40 mM TRIS-acetate pH 7.4
10 mM Mg-acetate

EMSA RUNNING BUFFER
50 mL 10X TA buffer
450 mL $ddH_2O$

MST BUFFER
40 mM TRIS-acetate pH 7.4
10 mM Mg-acetate
0.05 % NP40

MST "CELL CONDITIONS" BUFFER
10 mM HEPES pH 7.4
10 mM NaCl
140 mM KCl
10 mM $MgCl_2$
0.05 % NP40

10X ORANGE G LOADING DYE
50 % Glycerin
10 mM EDTA
0.05 % (w/v) Orange G

15% POLYACRILAMIDE GEL
Rotiphorese30 5 mL
10X TA buffer 1 mL

| | |
|---|---|
| 20% APS | 42 µL |
| MQ H2O | 3.95 mL |
| TEMED | 8 µL |

## 6.2.4 Experimental procedures

### Triplex sample preparation

*Annealing of oligonucleotides*

Annealing of fluorescently labeled oligonucleotides was performed with the unlabeled strand in approximately 10 % excess to the labeled strand (5' Cy5) to ensure complete annealing of the labeled oligonucleotide. This was done to avoid the background signal from the single-stranded fluorescently labeled oligonucleotide and to avoid the formation of R-loops (ssDNA + ssRNA) in the MST measurements. Sense and complementary antisense oligonucleotides were diluted to 10 µM in 1x OA buffer. The samples were incubated at 95°C for 5 min in a PCR machine. Afterward, the block was switched off, and the samples were left to cool down to room temperature for 2 h slowly. For the working stock solutions, samples were diluted to 1 µM with MQ $H_2O$.

*Preparing the TFO and TrTS solutions*

The desired concentration of the TFO (5 – 850 µM) was prepared in a 20 µL volume and 0.05% NP40 and 1x TA buffer. The desired concentration of the TrTS (12.5 – 50 nM) was prepared in a 200 µL volume and 0.05% NP40 and 1x TA buffer.

*Serial dilutions*

Sixteen serial dilutions of the TFO were prepared by diluting the initial TFO solution with a V/V amount of the MST buffer. An equal amount (V/V) of the TrTS solution was added to the serial dilutions of the TFO.

*Triplex formation*

Sixteen samples containing the TrTS and decreasing amount of TFO and one control sample containing only TrTS were incubated for 15 min at 37°C (with lid temperature at 95°C) in a PCR machine. Samples were allowed to cool down at room temperature for 10 min and mixed before taking an MST measurement.

### Microscale thermophoresis

From each of the sixteen serial dilutions, 10 µL were drawn into standard treated glass capillaries and inserted in the Monolith MST device. A binding affinity measurement was taken under the red excitation, medium MST-power, and temperature control at 22°C.

### Binding affinity calculation

Each TrTS has a unique binding affinity constant for a specific TFO. The equilibrium reaction describes how a target (TrTS) binds to its ligand (TFO) to create the triplex complex (TrTS + TFO):

TrTS + TFO ⇔ Triplex

First, a ratio between TrTS and TFO, where both unbound and bound states can be observed, is determined—followed by at least three independent MST measurements. Replicates are then plotted together in a binding curve graph, from which a binding affinity, $K_d$, is determined (Figure 33, B). $K_d$ is the equilibrium dissociation constant. Small values of $K_d$ indicate higher affinity.

The x-axis on the binding affinity graph represents the ligand concentration (TFO), and the y-axis represents the percentage of free TrTS molecules (Θ) not forming a triplex (Figure 33, B). The values of Θ range from 0 to 1 (corresponding to the range from 0 to 100%). For example, if Θ is 0.5, this means that 50% of the available TrTS molecules are in a triplex.

### Electromobility shift assay

The aforementioned triplex samples (dilutions 1-9) and a control sample containing only TrTs were additionally used for a triplex EMSA. Before sample loading, the gels were pre-run for 30 min at 90V. A 1x Orange G loading dye was added to all the samples for better visualization. 10 µL of each tube was loaded on a 15% polyacrylamide gel and run in 1x TA triplex buffer at room temperature. For DNA:DNA triplexes run was performed at 110V for 1.5 - 2h, whereas for RNA:DNA triplexes, the run was performed at 75V for 4 - 5h. Gels were then imaged in the FLA3000 fluorescent imager with the corresponding filter settings.

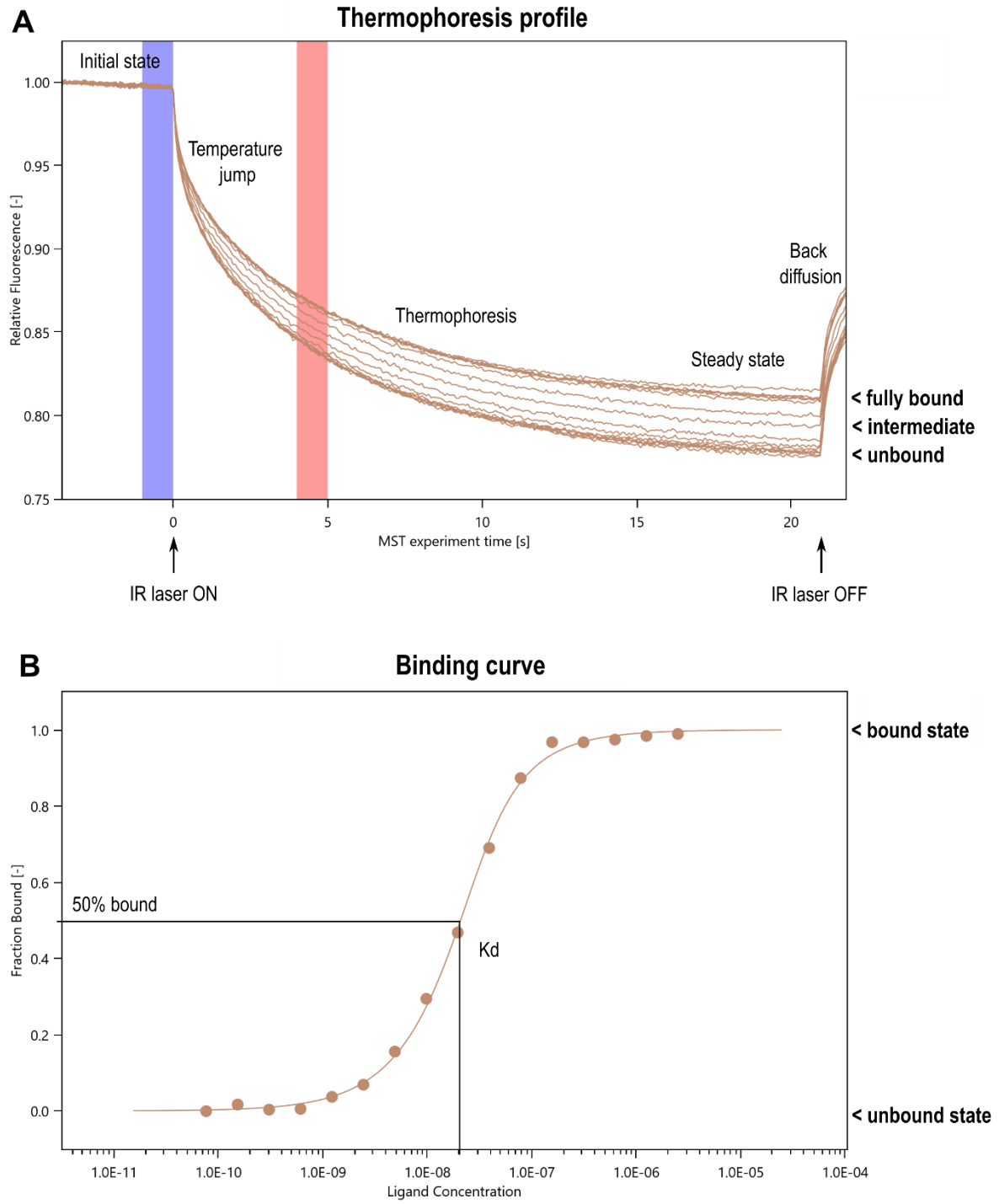



**Figure 34:** The principle of the microscale thermophoresis and binding affinity calculation. (A) Microscale thermophoresis experiments start with an IR laser induction, upon which molecules start to align along the temperature gradient (Temperature jump). Molecules then move along the temperature gradient in a manner dependent on their size, charge, and hydration shell (Thermophoresis). Finally, a steady-state is reached, and the laser is switched off. This results in an inverse temperature jump and back diffusion of the molecules. Initial fluorescence is indicated in the blue square, whereas the fluorescence change used for the binding affinity calculation is indicated in the pink square. (B) Thermophoresis and temperature jump signals are plotted against the concentration of the unlabeled molecule (in our case, the TFO), and a binding curve is generated. From the binding curve, a binding affinity - Kd value is calculated.

**TSE-seq**

*Oligo annealing*

Oligo annealing was done as before, with one difference. Here one self-TrTS forming oligonucleotide was used instead of two complementary oligonucleotides.

*Triplex-EMSA*

RNA:DNA triplex EMSA was performed as mentioned before. TrTS and TFO oligos were added to all the samples in 1:11 ratios. Final samples were prepared using 1 µg/µL BSA and 1x TA buffer in 20 µL solution. Positive and negative controls were used as pyrimidine and purine complementary sequences, respectively.

*Gel staining and sample excising*

Gels were stained with 10x SYBR Green I nucleic acid gel dye in 1x TA buffer for 40 min with shaking. A UV light was used to visualize the bands and using a clean scalpel, and bands were excised from the gel and processed.

*RNA isolation*

Triplexes were isolated from the gel slices using the ZR Small-RNA PAGE recovery kit and following the manufacturer's instructions. Triplexes were then treated with DNase to remove DNA nucleotides from the triplex structures, following the manufacturer's instructions. Afterward, sample concentration was measured using a Qubit ssRNA kit, following the manufacturer's instructions.

*Library preparation*

A small RNA-seq library was prepared using 10 ng of input material, following the manufacturer's instructions (Diagenode CATS Small RNA-seq kit). Final libraries were checked using the Tape station DNA ScreenTape Analysis.

*High throughput sequencing*

Samples were sequenced on an Illumina MiSeq device with 20 % PhiX to compensate for low nucleotide diversity in the input libraries. Samples were sequenced as single-end 35 nt reads.

### 6.2.2 Software

| Software | Version | Reference / Manufacturer |
|---|---|---|
| MO.Control | 1.6 | NanoTemper |
| MO.Affinity analysis | 2.3 | NanoTemper |
| fastQC | 0.11.9 | (Andrews, 2010) |
| multiQC | 1.7 | (Ewels *et al.*, 2016) |
| Trimmomatic | 0.38 | (Bolger, Lohse and Usadel, 2014) |
| CutAdapt | 1.16 | (Marcel, 2011) |
| SALMON | 1.3.0 | (Patro *et al.*, 2017) |
| R | 4.1.0 | (R Core Team, 2020) |
| Biostrings R package | 2.62.0 | (Pagès *et al.*, 2021) |
| DiffLogo R package | 2.18 | (Nettling *et al.*, 2015) |
| seqLogo R package | 1.60.0 | (Bembom and Ivanek, 2021) |
| seqinr R package | 4.2.12 | (Charif and Lobry, 2007) |

### 6.2.3 Data and code availability

All the raw MST data, EMSA figures, and TSE-seq samples are available in the following link: http://gofile.me/6HRGV/0B7YBd85h

The code used for the analysis and figure generation is available on my GitHub page: https://github.com/sarawernig/TheTriplexCode

### 6.2.4 Data analysis

Computational analysis was performed on an Apple Mac Pro Late 2013, macOS 10.15.

**Processing of TSE-seq data**

Raw data in .fastq format was first cleaned for adapter sequences using Trimmomatic v. 0.38 and trimmed to 27 nt. Reads with sequencing quality below XX and reads where an initial GGGCCT sequence was missing were removed. The GGG bases come from the CATS library kit, whereas the CCT is the start of all the reads in the input library.

**Replicate analysis**

Consistency between replicates is an important quality measure. Here I analyzed the consensus matrix of each replicated for the input TFO library and the triplex libraries. I additionally analyzed the similarities between the replicates using the Jensen–Shannon divergence.

**Motif analysis**

*In silico* input, a library was created using the merged input TFO library and SALMON tool. The merged triplex library was then quantified against the in silico TFO library. This ensured only reads present in the input library were quantified and avoided analyzing reads present only in the triplex library. Using reads absent in the input TFO library would create a bias in the downstream analysis. Furthermore, motifs with less than two reads in the input library were also removed to ensure a bias in the library does not skew the results. An enrichment score was calculated by dividing a normalized triplex motif count in the triplex library by the normalized read count in the input library on a log10 scale. An enrichment score ranked quantified reads, and the top 15 and 30 were used for downstream processing. Motifs with more than 1 and 2 log10 fold changes were analyzed as well as motifs with a negative enrichment lower than -0.5 and -1 log10 fold changes were analyzed.

A position-weighted matrix was created for each group of enrichments and compared to the input TFO library. Jensen–Shannon divergence matrix was used to compare the similarities between different enrichment groups and the input. For each comparison, 100 permutations were used to calculate a p-value for each base.

# 7 REFERENCES


Agazie, Y.M., Lee, J.S. and Burkholder, G.D. (1994) "Characterization of a new monoclonal antibody to triplex DNA and immunofluorescent staining of mammalian chromosomes.," *Journal of Biological Chemistry*, 269(9), pp. 7019–7023. doi:10.1016/S0021-9258(17)37476-8.

Alberti, P. *et al.* (2002) "A directional nucleation-zipping mechanism for triple helix formation," *Nucleic Acids Research*, 30(24), p. 5407. doi:10.1093/NAR/GKF675.

Anderson, J.D., Lowary, P.T. and Widom, J. (2001) "Effects of histone acetylation on the equilibrium accessibility of nucleosomal DNA target sites," *Journal of Molecular Biology*, 307(4), pp. 977–985. doi:10.1006/jmbi.2001.4528.

