THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)

Eukaryotic chromatin structure in the context of R-loops and histone modifications

by
László Halász

UNIVERSITY OF DEBRECEN
DOCTORAL SCHOOL OF MOLECULAR CELL AND IMMUNE BIOLOGY

DEBRECEN, 2018

Thesis for the degree of doctor of philosophy (PhD)

Eukaryotic chromatin structure in the context of R-loops and histone modifications

by
László Halász

Supervisor: Dr. Lóránt Székvölgyi

University of Debrecen
Doctoral School of Molecular Cell and Immune Biology

Debrecen, 2018

Table of Contents

1. Abbreviations

3C	Chromosome Conformation Capture
5’ UTR	Five prime untranslated region
AUC	Area under the curve
CGI	CpG island
CTCF	CCCTC-binding factor
ChIP	Chromatin Immunoprecipitation
DRIP	DNA-RNA Immunoprecipitation
DRIVE	DNA-RNA in vitro enrichment
DSB	DNA double strand break
FPR	False positive rate
GEO	Gene expression omnibus
H3K4me3	H3 lysine 4 trimethylation
HMM	Hidden Markov Model
MDS	Multidimensional scaling
MS	Mass spectrometry
NGS	Next generation sequencing
ORF	Open reading frame
ORI	Origin of replication
QGRS	Quadruplex forming G-rich Sequence
RE	Restriction enzyme
REZ	R-loop extension zone
RIZ	R-loop initiation zone
RLFS	R-loop forming sequence
ROC	Receiver operating characteristic
RPA	Replication Protein A
RPG	Ribosomal protein gene
RPKM	Reads per kilobase per million
SPM	Yeast sporulation medium
SPS	Yeast presporulation medium
SRA	Sequence read archive
SSB	Single-stranded binding protein
TPR	True positive rate
TSS	Transcription start site
TTS	Transcription termination site
circRNA	circular RNA
dsDNA	double-stranded DNA
lncRNA	long non-coding RNA
ncRNA	non-coding RNA
snoRNA	small nucleolar RNA
ssDNA	single-stranded DNA

2. Introduction

2.1. Canonical nucleic acid structures

All living cells on Earth stores their genetic information (required for growth, development, functioning and reproduction) in double-stranded deoxyribonucleic acid (DNA) molecules. DNA is a long, unbranched and paired polymer, assembled from monomers containing sugar-phosphate molecules and covalently bonded nucleotide subunits of four major types: adenine (A), thymine (T), cytosine (C) and guanine (G) (Alberts et al. 2014). Nucleotides are chained together covalently by phosphodiester bonds that join the 3'-hydroxy group of one sugar molecule to the 5'-hydroxyl group of the next sugar molecule, giving chemical polarity to the sugar-phosphate backbone (Alberts et al. 2014). The two strands are held together by hydrogen bonds formed by complementary base-pairing. Adenine (A) is always paired with thymine (T) through two hydrogen bonds, while guanine (G) is paired with cytosine (C) through three hydrogen bonds (Alberts et al. 2014). The stability of double stranded DNA is influenced by the number of hydrogen bonds formed between the bases of each strand (Alberts et al. 2014).

Since the seminal discovery of right-handed double-helical model for B-form DNA in 1953 (WATSON and CRICK 1953), more than 20 alternative helix configurations have been observed (Ghosh and Bansal 2003). Alternative DNA configurations can be functionally important (replication, recombination and transcription) and their existence is dependent on the local nucleotide sequence characteristics and environmental factors (salt concentration, pH, supercoiling, water content and interactions with various proteins) (Potaman and Sinden 2005). The canonical B-form DNA is believed to predominate in cells at neutral pH under physiological conditions (Ghosh and Bansal 2003). The B-form helix completes a full turn in every 10.5 base pairs (Alberts et al. 2014). The A-form helix is shorter and wider than the B-form and completes a full turn in every 11 base pairs (Alberts et al. 2014). The A-form helix usually occurs after protein binding (such as TATA-box binding protein) and has also been indicated in double stranded RNA duplexes (dsRNA) (Potaman and Sinden 2005). The Z-form is left-handed and usually forms at regions with alternating purine and pyrimidine sequences $(CG)_n$ and $(TG)_n$ (Alberts et al. 2014). The Z-form is thinner than the B-form and completes a whole turn in every 12 base pairs (Potaman and Sinden 2005).

Thus, local structural shift from B-form of the DNA into other alternative forms can be functionally important. External stimuli, environmental changes, superhelical tension and protein binding might be involved in the formation of these structural changes (Potaman and Sinden 2005).

2.2. Eukaryotic chromatin structure and organization

During development and throughout life, a large collection of cells must be generated to ensure the proper function of each tissue and organ. Since the length of the DNA is far greater than the size of the cell's nucleus, DNA must be spatially organized to fit in its compartment. For this reason, eukaryotic cells (yeast to higher level eukaryotes) have evolved molecular mechanisms allowing their DNA content to be packed at many scales during interphase (from chromosome territories to interacting chromatin loops) (**Figure 1**).

The genome of eukaryotic cells is organized into chromatin, which displays hierarchical levels. At the primary level, nucleosomes represent the fundamental repeating building blocks of the eukaryotic chromatin. The nucleosome is the smallest structural component of chromatin, consisting of 147 base pairs of DNA wrapped around an octamer of core histone proteins. The histone octamer is composed of a central heterotetramer of histones H3 and H4, flanked by two heterodimers of histones H2A and H2B. Nucleosome units are connected to the adjacent nucleosome by short DNA sequences known as the linker DNA, creating a nucleosome chain. The H1 linker histone binds the nucleosome at the entry and exit sites of the DNA, thus locking the DNA into place. The main functions of the chromatin are packaging DNA into a more compact form, prevention of DNA damage, and controlling gene expression programs and replication.

Nucleosomes are organized into 10-30 nm chromatin fibers that can form various higher-order structures allowing effective functional compartmentalisation of the genome (**Figure 1**). Recent studies have revealed two prominent features of higher-order genome organization; alternating active and inactive chromatin regions (1-10 Mb, A-B compartments) and topologically associated domains (TADs, <1 Mb) where intra-TAD interactions occur most frequently. TADs are usually referred as the fundamental structural and functional building blocks of the interphase chromosomes. However, the underlying mechanisms of TAD formation remain unexplored (Fudenberg et al. 2016). In the proposed loop extrusion model, cis-acting loop extrusion factors (e.g. cohesin) form progressively larger loops but stall at TAD boundaries due to interaction with structural proteins, like CCCTC-binding factor (CTCF).

Chromatin function is tightly linked to its three-dimensional structure. As mentioned above, based on the accessibility of DNA, chromatin is generally classified into two main compartments: euchromatin and heterochromatin. Euchromatin is gene-dense, active (A-compartment) while heterochromatin is condensed and indicated in the repression of gene expression (B-compartment). However, several genome-wide studies fine-tune this classification into multiple chromatin states, each with unique characteristics (Ernst and Kellis 2017). Thus, the genome functions like an information-retrieval machine in which the 3D chromatin structure is critical for selected information exposure and cell identity. Interestingly, the chromatin is dynamic and able to dramatically change conformations (condense/decondense) under special conditions locally or globally in processes like cell division, transcription, differentiation, recombination or in response to intrinsic or extrinsic stimulus. Perturbation of chromatin structure is often related to developmental and pathological diseases, due to the misregulation of specific genes or gene-networks.

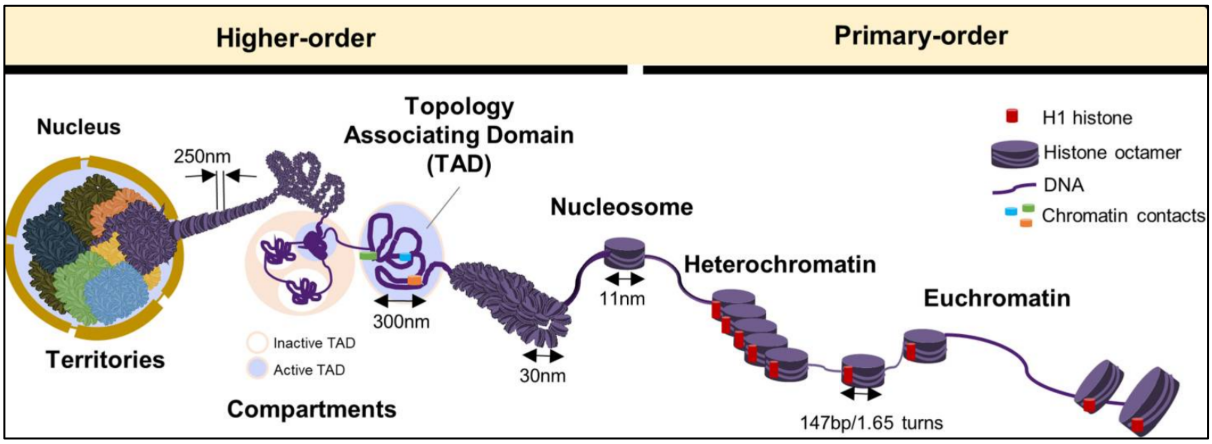


Figure 1. The levels of eukaryotic genome organization. Cartoon showing the hierarchical organization of euchariotic chromatin from higher- to primary order according to the folding complexity. (Chang et al. 2018)

Genome organization is a very complex multi-layered process with many players involved. Several architectural proteins are well characterized and orchestrate 3D chromatin looping and structure: CCCTC-binding factor (CTCF), YY1 and the cohesin complex (Beagan et al. 2017). Their binding is regulated by the local histone environment, secondary DNA structures and DNA sequence.

Two major approaches to study spatial organization of the chromosome can be categorized into microscopic and molecular assays. Light microscopy or fluorescent microscopy provide information about the distribution and shape of the chromosomes with low resolution (50-100 nm) in single cells. However, the high-resolution visualization of 3D chromosome and chromatin structure with microscopy is still difficult. Rather, there are novel techniques that are able to make it possible (Ou et al. 2017). More widely used methods to study chromatin structure are based on chromosome conformation capture (3C) technology, sequencing and bioinformatics. These assays (4C, 5C, 6C, ChIA-PET, ChIP-Loop, HiChIP and Hi-C) provide relative spatial-contact probabilities among two linearly distal loci for a population of cells at near 1 kb resolution. Of note, single-cell approaches are also available for Hi-C. Most recently, researchers developed tyramide signal amplification (TSA-seq), the first genomic method capable of estimating cytological distances of chromosome loci genome-wide relative to a particular nuclear compartment and even inferring chromosome trajectories from one compartment to another (Chen et al. 2018).

In conclusion, the 3D organization of the genome provides an important layer of how cells behave and express their information content. Thus, not surprisingly, studying 3D or 4D (Dekker et al. 2017) chromatin organization became a hot topic in the field of genomics.

2.2.1. Histone modifications

As mentioned in the previous section, the primary level of genome organization is the nucleosome structure composed of highly conserved histone proteins. The histone proteins have protruding N-terminal amino acid tails, that can be post-translationally modified (PTM), affecting key cellular events, including chromatin compaction, nucleosome dynamics, gene expression and recombination. PTMs provides enormous regulatory potential by providing modularity within core particles. The main modifications include: acetylation, phosphorylation and methylation. Yet, there are other known modifications exists such as deimination, ADP ribosylation, ubiquitination, crotonylation, SUMOylation and GlcNAcylation (**Table 1**). Recent studies showed that not only the histone tails, but the lateral surface of the core proteins, which is in direct contact with the DNA, can also be modified (Lawrence et al. 2016). Histone modifications and the protein machinery that adds, removes and recognizes these post-translational modifications (histone writers, erasers and readers), become central figures of how cells control physiological states and identities.

The ever-growing list of PTMs, their crosstalk and function are however not well understood. In the next few paragraphs as being relevant in the second section of this thesis, I review one of the most studied histone modifications: histone H3 lysine 4 trimethylation (H3K4me3).

Table 1. List of selected histone tail modifications. (Lawrence et al. 2016)

Histone	Modification	Role
H2A	H2AS1P	Mitosis; chromatin assembly
	H2AK4/5ac	Transcription activation
	H2AK7ac	Transcription activation
	H2AK119P	Spermatogenesis
	H2AK119uq	Transcription repression
H2B	H2BS14P	Apoptosis
	H2BS33P	Transcription activation
	H2BK5ac	Transcription activation
	H2BK11/12ac	Transcription activation
	H2BK15/16ac	Transcription activation
	H2BK20ac	Transcription activation
	H2BK120uq	Spermatogenesis/meiosis
	H2BK123uq	Transcription activation
H3	H3K4me2	Permissive euchromatin
	H3K4me3	Transcriptional elongation; active euchromatin; DSBs
	H3K9me3	Transcription repression; imprinting; DNA methylation
	H3K9ac	Histone deposition; transcription activation
	H3K14ac	Transcription activation; DNA repair
	H3K18ac	Transcription activation; DNA repair; DNA replication
	H3K23ac	Transcription activation; DNA repair
	H3K27ac	Transcription activation
	H3T3P	Mitosis
	H3S10P	Mitosis; meiosis; transcriptional activation
	H3T11/S28P	Mitosis
H4	H4R3me	Transcription activation
	H4K20me1	Transcriptional silencing
	H4K20me3	Heterochromatin
	H4K12ac	Histone deposition; telomeric silencing; DNA repair
	H4K16ac	Transcription activation; DNA repair
	H4S1P	Mitosis

2.2.2. H3 histone methylation

Lysine methylation of the H3 histone protein is balanced by the action of methyltransferases (“writer”) and demethylases (“eraser”). Three methylation states can be present on this lysine residue: mono-, di- and trimethylation resulting in distinct biological outcomes (Hyun et al. 2017). It is important to mention, however, that none of these modifications changes the charge of the amino acids and subsequently the structure of the nucleosome, but they serve as a docking site for other effector proteins. Unlike acetylation that has a half-life of several minutes, methylation is considered to be more stable.

In *Saccharomyces cerevisiae* H3K4me3 is commonly associated with the activation of nearby genes by recruiting nucleosome remodeling factors (CHD1 and BPTF), while blocking negative regulator binding (NuDR) and the H3K4me3 level correlates with the transcription rate within the interphase nucleus. During transcription H3K4me3 is rapidly generated at transcription start sites or promoters by RNAP-associated Set1 when genes are turned on and remains present even if Set1 is no longer there, leaving a memory mark of recent transcription. On the other hand, upon gene repression, H3K4me3 marks are lost. Apart from transcription, H3K4me3 also contributes to class-switch recombination, S-phase DNA damage checkpoint and meiotic recombination.

2.2.2.1. Set1C/COMPASS

Methylation is an evolutionally conserved mechanism. In yeasts, methylation is carried out by a SET domain-containing lysine-specific methyltransferase; Set1. *Drosophila melanogaster* have three Set1 homologs, while humans have six. Set1 alone is inactive since this protein is part of a larger complex with seven other proteins: Spp1, Bre2, Swd1, Swd2, Swd3, Sdc1 and Shg1. The complex is called Set1 complex (Set1C or COMPASS) (Hyun et al. 2017; Karányi et al. 2018). Set1 is essential for mono-, di-, and trimethylation of histone H3 at K4. The ability of COMPASS to mono-, di- and trimethylate K4 of histone H3 depends on its subunit composition. For example, COMPASS lacking Bre2 (Brefeldin-A; 58 KDa) cannot trimethylate K4 of histone H3 and have no effect on mono- and dimethylation, which subsequently plays a role in telomere length maintenance and transcription elongation regulation. Bre2 is known to interact with SET1 and SDC1. In addition, any alteration of the SET-domain of Set1 results in a complete loss of complex formation and activity of the enzyme. Each subunit is responsible for specific function in the assembly. Set1, Swd3 and Swd1 are essential for the stability and function of the complex as cells lacking any of these subunits are defective in H3K4 methylation. Swd2 subunit is required for optimal di- and trimethylation but not for monomethylation. Swd2 also facilitates the function of cleavage and polyadenylation factor (CPF), a complex involved in transcription termination. The PHD-domain containing COMPASS component Spp1 has been shown to promote the recruitment of potential DSB sites to the chromosome axis allowing the Spo11 to cleave and generate DNA double strand breaks. In addition, this subunit also regulates the catalytic activity of the Set1C (Acquaviva et al. 2013a). Similarly, to Swd2, Sdc1 and Bre2 subunits of Set1C appear to be required for proper H3K4 di- and trimethylation, but not monomethylation.

Taken together, apart from being the least abundant histone modification, H3K4me3 is a very important and conserved epigenetic marker for active transcription and recombination. The molecular mechanism of writing and erasing methylation has become an important field of research. Moreover, as identified for Swd2 and Spp1, other Set1C subunits may participate in diverse biological processes apart from the Set1C.

2.2.3. Meiotic DSB formation and its connection with H3K4me3

During prophase I in meiosis, recombination is initiated by the generation of programmed DNA double-strand breaks (DSBs) at non-random points in the genome by the meiotic nuclease Spo11. These DSBs can be subsequently repaired using homologous chromosomes resulting in crossovers or non-crossovers. Mechanistically, a DSB occurs in a highly organized chromatin structure. The distribution and frequency of DSBs vary along chromosomes and are often localized in ~1-2 kb hotspots. The hotspots are usually in close proximity to gene promoters with nucleosome-depleted regions flanked by H3K4me3 and rarely found within exons or gene terminal sites. However, the mechanism by specific sites became hotspots and anchored to the chromosomal axis is poorly understood. The tethered-loop axis model proposes that Spp1, the PHD finger domain containing H3K4me3 reader subunit of the Set1C interacts with both H3K4me3 marks and chromosome axis protein REC114-MEI4-MER2 complexes (RMM) (Acquaviva et al. 2013b). This interaction tethers distal DNA sequences to the chromosome axis, allowing the cleavage by Spo11 and subsequent repair. These results indicate that Spp1 is a multifaceted molecule and emerged as a key regulator of H3K4 trimethylation (Acquaviva et al. 2013a). Despite of the intense research that discovered many aspects of meiotic DSB formation, the nuclear dynamics of Set1C subunits is still unknown.

In the second part of this study, with the use of next-generation sequencing and bioinformatics, we showed how the Spp1 subunit was redistributed from transcribed genes to chromosome axis sites during meiosis independently from Set1C.

2.3. Non-canonical nucleic acid structures

Beyond the alternative DNA forms, there are numerous non-canonical nucleic acid structures affecting cellular homeostasis, plasticity and chromatin structure. Among them single-stranded DNA (ssDNA), G-quadruplexes and RNA-DNA hybrids (R-loops) are the most frequent. These structures are detailed in the following sections.

2.3.1. Single-stranded DNA

Single-stranded DNA (ssDNA) refers to a specific region in the genome, where the two strands of the double helix are not bound together by hydrogen bonds and transiently separated in the local space. ssDNA is an essential intermediate (template) in various cell functions, such as DNA replication, recombination, repair and transcription. Furthermore, ssDNA interacts with large number of proteins (Alberts et al. 2014). In 1979, Lindahl and colleagues showed that ssDNA represents 1-2% of total DNA in growing animal cells and the majority of them are formed during DNA synthesis (Bjursell et al. 1979). Although the majority of ssDNA molecules is transient with a short lifetime, there are discrete locations (certain promoter elements and telomeres) where ssDNA is present in a stable form (Dickey et al. 2013).

ssDNA is created when the internal base pairing of the original double-stranded DNA is broken, and the helix is unwound. Since double-stranded DNA typically has a stable structure, DNA unwinding requires a set of specific enzymes that overcome this thermodynamic barrier (using ATP hydrolysis) and allow other proteins to interact with the DNA (Bhattacharyya and Keck 2014). This process is carried out by a specific group of enzymes called DNA helicases. Helicases are categorized into six superfamilies based on their DNA recognition motif (Singleton et al. 2007). The human genome encodes 95 non-redundant helicase proteins; 64 RNA helicases and 31 DNA helicases (Umate et al. 2011). The opening sites of the double helix are generally AT-rich regions and can be denaturated easily due to their low helical stability (Coman and Russu 2005). AT-rich sequences can also be found at core promoter regions (e.g. TATA-box) and at the regions of the origin of replication (ORIs). If ssDNA is not stabilized by single-stranded DNA-binding proteins (SSBs), ssDNA can form various secondary structures such as hairpins (variable length, twists and turns) or G-quadruplexes (discussed later) by linking itself with hydrogen bonds.

In order to accurately coordinate these diverse molecular functions and to preserve genomic integrity, a variety of proteins has evolved to bind selectively and non-covalently to ssDNA. These groups of proteins are called SSBs. When a single strand of the DNA is exposed, single-stranded DNA binding proteins are essential for the isolation, stabilization and processing of ssDNA, since ssDNA are hypersensitive to both chemical and enzymatic degradation.

While SSBs are found in every organism, the proteins themselves share surprisingly little sequence similarity, subunit composition, and oligomerization states (**Figure 2**) (Marceau 2012). In humans, there are 52 experimentally validated single-stranded DNA binding proteins based on Gene Otology (Zerbino et al. 2017; Carbon et al. 2009). In SSB proteins a special protein domain is responsible for the interaction of single-stranded DNA and oligonucleotide/oligosaccharide-binding (OB)-fold domain. The length of the OB-fold domains can vary depending on different proteins consisting of 70 to 150 amino acids. Structurally, the OB-folds are β barrels consisting of 5 highly coiled, antiparallel β sheets (Flynn and Zou 2010). SSBs can be classified into two groups based on the number of OB-folds present in their structure. Simple SSBs (hSSB1, hSSB2, and mtSSB) contain one OB-fold domain, whereas higher order SSBs (RPA1, RPA2, and RPA3) contain two or more OB-folds (Wu et al. 2016). SSBs usually function in heterotrimeric protein complexes, consisting of multiple tandem repeat OB-folds.

Replication protein A (RPA) complex is a well-known ssDNA-binding protein. This complex plays an essential role in eukaryotic DNA metabolism. This complex is composed of three subunits RPA1 (71 kDa), RPA2 (32 kDa) and RPA3 (14 kDa) (Wu et al. 2016). It was first characterized as an important element for the DNA replication machinery by preventing premature rehybridization while it also protects ssDNA. Moreover, RPA has an important role in checkpoint signaling, DNA repair and R-loop homeostasis (discussed later in the thesis).

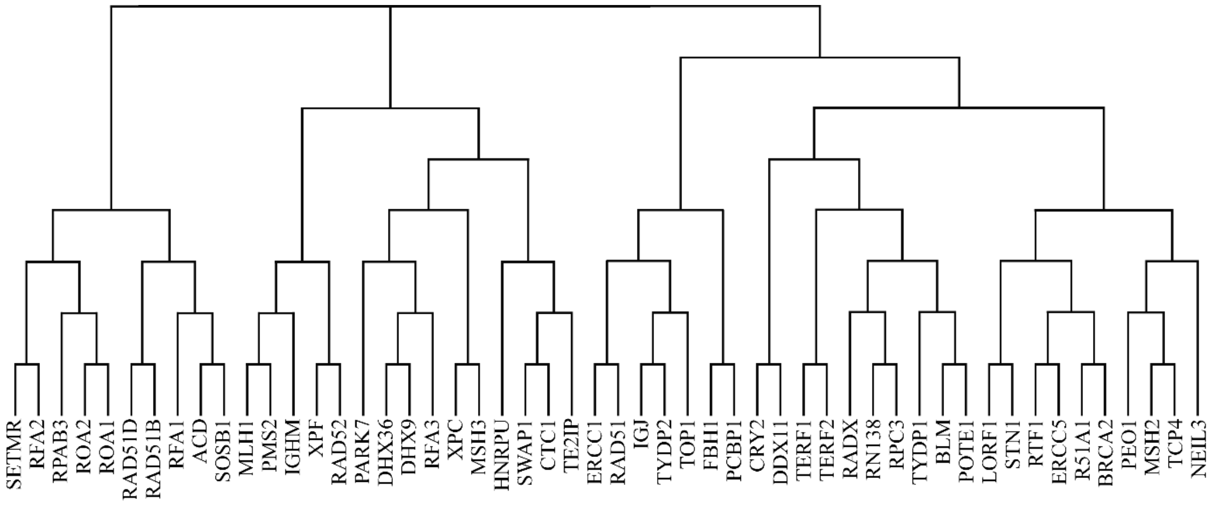


Figure 2. Sequence similarity of experimentally validated human single-strand binding proteins. Cladogram shows the classification of human SSBs. The tree was constructed after multiple sequence alignment of 52 human SSB protein sequences using Clustal Omega (Sievers et al. 2011). List of SSBs were gathered using Gene Ontology term: GO:0003697.

The measurement and characterization of ssDNA formation (*in vitro* and *in vivo*) is fundamental in understanding the inner life of cells. Antibody based methods such as immunolabeling or chromatin immunoprecipitation (ChIP) utilizes the specific recognition of SSBs by antibodies. With the currently available commercial SSB binding antibodies, one can selectively study the distribution and localization of a certain SSB. Alternatively, native (non-denaturing) bisulfite-treatment can be applied to detect single-stranded DNA footprints (discussed later).

Single-stranded DNA is hypersensitive for nuclease cleavage, secondary structure formation, and damage resulting in double-strand breaks and mutations. Hence, the presence of ssDNA acts as an initiator signal for many regulatory pathways. In this thesis, many of these pathways are described in the sections discussing RNA-DNA hybrids and their formation from ssDNA.

2.3.2. G-quadruplex

Guanin (G)-rich regions of the genome can self-assemble into functional inter- and intramolecular secondary structures. The DNA G-quadruplex, discovered in the late 1980s, is a stable four-stranded structure, composed of sets of guanine quartets (rings) held together by Hoogsteen hydrogen bonding (**Figure 3A-B**). The topology is further supported by a cation (strength of stabilization: $K^+ > Na^+ > NH_4{}^+ > Li^+$) located within the central channel formed by G-quartets. These structures are compact, highly stable and resistant to nuclease digestion under physiological conditions (Capra et al. 2010). G-quadruplexes have been demonstrated to be evolutionally conserved and occur in the DNA of human cells and other model systems (Capra et al. 2010).

The underlying DNA sequence motifs involved in the G-quartet determines how the quadruplex folds. The G4 motif or Quadruplex forming G-rich Sequence (QGRS), is a general sequence pattern, where G-quadruplexes can possibly form and can be represented by a regular expression: $G^+N^*G^+N^*G^+N^*G^+$. The motif contains four G-tracts separated by nucleotides with variable length, where '*N*' is an arbitrary base including guanine, '+' stands for one or more repetitions and '*' denotes zero or more repetitions (**Figure 3C**) (Rawal et al. 2006). Depending on the composition and length of individual G-tracts, G-quadruplexes can adopt several topologies with varying loop configurations.

Many important biological functions have been proposed for these structures, regulatory processes such as gene expression, replication and telomere maintenance. G-quadruplex binding proteins play a crucial role in mediating these functions. Based on literature data, there are approximately 60 proteins that can bind and interact with G-quadruplex structures with high affinity (Mishra et al. 2016). G-quadruplexes have also been shown to influence genomic instability, and they have been indicated in cancer, neurodegenerative- and other diseases (Wu and Brosh 2010). Their role in these processes is an active area of functional genomics research.

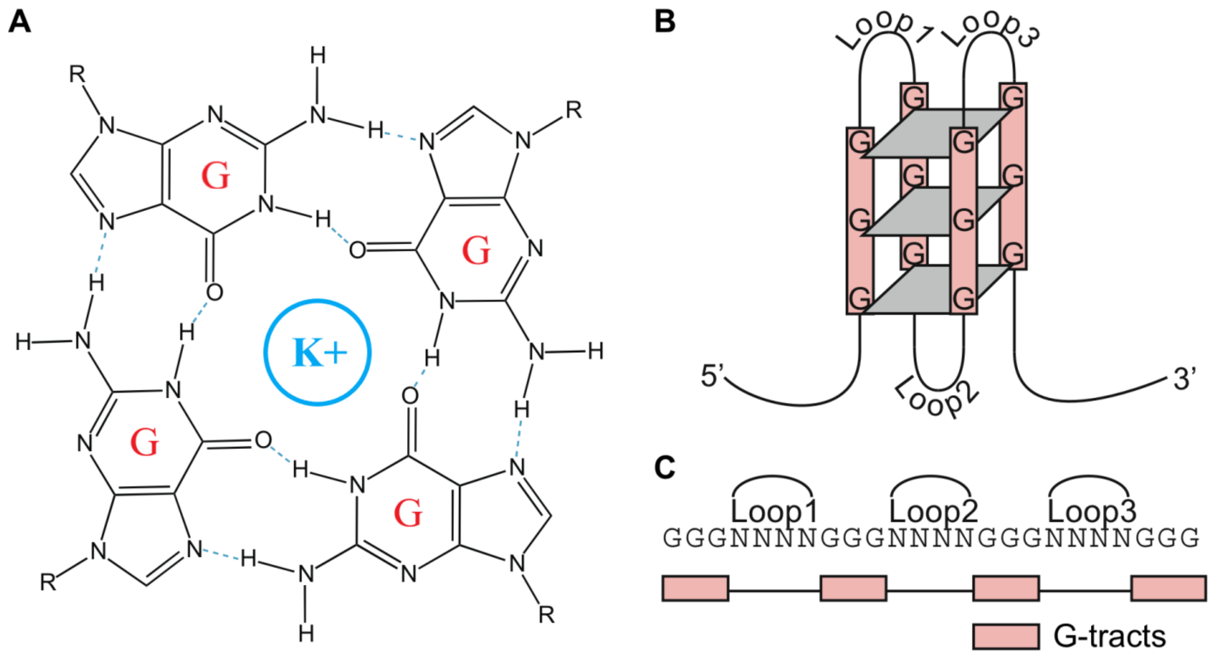


Figure 3. Structure and motif of the intramolecular G4 DNA quadruplex. (**A**) Structure of a G-quartet. The planar ring of four hydrogen-bonded (Hoogsteen base-pairing) guanines is formed by guanines from different G-tracts, which are separated by intervening loop regions in the intra-molecular G4 DNA structure. (**B**) Schematic of the four-stranded secondary structure consisting of three G-quartets. (**C**) The G4 DNA motif sequence with four G-tracts of three guanines separated by loop regions. (Capra et al. 2010)

Because early research showed difficulties identifying these structures *in vivo*, genome-wide computational methods have emerged to predict regions with G4 forming motifs. These *in silico* algorithms use regular expressions or advanced statistical models to predict potential sites for G-quadruplexes. Simple pattern matching programs (Mishra et al. 2016) search for exact matches of G4 motifs for a given nucleotide sequence using regular expressions. Although many putative sites can be found by exact matches, most of these sequences are pointed out to be false positive (Yano and Kato 2014). Computational methods, like G4HMM (Yano and Kato 2014) employ hidden Markov models (HMMs) to reduce the false positive rate of prediction. Other methods use experimental data (discussed later) to support G4 prediction by implementing supervised machine learning approaches (Extreme Gradient Boosting Tree) (Sahakyan et al. 2017). These algorithms predicted over 300.000 putative G4-

motifs in the human genome, enriched within promoters, 5’ UTRs, first introns and telomeric regions (Hänsel-Hertsch et al. 2017).

Recent methodological advances allow us to generate explicit experimental data about the localization and function of G-quadruplex structures. Structure-selective antibodies (BG4 and 1H6) have been generated to visualize these structures in living cells using immunofluorescence and immunohistochemistry (Hänsel-Hertsch et al. 2017). These studies showed punctuated distributions of G-quadruplex foci and cell cycle dependent dynamics (max in S-phase).

G4 chromatin immunoprecipitation followed by next-generation sequencing (G4 ChIP-seq or G4-seq) captures endogenous G-quadruplex sites at higher resolution and throughput. This method adapts the traditional ChIP-seq protocol using G-quadruplex specific antibody (BG4) (Hänsel-Hertsch et al. 2018). Using G4 ChIP-seq, ~10.000 G-quadruplex sites have been detected *in vivo* (Hänsel-Hertsch et al. 2016). These identified G-quadruplexes showed strong correlation with regulatory sites, such as nucleosome depleted regions (NDRs) and highly expressed genes (Hänsel-Hertsch et al. 2018). Alternatively, ChIP-seq can be used against G4-binding proteins, to capture specific G-quadruplexes set bound by a known protein. These experimental data support the biological significance of G-quadruplexes.

Overall, G-quadruplexes are potent regulators of genome functions and gaining increasing attention in the field of genomics.

2.4. RNA-DNA hybrids (R-loops)

RNA-DNA hybrids (R-loops) - a special type of nucleic acid structure - have been detected in various organisms from bacteria to mammals. These cellular structures have become increasingly popular in chromatin research during the past 10 years (**Figure 4**). This growing body of research can be originated from technological advancements, allowing scientists to study these structures in a previously unreachable resolution and scale. These studies unfolded numerous properties of RNA-DNA hybrids, marking these structures as important regulators of genome integrity and function. As being the primary focus of this dissertation, the following sections will explore the details of this structure.

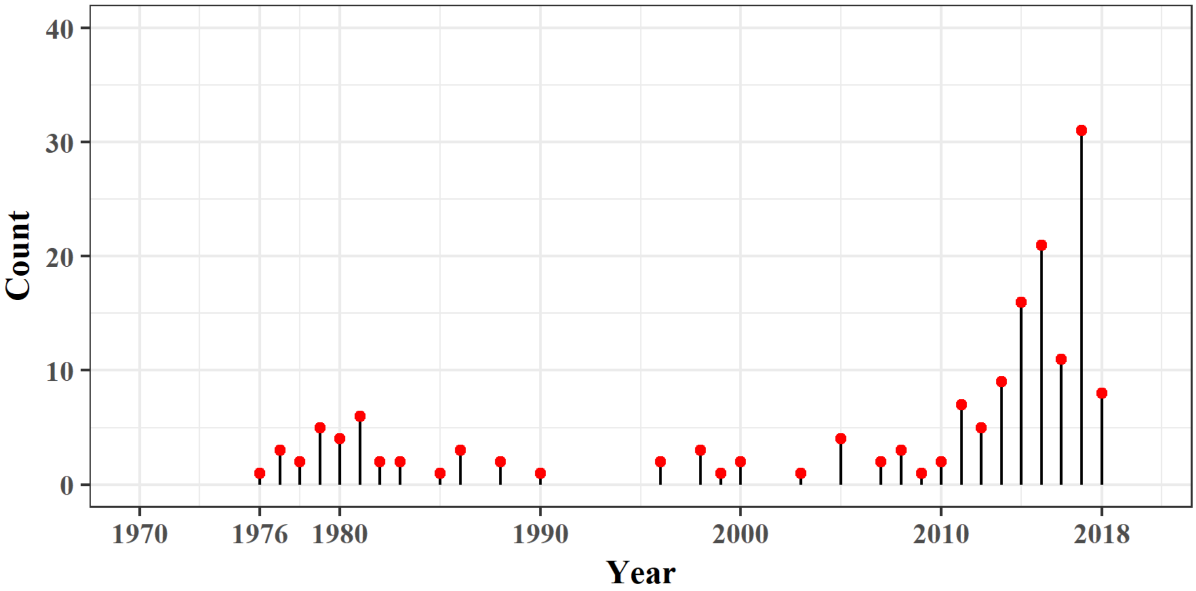


Figure 4. Number of published papers with keyword: R-loop & hybrid. Figure showing the number of R-loop publications during the period of 1976-2018. Data was obtained from the NCBI PubMed database.

2.4.1. Structure

An R-loop is an evolutionally conserved, three-stranded secondary DNA structure in which an RNA molecule partially or completely hybridize to a template strand of the DNA duplex via complementary base-pairing, generating an RNA-DNA hybrid, while displacing the non-template single-stranded DNA (**Figure 5**) (Aguilera and García-Muse 2012). The first study in 1976 demonstrated that RNA can hybridize to complementary DNA under near denaturation temperature in the presence of 70% formamide *in vitro*, indicating that RNA-DNA hybrids are thermodynamically more stable than duplex DNA (Thomas et al. 1976). The stability of the structure is further supported by Li^+, Na^+, K^+ and Cs^+ ions and protein binding (discussed later). Alternatively, the single-stranded part of the R-loop can give rise to G-quadruplex for stabilization, due to transcription and negative supercoiling (Zheng et al. 2017). However, the regulation of this G-quadruplex formation and stabilization is still unclear.

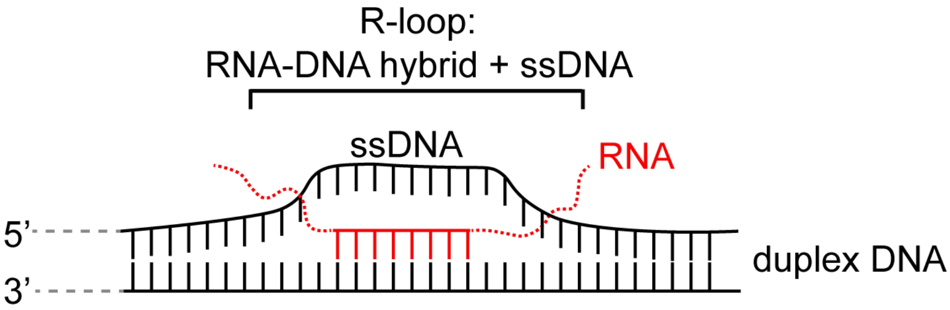


Figure 5. Structure of an R-loop. R-loops are characterized by the invasion of an RNA molecule into duplex DNA, generating an RNA-DNA hybrid and a displaced single-stranded DNA (ssDNA).

2.4.2. R-loop formation and functions

R-loops are abundant epigenetic features in mammalian systems. It is estimated, that ~5% of the human genome has the ability to form R-loops (Sanz et al. 2016b). Recently, several accepted models exist for R-loop formation (**Figure 6**). The first models proposed by Lieber and Roy are the 'thread back' and 'extended hybrid' model (Roy et al. 2008). In the thread back model, single-stranded nascent RNA reanneals to its complementary sequence in a short period of time. While in the extended hybrid model, an R-loop forms upon abortive transcription (e.g. 8 bp RNA-DNA hybrid at RNA polymerase active site). These mechanisms are also called cis-R-loop formations, due to their co-transcriptional associations. It has been generally accepted, that most R-loops form in a co-transcriptional manner. However, R-loops can form in both coding and non-coding parts of the genome. Recent models support the idea of trans-R-loop formation, in addition to the previous model. According to the trans-R-loop model, the RNA molecule is transcribed elsewhere in the genome (even other chromosome). These RNAs can be regulatory long non-coding RNAs (lncRNAs), circular RNAs (circRNA) or repetitive RNAs. In some cases, both cis- and trans-R-loops can be present in the region. This is the so-called mixed model.

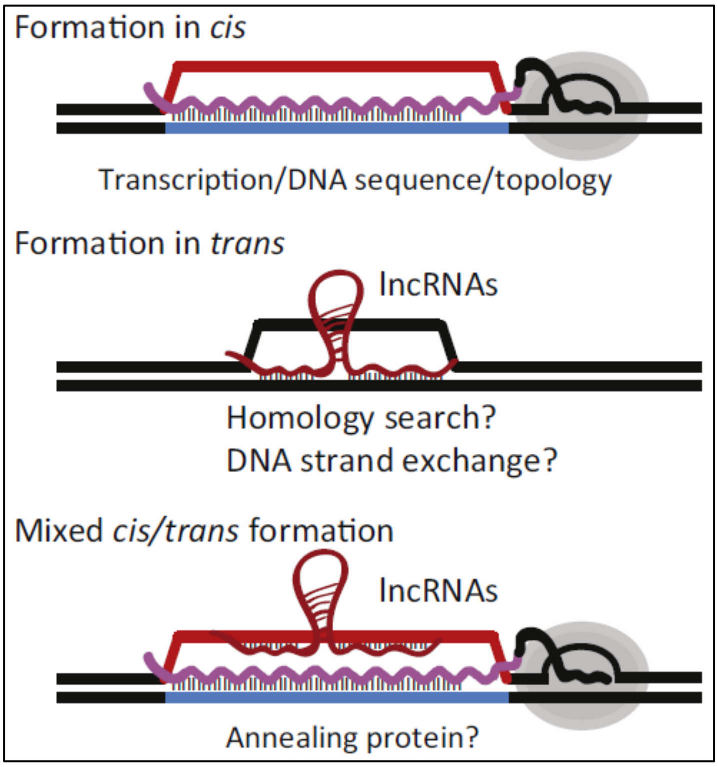


Figure 6. R-loop formation mechanisms. Three possible mechanisms of R-loop formation: cis-R-loop, trans-R-loop and mixed (Chédin 2016).

R-loop formation is not a random process. There are several basic determinants in the genome that facilitate or prevent R-loop formation (**Figure** 7) (Chédin 2016).

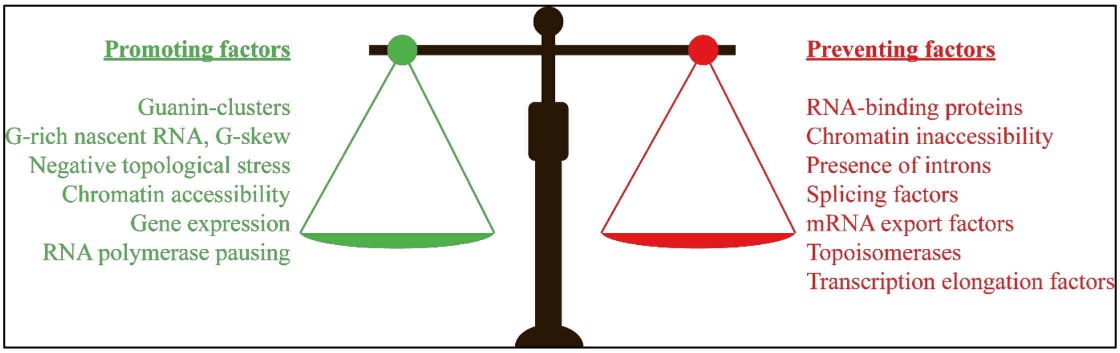


Figure 7. Schematic visualization of features regulating R-loop formation. List of R-loop promoting (green) and preventing (red) features (Chédin 2016).

Multiple features listed here are inter-related. Early *in vitro* studies revealed, that R-loop formation is highly related to their sequence environment. Efficient R-loop initiation requires G-rich nascent RNAs, particularly with guanine clusters. Even short sequences with only one G-cluster (G4) is more favourable for R-loop initiation than random sequences (Roy and Lieber 2009). R-loop elongation beyond the initiation sites is not reliant on G-clusters. Other studies demonstrated that transcription through unmethylated CpG island (CGI) promoters with GC-skew (strand asymmetry in the distribution of guanine versus cytosine residues) leads to R-loop formation (Ginno et al. 2012). Studies investigating DNA supercoiling showed, that negative topological stress is tightly linked to R-loop formation (Drolet et al. 1994). Other observations have indicated that promoter R-loops tend to form over DNA sequences where elevated RNA Polymerase II pausing happens (Chen et al. 2017). Open and active chromatin regions (marked by H3K36me3, H3K4me3/me2 and H3K9Ac) and high transcription rates also positively correlate with R-loop formation (Xu et al. 2017). Interestingly, genes with R-loops are expressed in a higher amount compared to genes without R-loops. With lesser extent, R-loops can also form within repressive chromatin states (marked by H3K27me3, H3K9me2 and H3K27me1). In a recent publication, Chédin and colleagues demonstrated, that gene associated R-loops undergo dynamic turnover with an average 10 minute half-life after transcription inhibition (Sanz et al. 2016b). R-loops can also form within intergenic regions of the genome. Experimental evidence demonstrated the existence of R-loops within repetitive elements, telomeric, pericentromeric regions (Nadel et al. 2015).

These observations indicate that R-loop formation is a complex interplay between nucleotide sequence, transcription, DNA topology and other chromatin characteristics. Despite of the increasing evidence of R-loop formation and functions, the mechanistic details of these processes are still lacking.

2.4.3. R-loop resolution

Since R-loops are associated with many biological processes it is important for the cell to coordinate and regulate R-loop formation to genomic location where needed. For this reason, several RNA and DNA metabolism factors prevent R-loop formation. The primary enzyme responsible for R-loop dissolution is ribonuclease H (RNase H), which is part of the ribonuclease superfamily and specialized to catalyse the cleavage of the RNA molecule within an RNA-DNA substrate in a non-sequence specific fashion. The human genome encodes two isoforms, with slightly different substrates; H1 and H2. Ribonuclease H1 treatment is often considered as a negative control in R-loop studies. Besides RNaseH, specific RNA-DNA hybrid helicases have been identified (e.g. Senataxin, Aquvarius, Pif1) that are able to resolve RNA-DNA hybrid structures.

2.4.4. Function

In recent publications, R-loops have been proposed to act as double-edged swords in the genome since they can mediate physiological and pathological events (Skourti-Stathaki and Proudfoot 2014). In this section, I highlight several R-loop mechanisms and biological processes they regulate.

For example, R-loops potentiates the binding of two key chromatin-regulatory complexes, Tip60-p400 (histone acetyltransferase) and polycomb repressive complex 2 (PRC2) in mouse embryonic stem cells (ESC) to promote their differentiation (Chen et al. 2015). Both factors are regulated by interactions with RNA molecules, although this mechanism is poorly understood. Interestingly, transcription is positively correlate with Tip60-p400 recruitment to promoters. Moreover, genes marked by promoter-proximal R-loops, have a higher level of Tip60-p400 levels and low PRC2 levels and this ratio is flipped if R-loops were disrupted. Taken together, these features clearly showed that R-loops can modulate the binding of key pluripotency regulators.

Secondly, the vimentin (*VIM*) gene expression positively correlates with the expression of its antisense transcript. Both transcripts are silenced in several tumors by promoter hypermethylation. Furthermore, it has been identified that *VIM-AS* transcription promotes the formation of an R-loop structure around the *VIM* promoter. Using *VIM-AS* knockdown and R-loop destabilization assays, local chromatin condensation around promoter regions and reduced binding of transcription factors such as NFκB have been showed. These results indicate that R-loop formation promotes transcriptional activation (Boque-Sastre et al. 2015).

R-loop levels can be modulated by external stimuli. Using breast cancer model, Stork and their colleagues (Stork et al. 2016) showed, that 2h of estrogenic treatment with E2 causes a dramatic increase in R-loops. Moreover, R-loop dependent DNA damage is associated with cell cycle, specifically with the S-phase.

R-loops can mediate heterochromatin formation and higher order chromatin organisation in fission yeast. Using RNA immunoprecipitation and immunofluorescent assays Nakama and their colleagues (Nakama et al. 2012) identified, that heterochromatic ncRNA was associated with chromatin via the formation of R-loops. This R-loop structure is further bound by the RNA-induced transcriptional silencing (RITS) complex. Interestingly, the overexpression or depletion of RNase H1 *in vivo* decreased or increased the number of R-loops and consequently the local heterochromatin.

In a very recent publication Cristini and their colleagues (Cristini et al. 2018) defined the molecular players in the RNA-DNA interactome in human cells using affinity purification and mass spectrometry. Moreover, they showed that one of the top interactome candidates is DXH9 which promotes R-loop suppression and regulates transcriptional termination.

Given the results of the MS analysis (**Table 2** and **Table 3**), they provided a huge amount of resources and hints for further studies of R-loop functions.

Table 2. List of RNA-DNA hybrid interactors identified by MS. (Wang et al. 2018; Cristini et al. 2018)

Gene	Protein name	p-value
Transcription		
DDX5	Probable ATP-dependent RNA helicase DDX5	1.32E-06
ZNF326	DBIRD complex subunit ZNF326	2.83E-06
CTCF	Transcriptional repressor CTCF	2.88E-06
MED19	Mediator of RNA polymerase II transcription subunit 19	8.20E-06
TTF1	Transcription termination factor 1	8.67E-06
Splicing and Processing		
SYNCRIP	Heterogeneous nuclear ribonucleoprotein Q	1.91E-06
SNRPE	Small nuclear ribonucleoprotein E	3.04E-06
PRPF19	Pre-mRNA-processing factor 19	3.45E-06
HNRNPA1	Heterogeneous nuclear ribonucleoprotein A1	4.03E-06
TRA2A	Transformer-2 protein homolog alpha	4.97E-06
SRPK1	SRSF protein kinase 1	1.10E-05
U2AF1	Splicing factor U2AF 65 kDa subunit	1.38E-05
SRSF9	Serine/arginine-rich splicing factor 9	2.60E-05
SNRNP70	U1 small nuclear ribonucleoprotein 70 kDa	1.66E-04
U2AF2	Splicing factor U2AF 65 kDa subunit	2.38E-04
Epigenetic gene regulation		
WHSC1	Histone-lysine N-methyltransferase NSD2	2.42E-06
HP1BP3	Heterochromatin protein 1-binding protein 3	3.04E-06
HDAC2	Histone deacetylase 2	5.17E-06
BAZ1B	Tyrosine-protein kinase BAZ1B	5.97E-06
MBD2	Methyl-CpG-binding domain protein 2	1.77E-05
NAT10	N-acetyltransferase 10	3.42E-05
KMT2A	Histone-lysine N-methyltransferase 2A	3.82E-05
CDYL	Chromodomain Y-like protein	3.87E-05
BRD7	Bromodomain-containing protein 7	6.04E-05
CBX3	Chromobox protein homolog 3	8.06E-05
RUVBL2	RuvB-like 2	1.58E-04
DNMT1	DNA (cytosine-5)-methyltransferase 1	1.71E-04
SUV39H1	Histone-lysine N-methyltransferase SUV39H1	1.25E-03
CBX5	Chromobox protein homolog 5	1.56E-03
SMARCA5	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A member 5	4.41E-03
DNA replication and repair		
TOP2A	DNA topoisomerase 2-alpha	7.76E-06
PRKDC	DNA-dependent protein kinase catalytic subunit	1.94E-05
PARP1	Poly [ADP-ribose] polymerase 1	3.42E-05
PARP2	Poly [ADP-ribose] polymerase 2	9.60E-05
PCNA	Proliferating cell nuclear antigen	1.42E-04
DDB1	DNA damage-binding protein 1	2.87E-04
XAB2	XPA Binding Protein 2	3.71E-04
MCM3	DNA replication licensing factor MCM3	2.81E-03

Table 3. RNA-DNA hybrid interactors identified by MS with known implications in R-loop biology in mammalian cells. (Cristini et al. 2018)

Gene	Protein name	p-value
Transcription		
DHX9	ATP-dependent RNA helicase A	1.11E-06
ILF3	Interleukin enhancer-binding factor 3	1.82E-06
ILF2	Interleukin enhancer-binding factor 2	2.27E-06
XRN2	5-3 exoribonuclease 2	4.10E-06
DDX1	ATP-dependent RNA helicase DDX1	4.94E-06
SUPT16H	FACT complex subunit SPT16	2.12E-04
SNW1	SNW domain-containing protein 1	2.39E-04
SSRP1	FACT complex subunit SSRP1	1.86E-03
RNA processing and export		
DDX21	Nucleolar RNA helicase	7.91E-07
HNRNPC	Heterogeneous nuclear ribonucleoproteins C1/C2	1.29E-06
SNRPD1	Small nuclear ribonucleoprotein Sm D1	1.32E-06
SNRPB	Small nuclear ribonucleoprotein-associated proteins B	1.87E-06
HNRNPU	Heterogeneous nuclear ribonucleoprotein U	3.62E-06
SNRPD3	Small nuclear ribonucleoprotein Sm D3	3.88E-06
SNRPA1	U2 small nuclear ribonucleoprotein A	4.03E-06
SNRNP40	U5 small nuclear ribonucleoprotein 40 kDa protein	8.88E-06
FUS	RNA-binding protein FUS	1.23E-05
TARDBP	TAR DNA-binding protein 43	1.85E-05
PRPF8	Pre-mRNA-processing- splicing factor 8	2.10E-05
DDX23	Probable ATP-dependent RNA helicase DDX23	2.11E-05
TARBP2	RISC-loading complex subunit TARBP2	3.24E-05
TAF15	TATA-binding protein- associated factor 2N	8.7E-05
CRNKL1	Crooked neck-like protein 1	1.52E-04
CDC40	Pre-mRNA-processing factor 17	2.08E-04
SRPK2	SRSF protein kinase 2	2.58E-04
SRSF1/2/3	Serine/arginine-rich splicing factor 3	3.49E-04
DNA Topology		
TOP1	DNA topoisomerase 1	1.83E-04
Replication		
MCM5	DNA replication licensing factor MCM5	3.71E-04
Mitosis		
BUB3	Mitotic checkpoint protein BUB3	3.29E-04
ZNF207	Zinc finger protein 207	4.14E-04

2.4.4.1. Role in disease

Abnormal R-loop structures or perturbation of normal R-loop homeostasis in the cells are being realized as a crucial contributor to human disease by generating hotspots for genomic instability (DNA double-strand or single-strand breaks, hypermutations or chromosome rearrangement).

Mutations in genes involved in R-loop homeostasis can result in increased or decreased R-loop levels at genes or regulatory regions leading to the perturbation of normal cell functions. R-loops are associated with different kind of cancers and neurodegenerative diseases. Several excellent reviews have been published in this topic, collecting the current knowledge about R-loops and their pathological context (Richard and Manley 2017; Groh and Gromak 2014; Sollier and Cimprich 2015; Skourti-Stathaki and Proudfoot 2014).

As a result, R-loops or R-loop processing/promoting factors can be potential targets for drug development.

2.4.5. Detection approaches of the RNA-DNA hybrids

In this part of the dissertation, I briefly review the currently used, different approaches to detect RNA-DNA hybrids (R-loops), with special attention to their individual strengths and weaknesses.

After the initial discovery of R-loops using electron microscopy (Thomas et al. 1976), several other techniques became available to identify these structures. The most important milestone of the field was the development of the R-loop monoclonal antibody: S9.6 in 1986. This antibody recognizes the RNA-DNA hybrid part of the R-loops with high affinity. The S9.6 antibody made it possible to study R-loops *in vivo* with many different molecular biology techniques, like immunofluorescence imaging or high throughput sequencing (**Figure 8**).

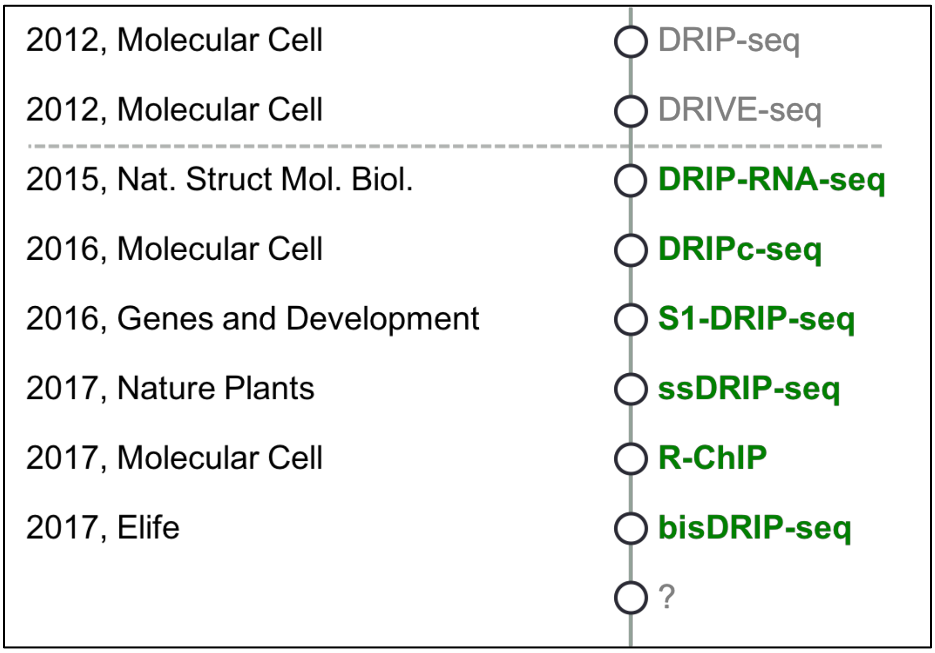


Figure 8. List of NGS based methods for R-loop detection. Schematic figure of the available R-loop detection methods, showing the timeline ordered by years. Dashed line represents the date of first publications.

The most commonly used methodology is DNA-RNA immunoprecipitation followed by quantitative PCR (DRIP-qPCR) or next-generation sequencing (DRIP-seq) (Ginno et al. 2012). Briefly, extracted genomic DNA is fragmented either by sonication or restriction enzymes. Next, S9.6 coated antibodies capture the DNA fragments with hybrid structures while removing any unwanted fragments. After eluting the fragments from the beads, the antibody-DNA-RNA hybrid connection is unlinked. In the last step of the experiment, the purified nucleic acid fragments are quantified by qPCR or NGS. Usually, RNase H1 treatment is used for negative control.

Numerous research groups have utilized the DRIP method since the initial publications (**Figure 9**), however, the potential limitations of the method are not completely understood. For example, the validation of the results is difficult without using the antibody. Moreover, short read sequencing produces uneven read coverage in GC-rich regions, thus, by its low complexity properties, aligners have difficulties during sequence alignment. Later in this thesis, we unravel the critical steps of the experiment in depth (Halász et al. 2017).