Andrews, S. (2010) "FastQC: A Quality Control Tool for High Throughput Sequence Data." Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arimura, Y. *et al.* (2012) "Structural analysis of the hexasome, lacking one histone H2A/H2B dimer from the conventional nucleosome," *Biochemistry*, 51(15), pp. 3302–3309. doi:10.1021/bi300129b.

Aviñó, A. *et al.* (2016) "The Effect of Small Cosolutes that Mimic Molecular Crowding Conditions on the Stability of Triplexes Involving Duplex DNA," *International Journal of Molecular Sciences Article* [Preprint]. doi:10.3390/ijms17020211.

Bacolla, A., Wang, G. and Vasquez, K.M. (2015) "New Perspectives on DNA and RNA Triplexes As Effectors of Biological Activity," *PLoS Genetics*, 11(12), pp. 1–12. doi:10.1371/journal.pgen.1005696.

Bao, Y. *et al.* (2004) "Nucleosomes containing the histone variant H2A.Bbd organize only 118 base pairs of DNA," *The EMBO Journal*, 23(16), p. 3314. doi:10.1038/SJ.EMBOJ.7600316.

Barski, A. *et al.* (2007) "High-Resolution Profiling of Histone Methylations in the Human Genome," *Cell*, 129(4), pp. 823–837. doi:10.1016/J.CELL.2007.05.009.

Beal, P.A. and Dervan, P.B. (1991) "Second structural motif for recognition of DNA by oligonucleotide-directed triple-helix formation," *Science*, 251(4999), pp. 1360–1363.

Bembom, O. and Ivanek, R. (2021) "seqLogo: Sequence logos for DNA sequence alignments. R package version 1.60.0."

Bolger, A.M., Lohse, M. and Usadel, B. (2014) "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, 30(15), pp. 2114–2120. doi:10.1093/bioinformatics/btu170.

Bowman, G.D. and Poirier, M.G. (2015) "Post-Translational Modifications of Histones That Influence Nucleosome Dynamics," *Chemical Reviews*, 115(6), p. 2274. doi:10.1021/CR500350X.

Brahma, S. and Henikoff, S. (2019) "RSC-Associated Subnucleosomes Define MNase-Sensitive Promoters in Yeast," *Molecular Cell* [Preprint]. doi:10.1016/j.molcel.2018.10.046.

Buenrostro, J.D. *et al.* (2013) "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position," *Nature Methods 2013 10:12*, 10(12), pp. 1213–1218. doi:10.1038/nmeth.2688.

Buske, F.A. *et al.* (2012) "Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data.," *Genome research*, 22(7), pp. 1372–81. doi:10.1101/gr.130237.111.

Buske, F.A. *et al.* (2013) "Triplex-Inspector: an analysis tool for triplex-mediated targeting of genomic loci," *Bioinformatics*, 29(15), pp. 1895–1897. doi:10.1093/bioinformatics/btt315.

Buske, F.A., Mattick, J.S. and Bailey, T.L. (2011) "Potential in vivo roles of nucleic acid triple-helices," *RNA Biology*, 8(3), p. 427. doi:10.4161/RNA.8.3.14999.

Cabili, M.N. *et al.* (2015) "Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution," *Genome Biology*, 16(1), pp. 1–16. doi:10.1186/S13059-015-0586-4.

Carlson, M. and Maintainer, B.P. (2015) "TxDb.Dmelanogaster.UCSC.dm3.ensGene: Annotation package for TxDb object(s)."

Carr, C.E. *et al.* (2018) "Effect of dC → d(m5C) substitutions on the folding of intramolecular triplexes with mixed TAT and C+GC base triplets," *Biochimie*, 146, pp. 156–165. doi:10.1016/j.biochi.2017.12.008.

Celniker, S.E. *et al.* (2009) "Unlocking the secrets of the genome," *Nature 2009 459:7249*, 459(7249), pp. 927–930. doi:10.1038/459927a.

Chakravarthy, S., Patel, A. and Bowman, G.D. (2012) "The basic linker of macroH2A stabilizes DNA at the entry/exit site of the nucleosome," *Nucleic Acids Research*, 40(17), p. 8285. doi:10.1093/NAR/GKS645.

Charif, D. and Lobry, J.R. (2007) "SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis," in Bastolla, U. et al. (eds) *Structural approaches to sequence evolution: Molecules, networks, populations*. New York: Springer Verlag, pp. 207–232.

Chen, C. *et al.* (2018) "LNMAT1 promotes lymphatic metastasis of bladder cancer via CCL2 dependent macrophage recruitment," *Nature Communications* [Preprint]. doi:10.1038/s41467-018-06152-x.

Chen, G. and Chen, S.J. (2011) "Quantitative analysis of the ion-dependent folding stability of DNA triplexes," *Physical biology*, 8(6), p. 066006. doi:10.1088/1478-3975/8/6/066006.

Chen, K. *et al.* (2013) "DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing," *Genome Research*, 23(2), pp. 341–351. doi:10.1101/gr.142067.112.

Chereji, R. v. *et al.* (2016) "Genome-wide profiling of nucleosome sensitivity and chromatin accessibility in Drosophila melanogaster," *Nucleic Acids Research*, 44(3), pp. 1036–1051. doi:10.1093/nar/gkv978.

Chereji, R. v., Bryson, T.D. and Henikoff, S. (2019) "Quantitative MNase-seq accurately maps nucleosome occupancy levels," *Genome biology*, 20(1), p. 198. doi:10.1186/S13059-019-1815-Z/FIGURES/10.

Core, L.J. *et al.* (2012) “Defining the Status of RNA Polymerase at Promoters,” *Cell Reports*, 2(4), pp. 1025–1035. doi:10.1016/J.CELREP.2012.08.034/.

Core Team, R. (2020) “R: A language and environment for statistical computing,” *R Foundation for Statistical Computing* [Preprint]. Available at: https://www.r-project.org/ (Accessed: April 15, 2022).

Culkin, J. *et al.* (2017) “The role of DNA sequence in nucleosome breathing.” doi:10.1140/epje/i2017-11596-2.

Dalal, Y. *et al.* (2007) “Tetrameric Structure of Centromeric Nucleosomes in Interphase Drosophila Cells,” *PLoS Biology*. Edited by J. Kadonaga, 5(8), p. e218. doi:10.1371/journal.pbio.0050218.

Davey, C.A. *et al.* (2002) “Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution,” *Journal of Molecular Biology*, 319(5), pp. 1097–1113. doi:10.1016/S0022-2836(02)00386-8.

Dimitriadis, E.K. *et al.* (2010) “Tetrameric organization of vertebrate centromeric nucleosomes,” *Proceedings of the National Academy of Sciences of the United States of America*, 107(47), pp. 20317–20322. doi:10.1073/pnas.1009563107.

Dobin, A. *et al.* (2013) “STAR: ultrafast universal RNA-seq aligner,” *Bioinformatics*, 29(1), pp. 15–21. doi:10.1093/bioinformatics/bts635.

Duca, M. *et al.* (2008) “The triple helix: 50 years later, the outcome,” *Nucleic Acids Research*, 36(16), pp. 5123–5138. doi:10.1093/nar/gkn493.

Dunham, I. *et al.* (2012) “An integrated encyclopedia of DNA elements in the human genome,” *Nature*, 489(7414), pp. 57–74. doi:10.1038/NATURE11247.

van Dyke MW (2005) *Do DNA Triple Helices or Quadruplexes Have a Role in Transcription?* Springer US.

Ewels, P. *et al.* (2016) “MultiQC: summarize analysis results for multiple tools and samples in a single report,” *Bioinformatics*, 32(19), pp. 3047–3048. doi:10.1093/BIOINFORMATICS/BTW354.

Fan, J.Y. *et al.* (2002) “The essential histone variant H2A.Z regulates the equilibrium between different chromatin conformational states,” *Nature structural biology*, 9(3), pp. 172–176. doi:10.1038/NSB767.

Farabella, I. *et al.* (2021) “Three-dimensional genome organization via triplex-forming RNAs,” *Nature Structural & Molecular Biology 2021 28:11*, 28(11), pp. 945–954. doi:10.1038/s41594-021-00678-3.

Felsenfeld, G. and Rich, A. (1957) “Studies on the formation of two- and three-stranded polyribonucleotides,” *Biochimica et Biophysica Acta*, 26(3), pp. 457–468. doi:10.1016/0006-3002(57)90091-4.

Fuda, N.J. *et al.* (2015) “GAGA factor maintains nucleosome-free regions and has a role in RNA polymerase II recruitment to promoters,” *PLoS genetics*, 11(3). doi:10.1371/JOURNAL.PGEN.1005108.

Gallego, L.D. *et al.* (2020) “Phase separation directs ubiquitination of gene-body nucleosomes,” *Nature*, 579(7800), pp. 592–597. doi:10.1038/s41586-020-2097-z.

Gautier, T. *et al.* (2004) “Histone variant H2ABbd confers lower stability to the nucleosome,” *EMBO Reports*, 5(7), pp. 715–720. doi:10.1038/sj.embor.7400182.

Goñi, J.R. *et al.* (2006) “Exploring the reasons for the large density of triplex-forming oligonucleotide target sequences in the human regulatory regions,” *BMC Genomics*, 7(1), pp. 1–10. doi:10.1186/1471-2164-7-63/.

Goñi, J.R., de la Cruz, X. and Orozco, M. (2004) “Triplex-forming oligonucleotide target sequences in the human genome,” *Nucleic acids research*, 32(1), pp. 354–360. doi:10.1093/NAR/GKH188.

González, P.J. and Palacián, E. (1989) “Interaction of RNA polymerase II with structurally altered nucleosomal particles,” *Journal of Biological Chemistry*, 264(31), pp. 18457–18462. doi:10.1016/s0021-9258(18)51488-5.

Gorab, E. and Pearson, P.L. (2018) “Thiazole Orange as an Alternative to Antibody Binding for Detecting Triple-helical DNA in Heterochromatin of Drosophila and Rhynchosciara,” *Journal of Histochemistry and Cytochemistry*, 66(3), pp. 143–154. doi:10.1369/0022155417745496.

Greifenstein, A.A., Jo, S. and Bierhoff, H. (2021) “RNA:DNA triple helices: from peculiar structures to pervasive chromatin regulators,” *Essays in Biochemistry*, 65, pp. 731–740. doi:10.1042/EBC20200089.

Grote, P. *et al.* (2013) “The Tissue-Specific lncRNA Fendrr Is an Essential Regulator of Heart and Body Wall Development in the Mouse,” *Developmental Cell*, 24(2), pp. 206–214. doi:10.1016/J.DEVCEL.2012.12.012/.

Hadley Wickham (2016) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag. Available at: https://ggplot2.tidyverse.org.

He, S. *et al.* (2015) “LongTarget: a tool to predict lncRNA DNA-binding motifs and binding sites via Hoogsteen base-pairing analysis,” *Bioinformatics*, 31(2), pp. 178–186. doi:10.1093/BIOINFORMATICS/BTU643.

Henikoff, S. and Furuyama, T. (2012) “The unconventional structure of centromeric nucleosomes,” *Chromosoma*, 121(4), pp. 341–352. doi:10.1007/S00412-012-0372-Y.

Henikoff, S. and Smith, M.M. (2015) “Histone Variants and Epigenetics,” *Cold Spring Harbor Perspectives in Biology*, 7(1). doi:10.1101/CSHPERSPECT.A019364.

Van Holde, K. (1985) *Physical biochemistry*. Englewood Cliffs, NJ: Prentice-Hall.

Hon, J. *et al.* (2013) “Triplex: an R/Bioconductor package for identification and visualization of potential intramolecular triplex patterns in DNA sequences,” *Bioinformatics*, 29(15), pp. 1900–1901. doi:10.1093/BIOINFORMATICS/BTT299.

Hörz, W. and Altenburger, W. (1981) “Sequence specific cleavage of DNA by micrococcal nuclease,” *Nucleic Acids Research*, 9(12), pp. 2643–2658. doi:10.1093/NAR/9.12.2643.

Hutcheon, T., Dixon, G.H. and Levy-Wilson, B. (1980) "Transcriptionally active mononucleosomes from trout testis are heterogeneous in composition.," *Journal of Biological Chemistry*, 255(2), pp. 681–685. doi:10.1016/S0021-9258(19)86231-2.

Ibrahim, M.M. *et al.* (2018) "Determinants of promoter and enhancer transcription directionality in metazoans," *Nature Communications 2018 9:1*, 9(1), pp. 1–15. doi:10.1038/s41467-018-06962-z.

Ishii, H., Kadonaga, J.T. and Ren, B. (2015) "MPE-seq, a new method for the genome-wide analysis of chromatin structure," *Proceedings of the National Academy of Sciences of the United States of America*, 112(27), pp. E3457–E3465. doi:10.1073/PNAS.1424804112.

Jain, A. *et al.* (2010) "Human DHX9 helicase unwinds triple-helical DNA structures," *Biochemistry*, 49(33), pp. 6992–6999. doi:10.1021/BI100795M.

Jalali, S. *et al.* (2017) "Genome-wide computational analysis of potential long noncoding RNA mediated DNA: DNA: RNA triplexes in the human genome," *Journal of Translational Medicine*, 15(1), pp. 1–17. doi:10.1186/S12967-017-1282-9/.

James, P.L., Brown, T. and Fox, K.R. (2003) "Thermodynamic and kinetic stability of intermolecular triple helices containing different proportions of C+·GC and T·AT triplets," *Nucleic Acids Research*, 31(19), pp. 5598–5606. doi:10.1093/NAR/GKG782.

Jeffers, T.E. and Lieb, J.D. (2017) "Nucleosome fragility is associated with future transcriptional response to developmental cues and stress in C. Elegans," *Genome Research*, 27(1), pp. 75–86. doi:10.1101/gr.208173.116.

Jenjaroenpun, P. and Kuznetsov, V.A. (2009) "TTS mapping: Integrative WEB tool for analysis of triplex formation target DNA sequences, G-quadruplets and non-protein coding regulatory DNA elements in the human genome," *BMC Genomics*, 10, pp. 1–18. doi:10.1186/1471-2164-10-S3-S9.

Jiang, H.-X. *et al.* (2015) "Divalent cations and molecular crowding buffers stabilize G-triplex at physiologically relevant temperatures." doi:10.1038/srep09255.

Jimenez-Useche, I. *et al.* (2013) "DNA Methylation Regulated Nucleosome Dynamics," *Scientific Reports 2013 3:1*, 3(1), pp. 1–5. doi:10.1038/srep02121.

Johnson, E.M., Sterner, R. and Allfrey, V.G. (1987) "Altered nucleosomes of active nucleolar chromatin contain accessible histone H3 in its hyperacetylated forms.," *Journal of Biological Chemistry*, 262(15), pp. 6943–6946. doi:10.1016/s0021-9258(18)48181-1.

Kalwa, M. *et al.* (2016) "The lncRNA HOTAIR impacts on mesenchymal stem cells via triple helix formation," *Nucleic Acids Research*, 44(22), pp. 10631–10643. doi:10.1093/nar/gkw802.

Kaushik Tiwari, M. *et al.* (2016) "Triplex structures induce DNA double strand breaks via replication fork collapse in NER deficient cells," *Nucleic Acids Research*, 44(16), pp. 7742–7754. doi:10.1093/NAR/GKW515.

Keenen, M.M. *et al.* (2021) “HP1 proteins compact dna into mechanically and positionally stable phase separated domains,” *eLife*, 10. doi:10.7554/ELIFE.64563.

Kelly, T.K. *et al.* (2012) “Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules,” *Genome research*, 22(12), pp. 2497–2506. doi:10.1101/GR.143008.112.

Kim, Y. *et al.* (2002) “Polyamines favor DNA triplex formation at neutral pH,” *Biochemistry*, 30(18), pp. 4455–4459. doi:10.1021/BI00232A012.

Kireeva, M.L. *et al.* (2002) “Nucleosome remodeling induced by RNA polymerase II: loss of the H2A/H2B dimer during transcription,” *Molecular cell*, 9(3), pp. 541–552. doi:10.1016/S1097-2765(02)00472-0.

Kunkler, C.N. *et al.* (2019) “Stability of an RNA•DNA–DNA triple helix depends on base triplet composition and length of the RNA third strand,” *Nucleic Acids Research*, 47(14), pp. 7213–7222. doi:10.1093/nar/gkz573.

Kuo, C.-C. *et al.* (2019) “Detection of RNA-DNA binding sites in long noncoding RNAs,” *Nucleic Acids Research* [Preprint], (1). doi:10.1093/nar/gkz037.

Kuo, C.C. *et al.* (2019) “Detection of RNA–DNA binding sites in long noncoding RNAs,” *Nucleic Acids Research*, 47(6), pp. e32–e32. doi:10.1093/NAR/GKZ037.

Langmead, B. and Salzberg, S.L. (2012) “Fast gapped-read alignment with Bowtie 2,” *Nature Methods*, 9(4), pp. 357–359. doi:10.1038/nmeth.1923.

Längst, G. *et al.* (1997) “Structural analysis of mouse rDNA: coincidence between nuclease hypersensitive sites, DNA curvature and regulatory elements in the intergenic spacer,” *Nucleic Acids Research*, 25(3), pp. 511–517. doi:10.1093/nar/25.3.511.

Längst, G. and Manelyte, L. (2015) “Chromatin remodelers: From function to dysfunction,” *Genes*. MDPI AG, pp. 299–324. doi:10.3390/genes6020299.

Lavelle, C. and Prunell, A. (2007) “Chromatin Polymorphism and the Nucleosome Superfamily: A Genealogy,” *Cell Cycle*, 6(17), pp. 2113–2119. doi:10.4161/cc.6.17.4631.

Lee, J.Y. and Lee, T.H. (2012) “Effects of DNA methylation on the structure of nucleosomes,” *Journal of the American Chemical Society*, 134(1), pp. 173–175. doi:10.1021/JA210273W.

Lee, M.S. and Garrard, W.T. (1991) “Transcription-induced nucleosome ‘splitting’: an underlying structure for DNase I sensitive chromatin.,” *The EMBO Journal*, 10(3), pp. 607–615. doi:10.1002/J.1460-2075.1991.TB07988.X.

Levendosky, R.F. *et al.* (2016) “The Chd1 chromatin remodeler shifts hexasomes unidirectionally,” *eLife*, 5. doi:10.7554/ELIFE.21356.

Li, G. *et al.* (2005) “Rapid spontaneous accessibility of nucleosomal DNA,” *Nature Structural and Molecular Biology*, 12(1), pp. 46–53. doi:10.1038/nsmb869.

Li, G. and Widom, J. (2004) “Nucleosomes facilitate their own invasion,” *Nature Structural and Molecular Biology*, 11(8), pp. 763–769. doi:10.1038/nsmb801.

Li, H. *et al.* (2009) “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, 25(16), pp. 2078–2079. doi:10.1093/bioinformatics/btp352.

Li, J. and Gilmour, D.S. (2013) “Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor,” *EMBO Journal*, 32(13), pp. 1829–1841. doi:10.1038/emboj.2013.111.

Li, S. *et al.* (2022) “DNA methylation cues in nucleosome geometry, stability and unwrapping,” *Nucleic Acids Research*, 50(4), pp. 1864–1874. doi:10.1093/NAR/GKAC097.

Li, X. and Fu, X.-D. (2019) “Chromatin-associated RNAs as facilitators of functional genomic interactions,” *Nature Reviews Genetics* [Preprint]. doi:10.1038/s41576-019-0135-1.

Li, Y., Syed, J. and Sugiyama, H. (2016) “RNA-DNA Triplex Formation by Long Noncoding RNAs,” *Cell Chemical Biology*, 23(11), pp. 1325–1333. doi:10.1016/j.chembiol.2016.09.011.

Liao, Y., Smyth, G.K. and Shi, W. (2014) “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features,” *Bioinformatics*, 30(7), pp. 923–930. doi:10.1093/bioinformatics/btt656.

de los Santos, C., Rosen, M. and Patel, D. (1989) “NMR studies of DNA (R+)n.(Y-)n.(Y+)n triple helices in solution: imino and amino proton markers of T.A.T and C.G.C+ base-triple formation,” *Biochemistry*, 28(18), pp. 7282–7289. doi:10.1021/BI00444A021.

Lubitz, I., Zikich, D. and Kotlyar, A. (2010) “Specific high-affinity binding of thiazole orange to triplex and g-quadruplex DNA,” *Biochemistry*, 49(17), pp. 3567–3574. doi:10.1021/bi1000849.

Maine, I.P. and Kodadek, T. (1994) “Efficient unwinding of triplex DNA by a DNA helicase,” *Biochemical and biophysical research communications*, 204(3), pp. 1119–1124. doi:10.1006/BBRC.1994.2578.

Maldonado, R. *et al.* (2017) “Purine- and pyrimidine-triple helix forming oligonucleotides recognize qualitatively different target sites at the ribosomal DNA locus,” *RNA*, (24), pp. 371–380. doi:10.1261/rna.063800.117.

Maldonado, R. *et al.* (2019) “Nucleosomes Stabilize ssRNA-dsDNA Triple Helices in Human Cells,” *Molecular Cell*, 73(6), pp. 1–12. doi:10.1016/j.molcel.2019.01.007.

Marcel, M. (2011) “Cutadapt removes adapters from high-throughput sequencing reads,” *EMBnet.journal*, 17(1), pp. 10–12. doi:https://doi.org/10.14806/ej.17.1.200.

Mieczkowski, J. *et al.* (2016) “MNase titration reveals differences between nucleosome occupancy and chromatin accessibility,” *Nature Communications*, 7(May), pp. 1–11. doi:10.1038/ncomms11485.

Mondal, T. *et al.* (2015) “MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA–DNA triplex structures,” *Nature Communications*, 6(1), p. 7743. doi:10.1038/ncomms8743.

Morgan, A.R. and Wells, R.D. (1968) “Specificity of the three-stranded complex formation between double-stranded DNA and single-stranded RNA containing repeating nucleotide sequences,” *Journal of molecular biology*, 37(1), pp. 63–80. doi:10.1016/0022-2836(68)90073-9.

Mueller, B. *et al.* (2017) “Widespread changes in nucleosome accessibility without changes in nucleosome occupancy during a rapid transcriptional induction,” *Genes & Development*, 31(5), pp. 451–462. doi:10.1101/gad.293118.116.

Nettling, M. *et al.* (2015) “DiffLogo: a comparative visualization of sequence motifs,” *BMC Bioinformatics*, 16(387). doi:10.1186/s12859-015-0767-x.

North, J.A. *et al.* (2012) “Regulation of the nucleosome unwrapping rate controls DNA accessibility,” *Nucleic Acids Research*, 40(20), pp. 10215–10227. doi:10.1093/nar/gks747.

Ohno, S. (1972) “So much ‘junk’ DNA in our genome.,” *Brookhaven symposia in biology*, 23, pp. 366–70.

Okonechnikov, K., Conesa, A. and García-Alcalde, F. (2016) “Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data,” *Bioinformatics*, 32(2), pp. 292–294. doi:10.1093/BIOINFORMATICS/BTV566.

O’Leary, V.B. *et al.* (2015) “PARTICLE, a triplex-forming long ncRNA, regulates locus-specific methylation in response to low-dose irradiation,” *Cell Reports*, 11(3), pp. 474–485. doi:10.1016/J.CELREP.2015.03.043.

Olivas, W.M. and Maher, L.J. (1995) “Competitive Triplex/Quadruplex Equilibria Involving Guanine-Rich Oligonucleotides,” *Biochemistry*, 34(1), pp. 278–284. doi:10.1021/bi00001a034.

Pagès, H. *et al.* (2021) “Biostrings: Efficient manipulation of biological strings. R package version 2.62.0.” Available at: https://bioconductor.org/packages/Biostrings.

Patro, R. *et al.* (2017) “Salmon provides fast and bias-aware quantification of transcript expression,” *Nature Methods 2017 14:4*, 14(4), pp. 417–419. doi:10.1038/nmeth.4197.

Paugh, S.W., Coss, D.R., Bao, J., Laudermilk, L.T., Grace, C.R., Ferreira, A.M., Waddell, M.B., Ridout, G., Naeve, D., Leuze, M., LoCascio, P.F., Panetta, J.C., Wilkinson, M.R., Pui, C.-H.H., *et al.* (2016) “MicroRNAs Form Triplexes with Double Stranded DNA at Sequence-Specific Binding Sites; a Eukaryotic Mechanism via which microRNAs Could Directly Alter Gene Expression,” *PLOS Computational Biology*. Edited by I. Ioshikhes, 12(2), p. e1004744. doi:10.1371/journal.pcbi.1004744.

Paugh, S.W., Coss, D.R., Bao, J., Laudermilk, L.T., Grace, C.R., Ferreira, A.M., Waddell, M.B., Ridout, G., Naeve, D., Leuze, M., LoCascio, P.F., Panetta, J.C., Wilkinson, M.R., Pui, C.H., *et al.* (2016) “MicroRNAs Form Triplexes with Double Stranded DNA at Sequence-Specific Binding Sites; a Eukaryotic Mechanism via which microRNAs Could Directly Alter Gene Expression,” *PLOS Computational Biology*, 12(2), p. e1004744. doi:10.1371/JOURNAL.PCBI.1004744.

Pepenella, S., Murphy, K.J. and Hayes, J.J. (2014) “Intra- and inter-nucleosome interactions of the core histone tail domains in higher-order chromatin structure,” *Chromosoma*, 123(1–2), pp. 3–13. doi:10.1007/S00412-013-0435-8.

Philip, P. *et al.* (2015) “CBP binding outside of promoters and enhancers in Drosophila melanogaster,” *Epigenetics & Chromatin*, 8(1), p. 48. doi:10.1186/S13072-015-0042-4.

Polach, K.J. and Widom, J. (1995) “Mechanism of protein access to specific DNA sequences in chromatin: A dynamic equilibrium model for gene regulation,” *Journal of Molecular Biology*, 254(2), pp. 130–149. doi:10.1006/jmbi.1995.0606.

Postepska-Igielska, A. *et al.* (2015) “LncRNA Khps1 Regulates Expression of the Proto-oncogene SPHK1 via Triplex-Mediated Changes in Chromatin Structure,” *Molecular Cell*, 60(4), pp. 626–636. doi:10.1016/J.MOLCEL.2015.10.001.

Quinlan, A.R. and Hall, I.M. (2010) “BEDTools: a flexible suite of utilities for comparing genomic features,” *Bioinformatics*, 26(6), pp. 841–842. doi:10.1093/bioinformatics/btq033.

Ramachandran, S., Ahmad, K. and Henikoff, S. (2017) “Transcription and Remodeling Produce Asymmetrically Unwrapped Nucleosomal Intermediates,” *Molecular Cell*, 68(6), pp. 1038-1053.e4. doi:10.1016/j.molcel.2017.11.015.

Ramírez, F. *et al.* (2016) “deepTools2: a next generation web server for deep-sequencing data analysis,” *Web Server issue Published online*, 44. doi:10.1093/nar/gkw257.

Rhee, H.S. *et al.* (2014) “Subnucleosomal structures and nucleosome asymmetry across a genome,” *Cell*, 159(6), pp. 1377–1388. doi:10.1016/j.cell.2014.10.054.

Ritchie, M.E. *et al.* (2015) “limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Research*, 43(7). doi:10.1093/nar/gkv007.

Robinson, M.D., Mccarthy, D.J. and Smyth, G.K. (2010) “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data,” *BIOINFORMATICS APPLICATIONS NOTE*, 26(1), pp. 139–140. doi:10.1093/bioinformatics/btp616.

Rougée, M. *et al.* (1992) “Kinetics and Thermodynamics of Triple-Helix Formation: Effects of Ionic Strength and Mismatches,” *Biochemistry*, 31(38), pp. 9269–9278. doi:10.1021/BI00153A021.

Sanulli, S. *et al.* (2019) “HP1 reshapes nucleosome core to promote phase separation of heterochromatin,” *Nature*, 575(7782), pp. 390–394. doi:10.1038/s41586-019-1669-2.

Schneider, I. (1972) “Cell lines derived from late embryonic stages of Drosophila melanogaster,” *Journal of embryology and experimental morphology*, 27(2), pp. 353–365.