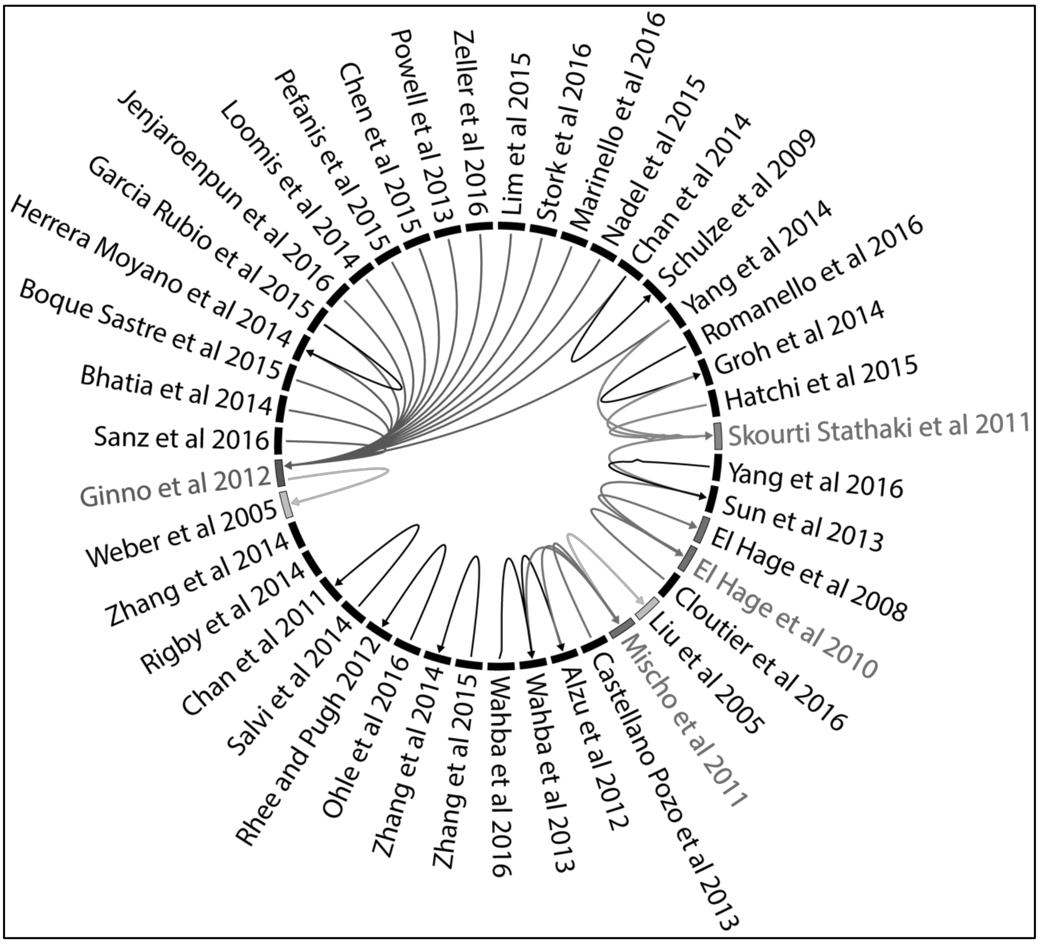


Figure 9. Reference network of DRIP experiments based on scientific literature. Circos plot shows the DRIP studies (nodes) and the DRIP methods referenced by each study (arrows). Base-nodes (in bold) point to studies that receive many citations; most DRIP experiments originate from 2-3 chief studies. Light-gray edge highlights the MeDIP approach (methylated DNA immunoprecipitation, Weber et al 2005) forming the basis of the original DRIP protocol.

2.4.5.1. Computational prediction

Given the sequence preferences of R-loops, putative R-loop forming sites for a given genome can be predicted. Kuznetsov and his co-authors (Kuznetsov et al. 2018; Wongsurawat et al. 2012) have developed a computational algorithm to map R-loop forming sequences (RLFS) in several organisms (**Table 4**). Briefly, RLFS regions can be partitioned into three bins: R-loop initiation zone (RIZ), linker and R-loop extension zone (REZ). The initiation is considered as few Gs (3-4 nt) in the region. The linker is between the RIZ and REZ regions up to 50 nucleotides. The extension zone requires high G density (at least 40% of G). The length of the REZ varies from 100 to 2000 nucleotides.

Table 4. Predicted R-loop forming sequences by organisms. (Kuznetsov et al. 2018)

Organism	#predicted RLFS	%RFLS in genic and proximal regions
Human	664.791	76.96
Mouse	575.403	52.88
Rat	454.018	42.12
Chimpanzee	530.728	57.04
Chicken	213.942	68.38
Frog	319.296	32.87
Fruitfly	5.947	77.30
Yeast	78	94.87

2.4.5.2. Alternative, complementary methods

Few years after the original DRIP protocol, several complementary methods have emerged (**Figure 8**). These methods can be grouped based on the immunoprecipitation target (DNA or protein), sequenced molecule (DNA or RNA) and library preparation.

S1-DRIP-seq (Wahba et al. 2016) is an improved methodology of the original method. It uses S1-nuclease treatment before immunoprecipitation which results in an improved signal-to-noise ratio.

Methods, like DRIP-RNA-seq (Chen et al. 2015) and DRIPc-seq (Sanz et al. 2016a) follows the steps of DRIP protocol up to immunoprecipitation. Purified and enriched RNA-DNA hybrids are denatured and treated with DNase I to remove any DNA contaminants from the samples. The remaining RNA molecules are subjected to strand-specific RNA-seq library preparation and sequencing. A clear advantage of these strategies is that we can gather information about the orientation of the hybrids.

A recent method applies single-strand DNA ligation-based library construction after DNA-RNA hybrid immunoprecipitation combined with high throughput sequencing (ssDRIP-seq) (Xu et al. 2017). DRIPed DNA samples are sonicated and denatured on 95 ^{o}C to obtain single-stranded DNA before library preparation and sequencing. Other methods make use of a catalytically-deficient but binding competent RNaseH1 mutant protein, like DNA-RNA *in vitro* enrichment (DRIVE-seq) (Ginno et al. 2012) and R-ChIP (Chen et al. 2017).

DRIVE-seq is prepared in affinity pulldown assays, while R-ChIP is a chromatin immunoprecipitation-based method.

The most recent, alternative method is the bisDRIP-seq (Dumelie and Jaffrey 2017). This is a bisulfite-based footprinting approach to map R-loops at a resolution of single base pair across whole-genomes. The concept behind this method is that bisulfite treatment selectively converts unmethylated cytosine residues into uracil at single-stranded DNA portion of the R-loop structure under non-denaturing conditions. Moreover, the RNA-DNA hybrid part of the R-loop is protected from the C-to-U conversions. Thus, this method provides a strand-specific and high-resolution R-loop mapping method. However, its main limitation is the uneven distribution of cytosines and methylation. Overall, huge effort has been made to improve the resolution, specificity and sensitivity to detect true positive R-loops. More technologies are expected to appear, like novel long-read and single molecule sequencing or other R-loop binding protein-based approach, like the ssDNA-binding, replication protein A (RPA-ChIP) can be envisaged. The first part of this thesis is focuses on the key experimental variables present in the DRIP protocol and how these variables affects the overall sensitivity and specificity of R-loop detection.

3. Aims of the study

Aim 1. Evaluation of the accuracy and sensitivity of DNA-RNA hybrid mapping method: DNA-RNA immunoprecipitation (DRIP).

RNA-DNA hybrids (R-loops) are prevalent epigenetic features existing in every level of the tree of life. R-loops form when an RNA molecule anneals to a homologous DNA molecule, creating an RNA-DNA hybrid and a displaced single stranded DNA. Early scientific research considered R-loops as potential hotspots for genome instability as their single stranded part is prone to damaging DNA and can introduce mutations or chromosomal rearrangements. However, a growing body of evidence suggests that R-loops massively form under physiological conditions affecting critical cellular processes such as transcription factor binding, gene expression or heterochromatin formation. Therefore, accurate identification and characterisation of these structures are of key importance. The work presented in this thesis aimed to:

- Systematically screen and determine the possible confounding effects related to the key experimental variables during R-loop detection, using DNA-RNA immunoprecipitation (DRIP)
- Determine the sensitivity and specificity of the DRIP method
- Do comparative functional analysis using whole-genome human R-loop datasets
- Draw attention to use optimal restriction enzyme combinations to avoid biased genome sampling
- Recommend an optimized DRIP protocol for the scientific community

Aim 2. Functional analysis of Spp1 chromatin binding during meiosis.

Meiosis and meiotic recombination are the essence of heredity and evolutional variability. Recombination occurs as a programmed event that results in new combinations of parental alleles via crossover and non-crossover pathways. It has been directly confirmed in the early 1990s, that meiotic recombination is initiated by the formation of DNA double-strand breaks (DSBs) during early prophase I. The process is catalysed at the chromosome axis by a conserved topoisomerase-like enzyme, Spo11. Recently, high throughput technologies mapped DSBs in several different organisms. These studies provided evidence that DSBs are discreetly scattered along the genome forming DSB hotspots and regulated at multiple levels. These hotspots are often accumulated within nucleosome-free and GC-rich intergenic regions flanked by H3K4me3 histone marks near promoter regions in *Saccharomyces cerevisiae (S. cerevisiae)*. In *S. cerevisiae*, all H3K4 methylation is catalysed by the Set1 complex (COMPASS/Set1C). However, the mechanism by specific sites became hotspots and anchored to the chromosomal axis is poorly understood. The tethered-loop axis model proposes that Spp1, the PHD finger domain containing H3K4me3 reader subunit of the Set1C interacts with both H3K4me3 marks and axis proteins enabling effective cleavage by Spo11. The work presented in this thesis investigates the chromatin binding characteristics of Spp1, using genome-wide assays and bioinformatic approaches in meiotic nuclei. More specifically, our aim was to:

- Investigate the chromosome binding kinetics of Spp1 during meiosis
- Characterize the functional relevance of binding sites with different binding kinetics

4. Materials and Methods

4.1. Methods

Cell cultures

The Jurkat human T lymphoblastoid cell line was cultured in RPMI-1640 medium (Sigma-Aldrich) supplemented with 10% fetal bovine serum albumin (BSA), L-glutamine and gentamycin at 37°C, in a humidified 5% CO_2 chamber. 100 million exponentially growing cells were washed twice with 1 x PBS and divided into equal aliquots for the twenty-four DRIP experiments. In DRIP experiments #5 and #13, the GM12878 B lymphoblastoid cell line was used for comparison with the Jurkat cell line. The GM12878 cells obtained from the Coriell Institute for Medical Research. Cells were grown at 37°C in a humidified 5% CO_2 chamber in vented 25 cm^2 cell culture flasks, containing 10-20 ml of RPMI-1640, L-glutamine and gentamycin.

Naive $CD4^+$ T-cell Isolation

Leukocyte enriched buffy coats were obtained from healthy blood donors (individual donations) drawn at the Regional Blood Center of the Hungarian National Blood Transfusion Service (Debrecen, Hungary) in accordance with the written approval of the Director of the National Blood Transfusion Service and the Regional and Institutional Ethics Committee of the University of Debrecen, Medical and Health Science Center (Hungary). PBMCs were separated by a standard density gradient centrifugation with Ficoll-Paque Plus (Amersham Biosciences, Uppsala, Sweden). Naive T-cells were separated from human blood mononuclear cells using the naive $CD4^+$ T-cell isolation kit based on negative selection according to the manufacturer's instruction (Miltenyi Biotec). Using the $CD4^+$ T Cell Isolation Kit, human $CD4^+$ T helper cells are isolated by negative selection. Non-target cells are labeled with a cocktail of biotin-conjugated monoclonal antibodies and the $CD4^+$ T Cell

MicroBead Cocktail. The magnetically labeled non-target T cells are depleted by retaining them on a MACS® Column in the magnetic field of a MACS Separator, while the unlabeled T helper cells pass through the column.

Detection of RNA-DNA hybrids by Dot Blot Assay

For dot blot immunoassays 6 µg of phenol-chloroform extracted and sonicated Jurkat nucleic acid were treated with 1 µl RNase A (10 mg/ml; UD-GenoMed Ltd.) in TE buffer with different salt concentration (from 25 mM to 500 mM) at 37°C for 1 hour. The RNase H digestion was performed using 1-8 µl of RNase H (5000 U/ml; NEB) at 37°C, overnight. For control samples, we used sonicated nucleic acid without any further treatment, alkali-treated sonicated nucleic acid (incubated with 1 µl of 50 mM NaOH at 65°C for 10 min) and sonicated nucleic acid resuspended in 1x RNase H Reaction Buffer (NEB) at incubated at 37°C without RNase H enzyme. Both the RNase A and RNase H digested samples were spotted on a nitrocellulose membrane (GE Healthcare) at two different concentrations (600 ng and 125 ng) in triplicates. After drying the spots at room temperature, the membrane was fixed with UV for 5 minutes and blocked by 5% bovine serum albumin (BSA) in PBST buffer (PBS containing 0.25% Tween) at room temperature for 20 minutes. The blocked membrane was then incubated with the S9.6 antibody in PBST buffer containing 5% bovine serum albumin at room temperature for 2 hours. The membrane was washed five times with PBST and incubated with goat anti-mouse IgG marked with HRP (Santa Cruz Biotechnology) at room temperature for 1 hour. After five washes with PBST, the signal was detected by the SuperSignal West Pico Chemiluminescent substrate (Thermo Fisher Scientific) and imaged using FlourChemQ (ProteinSimple).

Immunofluorescent Labeling of RNA-DNA hybrids

Jurkat cells were resuspended in pre-warmed (37°C) 1 x PBS to a density of $6x10^6$ ml and diluted 4-fold in 1% molten low melting point agarose dissolved in PBS. 22 µl of the cell suspension (~ 33.000 cells) was spread in each well of an 8 chamber Ibidi slide. After the gel set, agarose embedded cells were washed three times in PBS (500 µl/well) on ice. The permeabilization, lysis and nuclei preparation were performed in one step using a Lysis Buffer consisting of 1% (v/v) TritonX-100, and 2 M NaCl in PBS/EDTA (500 µl/well, 10 minutes on ice). The samples were blocked by 5 mg/ml BSA dissolved in PBS/5mM EDTA, on ice for 30 minutes. RNA-DNA hybrids were labeled by the S9.6 monoclonal mouse antibody and a rabbit anti-mouse Alexa647 secondary antibody. Imaging was carried out using a Zeiss Axiovert 200M confocal laser scanning microscope. Signal intensities were quantified by ImageJ.

Detection of RNA-DNA hybrids by DNA-RNA immunoprecipitation (DRIP)

I. DRIP classifiers 1-16

Crosslinking (Step 1)

Crosslinking of Jurkat cells (experiments 1-8) was done with 1% paraformaldehyde (UP) for 10 minutes, then quenched with 2.5 M glycine (pH 6, final concentration: 500 mM) for 5 minutes at room temperature. Crosslinking was omitted from experiments 9-16.

Cell lysis

Cells were lysed in 1 ml lysis buffer composed of 500 µl 2x lysis buffer (1% SDS, 20 mM Tris-HCl pH 7.5, 40 mM EDTA pH 8, 100 mM NaCl, ddH_2O) plus 500 µl TE buffer (100 mM Tris-HCl pH 8, 10 mM EDTA pH 8) per 5 million cells. Cell lysis was performed at two different temperatures: either at 65°C for 7 hours, or at 37°C overnight, as indicated in the text.

Phenol chloroform extraction of total nucleic acid (Step 2)

In experiments 1-4, and 9-12, total nucleic acid was prepared by phenol-chloroform extraction. Before the phenol-chloroform extraction step, the nucleic acid preps were treated with 10 µl of Proteinase K (20 mg/ml; Thermo Fisher Scientific) at 65°C for 7 hours, or at 37°C overnight, to remove the proteins. The extracted DNA was precipitated with 1/10 volume 3 M Na-acetate (pH 5.2) plus 1 volume of isopropanol. The DNA pellet was dissolved in 200 µl of 10 mM Tris-HCl pH 8.

Silica membrane-based (kit) extraction of total nucleic acid (Step 2)

In experiments 5-8 and 13-16, total nucleic acid was isolated by the NucleoSpin Tissue Kit (Macherey-Nagel) according to the manufacturer's protocol, except the cell lysis step that was performed either at 65°C for 7 hours (according to the kit protocol), or at 37°C overnight, where indicated in the text. Nucleic acids were eluted in 500 µl of elution buffer (5 mM Tris-HCl pH 8.5).

Removal of free RNA by RNase A treatment (Step 3)

In experiments 3-4, 7-8, 11-12, and 15-16, the DNA purification step was directly followed by the RNase A digestion of free ribonucleic acids. The purified DNA preps (from Step 2) were supplemented with 18 µl of 5 M NaCl and 2 µl of RNase A (10 mg/ml; UD-GenoMed Ltd.) in a buffer containing 10 mM Tris-HCl (pH 8) and 300 mM NaCl (V=300 µl) at 37°C for 1 hour. RNase A-treated samples were re-purified either by phenol-chloroform extraction (experiments 4, 12) or by the NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel) (experiments 8, 16). Phenol-chloroform extracted DNA was dissolved in 100 μl of 5 mM Tris-HCl pH 8.5. The DNA purified with the kit was eluted in 5 mM Tris-HCl pH 8.5.

Nucleic acid fragmentation by sonication (Step 4)

In experiments 1, 3, 5, 7, 9, 11, 13, 15, the purified nucleic acid preps were sonicated in a buffer of 10 mM Tris-HCl pH 8.5 supplemented with 300 mM NaCl (V=300 μl) for 2 x 5 min (30 sec ON, 30 sec OFF, LOW; Bioruptor, Diagenode) to yield an average DNA fragment size of ~300 bp.

Nucleic acid fragmentation by restriction enzyme digestion (Step 4)

In exp. 2, 4, 6, 8, 10, 12, 14, 16, purified DNA samples (~25 μg each) were fragmented using a restriction enzyme cocktail of 1 μl HindIII (20 U/μl), 1 μl EcoRI (20 U/μl), 2 μl BsrGI (10 U/μl), 1 μl XbaI (20 U/μl), 4 μl SspI (5 U/μl)) in NEB Buffer 2 (NEB) (V=300 μl) at 37°C, for 4 hours.

The fragmented DNA samples were re-purified either by phenol-chloroform extraction (experiments 1-4; 9-12) or by the NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel) (experiments 5-8; 13-16). The DNA was dissolved in 100 μl of 5 mM Tris-HCl pH 8.5.

Two percent (V/V%) of the DNA preps were kept as input DNA for the DRIP-qPCR measurement. Half of samples were treated by 8 μl of RNase H (5000 U/ml; NEB) in a total volume of 80 μl, at 37°C, overnight.

II. DRIP classifiers 17-24

Crosslinking (Step 1)

Crosslinking of Jurkat cells (experiments 17-20) was done with 1% paraformaldehyde (UP) for 10 minutes, then quenched with 2.5 M glycine (pH 6, final concentration: 500 mM) for 5 minutes at room temperature. Crosslinking was omitted from experiments 21-24.

Chromatin preparation (Step 2)

Cell lysis

Cells were lysed 750 µl of ChIP lysis buffer (50 mM HEPES-KOH at pH 7.5, 140 mM NaCl, 1 mM EDTA at pH 8, 1% Triton X-100, 0.1% Na-Deoxycholate, 1% SDS) per 10 million cells and homogenized using Fast Prep-24 5G (MP Biomedicals, speed: 6 m/s; time: 40 sec; 2 cycles; pause time: 120 sec; A lysing matrix).

Chromatin fragmentation by sonication (Step 3)

300 µl of chromatin preps were sonicated for 2 x 5 min (30 sec ON, 30 sec OFF, LOW, Bioruptor) to yield an average DNA fragment size of ~300 bp.

Removal of free RNA by RNase A treatment (Step 4)

In experiments 19, 20, 23, 24, the sonication step was directly followed by the RNase A digestion of free ribonucleic acids. The fragmented chromatin was supplemented with 270 µl of 5 M NaCl (300 mM) and 10 µl of RNase A (10 mg/ml; UD-GenoMed Ltd.) in 4500 µl of TE buffer (10 mM Tris-HCl pH 8, 10 mM EDTA pH 8) at 37°C for 1 hour.

Before Step 5, the chromatin preps were treated with 30 µl of Proteinase K (20 mg/ml; Thermo Fisher Scientific) at 65°C overnight to remove the proteins and reverse the cross-links.

Phenol chloroform extraction of total nucleic acid (Step 5)

In experiments 17, 19, 21, 23, total nucleic acid was prepared by phenol-chloroform extraction. The extracted DNA was precipitated with 1/10 volume 3 M Na-acetate (pH 5.2) plus 1 volume of isopropanol. The DNA pellet was dissolved in 100 µl of 5 mM Tris-HCl pH 8.5.

Silica membrane-based (kit) extraction of total nucleic acid (Step 5)

In experiments 18, 20, 22, 24, total nucleic acids were isolated by the NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel) according to the manufacturer's protocol. Nucleic acids were eluted in 100 µl of elution buffer (5 mM Tris-HCl pH 8.5).

Two percent (V/V%) of the DNA preps were kept as input DNA for the DRIP-qPCR measurement. Half of samples were treated by 8 µl of RNase H (5000 U/ml; NEB) in a total volume of 80 µl, at 37°C, overnight.

III. RNA-DNA hybrid Immunoprecipitation with the S9.6 antibody

Dynabeads Protein A magnetic beads (Thermo Fisher Scientific) were pre-blocked with PBS/EDTA containing 0.5% BSA. To immobilize the S9.6 antibody, 50 µl pre-blocked Dynabeads Protein A was incubated with 10 µg of S9.6 antibody in IP buffer (50 mM Hepes/KOH at pH 7,5; 0,14 M NaCl; 5 mM EDTA; 1% Triton X-100; 0,1% Na-Deoxycholate, ddH2O) at 4°C for 4 hours with rotation. Six micrograms of digested genomic DNA were added to the mixture and gently rotated at 4°C, overnight. Beads were recovered and washed successively with 1ml lysis buffer (low salt, 50 mM Hepes/KOH pH 7.5, 0.14 M NaCl, 5 mM EDTA pH 8, 1% Triton X-100, 0.1% Na-Deoxycholate), 1ml lysis buffer (high salt, 50 mM Hepes/KOH pH 7.5, 0.5 M NaCl, 5 mM EDTA pH 8, 1% Triton X-100, 0.1% Na-Deoxycholate), 1 ml wash buffer (10 mM Tris-HCl pH 8, 0.25M LiCl, 0.5% NP-40, 0.5%

Na-Deoxycholate, 1 mM EDTA pH 8) and 1ml TE (100 mM Tris-HCl pH 8, 10 mM EDTA pH 8) at 4°C, two times. Elution was performed in 100 µl of elution buffer (50 mM Tris-HCl pH 8, 10 mM EDTA, 1% SDS) for 15 min at 65°C. After purification by NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel), nucleic acids were eluted in 55 µl of elution buffer (5 mM Tris-HCl pH 8.5). The recovered DNA was then analyzed by quantitative real-time PCR (qPCR). qPCR was performed with LightCycler 480 SYBR Green I Master (Roche) and analyzed on QuantStudio 12K Flex Real-Time PCR System (Thermo Fisher Scientific). qPCR results were analyzed using the comparative C_T method. The RNA-DNA hybrid enrichment was calculated based on the IP/Input ratio.

Receiver Operating Characteristic (ROC) Analysis

ROC curves were obtained for each DRIP variables (DRIP experiments) by ranking the studied genomic loci having known RNA-DNA hybrid states (based on the training set) according to their DRIP-qPCR profile, starting from the lowest to the highest estimated DRIP scores and then calculating sensitivity and specificity. The ROC curves plotted the sensitivity or true positive rate (TPR) against the false-positive rate (FPR) or 1-specificity, estimated as follows: TPR=P(positive DRIP-qPCR result | R-loop present), FPR=P(positive DRIP-qPCR result | R-loop absent), where P means conditional probability. The AUC values were then calculated from the observed DRIP-qPCR (IP/input) yields using the pROC algorithm.

DNA-RNA immunoprecipitation (DRIP) sequencing

DRIP-seq libraries were prepared according to the Illumina’s TruSeq ChIP Sample Preparation protocol. Briefly, the enriched DRIP DNA was end-repaired and indexed adapters were ligated to the inserts. Purified ligation products were then amplified by PCR. Amplified libraries were prepared and sequenced in the Genomic Medicine and Bioinformatics Core

Facility of the University of Debrecen (1x50 bp read length, single-end, Illumina HiScan SQ) and at the EMBL Genomics Core Facility, Heidelberg (2x150 bp read length, paired-end, Illumina HiSeq 2500).

Sequenced reads were aligned to the Human reference genome (hg19) using default parameters of BWA MEM (Li and Durbin 2009) algorithm. Low mapping quality, supplementary alignments, reads mapped to blacklisted regions and redundant reads were omitted (Li et al. 2009; Quinlan and Hall 2010) from downstream analysis. Replicate experiments (rep1, rep2) were merged and then MACS2 (Zhang et al. 2008) was used to identify enriched regions (FDR 1%) of the genome normalized to input datasets.

Processed and merged alignments were subjected to bamCoverage (Ramírez et al. 2014) to generate signal files. RPKM (Reads Per Kilobase Million) values were calculated for 20 bp bins for each sample and smoothed using a 60 bp sliding window (--binSize 20 --smoothLength 60 –normalizeUsingRPKM --extendReads 300). The generated signal files were visualized in R 3.2.2, using the ggplot2 (Wickham 2009) and ggbio (Yin et al. 2012) packages.

Genomic Annotation of RNA-DNA hybrids

We used GenomicRanges (Lawrence et al. 2013) to determine the genomic distribution of DRIP peaks, allowing us to calculate the intersecting area between binding sites and the corresponding annotation categories. Areas occupied by the intersected regions were compared to a randomized peak coverage. Random peak sets were generated for each chromosome by permutation, considering the chromosomal distribution of chromatin states and omitting blacklisted regions.

In silico restriction enzyme digestion

To calculate the expected (theoretical) fragment length distribution generated by a combination of restriction enzymes (HindIII, EcoRI, BsrGI, XbaI and SspI), we cut the human (hg19) and yeast (sacCer3) genomes *in silico* with the DECIPHER R package (Wright 2016). From the cutting site positions, we calculated the length of restriction fragments. Statistical comparison of the resulting fragment length distributions was performed by the Wilcoxon Rank Sum test by randomly sampling 300 values 100 times. P-values were adjusted with Benjamini & Hochberg correction.

Step-by-step protocol of the best-performing DRIP experiment (exp. 5)

Crosslinking of cells was done with 1% paraformaldehyde (UP) for 10 minutes, then quenched with 2.5 M glycine (pH 6, final concentration: 500 mM) for 5 minutes at room temperature. Cells were lysed in lysis buffer provided by the NucleoSpin Tissue kit (Macherey-Nagel) at 65°C for 7 h (according to the kit protocol), or at 37°C overnight (where indicated in the main text). Total nucleic acid was isolated by a NucleoSpin Tissue Kit (Macherey-Nagel) and eluted in 100 µl of elution buffer (5 mM Tris-HCl pH 8.5). The purified nucleic acid prep was fragmented by sonication in 300 µl of Tris-HCl pH 8.5 (high salt concentration: 300 mM NaCl) for 2 x 5 min (30 sec ON, 30 sec OFF, LOW, Bioruptor) to yield an average DNA fragment size of ~500 bp. Fragment analysis was done by using 1% agarose gelelectophoresis. If it was necessary, further sonication was applied. The sonicated DNA sample was purified by a NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel) and eluted in 100 µl of elution buffer (5 mM Tris-HCl pH 8.5). Twelve micrograms of DNA were diluted with 5 mM Tris-HCl pH 8.5 to a total volume of 100 µl. Two percent of the sample was kept as input DNA. Half of the sample was treated with 8 µl of RNase H (5000 U/ml; NEB) in a total volume of 80 µl at 37°C, overnight. Dynabeads Protein A magnetic beads

(Thermo Fisher Scientific) were pre-blocked with PBS/5 mM EDTA containing 0.5% BSA. To immobilize the S9.6 antibody, 50 µl pre-blocked Dynabeads Protein A was incubated with 10 µg of S9.6 antibody in IP buffer (50 mM Hepes/KOH at pH 7,5; 0,14 M NaCl; 5 mM EDTA; 1% Triton X-100; 0,1% Na-Deoxycholate, ddH2O) at 4°C for 4 hours with rotation. Six micrograms of digested genomic DNA were added to the mixture and gently rotated at 4°C, overnight. Beads were recovered and washed successively with 1ml lysis buffer 1 (low salt, 50 mM Hepes/KOH pH 7.5, 0.14 M NaCl, 5 mM EDTA pH 8, 1% Triton X-100, 0.1% Na-Deoxycholate), 1ml lysis buffer 2 (high salt, 50 mM Hepes/KOH pH 7.5, 0.5 M NaCl, 5 mM EDTA pH 8, 1% Triton X-100, 0.1% Na-Deoxycholate), 1ml wash buffer (10 mM Tris-HCl pH 8, 0.25M LiCl, 0.5% NP-40, 0.5% Na-Deoxycholate, 1 mM EDTA pH 8) and 1ml TE (100 mM Tris-HCl pH 8, 10 mM EDTA pH 8) at 4°C, two times. Elution was performed in 100 µl of elution buffer (50 mM Tris-HCl pH 8, 10 mM EDTA, 1% SDS) for 15 min at 65°C. After purification by NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel), nucleic acids were eluted in 55 µl of elution buffer (5 mM Tris-HCl pH 8.5). The recovered DNA was then analyzed by quantitative real-time PCR (qPCR). qPCR was performed with LightCycler 480 SYBR Green I Master (Roche) and analyzed by a QuantStudio 12K Flex Real-Time PCR System (Thermo Fisher Scientific). The qPCR data were evaluated by the comparative C_T method. The RNA-DNA hybrid enrichment was calculated based on the IP/Input ratios.