Schones, D.E. *et al.* (2008) “Dynamic Regulation of Nucleosome Positioning in the Human Genome,” *Cell*, 132(5), pp. 887–898. doi:10.1016/j.cell.2008.02.022.

Schubert, T. *et al.* (2012) “Df31 Protein and snoRNAs Maintain Accessible Higher-Order Structures of Chromatin,” *Molecular Cell*, 48(3), pp. 434–444. doi:10.1016/j.molcel.2012.08.021.

Schwalb, B. *et al.* (2020) “LSD: Lots of Superior Depictions.” Available at: https://cran.r-project.org/package=LSD.

Schwartz, U. *et al.* (2019) “Characterizing the nuclease accessibility of DNA in human cells to map higher order structures of chromatin,” *Nucleic Acids Research*, 47(3), p. 1239. doi:10.1093/NAR/GKY1203.

Semerad, C.L. and James maher, L. (1994) “Exclusion of RNA strands from a purine motif triple helix,” *Nucleic Acids Research*, 22(24), pp. 5321–5325. doi:10.1093/nar/22.24.5321.

Sentürk Cetin, N. *et al.* (2019) “Isolation and genome-wide characterization of cellular DNA:RNA triplex structures,” *Nucleic Acids Research*, 47(5), pp. 2306–2321. doi:10.1093/nar/gky1305.

Soibam, B. and Zhamangaraeva, A. (2021) “LncRNA:DNA triplex-forming sites are positioned at specific areas of genome organization and are predictors for Topologically Associated Domains,” *BMC Genomics*, 22(1), pp. 1–10. doi:10.1186/S12864-021-07727-7.

Sugimoto, N. *et al.* (2001) “pH and cation effects on the properties of parallel pyrimidine motif DNA triplexes,” *Biochemistry*, 40(31), pp. 9396–9405. doi:10.1021/BI010666L.

Sun, Q., Hao, Q. and Prasanth, K. v. (2018) “Nuclear long noncoding RNAs: key regulators of gene expression,” *Trends in genetics : TIG*, 34(2), p. 142. doi:10.1016/J.TIG.2017.11.005.

Tachiwana, H. *et al.* (2011) “Crystal structure of the human centromeric nucleosome containing CENP-A,” *Nature 2011 476:7359*, 476(7359), pp. 232–235. doi:10.1038/nature10258.

Thuong, N.T. and Hélène, C. (1993) “Sequence-Specific Recognition and Modification of Double-Helical DNA by Oligonucleotides,” *Angewandte Chemie International Edition in English*, 32(5), pp. 666–690. doi:10.1002/anie.199306661.

Tims, H.S. *et al.* (2011) “Dynamics of nucleosome invasion by DNA binding proteins,” *Journal of Molecular Biology*, 411(2), pp. 430–448. doi:10.1016/j.jmb.2011.05.044.

Toscano-Garibay, J.D. and Aquino-Jarquin, G. (2014) “Transcriptional regulation mechanism mediated by miRNA-DNA•DNA triplex structure stabilized by Argonaute,” *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1839(11), pp. 1079–1083. doi:10.1016/j.bbagrm.2014.07.016.

Trembinski, D.J. *et al.* (2020) “Aging-regulated anti-apoptotic long non-coding RNA Sarrah augments recovery from acute myocardial infarction,” *Nature Communications 2020 11:1*, 11(1), pp. 1–14. doi:10.1038/s41467-020-15995-2.

Tsanev, R., and Petrov, P. (1976) “The substructure of chromatin and its variations as revealed by electron microscopy,” *J. Microsc. Biol. Cell.*, (27), pp. 11–19.

Voong, L.N. *et al.* (2016) “Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping,” *Cell*, 167(6), pp. 1555-1570.e15. doi:10.1016/j.cell.2016.10.049.

Whyte, W.A. *et al.* (2013) “Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes,” *Cell*, 153(2), pp. 307–319. doi:10.1016/J.CELL.2013.03.035.

Wu, P. *et al.* (2002) “Effect of divalent cations and cytosine protonation on thermodynamic properties of intermolecular DNA double and triple helices,” *Journal of Inorganic Biochemistry*, 91, pp. 277–285. doi:10.1016/s0162-0134(02)00444-0.

Wu, Q. *et al.* (2007) “High-affinity triplex-forming oligonucleotide target sequences in mammalian genomes,” *Molecular Carcinogenesis*, 46(1), pp. 15–23. doi:10.1002/MC.20261.

Xi, Y. *et al.* (2011) “Nucleosome fragility reveals novel functional states of chromatin and poises genes for activation,” *Genome Research*, 21(5), pp. 718–724. doi:10.1101/gr.117101.110.

Xiao, M. *et al.* (2017) “MicroRNAs activate gene transcription epigenetically as an enhancer trigger,” *RNA Biology*, 14(10), pp. 1326–1334. doi:10.1080/15476286.2015.1112487.

Yu, G. *et al.* (2012) “clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters,” *OMICS*, 16(5). doi:10.1089/omi.2011.0118.

Yu, G., Wang, L.-G. and He, Q.-Y. (2015) “ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization,” *Bioinformatics*, 31(14), pp. 2382–2383. doi:10.1093/bioinformatics/btv145.

Zeraati, M. *et al.* (2018) “I-motif DNA structures are formed in the nuclei of human cells,” *Nature Chemistry*, 10(6), pp. 631–637. doi:10.1038/s41557-018-0046-3.

Zhang, Y., Long, Y. and Kwoh, C.K. (2020) “Deep learning based DNA:RNA triplex forming potential prediction,” *BMC Bioinformatics*, 21(1), pp. 1–13. doi:10.1186/S12859-020-03864-0.

Zhong, J. *et al.* (2016) “Mapping nucleosome positions using DNase-seq,” *Genome Research*, 26(3), p. 351. doi:10.1101/GR.195602.115.

Zhu, L.J. *et al.* (2010) “ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data,” *BMC Bioinformatics*, 11(237). doi:10.1186/1471-2105-11-237.

# 8 ACKNOWLEDGEMENTS

I would firstly like to thank my supervisor**, prof. Dr. Gernot Längst**, who has given me this opportunity and provided a supportive environment that helped me grow as a scientist. I would also like to acknowledge my mentors**, prof. Dr. Gunter Meister,** and **prof. Dr. Michael Rehli**. Thank you for giving me excellent feedback through the years and for all the fruitful discussions.

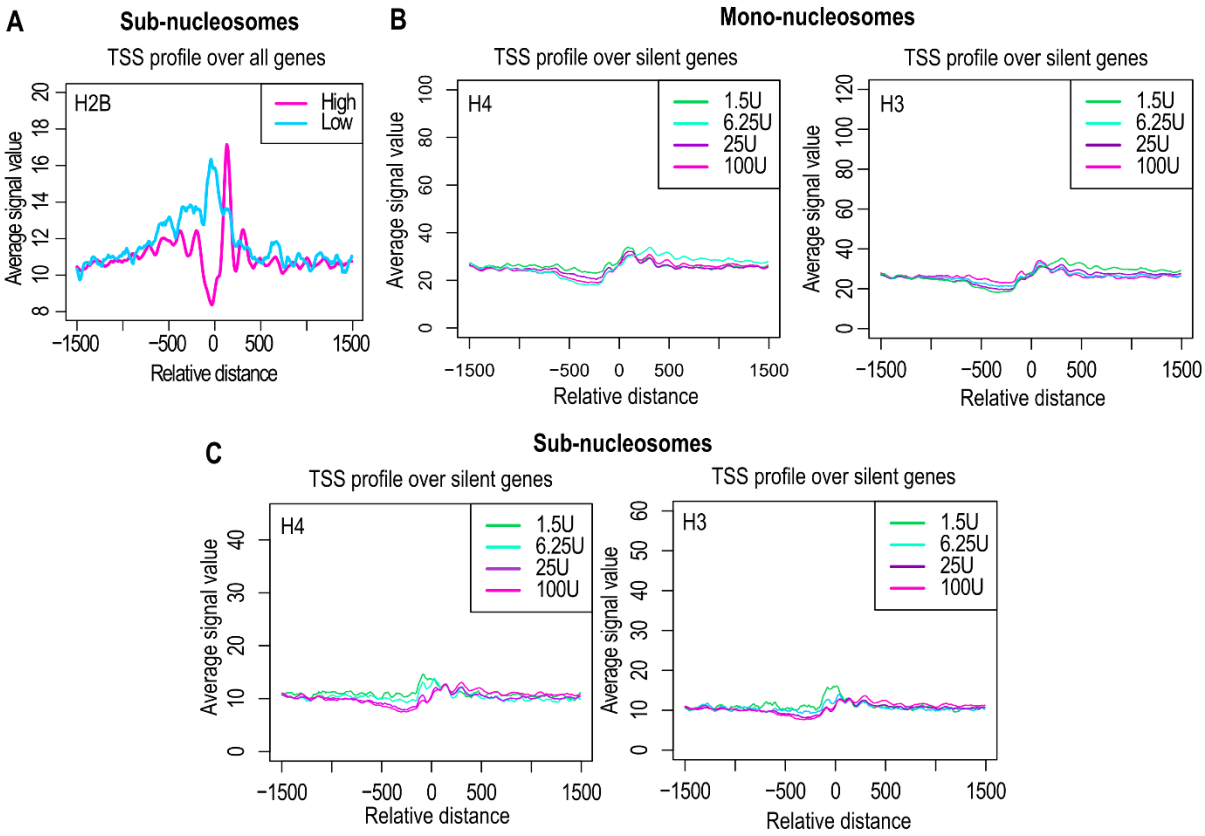
I want to thank current and former members of **AG Längst**, who have brightened my days and provided support during challenging times. Specifically, I would like to name the following individuals: **Caro and Eli**; thank you for being my friends for the past 4 years. You are and always will be my family. **Vicky**, thank you for all the beer Fridays, memes, laughter, and for being an excellent officemate, always ready for a discussion and providing instant feedback. I hope our friendship lasts a lifetime. **Rodrigo**, thank you for bringing fun to every experiment and reminding me that science is and should be fun. **Matthias**, thank you for all the lunches together, for all the laughter, and for the excellent coffee. **Sabrina and Nora**, thank you for the emotional support and fun times together. **Simon**, thank you for all the discussions and for the fun times we had organizing the BIG seminars. **Laura**, thank you for all the discussions and interesting talks we had together. **Uwe**, thank you for your valuable input.

Additionally, I want to thank my mother, **Renata Wernig**, for raising me on her own and sacrificing herself to give me a better life. She worked two jobs and spent every cent on language and career development courses for me.

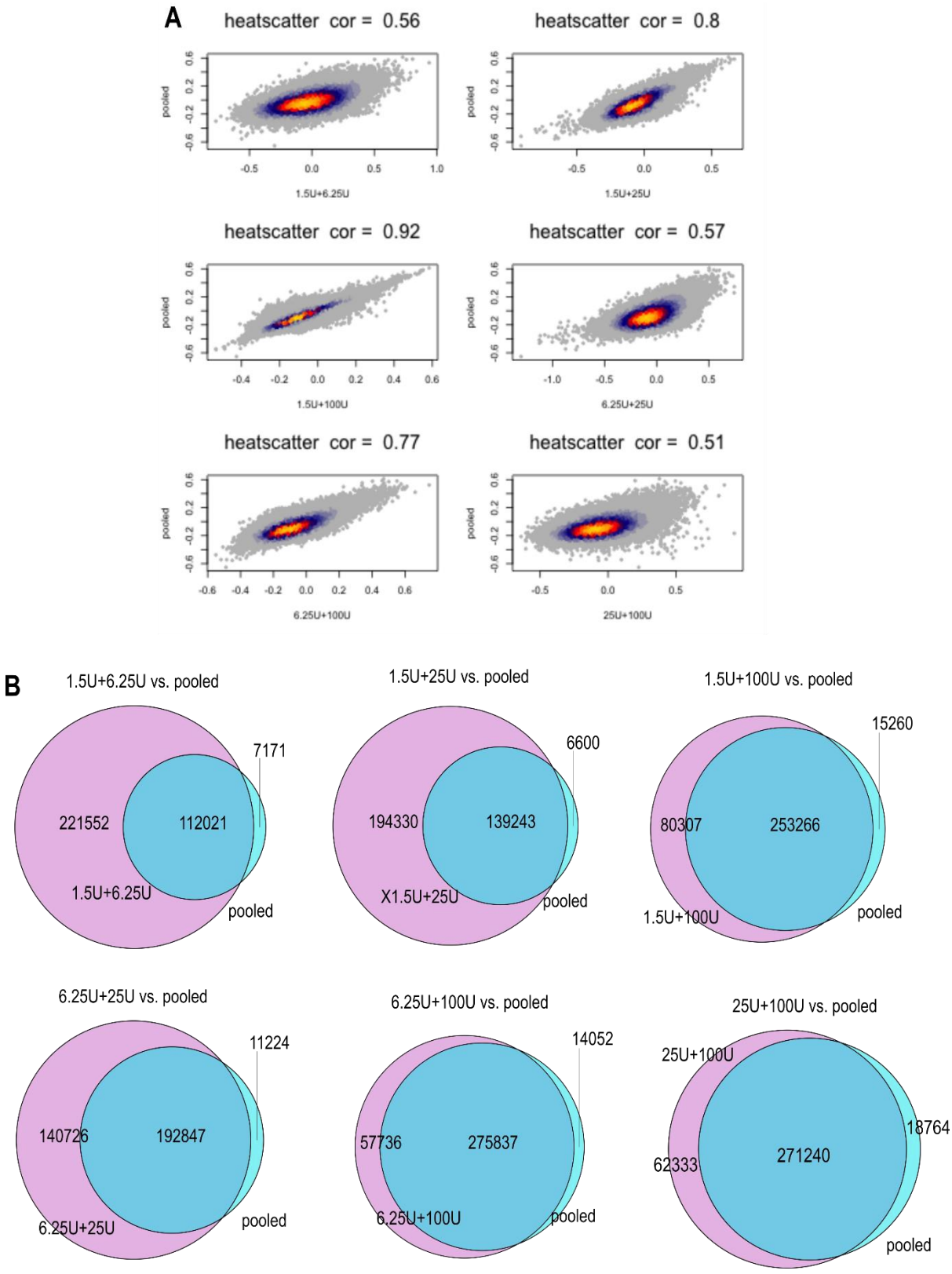
A special thanks goes to my husband, **Jernej Zorc**. I could not have done this without you. You are my biggest cheerleader and have always believed in me. Thank you for moving to three countries for me and putting your career aside to help me build mine. Thank you for your patience, understanding, love, and support.

I acknowledge that it is not only the people who have helped me achieve my dreams. I am grateful for growing up in a country (**Slovenija**) where all are given the same opportunities, regardless of the amount of money their family has. If it weren't for free school trips, lunches, school supplies, and most importantly, free and high-quality education at all levels, I would not be where I am today. We take this for granted, although it is a crucial element to achieving equality so that no child is left behind.

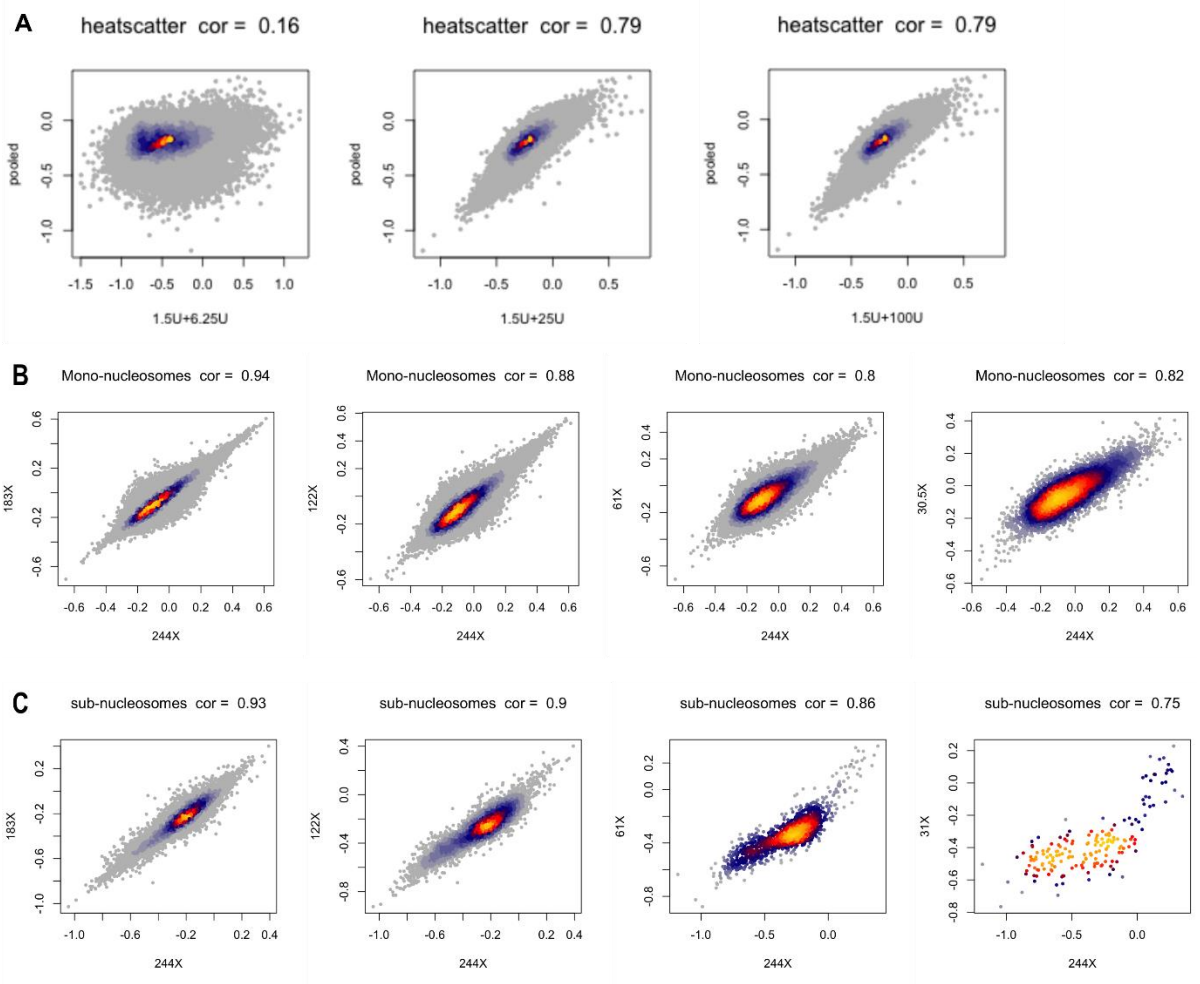
# **APPENDIX 1:** Nucleosome stability - supplemental figures



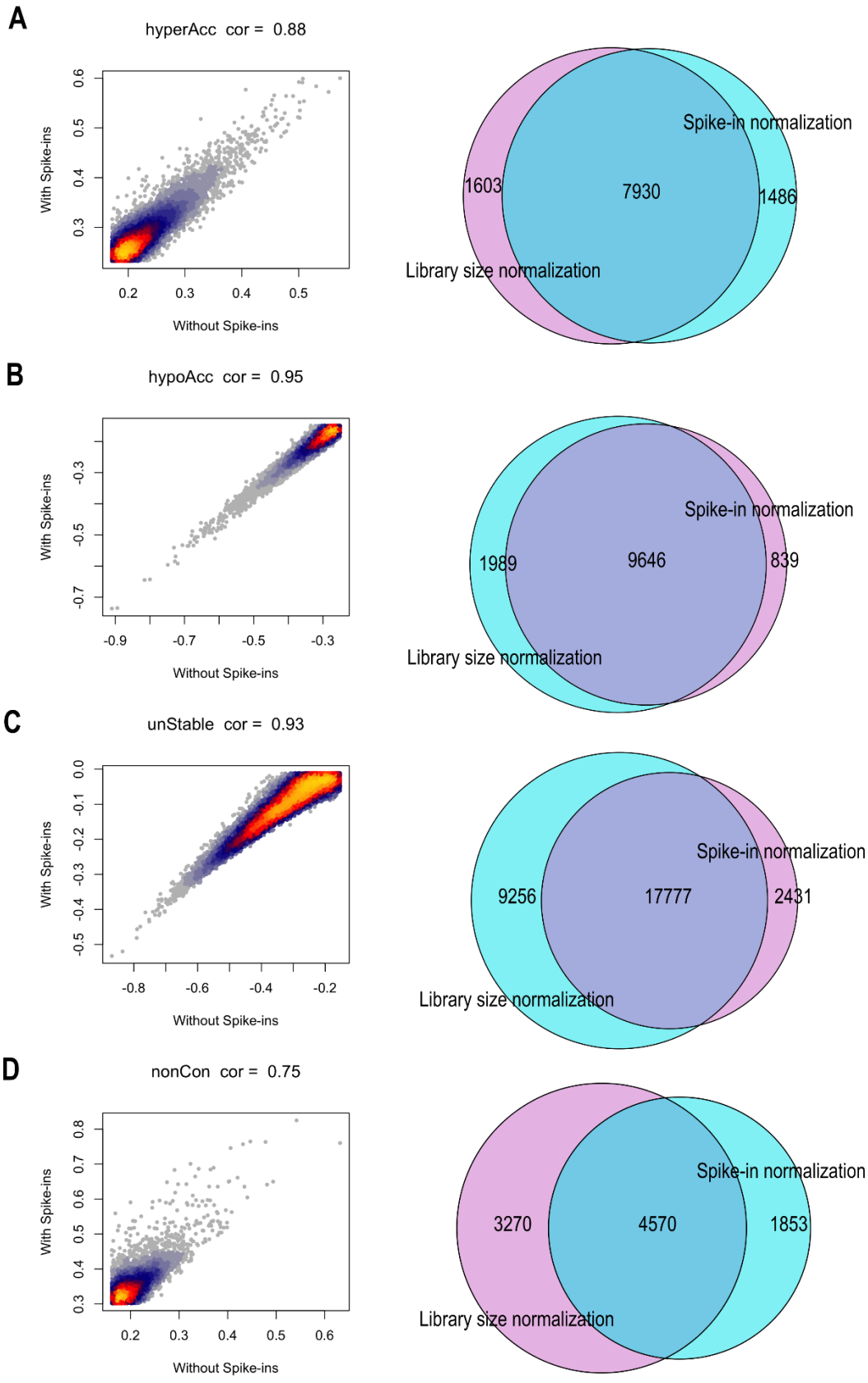

**Supplemental figure 1:** MNase-seq TSS profile. (A) MNase-H2B H3-ChIP-seq profile over the TSS of all genes. (B-C) MNase-H4 H3-ChIP (left) and MNase-H3-ChIP-seq (right) profile over the TSS of silent genes for mono-nucleosomes (B) and sub-nucleosomes (C).




**Supplemental figure 2:** Comparing the nucMACC scores between two or four MNase titrations for mono-nucleosomes. Correlation between nucMACC scores (A) and the overlap between calles nucleosome positions (B).




**Supplemental figure 3:** Comparing the nucMACC scores between two or four MNase titrations for sub-nucleosomes (A). Comparing the nucMACC scores for samples with reduced sequencing depth for mono- (B) and sub- (C) nucleosomes.




**Supplemental figure 4:** Comparing the nucMACC scores between samples where spike-ins are used or not. Correlation between nucMACC scores (left) and the overlap between calles positions (right) with or without spike-ins information for hyper-accessible (A), hypo-accessible (B), and un-stable (C) and non-canonical (D) nucleosomes.

## **APPENDIX 2:** Triplex oligonucleotides

| Oligo Name | Sequence [5' to 3'] | RNA DNA | Label |
|---|---|---|---|
| En3_D for Cy5 (3) | TCTTTTTTTTTTTTTTTCTTTTTTCCTCC | DNA | 5'Cy5 |
| En3_D_rev (4) | GGAGGAAAAAAGAAAAAAAAAAAAAAAGA | DNA | NL |
| En3_TFO DNA Pyr NL (6) | CCTCCTTTTTTCTTTTTTTTTTTTTTCT | DNA | NL |
| En3_TFO DNA Pur NL (7) | AGAAAAAAAAAAAAAAAGAAAAAAGGAGG | DNA | NL |
| En3_TFO RNA Pyr NL (10) | CCUCCUUUUUUCUUUUUUUUUUUUUUCU | RNA | NL |
| En3_TFO RNA Pur NL (11) | AGAAAAAAAAAAAAAAAGAAAAAAGGAGG | RNA | NL |
| En3_TFO RNA Pyr FAM (12) | CCUCCUUUUUUCUUUUUUUUUUUUUUCU | RNA | 5'FAM |
| En3_TFO DNA Pyr 12T NL (13) | CCTCCTTTTTTTTTTTTTTTTTTTTCT | DNA | NL |
| En3_TFO DNA Pyr 12A NL (14) | CCTCCTTTTTTATTTTTTTTTTTTTTCT | DNA | NL |
| En3_TFO DNA Pyr 12G NL (15) | CCTCCTTTTTTGTTTTTTTTTTTTTTCT | DNA | NL |
| CCCC motif_TTS For NL (23) | CCCCCCCCCCCCCCCCCCCCCCCCCCCCC | DNA | NL |
| CCCC motif_TTS For Cy5 (24) | CCCCCCCCCCCCCCCCCCCCCCCCCCCCC | DNA | 5'Cy5 |
| CCCC motif_TTS Rev NL (25) | GGGGGGGGGGGGGGGGGGGGGGGGGGGGG | DNA | NL |
| CCCT motif_TTS For Cy5 (27) | CCCTCCCTCCCTCCCTCCCTCCCTCCCTC | DNA | 5'Cy5 |
| CCCT motif_TTS Rev NL (28) | GAGGGAGGGAGGGAGGGAGGGAGGGAGGG | DNA | NL |
| CCCT motif_TFO DNA Pyr NL (29) | CCCTCCCTCCCTCCCTCCCTCCCTCCCTC | DNA | NL |
| CCCT motif_TFO DNA Pur NL (30) | GGGAGGGAGGGAGGGAGGGAGGGAGGGAG | DNA | NL |
| CCCT motif_TFO RNA Pyr NL (31) | CUCCCUCCCUCCCUCCCUCCCUCCCUCCC | RNA | NL |
| CCCT motif_TFO RNA Pur NL (32) | GGGAGGGAGGGAGGGAGGGAGGGAGGGAG | RNA | NL |
| CTTT motif_TTS For NL (33) | CTTTCTTTCTTTCTTTCTTTCTTTCTTTC | DNA | NL |
| CTTT motif_TTS For Cy5 (34) | CTTTCTTTCTTTCTTTCTTTCTTTCTTTC | DNA | 5'Cy5 |
| CTTT motif_TTS Rev NL (35) | GAAAGAAAGAAAGAAAGAAAGAAAGAAAG | DNA | NL |
| CTTT motif_TFO RNA Pyr NL (36) | CUUUCUUUCUUUCUUUCUUUCUUUCUUUC | RNA | NL |
| CTTT motif_TFO RNA Pur NL (37) | GAAAGAAAGAAAGAAAGAAAGAAAGAAAG | RNA | NL |
| TTTT motif_TTS For NL (38) | TTTTTTTTTTTTTTTTTTTTTTTTTTTT | DNA | NL |
| TTTT motif_TTS For Cy5 (39) | TTTTTTTTTTTTTTTTTTTTTTTTTTTT | DNA | 5'Cy5 |
| TTTT motif_TTS Rev NL (40) | AAAAAAAAAAAAAAAAAAAAAAAAAAAA | DNA | NL |
| CTCT motif_TTS For NL (41) | CTCTCTCTCTCTCTCTCTCTCTCTCTC | DNA | NL |
| CTCT motif_TTS For Cy5 (42) | CTCTCTCTCTCTCTCTCTCTCTCTCTC | DNA | 5'Cy5 |
| CTCT motif_TTS Rev NL (43) | GAGAGAGAGAGAGAGAGAGAGAGAGAG | DNA | NL |
| CTCT motif_TFO RNA Pyr NL (44) | CUCUCUCUCUCUCUCUCUCUCUCUCUC | RNA | NL |
| CTCT motif_TFO RNA Pur NL (45) | GAGAGAGAGAGAGAGAGAGAGAGAGAG | RNA | NL |
| TGFBR1_TTS For Cy5 (46) | CTCTCTCCCTCTCTCT | DNA | 5'Cy5 |
| TGFBR1_TTS Rev NL (47) | AGAGAGAGGGAGAGAG | DNA | NL |
| TGFB2_TTS For Cy5 (48) | TCTCTCTCTCTGCTCTCTG | DNA | 5'Cy5 |
| TGFB2_TTS Rev NL (49) | CAGAGAGCAGAGAGAGAGA | DNA | NL |
| SMAD2_TTS For Cy5 (50) | CTCTCCCTCTCT | DNA | 5'Cy5 |
| SMAD2_TTS Rev NL (51) | AGAGAGGGAGAG | DNA | NL |
| MEG3_TFO DNA Pur (52) | CGGAGAGCAGAGAGGGAGCG | DNA | NL |
| PCDH7_TTS For Cy5 (53) | TTTCTCTCTCCTCCCTCCTCTCTCTCCT | DNA | 5'Cy5 |
| PCDH7_TTS Rev NL (54) | AGGAGAGAGAGGAGGGAGGAGAGAGAAA | DNA | NL |
| PCDH7_TFO DNA Pur NL (55) | AGAGGAGGGAAGAGAG | DNA | NL |
| FG1_TTS For Cy5 (56) | GTTCGAATCCTTCCCCCCCCACCACCCCCTCCCCCTC | DNA | 5'Cy5 |
| FG1_TTS Rev NL (57) | GAGGGGGAGGGGGTGGTGGGGGGGGAAGGATTC GAAC | DNA | NL |
| AG30_TFO DNA Pur (58) | AGGAAGGGGGGGGTGGTGGGGGAGGGGGAG | DNA | NL |
| CCL2_TTS4 For Cy5 (71) | TCCGCCCTCTCTCCCTC | DNA | 5'Cy5 |
| CCL2_TTS4 Rev NL (72) | GAGGGAGAGAGGGCGGA | DNA | NL |
| LNMAT1_TFO4 NL (73) | AGGCTGGAGTGCAGTG | DNA | NL |