Yeast strains

All yeast strains are from the SK1 background (**Table 5**) and were constructed by Lóránt Székvölgyi and Laurent Acquaviva. For sporulation, cells were grown in rich medium (YPD) for 24 h, then transferred to SPS and grown overnight to a density of ~4 × 10^7 cells/ml. Cultures were harvested by centrifugation, washed with one volume of prewarmed 1% potassium acetate and resuspended in SPM (1% potassium acetate supplemented with amino acids and nucleotides according to auxotrophic requirements, and 0.0001% of polypropylene glycol 2000 as an anti-clumping agent) at a density of 2 × 10^7 cells/ml, at 30°C. Meiotic progression and sporulation efficiency was monitored by FACS and fluorescent microscopy of DAPI-stained nuclei. Aberrantly slow or asynchronous sporulation time courses were excluded from further experiments. Spore viability was assessed by tetrad dissection, and it was greater or equal to 90% for all the strains involved in this study.

Table 5. Genotype of yeast strains constructed and used in this study.

No	Strain ID	Genotype
1	Spp1-myc	Mata/alpha, SPP1-13xmyc-KanMX4/SPP1-13xmyc-KanMX4, leu2/LEU2, HIS4/his4, trp1/trp1, ura3/ura3
2	Bre2-myc	Mata/alpha, BRE2-13xmyc-NatMX4/BRE2-13xmyc-NatMX4, leu2/LEU2, HIS4/his4, trp1/trp1, ura3/ura3
3	Spp1PHDΔ-myc	Mata/alpha, spp1PHDΔ-13xmyc-KanMX4/spp1PHDΔ-13xmyc-KanMX4, leu2/LEU2, HIS4/his4, trp1/trp1, ura3/ura3
4	Spp1CXXCΔ-myc	Mata/alpha, spp1CXXCΔ-13xmyc-NatMX4/spp1CXXCΔ-13xmyc-NatMX4, leu2/LEU2, HIS4/his4, trp1/trp1, ura3/ura3
5	Spp1-HA, H3	Mata/alpha, lys2/lys2, ho::LYS2/ho::LYS2,trp1-1/trp1-1, ura3/ura3,leu2/LEU2,HIS4/his4, hht1Δ::HphMX/hht1Δ::HphMX, hht2Δ::KanMX4/hht1Δ::HphMX, SPP1-3xHA::KanMX4/SPP1-3xHA::KanMX4, pCEN-ARS(HHT2,HHF2,TRP1)
6	Spp1-HA, H3R2A	Mata/alpha, lys2/lys2, ho::LYS2/ho::LYS2, trp1-1/trp1-1, ura3/ura3, leu2/LEU2, HIS4/his4, hht1Δ::HphMX/hht1Δ::HphMX, hht2Δ::KanMX4/hht1Δ::HphMX, SPP1-3xHA::KanMX4/SPP1-3xHA::KanMX4, pCEN-ARS(HHT2R2A,HHF2,TRP1)
7	Spp1-HA, H3K4R	Mata/alpha, lys2/lys2, ho::LYS2/ho::LYS2, trp1-1/trp1-1, ura3/ura3, leu2/LEU2, HIS4/his4, hht1Δ::HphMX/hht1Δ::HphMX, hht2Δ::KanMX4/hht1Δ::HphMX, SPP1-3xHA::KanMX4/SPP1-3xHA::KanMX4, pCEN-ARS(HHT2K4R,HHF2,TRP1)

The mutants were created as described previously (Acquaviva et al. 2013b). Firstly, Spp1 C-terminal part is labelled with 9xmyc epitope using the plasmid pFA6a-9xmyc-KanMX6. Deletion of the PHD domain of SPP1 was performed by fusing two PCR fragments extending from -569 (relative to the ATG) to +3 and from +235 to +1062. Deletion of the CXXC motif of SPP1 (C263GYC266) was performed by deleting 12 base pairs in the integrative vector carrying SPP1-myc. The above mutations were introduced into histone point mutant strains (H3R2A and H3K4R) by crossing.

Spp1 ChIP experiments

50 ml of meiotic cells (4 × 10^7 cells/ml) were collected at the indicated time points and cross-linked with 1% formaldehyde for 20 min at room temperature. Formaldehyde was quenched with 125 mM glycine for 5 min at room temperature, and cells were washed three times with ice-cold 1× TBS at pH 7.5 (20 mM Tris-HCl at pH 7.5 and 150 mM NaCl). Cells were resuspended in 500 µl of lysis buffer (50 mM Hepes KOH at pH 7.5, 140 mM NaCl, 1mM EDTA, 1% Triton X-100, 0.1% Na-deoxycholate, and 1 tablet of complete inhibitor cocktail [Roche] in 50 ml solution) and lysed with acid-washed glass beads for 10 min in a FastPrep bead beater machine. Chromatin samples was fragmented to an average size of 300 bp by sonication (Bioruptor; Diagenode). To obtain whole-cell extract, a 50 µl pre-immunoprecipitation (IP) sample was removed and centrifuged at full speed for 10 s to pellet the cell debris (supernatant = whole-cell extract). The rest of the samples were also centrifuged at 12,000 rpm (4°C) for 20 s to pellet the cell debris. IP was performed by adding the 450-µl extract to a pellet of magnetic protein G dynabeads (Dynal), corresponding to 50 µl or 2 × 10^7 beads, which were preincubated with the 9E11 (monoclonal mouse anti-myc, ab56; Abcam) or anti-GFP (polyclonal rabbit, ab290; Abcam) antibodies overnight at 4°C. IP samples were washed twice with lysis buffer, twice with lysis buffer plus 360 mM NaCl, twice with washing buffer (10 mM Tris-HCl at pH 8.0, 250 mM LiCl, 0.5% NP-40, 0.5% Na-

deoxycholate, and 1 mM EDTA), and finally once with 1× TE at pH 7.5, using the magnetic device supplied by Dynal. After reversal of cross-linking by heating in TE-1% SDS overnight at 65°C, the proteins were digested with proteinase K (12 µl of 20 mg/ml stock) for 3h at 65°C. Nucleic acids were PCR clean up kit purified and RNA digestion (10 µg RNase) was performed for 1 h at 37°C. The DNA was finally resuspended in 50 µl nuclease-free dH_2O.

NGS library preparation and deep sequencing

Sequencing libraries were prepared according to the Illumina's TruSeq ChIP Sample Preparation protocol. In brief, the enriched ChIP DNA was end-repaired and indexed adapters were ligated to the inserts. Purified ligation products were then amplified by PCR. Amplified libraries were prepared in the Genomic Medicine and Bioinformatics Core Facility of the University of Debrecen, Debrecen, Hungary. The libraries were sequenced using 50 single-end reads with Illumina HiScan SQ (Genomic Medicine and Bioinformatics Core Facility of the University of Debrecen); or with Illumina HiSeq 2500 (EMBL Genomics Core Facility, Heidelberg, Germany).

Raw reads were aligned to the *S. cerevisiae* reference genome (SacCer3; SGD) using the default parameters of Burrows-Wheeler Aligner algorithm (Li and Durbin 2009) and 38–67% of the sequenced reads were retained after removing low mapping quality (MAPQ < 10) and PCR duplicate reads (Picard).

Enrichment analysis and Spp1 peak annotation

BayesPeak was used with default parameters to identify ChIP enriched regions (peaks) of the genome compared with input control (Cairns et al. 2011). Peaks sets identified at individual meiotic time points (SPS, 0, 2, 4, and 6 h in SPM) were concatenated and sorted by chromosomal position, and then merged. We used mergeBed (Quinlan and Hall 2010) to join the overlapping peak positions. The overlap of peak sets detected in different samples was represented by proportional Venn diagrams.

We used deepTools2 (Ramírez et al. 2014) bamCoverage to create Reads Per Kilobase per Million mapped reads (RPKM)–normalized bedgraph files. For each bedgraph we calculated the $\log_2$(IP/INPUT) ratios and used these coverage files for visualization and downstream analysis. Heat maps were generated with computeMatrix and plotHeatmap functions of deepTools2. Read density profiles were generated using HOMER and plotted in R.

To estimate the enrichment or depletion of Spp1 binding sites within genomic features we created 100–100 randomized peak sets with the shuffleBed function. Then, we calculated with intersectBed the coverage ratio of observed and randomized peak sets over the relevant annotation categories and over Mer2/Red1 ChIP binding sites. Differences in overlap ratios were then compared by the prop.test function of R.

Identification of dynamic Spp1 clusters

To classify Spp1 binding sites based on their binding dynamics, we first merged every Spp1 binding sites identified at all meiotic time points (union peak set). Next, we mapped the average $\log_2$(IP/INPUT) RPKM ratios of the ChIP samples back to the union peak set. Binding site coverage values were z-transformed across ChIP samples with the scale function in R. Dynamic clusters were identified using a k-means algorithm and plotted with pheatmap (Kolde 2015).

Multidimensional scaling (MDS)

Using the union peak set as described above, we applied the cmdscale MDS method in R to visualize the level of similarity between Spp1 datasets. Euclidean distance matrices generated from this table were readily used as an input for cmdscale. The resulting 2D coordinates were plotted in R as a scatter plot.

Statistical analysis

All statistical analyses were performed in R (version 3.4.4). Group comparisons were performed by ANOVA (aov() R function). Groups were compared with Tukey's post-hoc test (Tukey HSD R function). If the data did not fit the normal distribution, we used Kruskal-

Wallis's ANOVA (kruskal.test R function) and the Mann-Whitney *U* test (wilcox.test R function). Probability values of $P \leq 0.001$ were considered as statistically significant. Significance marks: not significant (ns). $P > 0.05$; *, $P \leq 0.05$; **, $P \leq 0.01$; ***, $P \leq 0.001$; ****, $P \leq 0.0001$. The number of cases (n) and P values were indicated in each figure legend.

External datasets

SacCer3 genome annotation files were obtained from *Saccharomyces* Genome Database. Promoter and downstream regions were defined as the arbitrary extension of TSSs with 500 bp and TTS by 200 bp. RNA-seq (Brar et al. 2012), H3K4me3 ChIP-chip (Borde et al. 2009), Mer2 and Red1 ChIP-Chip (Panizza et al. 2011; Sun et al. 2015), and Spo11-oligo DSB data (Mohibullah and Keeney 2017) were from the indicated publications.

4.2. Data and software availability

DRIP Sequencing data from this study have been submitted to the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession number SRP095885. For The raw sequencing data and processed data files related to the Spp1 nuclear dynamics have been deposited to Gene Expression Omnibus (GEO) with the accession: GSE107967.

5. Results

5.1. RNA-DNA hybrid (R-loop) immunoprecipitation mapping: an analytical workflow to evaluate inherent biases

As highlighted in the introduction of this thesis, R-loop mapping was largely dependent on a single approach, called DNA-RNA immunoprecipitation (DRIP). The original experimental protocol can be traced back to a few research groups. Yet, most of the studies are using the same protocol (with modifications or not), without paying attention to critical experimental variables that may account for some contradictory results in the field.

In the current dissertation, we present an analytical workflow in which we tested the possible confounding effects during R-loop detection with DRIP method. Briefly, we tested the effect of formaldehyde fixation, cell lysis temperature, fragmentation and removal of free RNA on the efficacy of RNA-DNA hybrid detection. We applied DRIP-sequencing, qPCR and ROC analysis, to rank each workflow based on their performance. We showed that some of the workflows performed poorly and generates random answers. Notably, the improper use of restriction enzymes results in lengthy DNA fragments, compromising mapping resolution and misinterpretation of many R-loop sites.

The work presented here, aims to establish an optimized workflow for R-loop detection using the objective criteria of ROC analysis.

5.1.1. Introducing DRIP classifiers to assess true and false R-loop associations

Based on the available workflows of published DRIP protocols and considering the main technical variables that might contribute to the observed heterogeneities, we designed 40 DRIP experimental schemes (binary classifiers), thus, we assessed how they rank different test loci according to their known RNA-DNA hybrid status (**Figure 10**). The classifiers ("DRIP experiments" or "dependent variables") were designed to systematically explore the main factors that might create experimental bias associated with the DRIP procedure. Experiments 1-16 considers the effect of *i.* formaldehyde (HCHO) fixation, *ii.* the method of nucleic acid isolation, *iii.* removal of free RNA, *iv.* the mode of nucleic acid fragmentation (**Figure 10A**), and *v.* cell lysis temperature (65°C as default vs. 37°C).

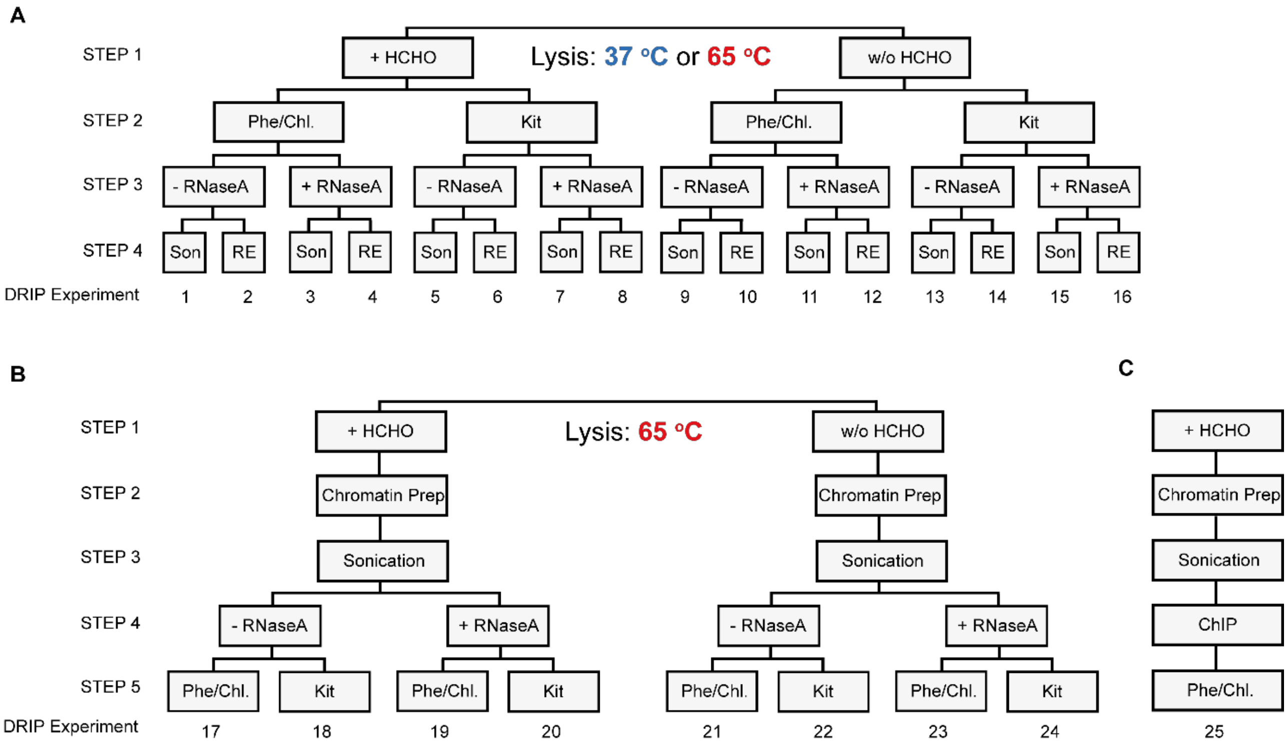


Figure 10. Experimental design: constructing DRIP schemes. (A) Experiments 1-16 explore the effect of formaldehyde-fixation (Step 1), nucleic acid isolation (Step 2), removal of free RNA (Step 3), and nucleic acid fragmentation (Step 4) on the outcome of RNA-DNA hybrid detection. Each experiment was performed at two parallel cell lysis temperatures (65°C and 37°C), respectively. (B) Experiments 17-24 test the impact of acoustic sharing performed on chromatin prep rather than on naked nucleic acid, similarly to the ChIP protocol. Each experiment was performed at 65°C cell lysis temperature. (C) Workflow of a ChIP experiment (shown only for comparison with the DRIP pipeline). Abbreviations: HCHO: Formaldehyde fixation; Phe/Chl: Phenol/Chloroform extraction; Kit: silica membrane-based nucleic acid purification; RNase A: Ribonuclease A digestion performed at high NaCl concentration (300 mM). Son: Sonication; RE: restriction enzyme cocktail digestion (HindIII, EcoRI, BsrGI, XbaI and SspI). As a negative control, RNase H digestion was applied in all DRIP experiments (not indicated in Figure 10).

5.1.1.1. The effect of formaldehyde fixation

The basic assumption behind HCHO-crosslinking is to maximize the DRIP yield while preserving biologically meaningful RNA-DNA hybrid interactions. But formaldehyde has some well-known adverse effects: *i.* the DNA accompanies a conformational change upon crosslinking, involving local denaturation or “breathing” of the double helix (McGhee and von Hippel 1977). This might create ectopic R-loop sites or abolish physiological R-loop contacts. *ii.* HCHO-treatment can reduce antigen accessibility or mask epitopes recognized by the antibody used for immunoprecipitation. This might prevent a fraction of R-loops from being detected. *iii*. HCHO-fixation elicits spurious localization of irrelevant proteins at highly expressed genes (Baranello et al. 2016), and induces massive poly(ADP)ribose polymer formation in live cells (Beneke et al. 2012). These examples warrant deeper investigation of the usage of HCHO-fixation in RNA-DNA hybrid mapping, therefore we classified our DRIP samples as HCHO-treated and non-treated categories (**Figure 10A-B**).

5.1.1.2. Comparison of nucleic acid purification

Two common methods were compared: organic (phenol/chloroform) extraction versus solid-phase (silica membrane) purification of total nucleic acids (**Figure 10A-B**).

5.1.1.3. Application of ribonucleolytic treatment

Most DRIP protocols do not treat the isolated nucleic acid with ribonucleases to remove free RNA, however the S9.6 antibody can recognize RNA duplexes with a ~5-fold reduced affinity compared to RNA-DNA hybrids (Phillips et al. 2013; Hartono et al. 2018). At this point, four kinds of ribonucleoleolytic digestion were incorporated into our DRIP pipelines: *i.* RNase H1 digestion that removes RNA-DNA hybrids (negative control #1), *ii.* alkaline hydrolysis by sodium hydroxide that degrades free RNA and RNA-DNA hybrids (negative control #2), *iii.* RNase A digestion at high NaCl concentration (300 mM) that removes free RNA, *iv.* RNase A digestion at low NaCl concentration (25 mM) that removes free RNA and RNA-DNA hybrids.

RNase H1 treatment is an accepted negative control for the DRIP procedure since it degrades the RNA strand in the hybrids preventing their recognition by the S9.6 antibody. Half of the nucleic acid prep was digested by RNase H1 before the DNA fragmentation step that allowed us to estimate the bulk level of RNA-DNA hybrids (dot blot setting; **Figure 11A**). The other half was digested just before the S9.6 immunoprecipitation step that let us obtain crucial information about the specificity of the IP signal (see DRIP-qPCR). As expected, RNA-DNA hybrids were sensitive to RNase H1 digestion *in vitro*. Similarly, to RNase H1, alkaline hydrolysis using 50 mM NaOH also eliminated the RNA-DNA hybrid signal efficiently (**Figure 11A**). Less is known about the salt-dependent RNase H-like activity of RNase A that is supposed to digest RNA-DNA hybrids as an efficient hybridase at low ionic strength. As shown in **Figure 11B**, the hybrids were indeed resistant to RNase A digestion at high ionic strength, but they became highly sensitive to RNase A as a function of decreasing monovalent concentration. The RNase H-like activity of RNase A at low salt condition was confirmed by an independent method (**Figure 11C-D**) applying fluorescent microscopic detection. Based on these experiments, RNase A digestion at high salt concentration (300 mM NaCl) was integrated into our DRIP protocol to test whether the removal of competing free RNA improves the specificity of the RNA-DNA hybrid signal. Also, RNase H1 digestion of the fragmented nucleic acid was kept as an obligatory negative control for immunoprecipitation.

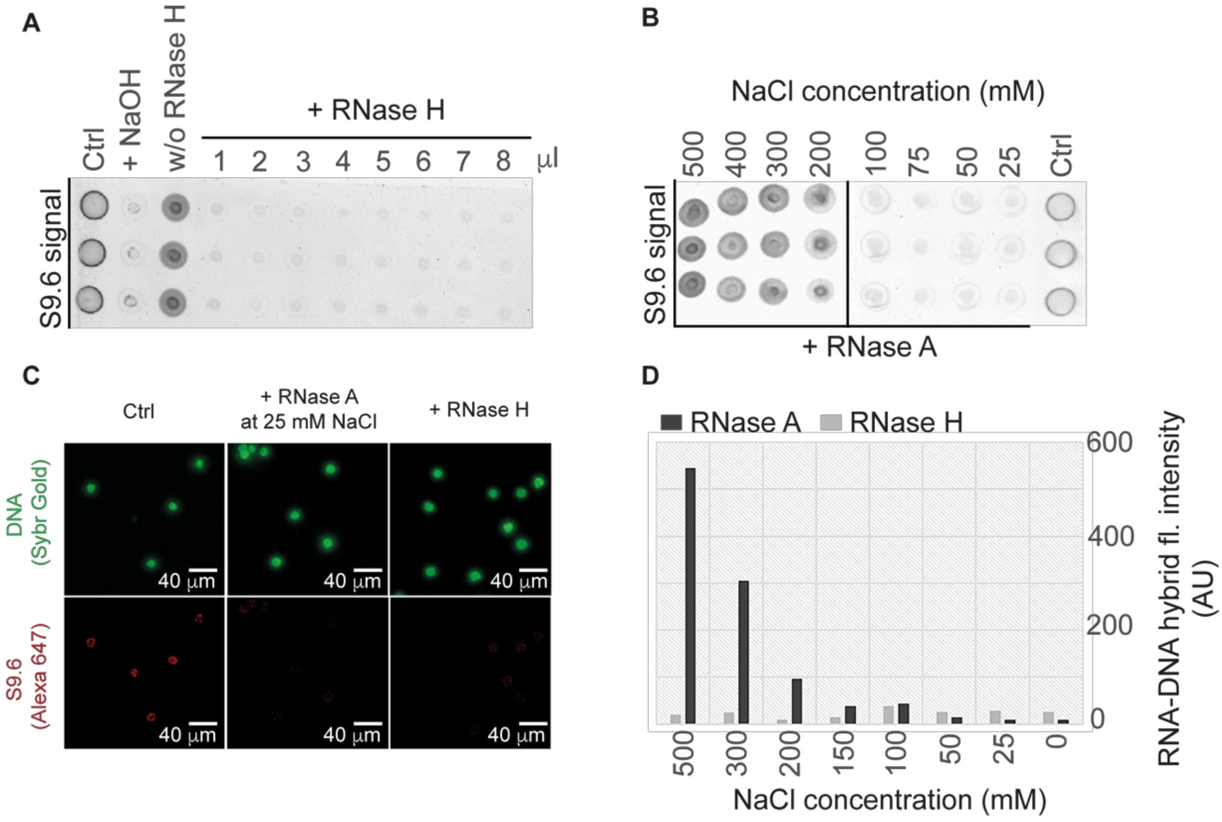


Figure 11. The effect of ribonucleolytic treatment on the level of R-loops. (A) S9.6 dot blot hybridization showing the decrease of RNA-DNA hybrid level due to RNase H digestion. Each spot contains 5 µg of sonicated nucleic acid pipetted in triplicates onto the membrane. The first three columns represent control assays: (1. no treatment; 2: alkaline hydrolysis of free RNA and RNA-DNA hybrids by 50 mM sodium hydroxide (NaOH); 3: buffer control (w/o RNase H). The remaining columns show the effect of RNase H added in increasing amount. (B) S9.6 dot blot hybridization showing the decrease of RNA-DNA hybrid level as a result of RNase A digestion. RNA-DNA hybrids become sensitive to RNase A as a function of decreasing monovalent (NaCl) concentration. Last column: negative (buffer only) control. (C) Confirmation of the salt-dependent RNase H-like hybridase activity of RNase A by fluorescent microscopy. Permeabilized Jurkat cells were treated with RNase A (at low ionic strength) or RNase H, and RNA-DNA hybrid were immunofluorescently labelled by the S9.6 antibody. Green channel: DNA stained by SybrGold. Red channel: rabbit anti-mouse Alexa647 secondary antibody. (D) Microscopic quantification of S9.6 signal intensities upon RNase A digestion performed at decreasing NaCl concentrations. Most RNA-DNA hybrids were not destroyed above 300 mM NaCl, but became efficiently digested below 100 mM NaCl, in line with the dot blot hybridization results. In parallel with each RNase A digestion reactions, nuclear preps were treated with RNase H (as a negative control).

5.1.1.4. Mode of nucleic acid fragmentation

The choice of restriction enzymes defines the cleavage pattern of DNA that is critical to achieve an optimal fragment length distribution and mapping resolution. Based on the original DRIP protocol (Ginno et al. 2012), we combined five enzymes (HindIII, EcoRI, BsrGI, XbaI and SspI) for *in silico* digestion, resulting in a median restriction fragment length of 314 bp (**Figure 12A**). In contrast to the theoretical fragment size distribution, we observed a broad DNA size range in a real digestion reaction (between 100-10.000 bp; **Figure 13A**). As a control, we repeated the restriction enzyme cleavage under different reaction conditions, without detecting any improvement in the digestion efficacy (**Figure 12B**). When a budding yeast genomic DNA was digested in a parallel experiment, we managed to obtain the expected (*in silico*) fragment size distribution (**Figure 12C**). These observations necessitate for the proper control of DNA fragment length distribution in DRIP samples that derive from restriction enzyme fragmented nucleic acid.

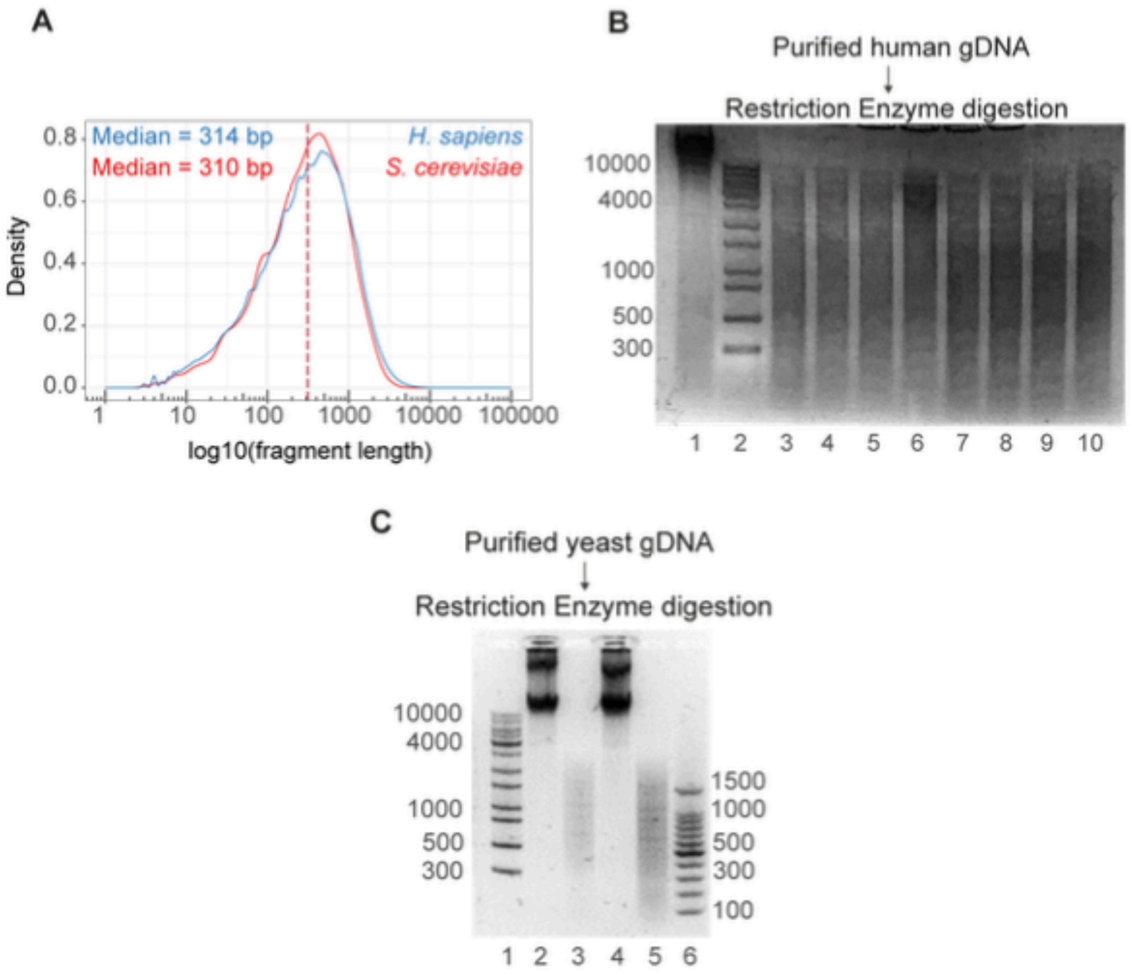


Figure 12. Distribution of DNA Fragment Length of Homo Sapiens and Saccharomyces cerevisiae After Restriction Enzyme Cleavage. (A) Theoretical DNA fragment size distribution as a result of an *in silico* restriction enzyme cocktail fragmentation. The obtained fragment length distributions are similar in both species (Wilcoxon rank sum test: p=0.944, not significant) with a median size of 310-314 bp. (B-C) Restriction fragment length distribution obtained as a result of restriction enzyme cocktail fragmentation in a real digestion reaction. Digestions were performed by a mix of HindIII, EcoRI, BsrGI, XbaI and SspI on genomic DNA purified from human and budding yeast cells, respectively. The observed DNA fragment length distribution of yeast DNA matches with the theoretical distribution, which is not the case for human DNA samples.

Lanes of the agarose gel in panel B:

1. Undigested gDNA
2. 1 Kb Plus DNA ladder
3. 5 µg gDNA + 20 U for each restriction enzyme in NEB2.1 buffer
4. 5 µg gDNA + 20 U for each restriction enzyme in NEB2.1 buffer + 0.1 mg/ml BSA
5. 5 µg gDNA + 20 U for each restriction enzyme in Tango buffer
6. 5 µg gDNA + 20 U for each restriction enzyme in CS buffer
7. 5 µg gDNA + 40 U for each restriction enzyme in NEB2.1 buffer
8. 5 µg gDNA + 40 U for each restriction enzyme in NEB2.1 buffer + 0.1 mg/ml BSA
9. 5 µg gDNA + 40 U for each restriction enzyme in Tango buffer
10. 5 µg gDNA + 40 U for each restriction enzyme in CS buffer

Lanes of the agarose gel in panel C:

1. 1 Kb Plus DNA ladder
2. Undigested DNA (BY4741)
3. 2 µg BY4741 gDNA + 20 U for each restriction enzyme in NEB2.1 buffer
4. Undigested gDNA (BY4742)
5. 2 µg BY4742 gDNA + 20 U for each restriction enzyme in NEB2.1 buffer
6. 100 bp DNA ladder

In contrast with restriction enzyme digestion, sonication creates random DNA fragments with a typical size of 150-500 bp which guides the spatial resolution of the DRIP assay (**Figure 13B**). However, excessive sonication can introduce strand breaks in the DNA or simply shake off a subset of R-loops from the chromosomes, potentially compromising their detectability by qPCR. Taken together, the mode of DNA fragmentation (restriction enzymes and sonication) was introduced as an important parameter in our DRIP pipeline (**Figure 10A**).

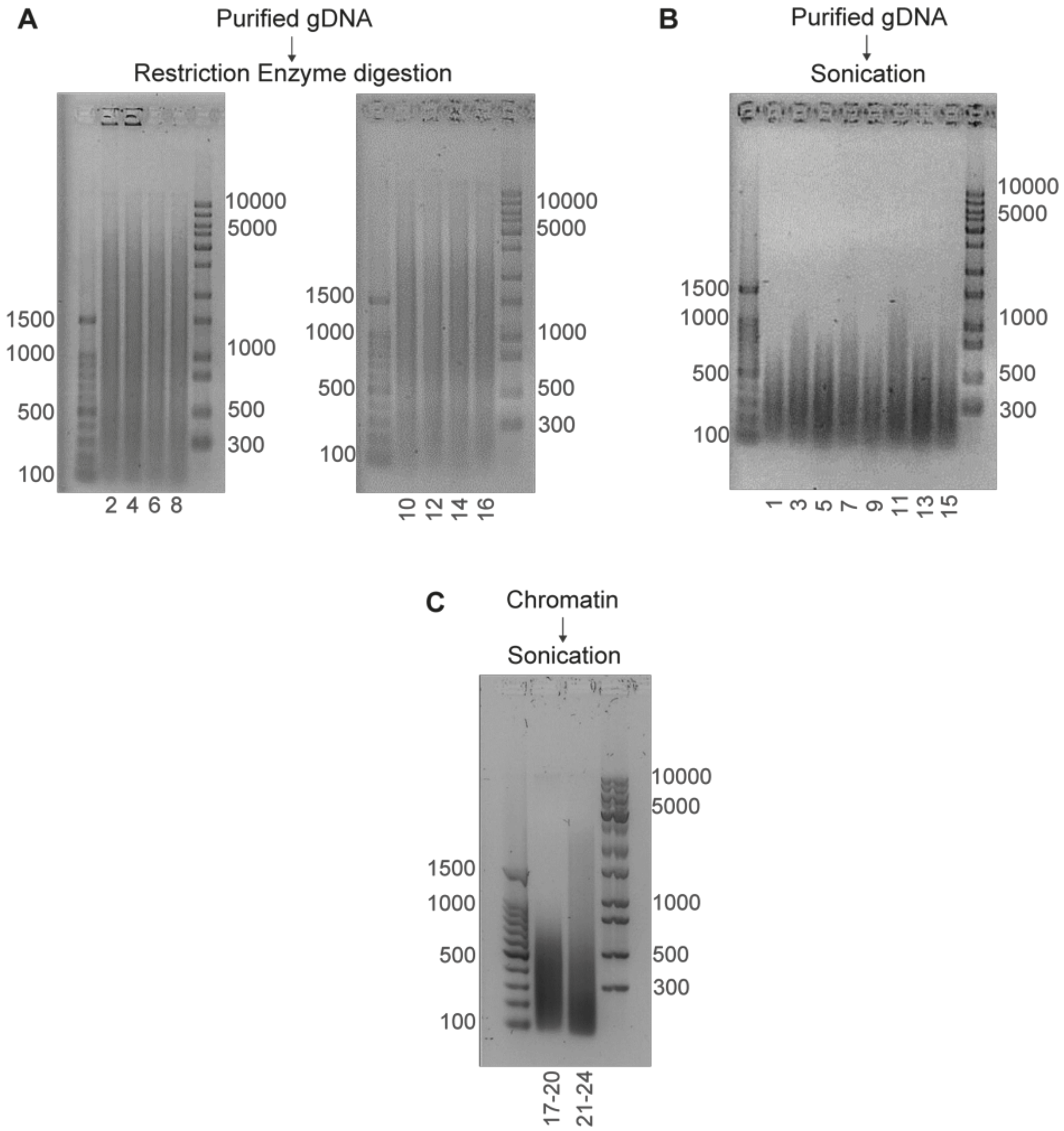


Figure 13. Comparison of the Effect of Restriction Enzyme Cocktail Fragmentation and Sonication to Homo sapiens Genomic DNA. (A-B) DNA fragment length distribution as a result of restriction enzyme cocktail fragmentation (A) and sonication (B). Both restriction enzyme digestions and sonication were performed on naked genomic DNA purified from Jurkat cells. (C) Fragmentation of chromatin by sonication rather than naked DNA. The numbers below the agarose gels indicate the relevant DRIP scheme IDs.

5.1.1.5. Fragmenting chromatin rather than purified genomic DNA

For experiments 17-24, in comparison to the original DRIP protocol, classical chromatin immunoprecipitation (ChIP) was applied to capture RNA-DNA hybrids by immunoprecipitation of cross-linked and sonicated chromatin (rather than naked DNA) followed by phenol/chloroform purification (**Figure 10C**). Since sonication, performed on purified genomic DNA, led to loss of ~80% of the DRIP signal in yeast (Wahba et al. 2016), we tested whether acoustic shearing performed on chromatin prep rather than on naked nucleic acid (**Figure 13C**) could improve the signal to noise ratio of the DRIP measurement (**Figure 10B**).

5.1.1.6. Effect of cell lysis temperature

Published DRIP protocols apply various cell lysis temperatures, ranging from 37°C to 65°C and lasting from a couple of hours to overnight. To test the effect of temperature on the specificity of RNA-DNA hybrid detection, we lysed the samples at 65°C for 7 hrs, or at 37°C overnight. Experiments 1-16 were processed in parallel at both temperatures, while experiments 17-24 were omitted from the temperature analysis since crosslink reversal typically occurs at 65°C.

Taken together, the above-mentioned experimental variables resulted in forty (16x2+8) autonomous DRIP classifiers (schemes) for which RNA-DNA hybrid enrichment scores were determined at several test loci. This allowed us to assess whether the S9.6 signal represented true or false R-loop associations under the applied conditions.

5.1.2. Making a reference R-loop set for benchmarking the DRIP classifiers

In order to derive the parameters of the DRIP classifiers, known positive and negative examples (genomic sites) could be chosen from the scientific literature based on their known R-loop profiles; however, the heterogeneity of the available DRIP-qPCR and DRIP-seq datasets prompted us to establish our independent R-loop training set. We performed DNA-RNA hybrid mapping (DRIP-seq) in two closely related human cell types (Jurkat T cell leukemia cell line and naive CD4$^+$ T lymphocytes) and identified 88.830 and 99.337 R-loop enriched regions, respectively (**Figure 14A**). A high-confidence R-loop peak set was generated from the identified binding sites and their chromosomal distribution was characterized. The peaks were significantly enriched at gene promoters and repetitive elements (**Figure 14B**), which is consistent with previously published DRIP-seq results (Ginno et al. 2012; Nadel et al. 2015). R-loop sites were underrepresented at protein coding exons, similarly to earlier DRIP experiments performed with sonicated nucleic acid, however restriction enzyme fragmented DRIP samples were positively biased towards exons. Sonicated and restriction enzyme digested samples were strikingly different in their R-loop length distributions (narrow: 179-2.369 bp *vs.* wide: 178-22.479 bp; **Figure 14C**), and the identified R-loop binding sites significantly overlapped within each group, but sharply stood apart between the two groups (**Figure 14D**). We attributed these differences to the extensive variation of R-loop lengths and heterogeneities of the studied cell types. Biological implications of having too wide peak sizes will be discussed later. With the observed variances in mind, our consensus R-loop set was regarded as an amenable reference to benchmark the DRIP classifiers.

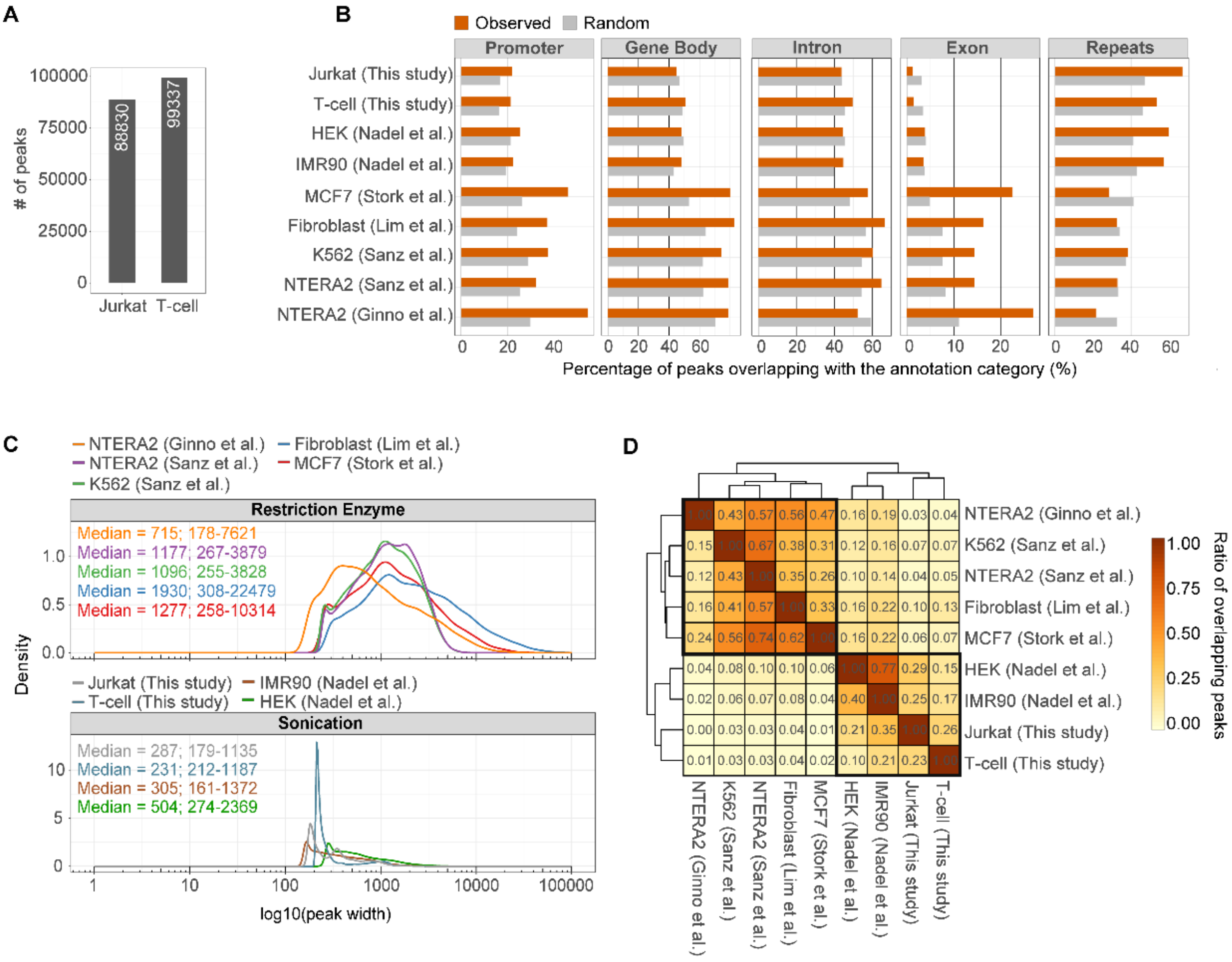


Figure 14. Summary of Available Human DRIP-seq Experiments. (A) Bar chart showing the number of identified R-loop peaks in human Jurkat cells and naive T cells (this study). (B) Annotation of R-loop binding sites over functional genomic elements. DRIP-seq peaks were determined in Jurkat cells and naive T cells, and in other published cell types (NTERA2, K562, Fibroblast, MCF7, IMR90, HEK293T). The upper four rows represent DRIP experiments fragmenting the nucleic acid by sonication, while the lower five rows highlight restriction enzyme-digested DRIP samples. The difference between the two groups is especially remarkable over exons (associated to 14%-27% and 1%-3.5% of R-loops, respectively) and repeat elements (SINEs, LINEs, LTRs, simple and low complexity repeats) that involve 22%-38% and 54%-67% of the R-loop peaks, respectively. At other annotation categories (gene body, introns and promoters) the difference was not significant between the two groups. (C) Density plots showing the distribution of R-loop peaks sizes, classified by fragmentation method (restriction enzyme vs. sonication). Median peak length and 2.5%-97.5% quantiles are indicated. Peak length distributions differ significantly between the two fragmentation methods. (D) Heatmap showing the overlap of R-loop binding sites between independent DRIP-seq experiments. Values and cell colours represent pairwise and unique overlap ratios between each peak set. The difference between the two nucleic acid fragmentation methods is clearly apparent, as peak sets from the same fragmentation process better resemble to each other (highlighted in black).

5.1.3. Measuring RNA-DNA hybrid enrichment over the DRIP classifiers

Positive and negative test regions were selected from the identified R-loop set (**Figure 15**) and were systematically probed for RNA-DNA hybrid enrichment across the DRIP classifiers (**Figure 16**). Five test regions were frequently used as positive and negative controls in various published DRIP studies (*SNRPN, ZNF554, MYADM, FMR1, APOE*; (Yang et al. 2014; Marinello et al. 2016; Groh et al. 2014; García-Rubio et al. 2015; Loomis et al. 2014; Bhatia et al. 2014; Herrera-Moyano et al. 2014; Boque-Sastre et al. 2015; Ginno et al. 2012), while the remaining sites were picked randomly from the consensus R-loop set (*PRR5L, LOC440704, NOP58, VIM, ING3*). The reference DRIP-seq signal (benchmarking the classifiers) is shown over selected test regions along with DRIP-seq patterns taken from published studies (**Figure 15**). DRIP-qPCR yields were measured in control and RNase H-treated samples for forty (16x2+8) DRIP classifiers, at ten test regions, in five independent experiments. The resulting 4000 (40x2x10x5) DRIP enrichment scores were then readily used as an input parameter of receiver operator characteristics (ROC) calculation.

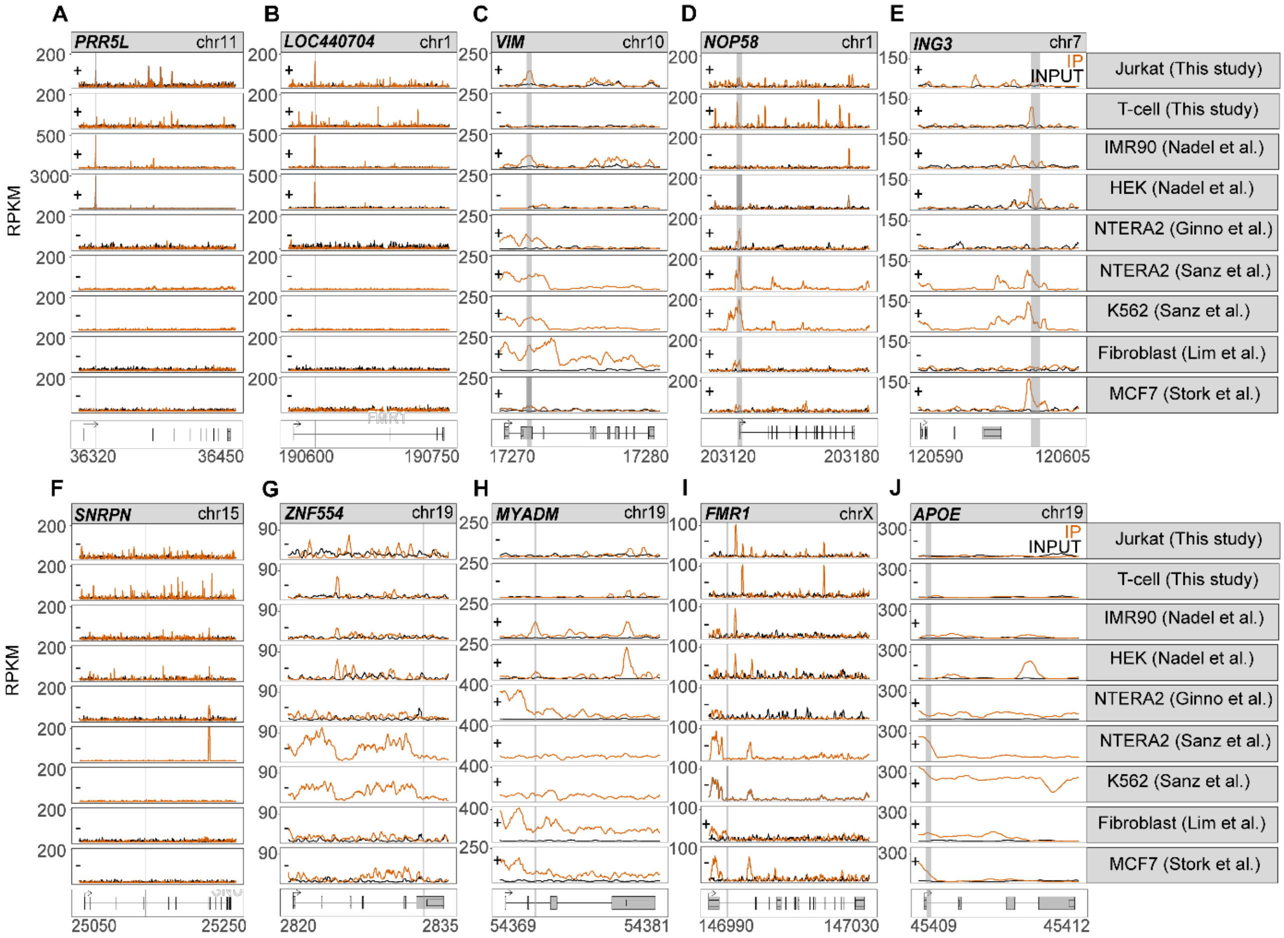


Figure 15. Making a Reference R-loop Set by DRIP-Seq Mapping. DNA-RNA hybrid mapping (DRIP-seq) was performed in two closely related human cell types (Jurkat T-cell leukaemia cell line and T CD4+ lymphocytes). Genome browser tracks show the IP (orange) and input (gray) signals over the selected test regions. (A-J) Test loci used for benchmarking the DRIP classifiers. DRIP profiles over the same test regions, obtained in other cell types, are also displayed. Test regions that are positive or negative for the presence of an R-loop (gray shading) are indicated by + and -, respectively. RPKM: reads per kilobase per million reads. Locus names and chromosome numbers are indicated on the top of each panel. Vertical light-gray boxes highlight the regions tested by DRIP-qPCR.

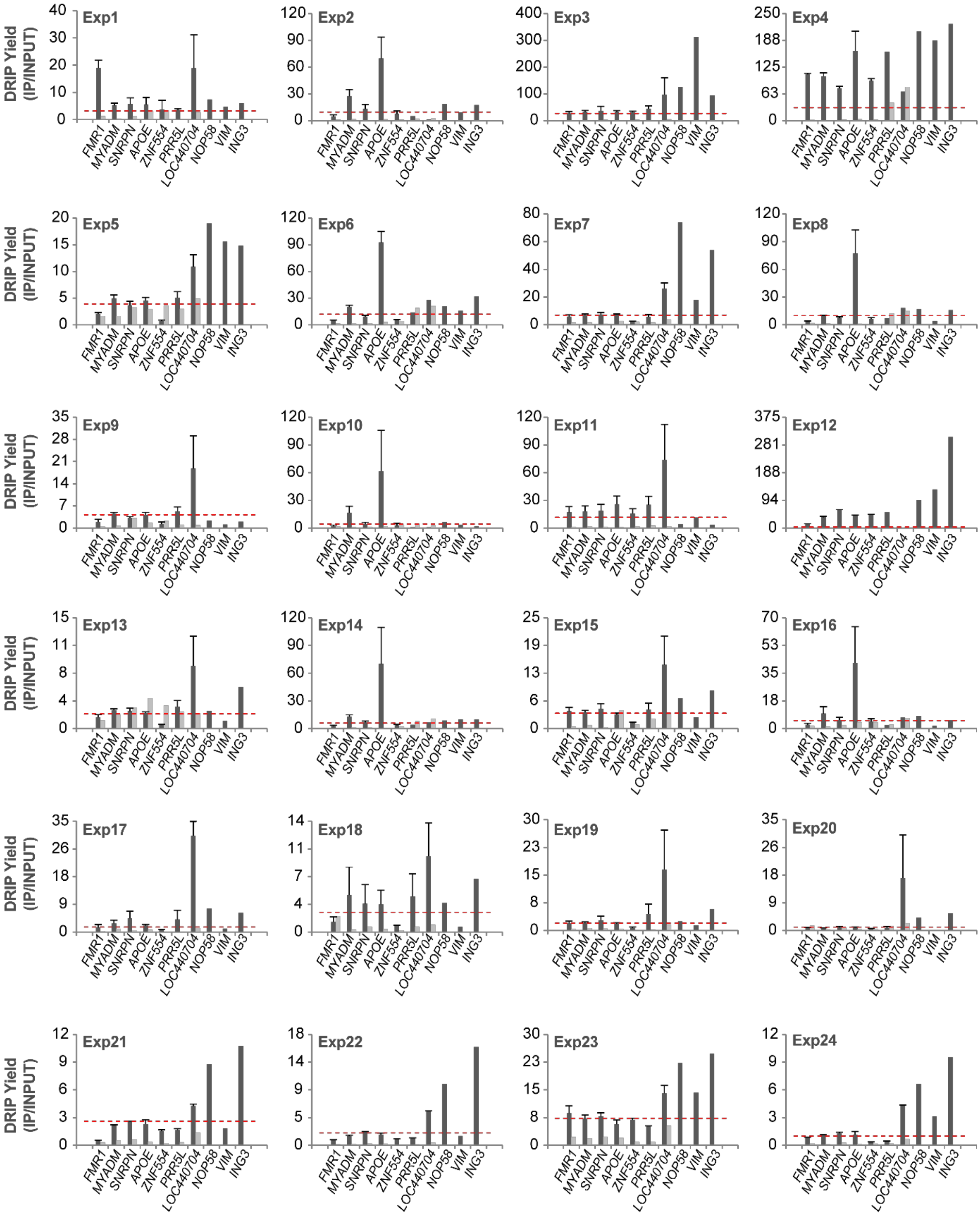


Figure 16. DRIP Enrichment Scores Determined by qPCR Over the Test Loci. DRIP-qPCR yield is shown for the twenty-four DRIP experiments over the selected reference loci. Black and grey bars represent DRIP yields from control and RNase H treated nucleic acid samples, respectively. The first five loci are negative controls (based on the lack of DRIP-seq enrichment), while the remaining five loci are positive controls (showing significant R-loop enrichment by DRIP-Seq). Horizontal dotted line represents the cut-off separating the real R-loop signal from background (extracted from the ROC curves). Optimal separation of negative and positive test loci is obtained in exp5, exp13, exp17 and exp18. We highlight these methods as “preferred”. On the contrary, positive and negative test loci are not properly distinguished by exp2, exp10, exp11 and exp19. We highlight these methods as “not preferred”.

5.1.4. Determining the sensitivity and specificity of RNA-DNA hybrid detection

We quantified the relative trade-offs between true positive hits and experimental errors (false R-loop associations) by performing ROC analysis (Robin et al. 2011) on the DRIP-qPCR screen characterizing the classifiers (**Figure 15-20**). The sensitivity, specificity and the area under the curve (AUC) values were extracted from the ROC plots and used as an objective measure of the robustness of the forty experiments. High (>0.7) AUC values were obtained for ten DRIP classifiers (exp. 5, 6, 13, 15, 17, 18, 19, 21, and 24), implying that those experiments could predict the presence or absence of an RNA-DNA hybrid with high efficacy (**Figure 17A**). AUC values close to 0.5 were obtained in four experiments (exp. 2, 10, 11, and 16), implying that the classifiers gave random answers without any predictive power as to the presence of an R-loop. Based on these considerations, the top four DRIP classifiers were: exp. 5, 13, 17, and 19 (**Figure 17B-C**) with a sensitivity of 68.5-75% and specificity of 68-79%. Similar (or even higher) ROC parameters were obtained in a repeated experiment using a B lymphoblastoid cell line (**Figure 19**), demonstrating the reliability of the tested DRIP protocols in other cell types.

Pairwise comparison of the main experimental variables (**Figure 17D**) revealed no significant difference between *i.* formaldehyde-fixed *vs.* unfixed samples, *ii.* phenol-chloroform extracted *vs*. silica membrane purified nucleic acid samples, and *iii.* DNA-fragmented (exp. 1-16) *vs*. chromatin-fragmented DRIP samples (exp. 17-24). Cell lysis temperature (65°C *vs.* 37°C) did not change the specificity and sensitivity of the DRIP assay (**Figure 19-20**). Statistically significant difference was obtained for RNase A-treated *vs.* untreated samples (p=0.03), suggesting that addition of RNase A does not improve the efficacy of RNA-DNA hybrid detection (Step 3, **Figure 17D**). We explain the adverse effect of RNase A by its reported DNA binding activity (Benore-Parsons and Ayoub 1997; Dona and Houseley 2014) that selectively eliminates a vast amount (micrograms) of melted DNA regions upon nucleic acid purification (Dona and Houseley 2014). We confirmed the strong DNA binding of RNase A as migration defects on DNA gels, when a plasmid DNA was incubated with the enzyme (**Figure 22**). The observed electrophoretic mobility shift was prevalent on supercoiled, nicked-circular and linearized DNA templates.

Finally, by comparing sonicated and restriction enzyme fragmented DRIP samples (Step 4, **Figure 17D**) we found a statistically significant difference (p=0.0002) in the ROC parameters, suggesting that sonication is more efficient in discriminating true positive signals from false positives, at least within the tested conditions.