| | | | |
|---|---|---|---|
| PITX2_TTS For Cy5 (74) | TTCTCCCTCCCCTCCATCCTCTTCCTTCTCCTCCTCCTC TT | DNA | 5'Cy5 |
| PITX2_TTS Rev NL (75) | AAGAGGAGGAGGAGAAGGAAGAGGATGGAGGGG AGGGAGAA | DNA | NL |
| FENDRR_TFO NL (76) | GAGGAGGAGGAGGAGGAGGAGGAGGAGGA GGAGGAGGA | DNA | NL |
| KIAA1324_TTS For Cy5 (77) | GAGAGAGAGAGAGAGAGAGAGAGAGAGAG AGAGAG | DNA | 5'Cy5 |
| KIAA1324_TTS Rev NL (78) | CTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTC | DNA | NL |
| PROX1_TTS For Cy5 (79) | GAGAGAGATAGAGAGAGAGAGAGAGAGAGAG AGAGAG | DNA | 5'Cy5 |
| PROX1_TTS Rev NL (80) | CTCTCTCTCTCTCTCTCTCTCTCTCTATCTCTCTC | DNA | NL |
| BCL9_TTS For Cy5 (81) | GAGAGACAGAGAGAGAGAGAGAGAGAGAGAG AGAGAG | DNA | 5'Cy5 |
| BCL9_TTS Rev NL (82) | CTCTCTCTCTCTCTCTCTCTCTCTCTCTGTCTCTC | DNA | NL |
| RP11-84A10_TFO DNA NL (83) | GAGAGAGAGAGAGAGAGAGAGAGAGAGAG AGAGAG | DNA | NL |
| WWOX1_TTS For Cy5 (84) | TCCCCCCTCCCTCCTT | DNA | 5'Cy5 |
| WWOX1_TTS Rev NL (85) | AAGGAGGGAGGGGGGA | DNA | NL |
| PARTICLE_TFO1 DNA NL (86) | TTGGTGGGTGGGGGCT | DNA | NL |
| WWOX2_TTS For Cy5 (87) | TTCTTTCCTCCCTCCC | DNA | 5'Cy5 |
| WWOX2_TTS Rev NL (88) | GGGAGGGAGGAAAGAA | DNA | NL |
| PARTICLE_TFO2 DNA NL (89) | TTTTTTGGTGGGTGGG | DNA | NL |
| miR-483-5p_TTS For Cy5 (90) | CTGCTAGCTACTGGGGGAAGAAGAGGGGGCAGAG CTGCTAGCTACT | DNA | 5'Cy5 |
| miR-483-5p_TTS Rev NL (91) | AGTAGCTAGCAGCTCTGCCCCCTCTTCTTCCCCCAGT AGCTAGCAG | DNA | NL |
| miR-483-5p_TFO (92) | AAGACGGGAGGAAAGAAGGGAG | DNA | NL |
| CTCT motif_TFO-2 DNA Pyr (59) | C TCT CTC TCT CTC TCT CTC TCT CTC TC | DNA | NL |
| CTCT motif_TFO-4 DNA Pyr (60) | CT CTC TCT CTC TCT CTC TCT CTC TC | DNA | NL |
| CTCT motif_TFO-6 DNA Pyr (61) | CTC TCT CTC TCT CTC TCT CTC TC | DNA | NL |
| CTCT motif_TFO-8 DNA Pyr (62) | C TCT CTC TCT CTC TCT CTC TC | DNA | NL |
| CTCT motif_TFO-10 DNA Pyr (63) | CT CTC TCT CTC TCT CTC TC | DNA | NL |
| CTCT motif_TFO-12 DNA Pyr (64) | CTC TCT CTC TCT CTC TC | DNA | NL |
| CTCT motif_TFO-2 DNA Pur (65) | GAG AGA GAG AGA GAG AGA GAG AGA GAG | DNA | NL |
| CTCT motif_TFO-4 DNA Pur (66) | GAG AGA GAG AGA GAG AGA GAG AGA G | DNA | NL |
| CTCT motif_TFO-6 DNA Pur (67) | GAG AGA GAG AGA GAG AGA GAG AG | DNA | NL |
| CTCT motif_TFO-8 DNA Pur (68) | GAG AGA GAG AGA GAG AGA GAG | DNA | NL |
| CTCT motif_TFO-10 DNA Pur (69) | GAG AGA GAG AGA GAG AGA G | DNA | NL |
| CTCT motif_TFO-12 DNA Pur (70) | GAG AGA GAG AGA GAG AG | DNA | NL |
| CTCT motif_TFO-14 DNA Pur (93) | GAGAGAGAGAGAGAG | DNA | NL |
| CTCT motif_TFO-18 DNA Pur (94) | GAGAGAGAGAG | DNA | NL |
| CTCT motif_TFO-22 DNA Pur (95) | GAGAGAG | DNA | NL |
| miR-483-5p_TTS-2_rev_NL (150) | TGGGGGAAGAAGAGGGGGCAGAGA | DNA | NL |
| miR-483-5p_TTS-2_for_Cy5 (151) | TCTCTGCCCCCTCTTCTTCCCCCA | DNA | 5'Cy5 |
| miR-24-1-TTS_for_Cy5 (152) | GTGACCGAGTCAAGTCGTCCTTGTCCTC | DNA | NL |
| miR-24-1-TTS_rev_NL (153) | GAGGACAAGGACGACTTGACTCGGTCAC | DNA | NL |
| miR-24-1-TFO_NL (154) | TGGCTCAGTTCAGCAGGAACAG | DNA | 5'Cy5 |
| CTCT motif_TFO-2 DNA Mix NL (158) | GTGTGTGTGTGTGTGTGTGTGTGTGTG | DNA | NL |
| CTCT motif_TFO-4 DNA Mix NL (159) | GTGTGTGTGTGTGTGTGTGTGTGTG | DNA | NL |
| CTCT motif_TFO-6 DNA Mix NL (160) | GTGTGTGTGTGTGTGTGTGTGTG | DNA | NL |
| CTCT motif_TFO-8 DNA Mix NL (161) | GTGTGTGTGTGTGTGTGTGTG | DNA | NL |
| CTCT motif_TFO-10 DNA Mix NL (162) | GTGTGTGTGTGTGTGTGTG | DNA | NL |
| CTCT motif_TFO-12 DNA Mix NL (163) | GTGTGTGTGTGTGTGTG | DNA | NL |
| CTCT motif_TFO-14 DNA Mix NL (164) | GTGTGTGTGTGTGTG | DNA | NL |

| | | | |
|---|---|---|---|
| En3_TFO_D_Y_28G (166) | CCTCCTTTTTTCTTTTTTTTTTTTTTTGT | DNA | NL |
| En3_TFO_D_Y_28T (167) | CCTCCTTTTTTCTTTTTTTTTTTTTTTT | DNA | NL |
| En3_TFO_D_Y_28A (168) | CCTCCTTTTTTCTTTTTTTTTTTTTTTAT | DNA | NL |
| En3_TFO_D_Y_19G (169) | CCTCCTTTTTTCTTTTTTGTTTTTTTTCT | DNA | NL |
| En3_TFO_D_Y_19C (170) | CCTCCTTTTTTCTTTTTTCTTTTTTTTCT | DNA | NL |
| En3_TFO_D_Y_19A (171) | CCTCCTTTTTTCTTTTTTATTTTTTTTCT | DNA | NL |
| En3_TFO_D_Y_5G (172) | CCTCGTTTTTTCTTTTTTTTTTTTTTTCT | DNA | NL |
| En3_TFO_D_Y_5T (173) | CCTCTTTTTTTCTTTTTTTTTTTTTTTCT | DNA | NL |
| En3_TFO_D_Y_5A (174) | CCTCATTTTTTCTTTTTTTTTTTTTTTCT | DNA | NL |
| CTCT motif_TFO DNA Mix NL (182) | GTGTGTGTGTGTGTGTGTGTGTGTGTGTG | DNA | NL |
| CCCT motif_TFO DNA Mix NL (183) | GGGTGGGTGGGTGGGTGGGTGGGTGGGTG | DNA | NL |
| CTTT motif_TFO DNA Mix NL (184) | GTTTGTTTGTTTGTTTGTTTGTTTGTTTG | DNA | NL |
| sf-TTS En3 20bp Cy5 (189) | GGAGGAAAAAAGAAAAAAAATTTTTTTTTTTTCTTTTTTCCTCC | DNA | 5’Cy5 |
| sf-TTS En3 29bp Cy5 (190) | GGAGGAAAAAAGAAAAAAAAAAAAAAGATTTTTCTTTTTTTTTTTTTTCTTTTTTCCTCC | DNA | 5’Cy5 |
| RNA En3 TFO 20nt (198) | CCUCCUUUUUUCUUUUUUUU | RNA | NL |
| CTCT motif_TFO RNA Mix NL (199) | GUGUGUGUGUGUGUGUGUGUGUGUGUG | RNA | NL |
| CCCT motif_TFO RNA Mix NL (200) | GGGUGGGUGGGUGGGUGGGUGGGUGGGUG | RNA | NL |
| CTTT motif_TFO RNA Mix NL (201) | GUUUGUUUGUUUGUUUGUUUGUUUGUUUG | RNA | NL |
| CTCT motif_TFO-12 RNA Mix NL (202) | GUGUGUGUGUGUGUGUG | RNA | NL |
| En3 TFO_20nt Core + 7nt variable (204) | CCUCCUUUUUUCUUUUUUUUNNNNNNN | RNA | NL |
| En3 sf-TTS 27nt core + 20nt ext NL (205) | GTATCGTACTACGATGCGCTGGAGGAAAAAAGAAAAAAAAAAAAATTTTTTTTTTTTTTTTTTTCTTTTTTCCTCCAGCGCATCGTAGTACGATAC | DNA | NL |
| En3 sf-TTS 27core+ 20extension (207) | GTATCGTACTACGATGCGCTGGAGGAAAAAAGAAAAAAAAAAAAATTTTTTTTTTTTTTTTTTTCTTTTTTCCTCCAGCGCATCGTAGTACGATAC | DNA | 5'Cy5 |
| SELEX_exp01_22939 (214) | CCUCCUUUUUUCUUUUUUUUUUUUACC | RNA | NL |
| SELEX_exp01_7616 (215) | CCUCCUUUUUUCUUUUUUUUUUUUGUC | RNA | NL |
| CTCT motif_TFO-10 RNA Mix NL (216) | GUGUGUGUGUGUGUGUG | RNA | NL |
| CTCT motif_TFO-8 RNA Mix NL (217) | GUGUGUGUGUGUGUGUGUG | RNA | NL |
| CTCT motif_TTS For Cy5 (219) | CTCTCTCTCTCTCTC | DNA | 5'Cy5 |
| CTCT motif_TTS Rev NL (220) | GAGAGAGAGAGAGAG | DNA | NL |
| CTCT motif_TTS For Cy5 (221) | CTCTCTCTCTCTCTCTC | DNA | 5'Cy5 |
| CTCT motif_TTS Rev NL (222) | GAGAGAGAGAGAGAGAG | DNA | NL |
| CTCT motif_TTS For Cy5 (223) | CTCTCTCTCTCTCTCTC | DNA | 5'Cy5 |
| CTCT motif_TTS Rev NL (224) | GAGAGAGAGAGAGAGAGAG | DNA | NL |
| SELEX_exp01_3308 (225) | CCUCCUUUUUUCUUUUUUUUUUUUAUG | RNA | NL |
| SELEX_exp01_26346 (226) | CCUCCUUUUUUCUUUUUUUUUUUUACG | RNA | NL |
| SELEX_exp01_16746 (227) | CCUCCUUUUUUCUUUUUUUUUUUGCUC | RNA | NL |
| CTCT motif_TFO-10 RNA Pur (228) | GAGAGAGAGAGAGAGAG | RNA | NL |
| CTCT motif_TFO-12 RNA Pur (229) | GAGAGAGAGAGAGAG | RNA | NL |
| CTCT motif_TFO-14 RNA Pur (230) | GAGAGAGAGAGAG | RNA | NL |
| CTCT motif_TFO-16 RNA Pur (234) | GAGAGAGAGAG | RNA | NL |
| CTCT motif_TFO-18 RNA Pur (235) | GAGAGAGAG | RNA | NL |
| CTCT motif_TFO-20 RNA Pur (236) | GAGAGAG | RNA | NL |
| CTCT motif_TFO-20 DNA Pur (237) | GAGAGAG | DNA | NL |
| CCCT motif_TFO-2 RNA Mix NL (238) | GGGUGGGUGGGUGGGUGGGUGGGUGGG | RNA | NL |
| CCCT motif_TFO-4 RNA Mix NL (239) | GGGUGGGUGGGUGGGUGGGUGGGUG | RNA | NL |
| CCCT motif_TFO-6 RNA Mix NL (240) | GGGUGGGUGGGUGGGUGGGUGGG | RNA | NL |
| CCCT motif_TFO-8 RNA Mix NL (241) | GGGUGGGUGGGUGGGUGGGUG | RNA | NL |
| CCCT motif_TFO-10 RNA Mix NL (242) | GGGUGGGUGGGUGGGUGGG | RNA | NL |
| CCCT motif_TFO-12 RNA Mix NL (243) | GGGUGGGUGGGUGGGUG | RNA | NL |