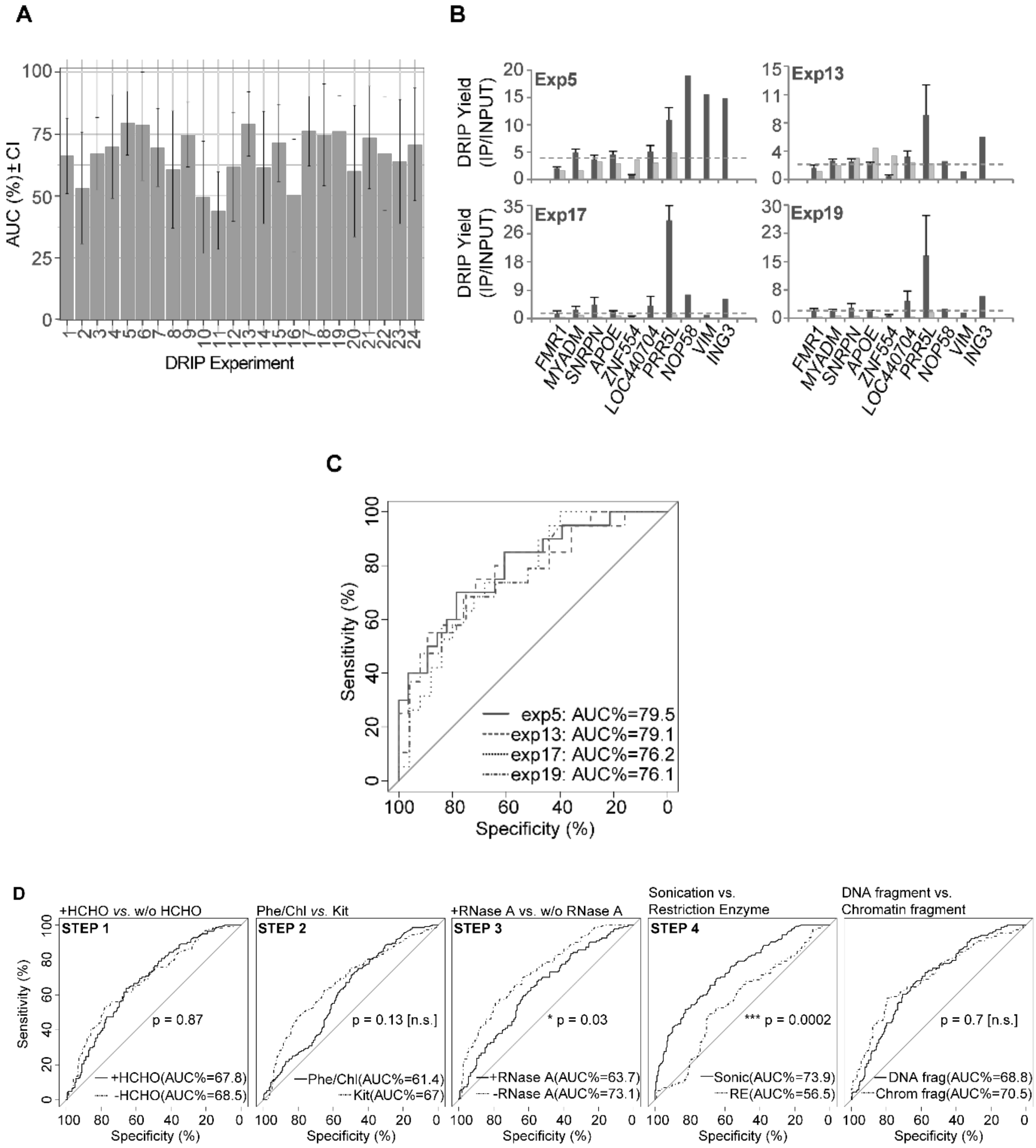


Figure 17. Good DRIP Practice. (A) Bar charts showing the distribution of AUC (area under the curve) values of ROC plots for twenty-four DRIP classifiers. Error bars represent the confidence interval of AUCs. High (>0.7) AUC values were obtained for ten DRIP classifiers (exp. 5, 6, 13, 15, 17, 18, 19, 21, and 24). Low (~0.5) AUC values were obtained in four DRIP experiments (exp. 2, 10, 11, and 16). We highlight these groups as "preferred" and "not preferred", respectively. (B-C) The top four DRIP experiments ranked by AUCs (exp 5, 13, 17, 19). (B) DRIP-qPCR enrichment scores are displayed over the test regions. Horizontal dotted lines represent the cut-off value (calculated from the ROC curves) separating the true R-loop signal from background. (C) ROC curves of the top four experiments. (D) Paired-ROC plots, comparing the main variables (steps) of the DRIP experiments. The level of statistical significance was 0.05.

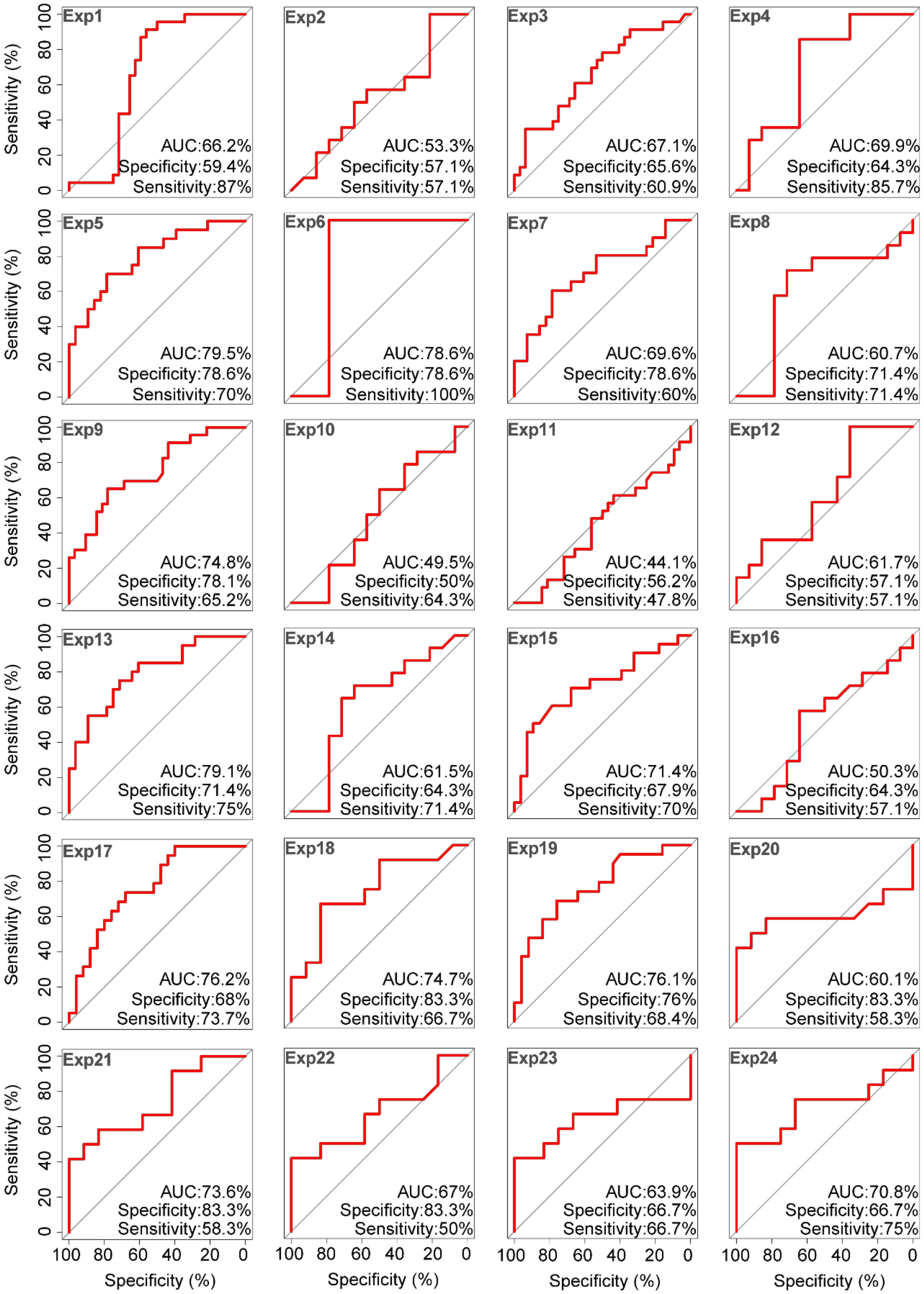


Figure 18. ROC Analysis of the DRIP Classifiers. ROC plots illustrating the efficacy of the twenty-four DRIP protocols. Area under the curve (AUC), specificity and sensitivity are labelled within each plot. The diagonal indicates an AUC of 0.5, corresponding to random answers obtained from the experiments.

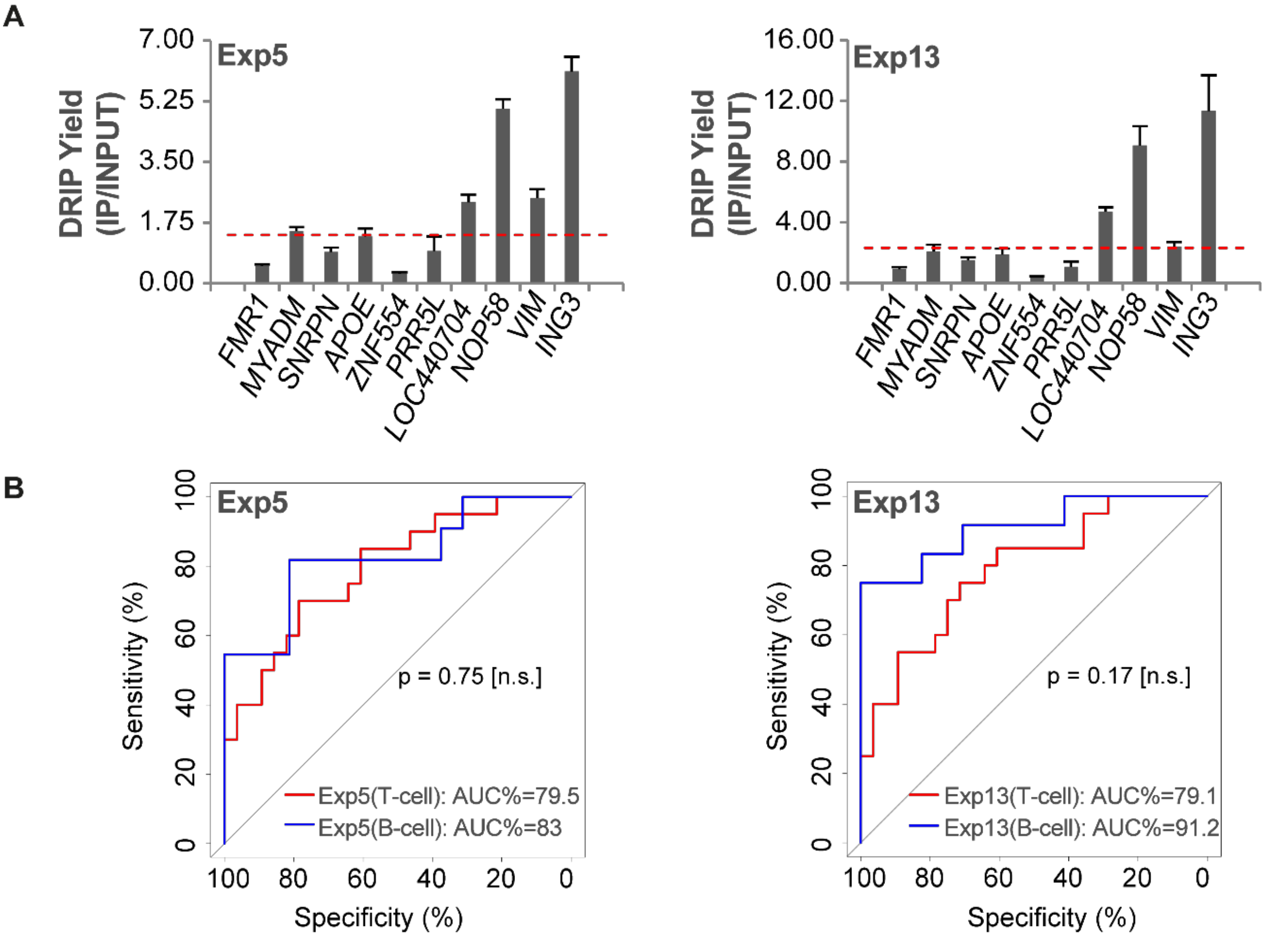


Figure 19. The best-performing DRIP protocols work equally well in other cell types. T and B lymphoblastoid cell lines (Jurkat and GM12878) were compared in exp5 and exp13, respectively. (A) DRIP-qPCR enrichment scores of the GM12878 cell line displayed over the ten test regions. Horizontal dotted line represents the cut-off value (calculated from the ROC curves) separating the true R-loop signal from background. (B) Paired ROC plots comparing the efficacy of the DRIP protocols in Jurkat (red line) and GM12878 (blue line) cells. No significant difference was observed between the cell lines.

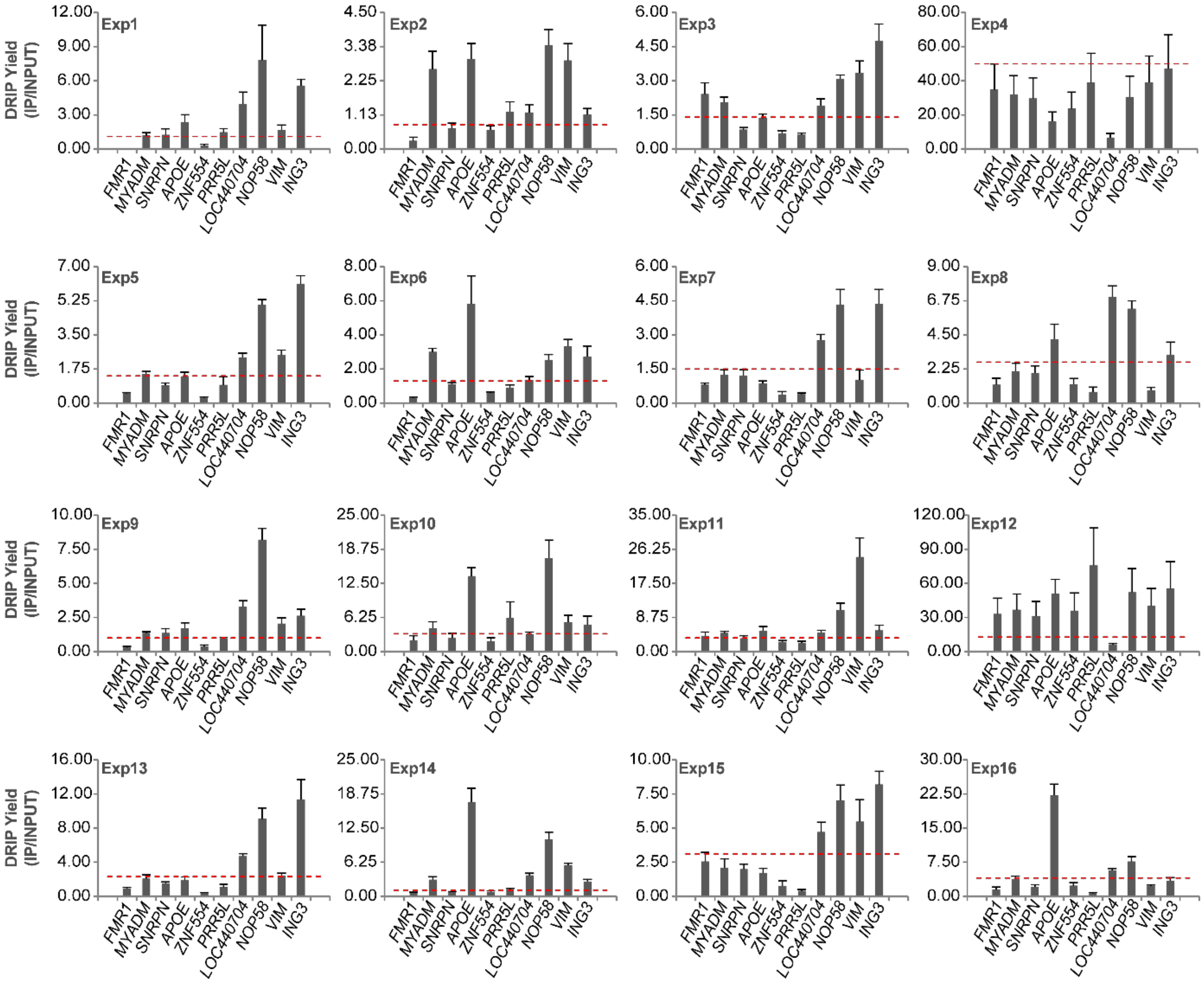


Figure 20. The effect of cell lysis performed at 37 ºC. DRIP yields were measured by qPCR in sixteen DRIP experiments (with the cell lysis step performed at 37 ºC) over the selected reference loci. Horizontal dotted line represents the cut-off separating the real RNA-DNA hybrid signal from background.

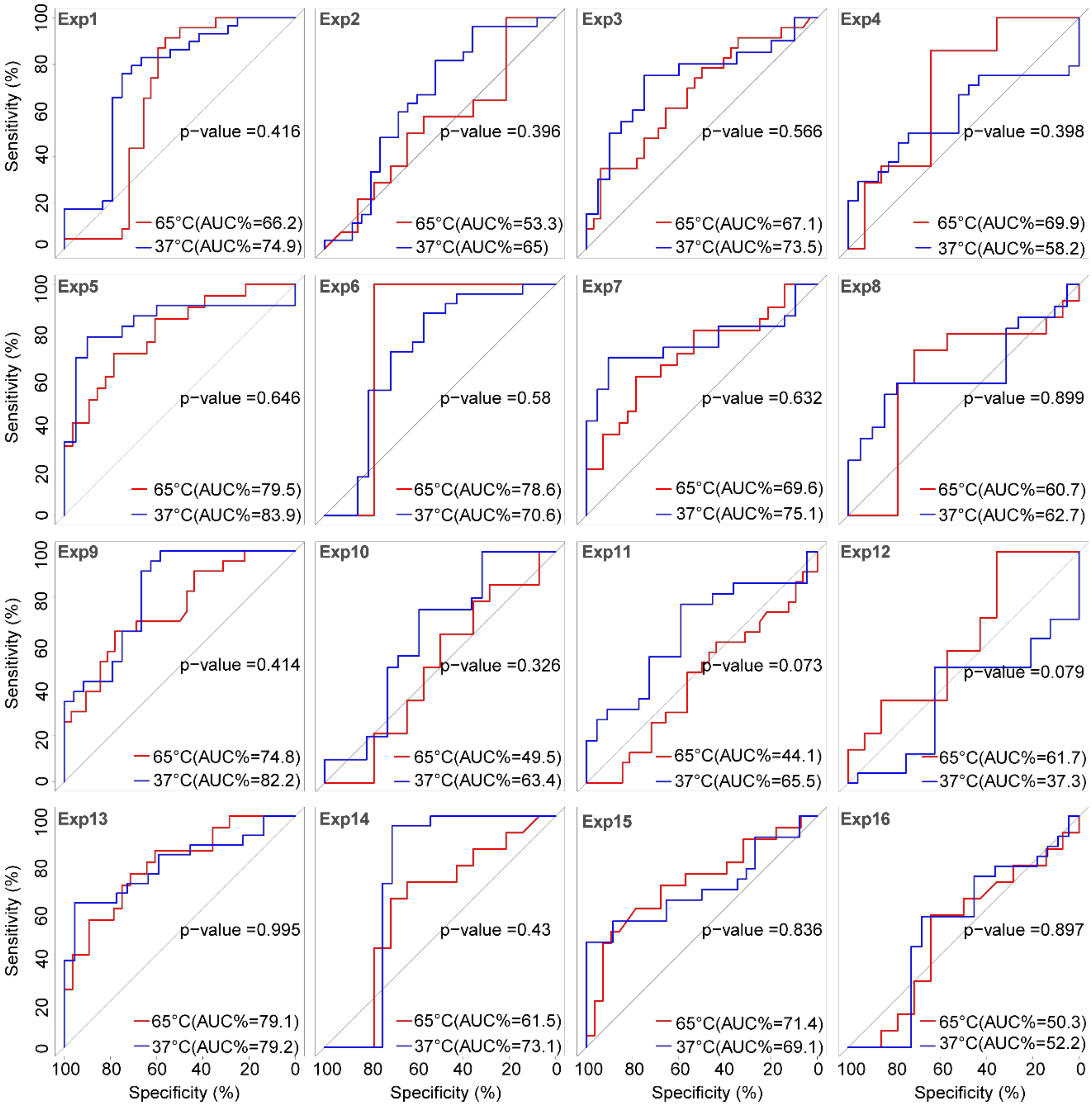


Figure 21. The effect of cell lysis at 65°C and 37°C on the specificity and sensitivity of the DRIP experiment. Paired ROC plots compare the efficacy of sixteen DRIP protocols performed at 65°C and 37°C, respectively. None of the tested conditions resulted in a significant difference between the two temperatures. Area under the curve (AUC), specificity and sensitivity are labelled on each plot.

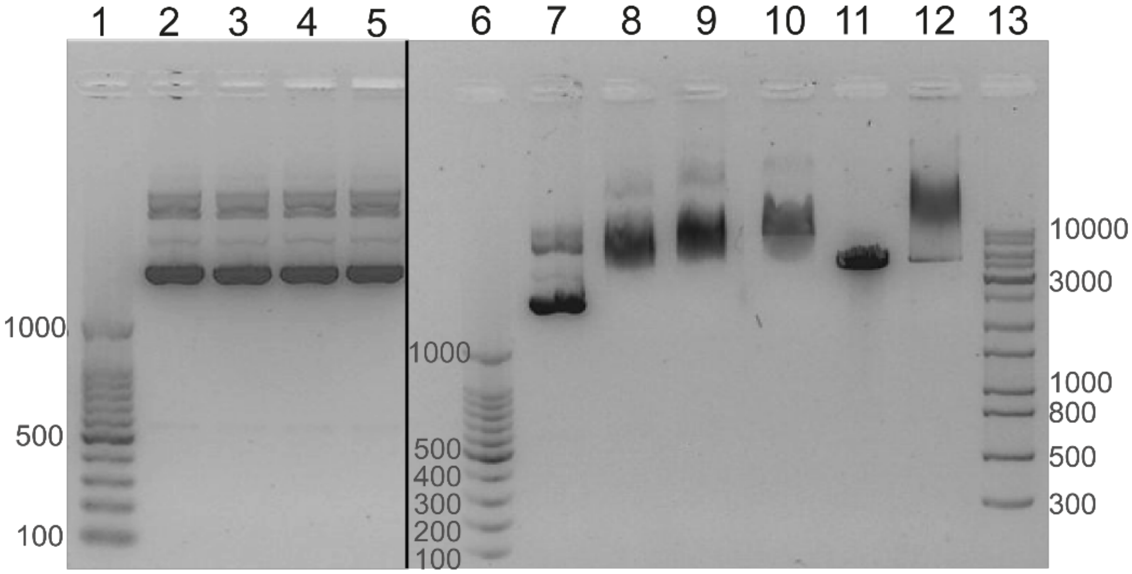


Figure 22. Evidence for the DNA binding of RNase A. A plasmid DNA incubated with DRIP samples (lanes 2-5) do not show any change in its electrophoretic mobility. Incubation of a plasmid DNA with RNase A (lanes 7-12) significantly changes the electrophoretic mobility via the DNA binding activity of the ribonuclease. The band shift occurs on supercoiled, nicked circular and linearized plasmid templates.

Lanes of the gel:

1. 100 bp marker
2. pmCherry-N1 plasmid (supercoiled form) incubated with DRIP IP sample 5 (65 ºC)
3. pmCherry-N1 plasmid (supercoiled form) incubated with DRIP IP sample 5 (37 ºC)
4. pmCherry-N1 plasmid (supercoiled form) incubated with DRIP IP sample 7 (65 ºC)
5. pmCherry-N1 plasmid (supercoiled form) incubated with DRIP IP sample 7 (37 ºC)
6. 100 bp marker
7. pmCherry-N1 plasmid (supercoiled form) in 300 mM NaCl/10 mM Tris-Cl pH 7.5
8. pmCherry-N1 plasmid (supercoiled form) + 2 µl RNase A (10 mg/ml, UDG) in 300 mM NaCl/10 mM Tris-Cl pH 7.5
9. pmCherry-N1 plasmid (supercoiled form) + 2 µl RNase A (10 mg/ml, NEB) in 300 mM NaCl/10 mM Tris-Cl pH 7.5
10. pmCherry-N1 plasmid (nicked circular form) + 2 µl RNase A (10 mg/ml, UDG) in 300 mM NaCl/10 mM Tris-Cl pH 7.5. Nicking was achieved by 30 min UV treatment.
11. pmCherry-N1 plasmid (linear) in 300 mM NaCl/10 mM Tris-Cl pH 7.5. Linearization was achieved by BamHI digestion of the plasmid. The digested plasmid was PCR clean up purified.
12. pmCherry-N1 plasmid (linear) + 2 µl RNase A (10 mg/ml, UDG) in 300 mM NaCl/10 mM Tris-Cl pH 7.5.
13. 1 kb marker

5.1.5. Impact on the annotation and basic biological function of R-loops

Suboptimal DRIP conditions might prevent the assignment of precise biological function to a significant fraction of R-loops. Although the average DNA fragment size resulting from restriction enzyme digestion fits the requirements of the DRIP assay, we found that the frequency of cutting sites was significantly higher within intergenic regions, producing lengthy restriction fragments over protein coding ORFs (**Figure 23**). Biased genome sampling, related to the non-random distribution of restriction enzyme recognition sequences, was even more pronounced over exons (**Figure 23C**), especially over the first exons (**Figure 23D**). In 82% of first exons there were only 0-1 suitable restriction sites compared to intergenic regions (59%). We estimated the digestion efficiency of restriction enzyme cutting sites to ~50% over intergenic regions (based on the proportion of zero reads over restriction enzyme cutting sequences, representing cleaved sites), which was significantly reduced over gene coding regions (**Figure 23E-F**). Consequently, genic regions void of suitable restriction sites appear as long DRIP fragments that potentially compromise mapping resolution. The *MYC*, *BCL6*, and *VIM* genes are shown as representative examples for large, restriction fragment-sized DRIP peaks (**Figure 24**). Precise genomic position of R-loops could be resolved by sonication.

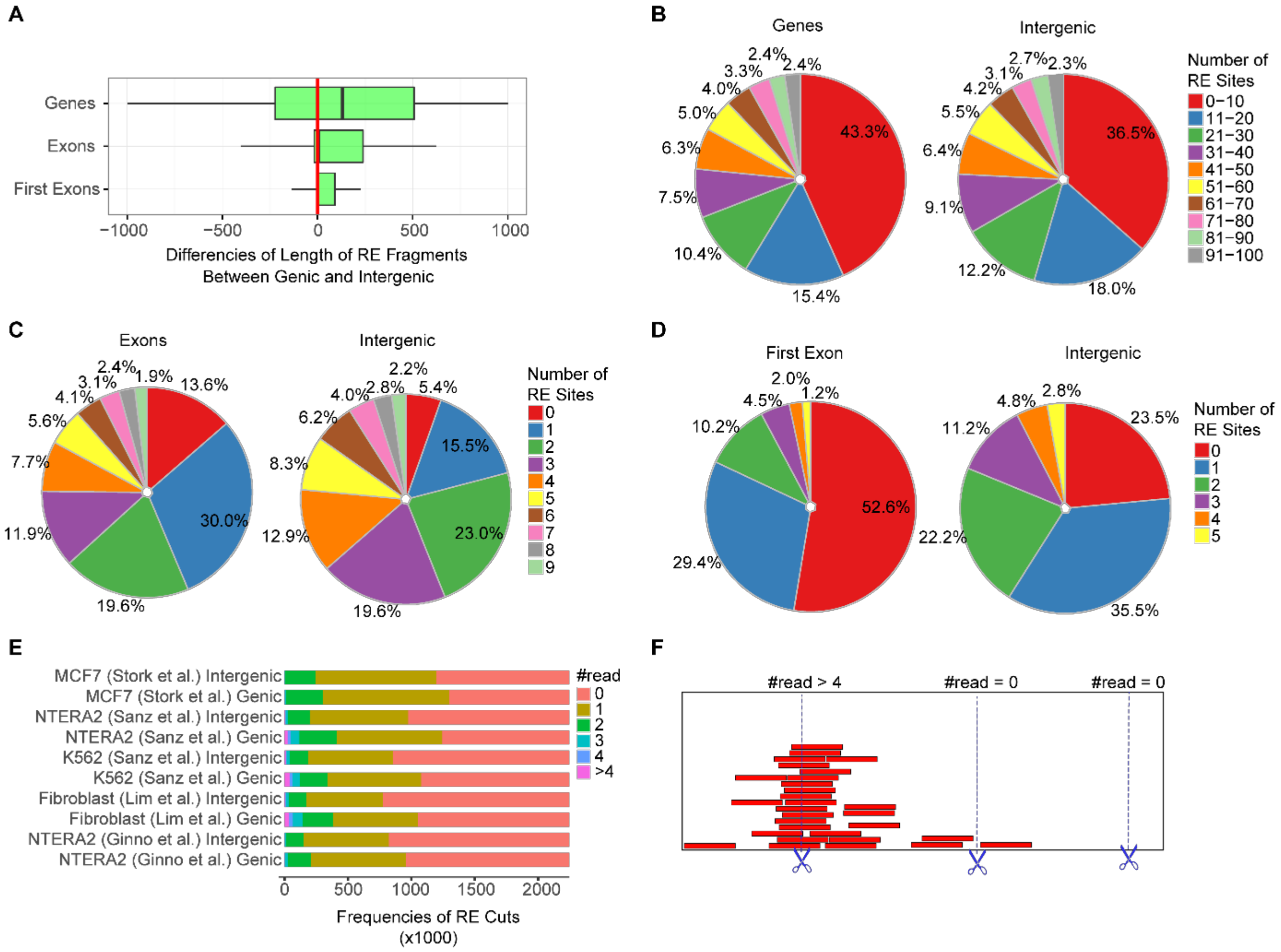


Figure 23. Analysis of restriction sites over genic and intergenic regions. (A) Restriction fragment lengths over genic regions (gene bodies, exons, first exons) are significantly larger compared to intergenic regions. The plot shows the difference of genic (observed) and intergenic (expected) fragment sizes in base pairs. The following enzymes were applied in combination: HindIII, EcoRI, BsrGI, XbaI and SspI. (B-C-D) The number of restriction sites over genic regions is significantly lower compared to intergenic regions. Colours indicate the proportion of cutting sites in each category. Red and blue slices, marking the rarest restriction site frequencies, are prevalent over genic elements in each pie chart. (E) Cutting efficiency of restriction enzymes applied in the indicated DRIP-seq experiments. Zero read: the restriction site was cut. Greater equal than one read: the restriction site was uncut in a fraction of cells. There were uncut reads (sites) over half of the theoretical restriction sites. The proportion of uncut reads was even higher within gene coding regions compared to intergenic regions. See the model of cutting efficiency in panel (F).

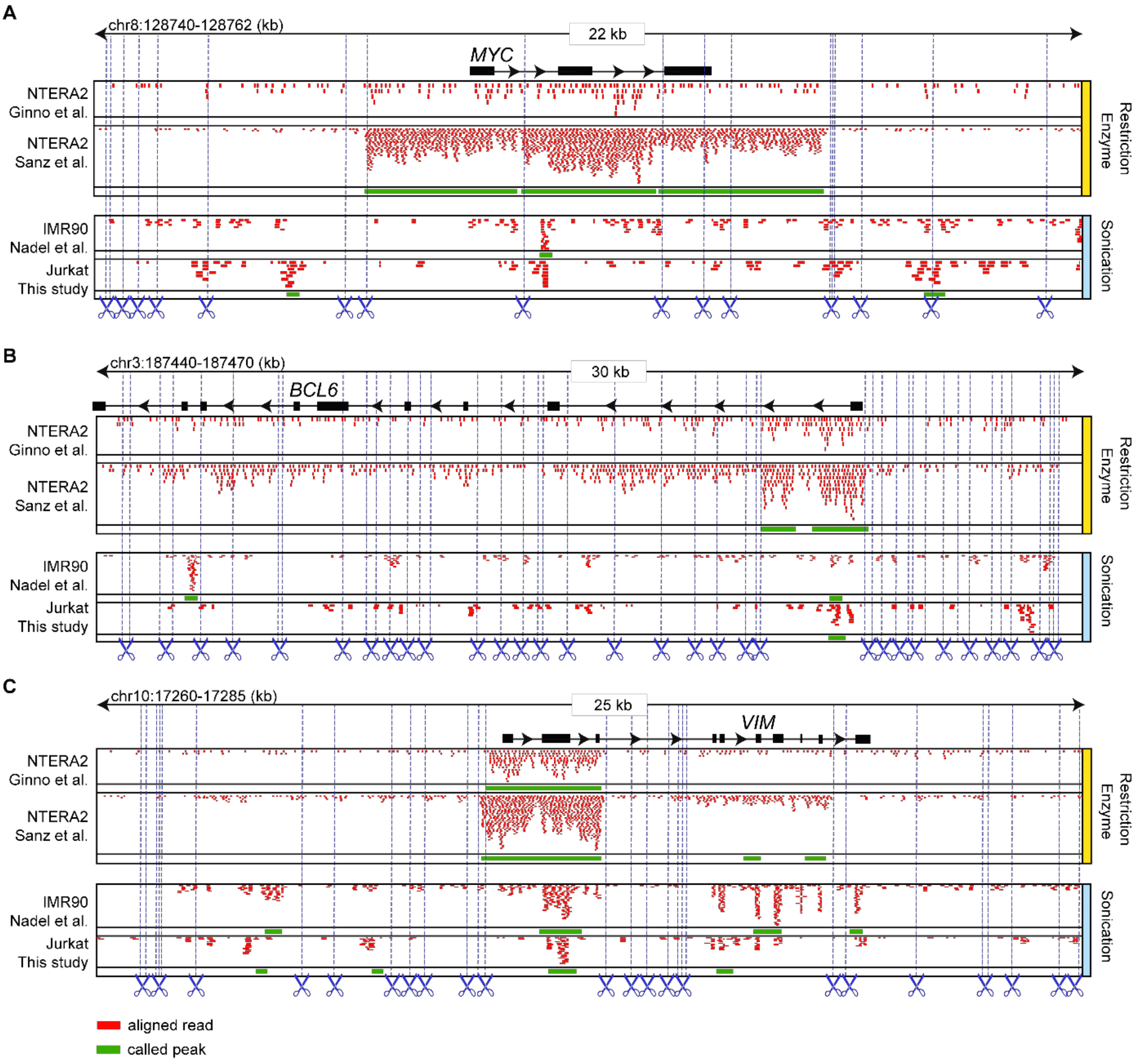


Figure 24. Large restriction fragments over gene bodies causes uncertainty in the precise localization of R-loops, potentially impeding their functional annotation. (A-C) Genome browser tracks showing three representative examples (MYC, BCL6 and VIM). Upper two tracks: restriction fragment-sized R-loops are prevalent over the 5' prime end of genes, vastly exceeding the gene borders in the case of MYC. Lower two tracks: the precise genomic position of R-loops was resolved in the sonicated group of samples. Green boxes represent R-loop enriched regions predicted by the peak callers. Blue dashed lines represent cutting sites for restriction enzymes (HindIII, EcoRI, BsrGI, XbaI and SspI).

5.2. Nuclear dynamics of the Set1C subunit Spp1 prepares meiotic recombination sites for break formation

In the second part of the dissertation I have described how the Set1C subunit Spp1 recruits H3K4me3 enriched DNA segments to the chromosomal axis where DNA double strand breaks are generated and subsequently repaired during meiosis. Specifically, we studied the nuclear dynamics of the Spp1 protein using time-resolving ChIP-seq experiments during meiosis. In addition, we generated several mutant yeast strains to decipher the molecular mechanisms driving this process.

5.2.1. Spp1 exhibits static and dynamic chromosome binding kinetics during meiosis

To gain insights about the chromatin dynamics of Spp1 during the progression of meiotic prophase, we mapped the chromosomal binding sites of epitope-tagged Spp1 and Bre2 by ChIP sequencing in synchronously sporulating yeast cultures. The distribution of Bre2 was used as a proxy to mark the chromosomal position of Set1C. Peak sets identified at individual meiotic time points (SPS, 0-2-4-6 hrs in SPM) were concatenated and sorted by chromosomal position, and then merged to create a consensus binding site set. Venn diagram analysis of chromatin binding sites shows that ~46% of the Spp1 peaks coincide with Bre2 (**Figure 25A**), indicating a group of Spp1 molecules associated with Set1C during meiosis.

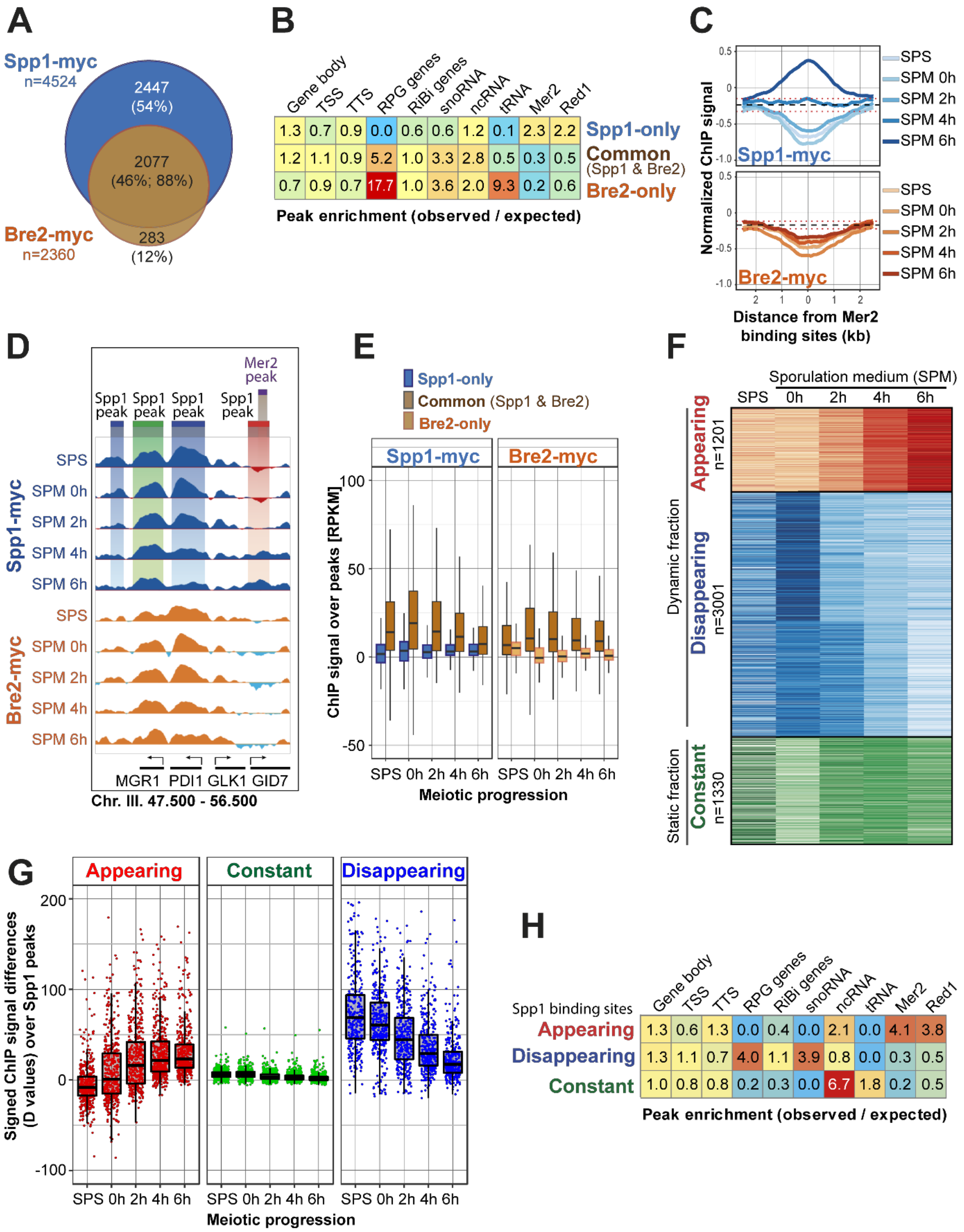
Overall, Spp1 & Bre2 (common) peaks and Bre2-only peaks show strong enrichment on ribosomal protein genes (RPGs), snoRNA/ncRNA genes and transcription start sites (TSS), but they are absent from Mer2/Red1 axial sites (**Figure 25B**). In contrast, Spp1-only peaks are significantly overrepresented at Mer2/Red1 sites. Strikingly, Bre2-only peaks are highly enriched at RPG and tRNA genes compared with common peaks of Spp1 and Bre2, indicating the presence of Spp1-free Set1C on these genes during meiosis.

Importantly, Spp1 showed a progressive loading onto Mer2 binding sites during meiotic prophase, while Bre2 remained depleted throughout the sporulation process (**Figure 25C**). Although Spp1 binding sites appear to be more dynamic than common (Spp1 & Bre2) sites (representative JBrowse example for dynamic Spp1 peaks is shown in **Figure 25D**), the latter peaks show much higher ChIP signal compared to Spp1-only or Bre2-only sites (ANOVA with TukeyHSD, p<0.0001, **Figure 25E**). We explain these differences with the differential turnover rate characteristics of the sites (Karányi et al. 2018).

To gain more mechanistic insights into the dynamics of Spp1, we performed unsupervised clustering analysis (k-means) on the time-resolved Spp1 ChIP signals, classifying the identified binding sites based on their similarity. Two kinetic groups were readily revealed based on the relative change of Spp1 peak signals over time (**Figure 25F**); dynamic sites, which gradually appeared (red) or disappeared (blue) as meiosis progressed, and static sites (green) showing permanent association with Spp1. These separate classes were reproduced by a clustering-independent approach that relied on the absolute change of Spp1 signal intensities in terms of time (**Figure 25G**).

Functional annotation revealed that i) appearing Spp1 peaks are strongly enriched at chromosome axial sites (Red1, Mer2) ii) disappearing Spp1 sites are enriched at RPG and snoRNA genes and iii) constant Spp1 peaks show strong association with ncRNAs (**Figure 25H**).

We conclude that the dynamic properties of Spp1 correlate with its non-canonical (Set1C independent) functions and the remodelling of Set1C at RPG and snoRNA genes during the meiotic process.



	Gene body	TSS	TTS	RPG genes	RiBi genes	snoRNA	ncRNA	tRNA	Mer2	Red1
Spp1-only	1.3	0.7	0.9	0.0	0.6	0.6	1.2	0.1	2.3	2.2
Common (Spp1 & Bre2)	1.2	1.1	0.9	5.2	1.0	3.3	2.8	0.5	0.3	0.5
Bre2-only	0.7	0.9	0.7	17.7	1.0	3.6	2.0	9.3	0.2	0.6

Peak enrichment (observed / expected)

Spp1 binding sites	Gene body	TSS	TTS	RPG genes	RiBi genes	snoRNA	ncRNA	tRNA	Mer2	Red1
Appearing	1.3	0.6	1.3	0.0	0.4	0.0	2.1	0.0	4.1	3.8
Disappearing	1.3	1.1	0.7	4.0	1.1	3.9	0.8	0.0	0.3	0.5
Constant	1.0	0.8	0.8	0.2	0.3	0.0	6.7	1.8	0.2	0.5

Peak enrichment (observed / expected)

Figure 25. Chromosomal distribution of Spp1 and Bre2 binding sites during meiotic prophase. (A) Proportional Venn diagram showing the overlap of Spp1-myc and Bre2-myc binding sites identified in a meiotic time course (SPS; 0, 2, 4, and 6 h in SPM) by ChIP-seq. 54% of Spp1 peaks show no overlap with Bre2, while 88% of Bre2 peaks coincide with Spp1 binding sites. (B) Functional annotation of Spp1 and Bre2 sites show differential enrichment over several genomic regions. Spp1-only peaks are highly enriched at chromosome axial sites (Mer2, Red1); common peaks (Spp1 and Bre2) are associated with RPG, snoRNA, and ncRNA genes and depleted over Mer2/Red1 sites; and Bre2-only peaks are enriched at RPG, snoRNA, ncRNA, and tRNA genes and depleted over Mer2/Red1 sites. Heat map shows the overlap ratio of observed and computer randomized binding sites (observed/expected) with the indicated annotation category. (C) Spp1-myc is progressively loaded

to Mer2 binding sites during meiotic prophase, while the Bre2-myc signal remains depleted throughout the sporulation process. Horizontal dashed line and red dotted lines show the genome-wide average ChIP signal ± SD. (D) Representative genome browser snapshot showing the chromosomal distribution of Spp1-myc (blue) and Bre2-myc (orange) ChIP signal. Tracks represent meiotic time points. Disappearing, constant, and appearing Spp1 peaks are highlighted in blue, green and red, respectively. A Mer2 site is also shown in purple. (E) Common (Spp1 & Bre2) binding sites show increased chromatin association compared with Spp1-only and Bre2-only sites (ANOVA with Tukey HSD; $P < 0.0001$). Box plots show the distribution of ChIP signals over the three categories (Spp1-only, common, and Bre2-only). Left, Spp1-myc enrichment. Right, Bre2-myc enrichment. (F) Temporal classes of Spp1 binding sites identified by cluster analysis. Appearing (red) and disappearing (blue) sites show dynamically increasing/decreasing ChIP enrichment, while constant sites (green) do not show significant temporal changes. Heat maps show the relative changes of ChIP enrichment over time (normalized by rows). (G) Confirming the kinetic classes of Spp1 binding sites by an independent approach, based on the absolute values of ChIP enrichments. Spp1 peaks were rank-ordered by their signed ChIP signal differences (D values) between 0 and 6 h in SPM. Sampling the bottom (<q20), middle (q40–q60) and top (>q80) quantiles of the D values recapitulated the dynamic classes of Spp1 sites visualized by cluster analysis (in panel F). (H) Functional annotation of the dynamic classes of Spp1 binding sites. Appearing Spp1 peaks are strongly enriched at chromosome axial sites (Red1 and Mer2). Disappearing Spp1 sites are enriched at RPG and snoRNA genes and depleted at Mer2/Red1 sites. Constant Spp1 peaks show strong association with ncRNAs and depletion over Mer2 binding sites. The data are representative of two independent biological replicate experiments. Sample size (n, number of peaks analyzed in each category) is indicated in panels A and F.

5.2.2. Functional analysis of Spp1 chromatin binding during meiosis

To further shed light on the molecular determinants of Spp1 chromatin binding, we also examined the binding sites of Spp1PHDΔ and Spp1CxxCΔ mutants and that of H3R2A and H3K4R mutants. Mutation of lysine 4 prevents H3K4 methylation while substitution of arginine 2 by alanine inhibits the deposition of H3K4me3 (Kirmizis et al. 2007; Yuan et al. 2012). Both modifications are expected to phenocopy the meiotic phenotype of the Spp1PHDΔ mutation (**Figure 26**).

We performed time-resolved meiotic ChIP-seq and mapped the binding of Spp1PHDΔ, Spp1CxxCΔ, and Spp1 in H3R2A/H3K4R mutants. As shown in Venn diagrams (**Figure 26B**), all four mutations eliminate about 50% of Spp1 binding sites during the meiotic time-course identified in the wild type strain. Interestingly, some new Spp1 sites (~10%) are also generated in each mutant (**Figure 26B**).

Moreover, multidimensional scaling (MDS) analysis on the identified binding sites has been performed to highlight temporal and cell type-specific differences in Spp1 chromosomal localization (**Figure 26C**). For all cell types and meiotic timepoints, exact chromosomal position and enrichment of all the identified Spp1 ChIP peaks were assigned to Ndimensional coordinates, defining Spp1 'states' by cell type and meiotic stage. All Spp1 states were then projected to a 2D plane (highlighted as dots in the MDS maps, **Figure 26C**) such that the closer is the difference between any two data points the more similar the Spp1 states are. As shown in the upper panel of **Figure 26C**, wild type cells and Spp1 PHD- and CxxC-domain mutants behave quite differently at the beginning of sporulation. Then, in the first two hours, there is a large, rapid and identical change in both wild type and mutant cells. By the end of the process, each cell type converges to a similar Spp1 state, which is shown by the small distance of dots at the 6-hour timepoint.

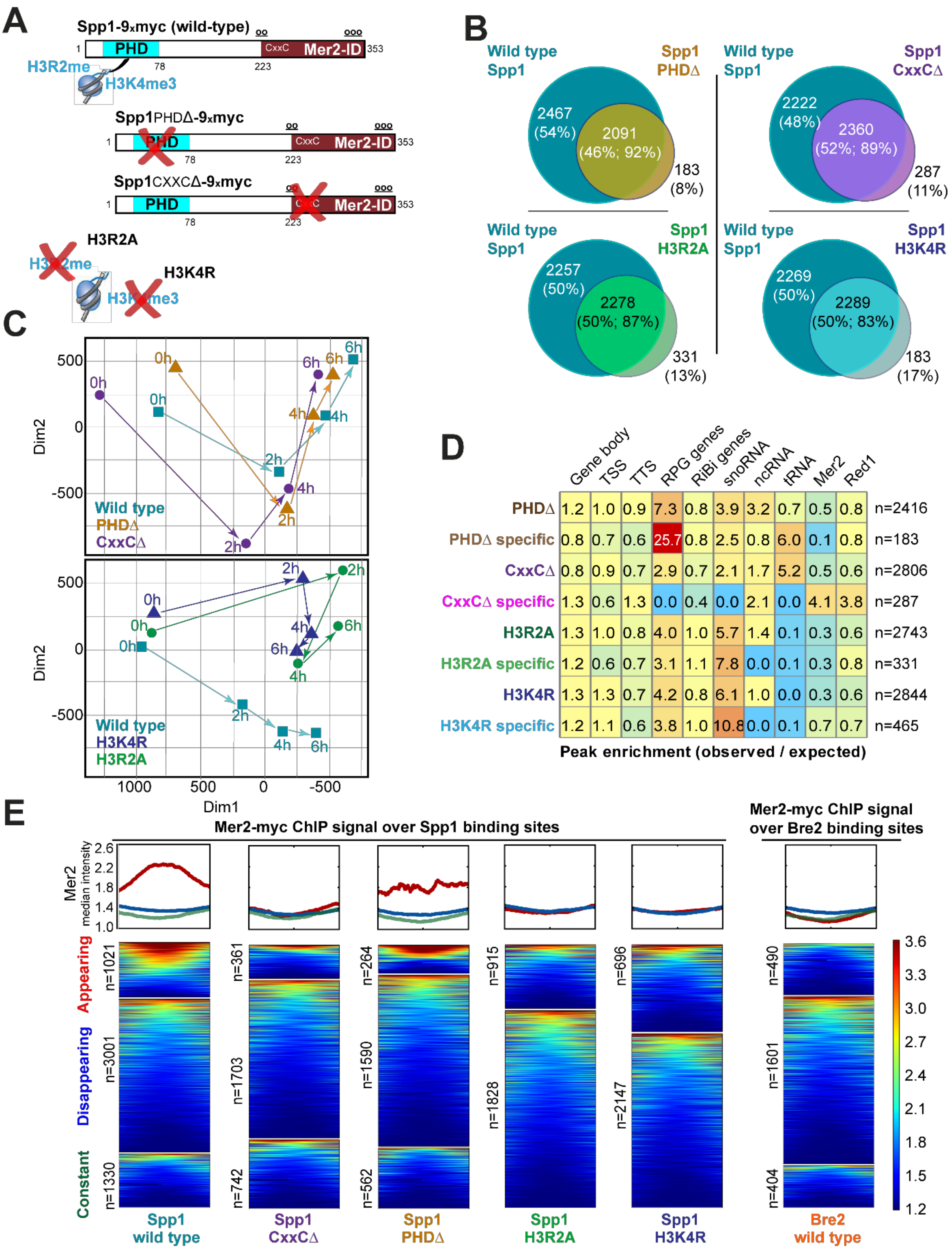
In the histone mutant backgrounds (lower panel in **Figure 26C**), Spp1 binding sites are more similar to the wild type at the beginning of sporulation (0 h in SPM). Thereafter, fast and dynamic changes occur in the first few hours, and both mutants quickly move away from the wild type. By the end of the process all of the three cell types can be characterized by a different Spp1 state.

We next analysed the overlap of Spp1 binding sites with annotated functional genomic elements in each mutant (Spp1PHDΔ, Spp1CxxCΔ, H3R2A, and H3K4R). As shown in **Figure 26D**, the resulting peaks are differentially enriched over several genomic elements and show variable overlap with each other.

Importantly, all mutations reduce the binding of Spp1 to axis sites (**Figure 26D**) and abrogate the association of Mer2 with the dynamic clusters of Spp1 peaks (**Figure 26E**). The PHDΔ mutant shows a very high enrichment of Spp1 at RPG genes, which highlights the role of the PHD domain in the removal of Spp1 from RPG genes (**Figure 26D**). Similarly, H3R2A and H3K4R mutants exhibit specific Spp1 enrichment at snoRNA genes, indicating that H3R2 and H3K4 methylation promotes the disappearance of Spp1 from snoRNAs.

The heatmaps shown in **Figure 26E** reveal that enrichment of Mer2 at appearing Spp1 peaks is abolished in the Spp1CxxCΔ, H3R2A, and H3K4R mutants. Deleting the PHD finger domain of Spp1 eliminates approximately 75% of the appearing Spp1 peaks (264/1021) detected in wild type cells, however, about half of the remaining Spp1PHDΔ sites (130 peaks) still exhibit significant Mer2 enrichment. These results are in contrast with the Spp1CxxCΔ binding sites and the effects of H3R2A/K4R mutations that apparently prevent Mer2 enrichment. For comparison, we also analysed Mer2's association with the dynamic clusters of Bre2 binding sites defined by cluster analysis (similarly to Spp1 sites). Remarkably low Mer2 signal was detected over the appearing Bre2 binding sites (**Figure 26E**, **right panel**).

Taken together, these results further strengthen the tethered loop axis model of meiotic DSB formation, proposing that proper localization of Spp1 to chromosome axial sites requires *i)* the Mer2-binding (CxxC) motif of Spp1, *ii)* to a lesser extent the PHD finger domain, and *iii)* the presence of histone modifications and modifiable residues (H3K4me3, H3R2me2s).



	Gene body	TSS	TTS	RPG genes	RiBi genes	snoRNA	ncRNA	tRNA	Mer2	Red1	
PHDΔ	1.2	1.0	0.9	7.3	0.8	3.9	3.2	0.7	0.5	0.8	n=2416
PHDΔ specific	0.8	0.7	0.6	25.7	0.8	2.5	0.8	6.0	0.1	0.8	n=183
CxxCΔ	0.8	0.9	0.7	2.9	0.7	2.1	1.7	5.2	0.5	0.6	n=2806
CxxCΔ specific	1.3	0.6	1.3	0.0	0.4	0.0	2.1	0.0	4.1	3.8	n=287
H3R2A	1.3	1.0	0.8	4.0	1.0	5.7	1.4	0.1	0.3	0.6	n=2743
H3R2A specific	1.2	0.6	0.7	3.1	1.1	7.8	0.0	0.1	0.3	0.8	n=331
H3K4R	1.3	1.3	0.7	4.2	0.8	6.1	1.0	0.0	0.3	0.6	n=2844
H3K4R specific	1.2	1.1	0.6	3.8	1.0	10.8	0.0	0.1	0.7	0.7	n=465

Peak enrichment (observed / expected)

Figure 26. Functional analysis of Spp1 chromosomal binding in meiosis. (A) Schematic structure of the Spp1 mutant proteins studied in ChIP-seq experiments. C-terminal tags are not illustrated. Mutations (highlighted by red Xs) were introduced into the PHD finger domain (blue) and Mer2-interacting motif (brown) of Spp1 and into histone H3R2 and H3K4 (H3R2A and H3K4R). (B) Proportional Venn diagram showing the overlap of Spp1 binding sites identified in wild type and mutant cells during the meiotic time course (0–6 h in SPM). Reduction of Spp1 binding sites for each mutation is indicated in the diagrams. About 90% of Spp1 peaks observed in the mutants overlapped with wild type Spp1 sites. About 10% of Spp1 peaks formed *de novo* in the mutants. (C) MDS plots visualizing the similarities and differences of Spp1 binding sites identified in wild type and mutant

cells during the meiotic time course (0–6 h in SPM). Each data point represents a characteristic Spp1 state specified by cell type and temporal stage in meiosis. Distance of any two data points in the MDS map is proportional to the variability of Spp1 states (i.e., Spp1 peak sets). The upper map compares wild type, Spp1PHDΔ, and Spp1CxxCΔ cells at four meiotic time points (0, 2, 4, and 6 h in SPM). The lower map depicts wild type, Spp1 H3R2A, and Spp1 H3K4R cells at the same time points. (D) Functional annotation of Spp1 binding sites identified in the mutants. Color scale indicates enrichment or depletion within the annotation category. (E) Analysis of Mer2 enrichment over the dynamic classes of Spp1 binding sites identified by cluster analysis. Left, Mer2-myc signal enrichment shown on metaplots, centered to Spp1 peak positions identified in wild type and mutant cells. In wild type cells, the appearing class of Spp1 binding sites show strong enrichment in Mer2. Dynamic Spp1 clusters are also revealed in the mutants by cluster analysis, however, none of these dynamic sites are associated with Mer2. Right, Mer2-myc signal enrichment over clustered Bre2 chromatin binding sites. The data are representative of two independent biological replicate experiments. Sample size (n, number of peaks analyzed in each category) is indicated in panels B, D, and E.

6. Discussion

6.1. RNA-DNA hybrid (R-loop) immunoprecipitation mapping: an analytical workflow to evaluate inherent biases

Considering the increasing attention of RNA-DNA hybrid structures in the physiology and pathology of chromosomes, here we present an analytical framework to estimate the inherent biases of existing DRIP protocols and to assess the power of the technology. The ROC parameters (AUC, sensitivity, specificity, threshold) served as an objective measure for the efficacy of predicting the presence or absence of RNA-DNA hybrids.

First, we measured the DRIP enrichment scores for experimental schemes across several genomic regions. This allowed us to rank DRIP workflows based on the ability to distinguish complex or weak DRIP-qPCR signals from background noise with high confidence. The top performing experiments were: 5, 13, 17 and 19. However, we found several conditions where some DRIP workflows performed unreliably and generated random answers: 2, 10, 11 and 16. Under our experimental conditions, we highlight these groups as “preferred” and “not preferred”.

By testing the main parameters of the DRIP experimental scheme - including formaldehyde fixation, cell lysis temperature, nucleic acid isolation, free RNA removal, and DNA fragmentation - we found that fragmenting the nucleic acid by sonication and omitting RNase A digestion could improve the precision and specificity of RNA-DNA hybrid detection.

Next, we showed that restriction enzyme fragmentation led to the overrepresentation of large DRIP fragments, over coding regions, which is especially over the first exons (**Figure 22-23**). This phenomenon severely compromised mapping resolution and therefore, the assignment of clear biological function to a fraction of R-loops. For instance, correct estimation of evolutionary conservation between R-loop binding sites, relying on sequence homologies of exons that are associated with R-loops (Sanz et al. 2016a), becomes uncertain. In addition, biased genome sampling affects many molecular biology techniques that utilize restriction enzyme genome fragmentation, such as 3C/4C/5C, Hi-C and reduced-representation bisulfite sequencing (RRBS) (**Figure 27**).