| | | | |
|---|---|---|---|
| CCCT motif_TFO-14 RNA Mix NL (244) | GGGUGGGUGGGUGGG | RNA | NL |
| CCCT motif_TFO-16 RNA Mix NL (245) | GGGUGGGUGGGUG | RNA | NL |
| Sarrah TFO RNA NL (265) | CCCCCUUCUCUUCUC | DNA | NL |
| Sarrah TTS for Cy5 (266) | CACCTCCCTTCCCCC | DNA | 5'Cy5 |
| Sarrah TTS rev NL (267) | GGGGGAAGGGAGGTG | DNA | NL |
| CTTT motif_TFO RNA-6 Pyr NL (272) | CUUUCUUUCUUUCUUUCUUUCUU | DNA | 5'Cy5 |
| CTTT motif_TFO RNA-8 Pyr NL (273) | CUUUCUUUCUUUCUUUCUUUC | DNA | NL |
| CTTT motif_TFO RNA-10 Pyr NL (274 | CUUUCUUUCUUUCUUUCUU | DNA | 5'Cy5 |
| CTTT motif_TFO RNA-12 Pyr NL (275) | CUUUCUUUCUUUCUUUC | DNA | NL |
| CCCT motif TTS-8 For Cy5 (276) | CCCTCCCTCCCTCCCTCCCTC | DNA | 5'Cy5 |
| CCCT motif TTS-8 Rev NL (277) | GAGGGAGGGAGGGAGGGAGGG | DNA | NL |
| CCT motif RNA Mix NL (280) | GGUGGUGGUGGUGGUGGUGGUGGUGG | RNA | NL |
| CCT motif TTS For Cy5 (281) | CCTCCTCCTCCTCCTCCTCCTCCTCC | DNA | 5'Cy5 |
| CCT motif TTS Rev NL (282) | GGAGGAGGAGGAGGAGGAGGAGGAGG | DNA | NL |
| CCCCCCT motif RNA mix NL (283) | GGGUGGGGGGUGGGGGGUGGGGGGUGGGG | RNA | NL |
| CCCCCCT motif TTS For Cy5 (284) | CCCTCCCCCCTCCCCCCTCCCCCCTCCCC | DNA | 5'Cy5 |
| CCCCCCT motif TTS Rev NL (285) | GGGGAGGGGGGAGGGGGGAGGGGGGAGGG | DNA | NL |

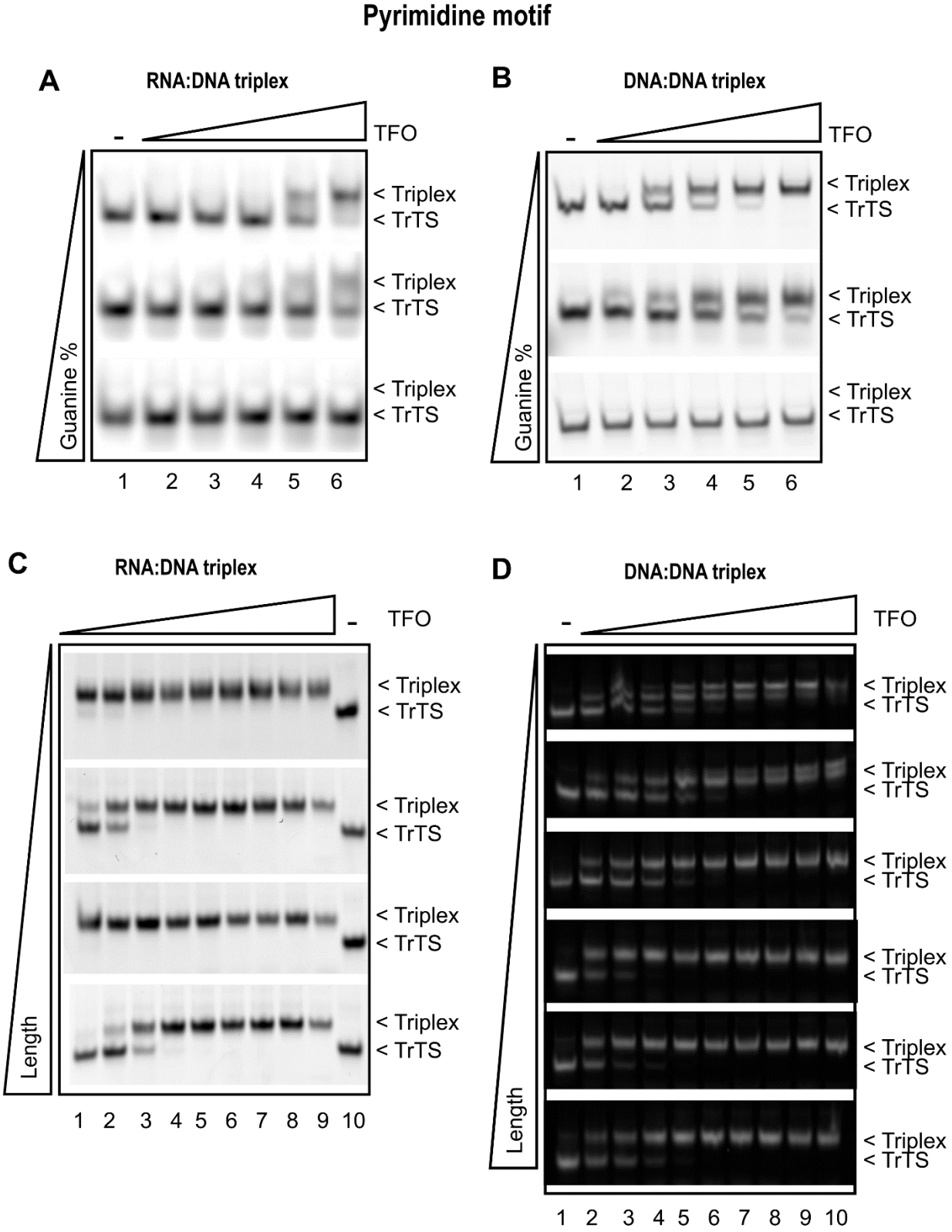
**Supplemental table 1:** Sequences of the triplex oligonucleotides used in this study. NL = non-labelled.

# **APPENDIX 3:** Published triplex pairs

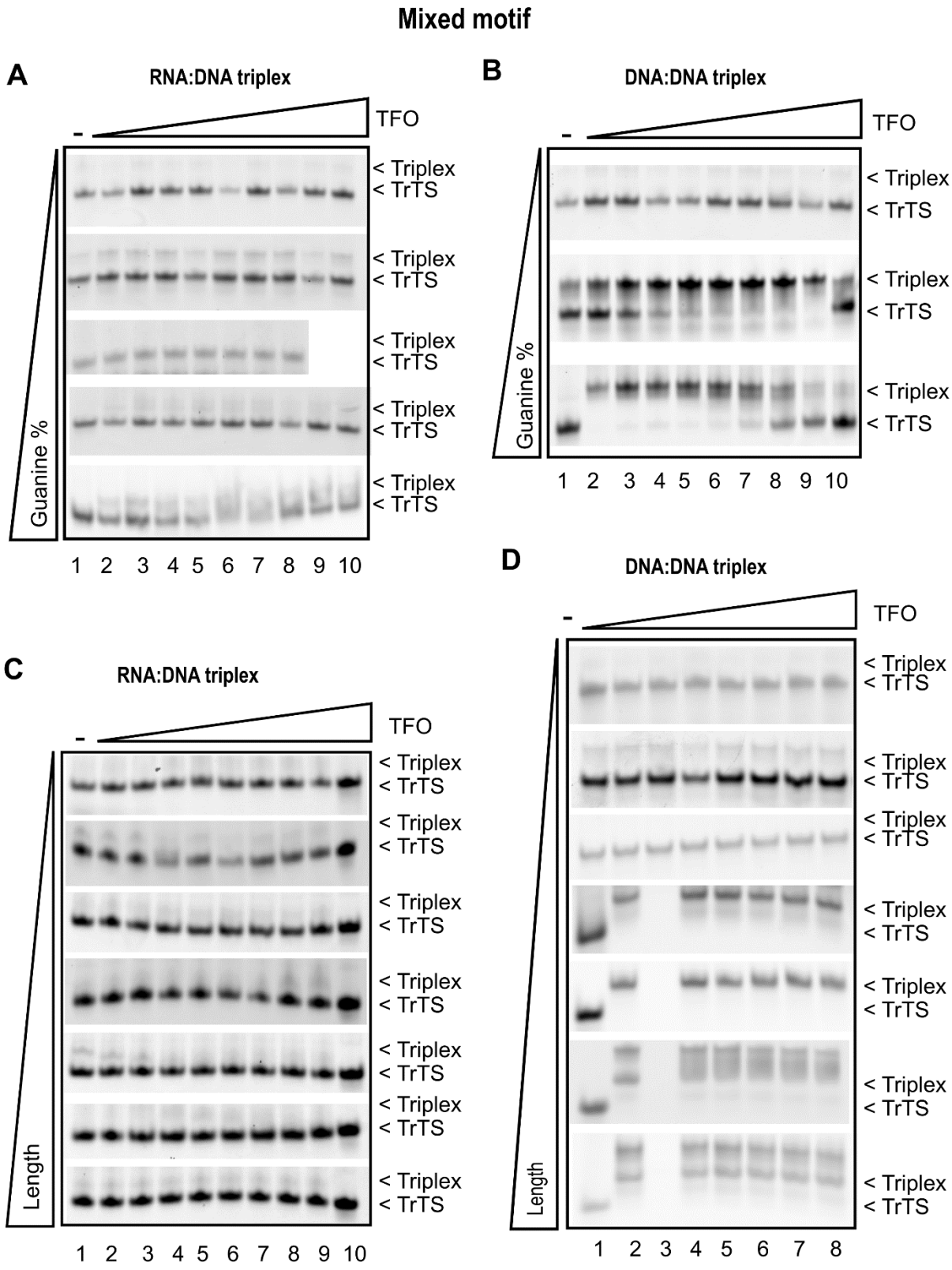
| **Triplex name and publication** | **Oligo** | **Sequence [5' to 3']** |
|---|---|---|
| CCL2 + LNMAT (Chen *et al.*, 2018) | TrTS_R | GAGGGAGAGAGGGCGGA |
| | TrTS_Y | TCCGCCCTCTCTCCCTC |
| | TFO | AGGCTGGAGTGCAGTG |
| PITX2 + FENDRR (Grote *et al.*, 2013) | TrTS_R | AAGAGGAGGAGGAGAAGGAAGAGGATGGAGGGGAGGGAGAA |
| | TrTS_Y | TTCTCCCTCCCCTCCATCCTCTTCCTTCTCCTCCTCCTCTT |
| | TFO | GAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGA |
| FG1 + AG30 (Kaushik Tiwari *et al.*, 2016) | TrTS_R | GAGGGGGAGGGGGTGGTGGGGGGGGAAGGATTCGAAC |
| | TrTS_Y | GTTCGAATCCTTCCCCCCCCACCACCCCCTCCCCCTC |
| | TFO | AGGAAGGGGGGGGTGGTGGGGGAGGGGGAG |
| miR-24-1 + TrTS (Xiao *et al.*, 2017) | TrTS_R | GAGGACAAGGACGACTTGACTCGGTCAC |
| | TrTS_Y | GTGACCGAGTCAAGTCGTCCTTGTCCTC |
| | TFO | TGGCTCAGTTCAGCAGGAACAG |
| miR-483-5p + TrTS (Paugh et al., 2016) | TrTS_R | TGGGGGAAGAAGAGGGGGCAGAGA |
| | TrTS_Y | TCTCTGCCCCCTCTTCTTCCCCCA |
| | TFO | AAGACGGGAGGAAAGAAGGGAG |
| PCDH7 + TrTS (Kalwa *et al.*, 2016) | TrTS_R | AGGAGAGAGAGGAGGGAGGAGAGAGAAA |
| | TrTS_Y | TTTCTCTCTCCTCCCTCCTCTCTCTCCT |
| | TFO | AGAGGAGGGAAGAGAG |
| KIAA1324 + RP11-84A10 (Jalali *et al.*, 2017) | TrTS_R | GAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG |
| | TrTS_Y | CTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTC |
| | TFO | GAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG |
| BCL9 + RP11-84A10 (Jalali *et al.*, 2017) | TrTS_R | GAGAGACAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG |
| | TrTS_Y | CTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTGTCTCTC |
| | TFO | GAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG |
| PROX1 + RP11-84A10 (Jalali *et al.*, 2017) | TrTS_R | GAGAGAGATAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG |
| | TrTS_Y | CTCTCTCTCTCTCTCTCTCTCTCTCTCTATCTCTCTC |
| | TFO | GAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG |
| SMAD2 + MEG3 (Mondal *et al.*, 2015) | TrTS_R | AGAGAGGGAGAG |
| | TrTS_Y | CTCTCCCTCTCT |
| | TFO | CGGAGAGCAGAGAGGGAGCG |
| TGFBR1 + MEG3 (Mondal *et al.*, 2015) | TrTS_R | AGAGAGAGGGAGAGAG |
| | TrTS_Y | CTCTCTCCCTCTCTCT |
| | TFO | CGGAGAGCAGAGAGGGAGCG |
| TGFBR2 + MEG3 (Mondal *et al.*, 2015) | TrTS_R | CAGAGAGCAGAGAGAGAGA |
| | TrTS_Y | TCTCTCTCTCTGCTCTCTG |
| | TFO | CGGAGAGCAGAGAGGGAGCG |
| WWOX1 + PARTICLE_1 (O'Leary *et al.*, 2015) | TrTS_R | AAGGAGGGAGGGGGGA |
| | TrTS_Y | TCCCCCCTCCCTCCTT |
| | TFO | TTGGTGGGTGGGGGCT |
| WWOX2 + PARTICLE_2 (O'Leary *et al.*, 2015) | TrTS_R | GGGAGGGAGGAAAGAA |
| | TrTS_Y | TTCTTTCCTCCCTCCC |
| | TFO | TTTTTTGGTGGGTGGG |
| SARRAH + TrTS (Trembinski *et al.*, 2020) | TrTS_R | GGGGGAAGGGAGGTG |
| | TrTS_Y | CACCTCCCTTCCCCC |
| | TFO | CCCCCUUCUCUUCUC |

**Supplemental table 2:** Sequences of the published triplex pairs used in this study.

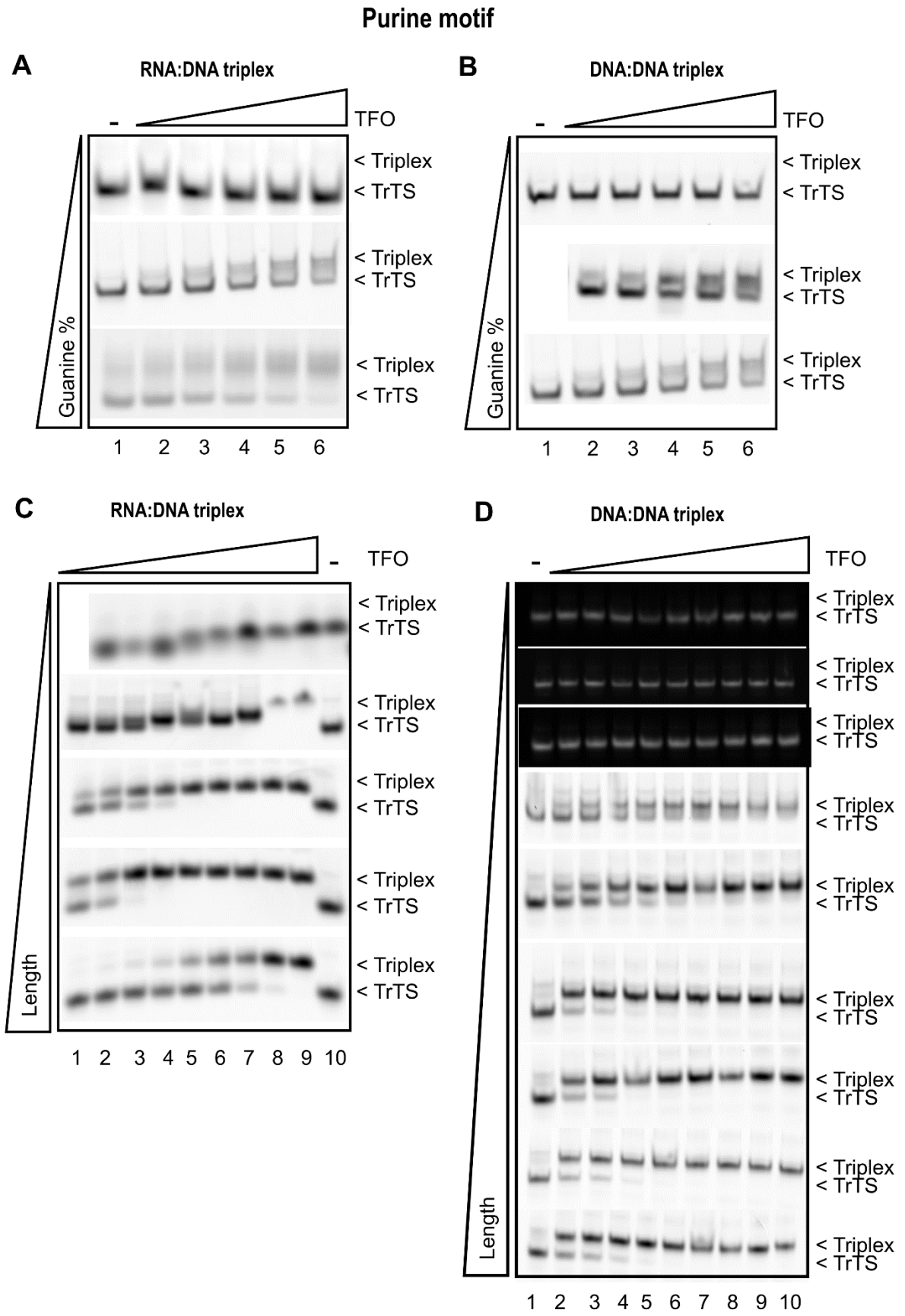
## **APPENDIX 4:** The triplex code – supplemental figures



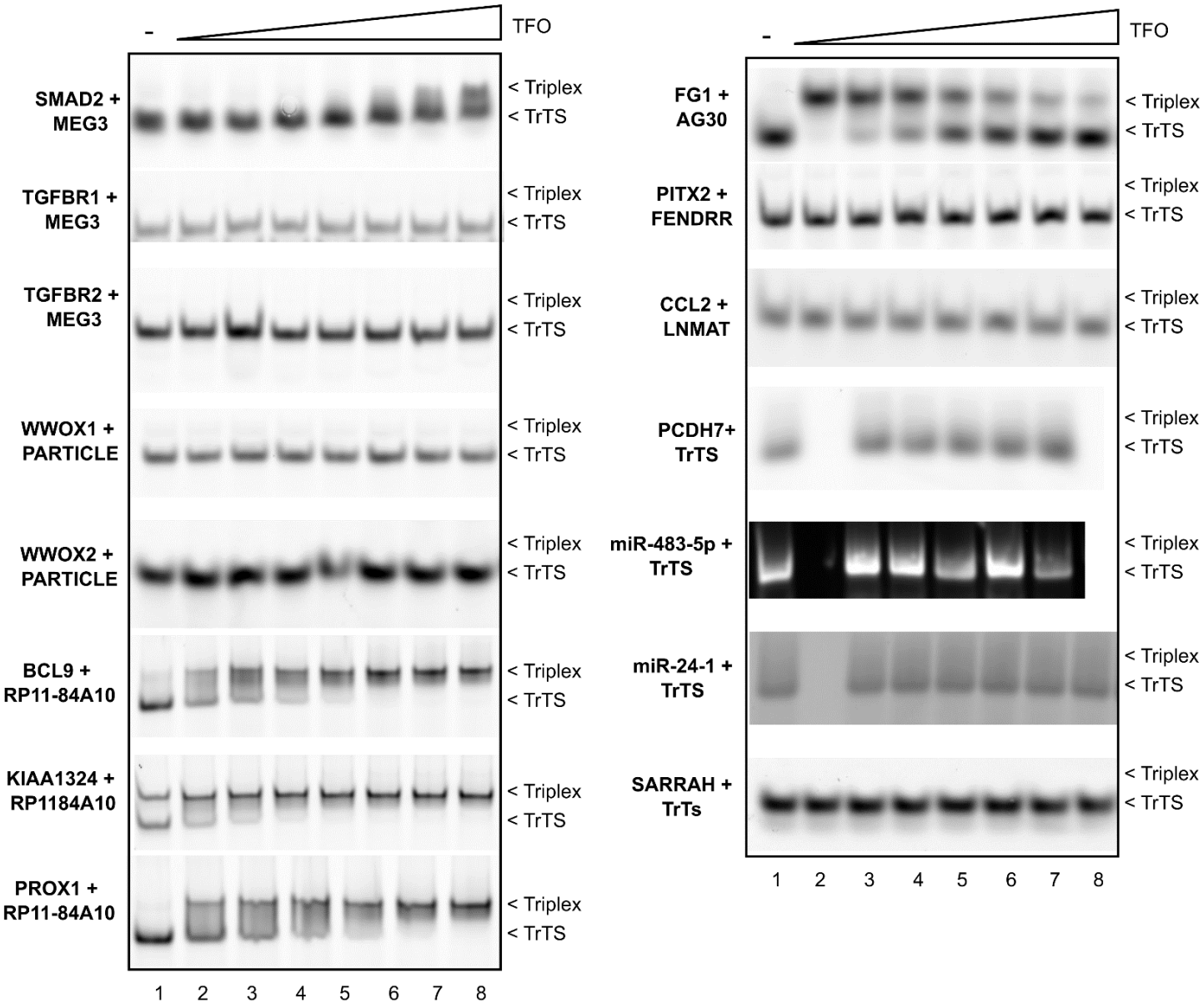

**Supplemental figure 5:** EMSA figures for the pyrimidine motif. (A-B) Guanine dependence for RNA:DNA (A) and DNA:DNA (B) triplexes. (C-D) Length dependence for RNA:DNA (C) and DNA:DNA (D) triplexes.



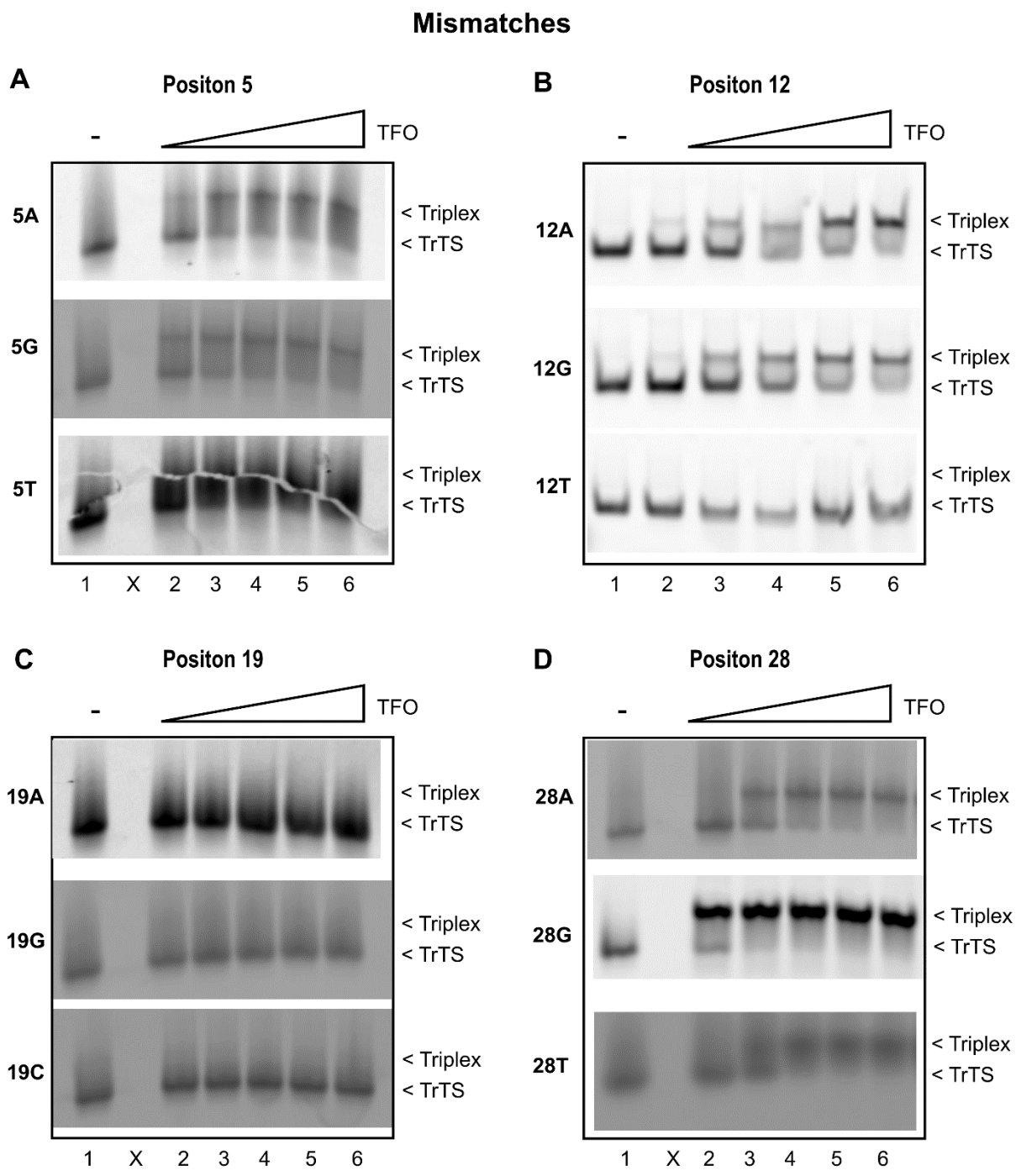

**Supplemental figure 6:** EMSA figures for the mixed motif. (A-B) Guanine dependence for RNA:DNA (A) and DNA:DNA (B) triplexes. (C-D) Length dependence for RNA:DNA (C) and DNA:DNA (D) triplexes.




**Supplemental figure 7:** EMSA figures for the purine motif. (A-B) Guanine dependence for RNA:DNA (A) and DNA:DNA (B) triplexes. (C-D) Length dependence for RNA:DNA (C) and DNA:DNA (D) triplexes.




**Supplemental figure 8:** EMSA figures for the published triplex pairs.




**Supplemental figure 9:** EMSA figures for the En3 DNA:DNA triplex with mismatches. (A-D) Mismatches at positions 5/29 (A), 12/29 (B), 19/29 (C) and 28/29 (D).