Based on the above-mentioned experience, we suggest the following refinements of DRIP workflows to obtain accurate estimates of RNA-DNA hybrid occupancies: 1. Omission of HCHO-fixation and RNase A treatment, isolation of nucleic acid by silica membrane (kit) purification, nucleic acid fragmentation by sonication, followed by immunoprecipitation with the S9.6 antibody. 2. If formaldehyde-fixation is applied, we recommend preparing soluble chromatin and fragmenting the prep by sonication (similarly to the ChIP protocol), followed by organic extraction and immunoprecipitation with the S9.6 antibody. 3. If restriction enzyme fragmentation needs to be applied (e.g. in some cases sonication might be too harsh to capture transient or very week RNA-DNA hybrid interactions), we advise the careful control of DNA fragment size distribution before immunoprecipitation.

An important premise is that our recommendations apply to the experimental conditions investigated by this study. Generalization should be avoided since altering critical parameters in the experiment (e.g. incorporating S1 nuclease (S1-DRIP) (Wahba et al. 2016) or lambda exonuclease digestion (DRIP-exo) (Ohle et al. 2016), or changing the model organism) might significantly affect the outcome of RNA-DNA hybrid detection.

In conclusion, the DRIP method remains a gold-standard for identifying *bona fide* R-loop binding sites across individual chromosomes, but a continued effort is needed to find alternatives and test complementary protocols. We hope that this aim has been achieved, at least in part by this study that will help recognize real R-loop binding events and enable a better interpretation of DRIP-seq mapping data.

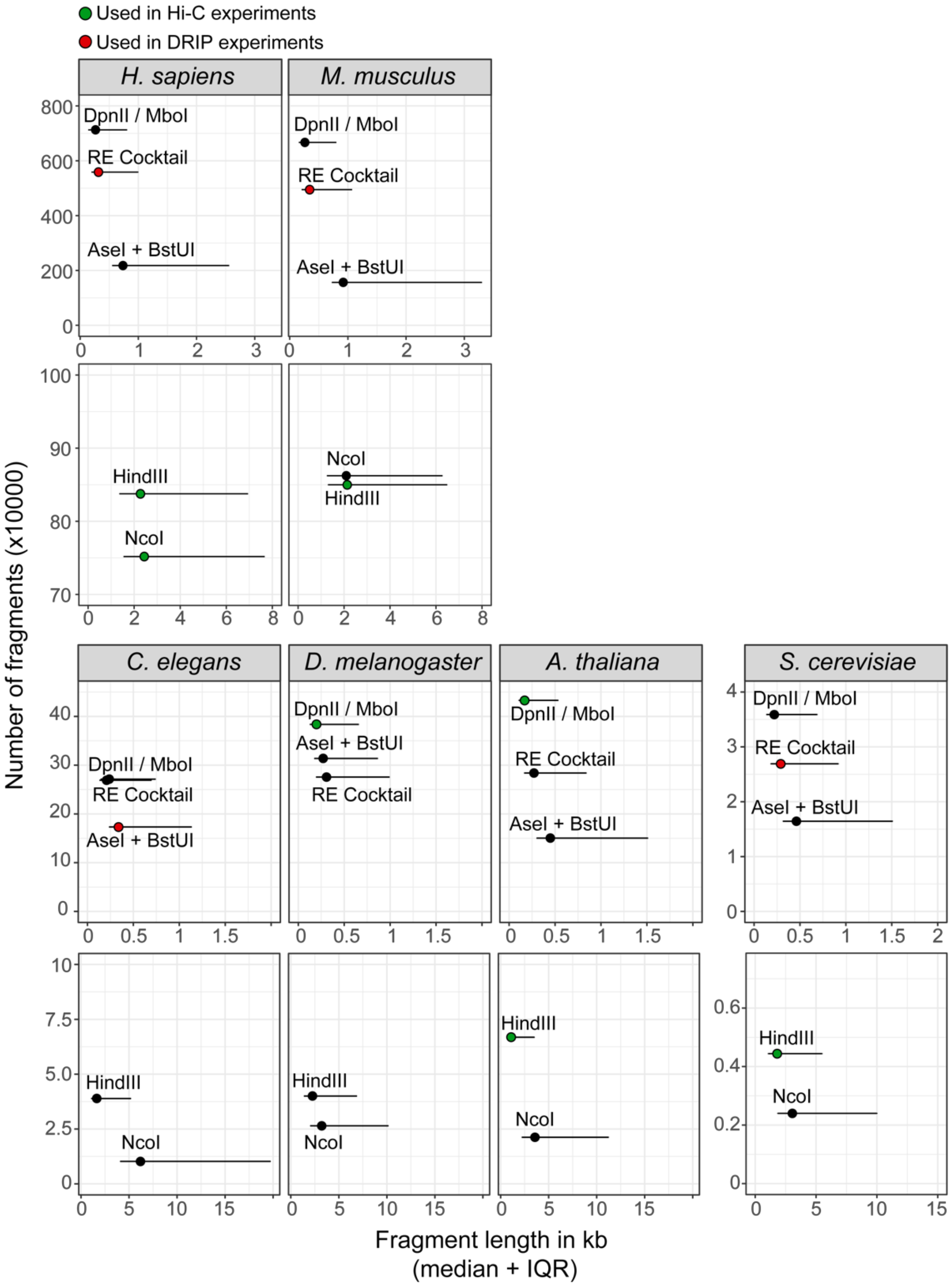


Figure 27. Genome Fragmentation by *In Silico* Restriction Enzyme Digestion in Species That Were Analyzed by DRIP-seq or HI-C. The absolute number of restriction fragments is shown in terms of the average fragment lengths (median + interquartile range [IQR]) obtained by the indicated restriction enzymes applied alone or in combination. RE cocktail denotes the HindIII, EcoRI, BsrGI, XbaI, and SspI enzymes.

6.2. Nuclear dynamics of the Set1C subunit Spp1 prepares meiotic recombination sites for break formation

In the loop-axis model, Spp1 has been reported to tether putative meiotic DSB sites to Mer2 bound chromosomal axis, thereby enabling Spo11 to cut nucleosome free axis-proximal regions (Acquaviva et al. 2013a, 2013b). However, it remained unexplored whether Spp1 was still linked to Set1C during this process and whether a specific subpopulation of Spp1 was relocalized from actively transcribed genes to chromosome axial sites. Additionally, the spatial and temporal dynamics of Spp1 redistribution have not been studied so far.

By capturing chromosomal binding sites of Spp1 while tracking Set1C with Bre2, we revealed that a specific subpopulation of Spp1 acts independently from the Set1C during meiotic progression. In addition, we found three Spp1 subclasses with different binding affinity and dynamics: appearing, disappearing and static. The appearing class of Spp1 is progressively loaded to Mer2/Red1 bound regions during meiosis, indicating *de novo* interaction with the chromosomal axis. These axis-proximal loops in turn enable Spo11 to generate DSBs. Disappearing Spp1 sites have been associated with downregulated genes, suggesting that Spp1 might be released from repressed or poised genes and show low enrichment over Mer2/Red1 sites. Interestingly, we found that disappearing sites are associated with RPG and snoRNA genes that are transiently repressed in the first hours after transfer to sporulation medium (Brar et al. 2012). The mechanism of dissociation of Spp1 from these sites remained an open question. Moreover, the strong association between constant Spp1 peaks and ncRNAs may reflect an unexplored role of Spp1 in regulating non-coding RNA expression.

We further explored the importance of specific protein motifs of Spp1 and their role during the loop-axis tethering. Specifically, we performed time-resolved meiotic ChIP-seq and mapped the binding of Spp1PHDΔ, Spp1CxxCΔ, and Spp1 in H3R2A/H3K4R mutants.

Our results showed that Mer2 enrichment in the Spp1CxxCΔ mutant is prevented over newly formed Spp1 peaks and is strongly reduced in the Spp1PHDΔ mutant. These functional data point towards the importance of the PHD and CxxC motifs for the relocation of Spp1. Interestingly, when the H3R2 and H3K4 side chains were mutated to H3R2A and H3K4R, binding of Spp1 to axial sites was compromised while Spp1 was still able to colocalize with Mer2.

Taken together, our findings presented in this thesis identify Spp1 as a multifaceted protein with dynamic chromatin binding characteristics and further support the tethered loop axis model (**Figure 28**) in the framework of meiotic chromatin structure (Adam et al. 2018; Sommermeyer et al. 2013; Acquaviva et al. 2013b).

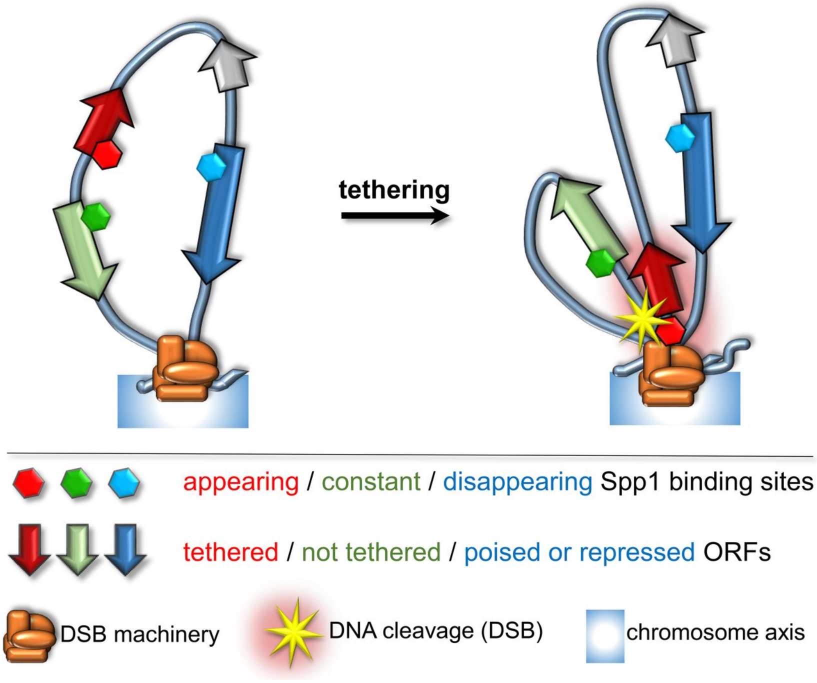

tethering
appearing / constant / disappearing Spp1 binding sites
tethered / not tethered / poised or repressed ORFs
DSB machinery
DNA cleavage (DSB)
chromosome axis

Figure 28. Loop-axis model showing the dynamic behaviour of Spp1 upon tethering. A subset of Spp1 binding sites (red hexagons) becomes tethered to the chromosome axis (ORFs in red). These tethered sites correspond to the dynamic fraction of Spp1 peaks identified by time resolved ChIP-Seq. Appearing Spp1 sites have the ability to interact with Mer2 and to tether DSB sites to the axis where they undergo Spo11-mediated DNA cleavage (yellow star). These properties depend on the PHD domain and the Mer2 binding motif of Spp1, as well as H3K4 and H3R2 methylation. Constant Spp1 sites (green hexagons) remain unchanged during the meiotic time course. Constant Spp1 sites do not interact with Mer2 and remain colocalized with Set1C (ORFs in green). Disappearing Spp1 sites (blue hexagons) are mainly associated with RPG and SnoRNA genes (ORFs in blue). Spp1 tends to be released from the Set1 holocomplex at the latter sites, reflecting the plasticity of Set1C.

7. Summary

- Considering the main experimental variables (formaldehyde fixation, cell lysis temperature, nucleic acid isolation, free RNA removal, and DNA fragmentation), we tested the sensitivity and specificity of 40 DRIP schemes across several genomic regions. Overall, we found that fragmenting nucleic acid by sonication and omitting RNase A digestion could improve the detection specificity of RNA-DNA hybrid detection.

- Comparative analysis of DRIP-seq datasets revealed that restriction enzyme digestion leads to overrepresentation of lengthy DRIP-fragments, especially in the first exons. This biased genome sampling compromises mapping resolution and effects the precise annotation of a subset of RNA-DNA hybrids. In use, we advise to check the fragment size distribution both *in silico* and *in vitro*.

- We identified a Set1C independent Spp1 subpopulation during meiotic progression.

- Using time-resolved meiotic ChIP-seq, we revealed three Spp1 subclasses each with different chromatin binding kinetics (appearing, disappearing and static) and biological functions.

- By analysing loss of function mutants; Spp1PHDΔ, Spp1CxxCΔ, H3R2A and H3K4R mutants, we revealed that proper localization of Spp1 to chromosome axial sites requires: (1) the Mer2-binding (CxxC) motif of Spp1; (2) to a lesser extent, the PHD finger domain; and (3) the presence of histone modifications and modifiable residues (H3K4me3 and H3R2me2s).

Összefoglalás

- Figyelembe véve a főbb kísérleti változókat (formaldehid keresztkötés, sejtlízis hőmérséklet, nukleinsav izolálás, szabad RNS eltávolítás és DNS fragmentálás), teszteltük 40 DRIP kísérleti séma szenzitivitását és specifitását több genomi régión. Összességében, azt találtuk, hogy a nukleinsav fragmentálása szonikálással, valamint az RNase A kezelés kihagyása javíthat az RNS-DNS hibridek detektálásának specifitásán.

- Az elérhető DRIP-seq adatok összehasonlítása felfedte, hogy a restrikciós enzim emésztés a hosszabb DRIP fragmentumok felül-reprezentációjához vezet. Ez a torzított mintavételezés befolyásolja a módszer felbontóképességét, valamint az RNS-DNS hibridek egy részének precíz annotációját. Ha restrikciós enzim emésztést alkalmaznak, azt javasoljuk, hogy *in silico* és *in vitro* is ellenőrizzék a fragmentumok hosszának eloszlását.

- Egy Set1C független Spp1 szubpopulációt azonosítottunk a meiotikus progresszió során.

- Időfüggő meiotikus ChIP-seq módszer alkamazásával három eltérő kötési dinamikával és biológiai funkcióval rendelkező Spp1 csoportot mutattunk ki: megjelenő, eltűnő és statikus.

- Funkcióvesztett mutánsok elemzésével - Spp1PHDΔ, Spp1CxxCΔ, H3R2A és H3K4R - kimutattuk, hogy az Spp1 a kromoszóma tengelyhez való helyes lokalizációjához szükséges: (1) az Spp1 Mer2-kötő cink-ujj motívuma (CxxC); (2) kisebb mértékben a fehérje PHD doménje; és (3) a hisztonmódosítások valamint a módosítható oldalláncok (H3K4me3 és H3R2me2s) jelenléte.

8. References

8.1. References related to the dissertation

Acquaviva L, Drogat J, Dehé PM, de La Roche Saint-André C, Géli V. 2013a. Spp1 at the crossroads of H3K4me3 regulation and meiotic recombination. *Epigenetics* **8**: 355–360.

Acquaviva L, Szekvolgyi L, Dichtl B, Dichtl BS, Andre de LR Saint, Nicolas A, Geli V. 2013b. The COMPASS subunit Spp1 links histone methylation to initiation of meiotic recombination . *Science (80-)* **339**: 215–218.

Adam C, Guérois R, Citarella A, Verardi L, Adolphe F, Béneut C, Sommermeyer V, Ramus C, Govin J, Couté Y, et al. 2018. The PHD finger protein Spp1 has distinct functions in the Set1 and the meiotic DSB formation complexes. *PLoS Genet* **14**: e1007223.

Aguilera A, García-Muse T. 2012. R Loops: From Transcription Byproducts to Threats to Genome Stability. *Mol Cell* **46**: 115–124.

Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, Walter P. 2014. *Molecular Biology of the Cell 6e*.

Baranello L, Kouzine F, Sanford S, Levens D. 2016. ChIP bias as a function of cross-linking time. *Chromosom Res* **24**: 175–181.

Beagan JA, Duong MT, Titus KR, Zhou L, Cao Z, Ma J, Lachanski C V., Gillis DR, Phillips-Cremins JE. 2017. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res* **27**: 1139–1152.

Beneke S, Meyer K, Holtz A, Hüttner K, Bürkle A. 2012. Chromatin composition is changed by poly(ADP-ribosyl)ation during chromatin immunoprecipitation. *PLoS One* **7**: 1–10.

Benore-Parsons M, Ayoub MA. 1997. Presence of RNase a causes aberrant DNA band shifts. *Biotechniques* **23**: 128–131.

Bhatia V, Barroso SI, García-Rubio ML, Tumini E, Herrera-Moyano E, Aguilera A. 2014. BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. *Nature*.

Bhattacharyya B, Keck JL. 2014. Grip it and rip it: Structural mechanisms of DNA helicase substrate binding and unwinding. *Protein Sci* **23**: 1498–1507.

Bjursell G, GUSSANDER E, LINDAHL T. 1979. Long regions of single-stranded DNA in human cells. *Nature*. http://www.nature.com/nature/journal/v280/n5721/abs/280420a0.html.

Boque-Sastre R, Soler M, Oliveira-Mateos C, Portela A, Moutinho C, Sayols S, Villanueva A, Esteller M, Guil S. 2015. Head-to-head antisense transcription and R-loop formation promotes transcriptional activation. *Proc Natl Acad Sci* **112**: 201421197.

Borde V, Robine N, Lin W, Bonfils S, Géli V, Nicolas A. 2009. Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. *EMBO J*.

Brar G a., Yassour M, Friedman N, Regev A, Ingolia NT, Weissman JS. 2012. High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science (80-)* **335**: 552–557.

Cairns J, Spyrou C, Stark R, Smith ML, Lynch AG, Tavaré S. 2011. BayesPeak - An R package for analysing ChIP-seq data. *Bioinformatics*.

Capra JA, Paeschke K, Singh M, Zakian VA. 2010. G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in Saccharomyces cerevisiae. *PLoS Comput Biol* **6**: 9.

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, Lomax J, Mungall C, Hitz B, Balakrishnan R, et al. 2009. AmiGO: Online access to ontology and annotation data. *Bioinformatics* **25**: 288–289.

Chang P, Gohain M, Yen MR, Chen PY. 2018. Computational Methods for Assessing Chromatin Hierarchy. *Comput Struct Biotechnol J* **16**: 43–53.

Chédin F. 2016. Nascent Connections: R-Loops and Chromatin Patterning. *Trends Genet* **32**: 828–838.

Chen L, Chen JY, Zhang X, Gu Y, Xiao R, Shao C, Tang P, Qian H, Luo D, Li H, et al. 2017. R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters. *Mol Cell* **68**: 745–757.e5.

Chen PB, Chen H V., Acharya D, Rando OJ, Fazzio TG. 2015. R loops regulate promoter-proximal chromatin architecture and cellular differentiation. *Nat Struct Mol Biol*.

Chen Y, Zhang Y, Wang Y, Zhang L, Brinkman EK, Adam SA, Goldman R, van Steensel B, Ma J, Belmont AS. 2018. Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *J Cell Biol* jcb.201807108. http://www.ncbi.nlm.nih.gov/pubmed/30154186 (Accessed August 30, 2018).

Coman D, Russu IM. 2005. Base pair opening in three DNA-unwinding elements. *J Biol Chem* **280**: 20216–20221.

Consortium EP. 2013. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Cristini A, Groh M, Kristiansen MS, Gromak N. 2018. RNA/DNA Hybrid Interactome Identifies DXH9 as a Molecular Player in Transcriptional Termination and R-Loop-Associated DNA Damage. *Cell Rep* **23**: 1891–1905.

Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, Mirny LA, O'Shea CC, Park PJ, Ren B, et al. 2017. The 4D nucleome project. *Nature*.

Dickey TH, Altschuler SE, Wuttke DS. 2013. Single-stranded DNA-binding proteins: Multiple domains for multiple functions. *Structure* **21**: 1074–1084.

Dona F, Houseley J. 2014. Unexpected DNA loss mediated by the DNA binding activity of ribonuclease A. *PLoS One* **9**: 1–11.

Drolet M, Bi X, Liu LF. 1994. Hypernegative supercoiling of the DNA template during transcription elongation in vitro. *J Biol Chem* **269**: 2068–2074.

Dumelie JG, Jaffrey SR. 2017. Defining the location of promoter-associated R-loops at near-nucleotide resolution using bisDRIP-seq. *Elife*.

Ernst J, Kellis M. 2017. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* **12**: 2478–2492.

Flynn RL, Zou L. 2010. Oligonucleotide/oligosaccharide-binding fold proteins: A growing family of genome guardians. *Crit Rev Biochem Mol Biol* **45**: 266–275.

Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. 2016. Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep* **15**: 2038–2049.

García-Rubio ML, Pérez-Calero C, Barroso SI, Tumini E, Herrera-Moyano E, Rosado I V., Aguilera A. 2015. The Fanconi Anemia Pathway Protects Genome Integrity from R-loops. *PLoS Genet* **11**: 1–17.

Ghosh A, Bansal M. 2003. A glossary of DNA structures from A to Z. *Acta Crystallogr - Sect D Biol Crystallogr* **59**: 620–626.

Ginno PA, Lott PL, Christensen HC, Korf I, Chédin F. 2012. R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters. *Mol Cell* **45**: 814–825.

Groh M, Gromak N. 2014. Out of Balance: R-loops in Human Disease. *PLoS Genet.*

Groh M, Lufino MMP, Wade-Martins R, Gromak N. 2014. R-loops Associated with Triplet Repeat Expansions Promote Gene Silencing in Friedreich Ataxia and Fragile X Syndrome. *PLoS Genet* **10**: e1004318.

Halász L, Karányi Z, Boros-Oláh B, Kuik-Rózsa T, Sipos É, Nagy É, Mosolygó-L Á, Mázló A, Rajnavölgyi É, Halmos G, et al. 2017. RNA-DNA hybrid (R-loop) immunoprecipitation mapping: An analytical workflow to evaluate inherent biases. *Genome Res* **27**: 1063–1073.

Hänsel-Hertsch R, Beraldi D, Lensing S V., Marsico G, Zyner K, Parry A, Di Antonio M, Pike J, Kimura H, Narita M, et al. 2016. G-quadruplex structures mark human regulatory chromatin. *Nat Genet* **48**: 1267–1272.

Hänsel-Hertsch R, Di Antonio M, Balasubramanian S. 2017. DNA G-quadruplexes in the human genome: Detection, functions and therapeutic potential. *Nat Rev Mol Cell Biol* **18**: 279–284.

Hänsel-Hertsch R, Spiegel J, Marsico G, Tannahill D, Balasubramanian S. 2018. Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nat Protoc* **13**: 551–564. http://www.nature.com/doifinder/10.1038/nprot.2017.150.

Hartono SR, Malapert A, Legros P, Bernard P, Chédin F, Vanoosthuyse V. 2018. The Affinity of the S9.6 Antibody for Double-Stranded RNAs Impacts the Accurate Mapping of R-Loops in Fission Yeast. *J Mol Biol*.

Herrera-Moyano E, Mergui X, García-Rubio ML, Barroso S, Aguilera A. 2014. The yeast and human FACT chromatin-reorganizing complexes solve R-loop-mediated transcription-replication conflicts. *Genes Dev* **1**: 735–748.

Hyun K, Jeon J, Park K, Kim J. 2017. Writing, erasing and reading histone lysine methylations. *Exp Mol Med.*

Karányi Z, Halász L, Acquaviva L, Jónás D, Hetey S, Boros-Oláh B, Peng F, Chen D, Klein F, Géli V, et al. 2018. Nuclear dynamics of the Set1C subunit Spp1 prepares meiotic recombination sites for break formation. *J Cell Biol*.

Kirmizis A, Santos-Rosa H, Penkett CJ, Singer MA, Vermeulen M, Mann M, Bähler J, Green RD, Kouzarides T. 2007. Arginine methylation at histone H3R2 controls deposition of H3K4 trimethylation. *Nature*.

Kolde R. 2015. pheatmap : Pretty Heatmaps. *R Packag version 108*.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information esthetic for comparative genomics. *Genome Res* **19**: 1639–1645.

Kuznetsov VA, Bondarenko V, Wongsurawat T, Yenamandra SP, Jenjaroenpun P, Name A. 2018. Toward predictive R-loop computational biology: genome-scale prediction of R-loops reveals their association with complex promoter structures, G-quadruplexes and transcriptionally active enhancers. *Nucleic Acids Res*.

Lawrence M, Daujat S, Schneider R. 2016. Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends Genet* **32**: 42–56.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol* **9**: 1–10.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lim YW, Sanz LA, Xu X, Hartono SR, Chédin F. 2015. Genome-wide DNA hypomethylation and RNA:DNA hybrid accumulation in Aicardi–Goutières syndrome. *Elife* **4**.

Loomis EW, Sanz L a, Chédin F, Hagerman PJ. 2014. Transcription-Associated R-Loop Formation across the Human FMR1 CGG-Repeat Region. *PLoS Genet* **10**: e1004294.

Marceau AH. 2012. Functions of single-strand DNA-binding proteins in DNA replication, recombination, and repair. *Methods Mol Biol* **922**: 1–21. http://www.ncbi.nlm.nih.gov/pubmed/22976174%5Cnhttp://link.springer.com/10.1007/978-1-62703-032-8.

Marinello J, Bertoncini S, Aloisi I, Cristini A, Tagliazucchi GM, Forcato M, Sordet O, Capranico G. 2016. Dynamic effects of topoisomerase i inhibition on R-loops and short transcripts at active promoters. *PLoS One* **11**: 1–18.

McGhee JD, von Hippel PH. 1977. Formaldehyde as a Probe of DNA Structure. 4. Mechanism of the Initial Reaction of Formaldehyde with DNA. *Biochemistry* **16**: 3276–3293.

Mishra SK, Tawani A, Mishra A, Kumar A. 2016. G4IPDB: A database for G-quadruplex structure forming nucleic acid interacting proteins. *Sci Rep* **6**.

Mohibullah N, Keeney S. 2017. Numerical and spatial patterning of yeast meiotic DNA breaks by Tel1. *Genome Res*.

Nadel J, Athanasiadou R, Lemetre C, Wijetunga NA, Ó Broin P, Sato H, Zhang Z, Jeddeloh J, Montagna C, Golden A, et al. 2015. RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. *Epigenetics Chromatin* **8**: 46. http://www.epigeneticsandchromatin.com/content/8/1/46.

Nakama M, Kawakami K, Kajitani T, Urano T, Murakami Y. 2012. DNA-RNA hybrid formation mediates RNAi-directed heterochromatin formation. *Genes to Cells*.

Ohle C, Tesorero R, Schermann G, Dobrev N, Sinning I, Fischer T. 2016. Transient RNA-DNA Hybrids are Required for Efficient Double-Strand Break Repair. *Cell* **167**: 1001–1013.

Ou HD, Phan S, Deerinck TJ, Thor A, Ellisman MH, O'Shea CC. 2017. ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science (80-)* **357**.

Panizza S, Mendoza MA, Berlinger M, Huang L, Nicolas A, Shirahige K, Klein F. 2011. Spo11-accessory proteins link double-strand break sites to the chromosome axis in early meiotic recombination. *Cell*.

Phillips DD, Garboczi DN, Singh K, Hu Z, Leppla SH, Leysath CE. 2013. The sub-nanomolar binding of DNA-RNA hybrids by the single-chain Fv fragment of antibody S9.6. *J Mol Recognit* **26**: 376–381.

Potaman VN, Sinden RR. 2005. CHAPTER 1 DNA: Alternative Conformations and Biology. *DNA Conform Transcr* 1–16. https://www.landesbioscience.com/curie/chapter/2078/.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. 2014. DeepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**: 1–5.

Rawal P, Kummarasetti VBR, Ravindran J, Kumar N, Halder K, Sharma R, Mukerji M, Das SK, Chowdhury S. 2006. Genome-wide prediction of G4 DNA as regulatory motifs: Role in Escherichia coli global regulation. *Genome Res* **16**: 644–655.

Richard P, Manley JL. 2017. R Loops and Links to Human Disease. *J Mol Biol*.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**: 77.

Roy D, Lieber MR. 2009. G Clustering Is Important for the Initiation of Transcription-Induced R-Loops In Vitro, whereas High G Density without Clustering Is Sufficient Thereafter. *Mol Cell Biol* **29**: 3124–3133. http://mcb.asm.org/cgi/doi/10.1128/MCB.00139-09.

Roy D, Yu K, Lieber MR. 2008. Mechanism of R-Loop Formation at Immunoglobulin Class Switch Sequences. *Mol Cell Biol* **28**: 50–60. http://mcb.asm.org/cgi/doi/10.1128/MCB.01251-07.

Sahakyan AB, Chambers VS, Marsico G, Santner T, Di Antonio M, Balasubramanian S. 2017. Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci Rep* **7**.

Sanz LA, Hartono SR, Lim YW, Ginno PA, Sanz LA, Hartono SR, Lim YW, Steyaert S, Rajpurkar A, Ginno PA, et al. 2016a. Prevalent , Dynamic , and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Mol Cell* **63**: 167–178.

Sanz LA, Hartono SR, Lim YW, Steyaert S, Rajpurkar A, Ginno PA, Xu X, Chédin F. 2016b. Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Mol Cell* **63**: 167–178.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**.

Singleton MR, Dillingham MS, Wigley DB. 2007. Structure and Mechanism of Helicases and Nucleic Acid Translocases. *Annu Rev Biochem* **76**: 23–50. http://www.annualreviews.org/doi/10.1146/annurev.biochem.76.052305.115300.

Skourti-Stathaki K, Proudfoot NJ. 2014. A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes Dev*.

Sollier J, Cimprich KA. 2015. Breaking bad: R-loops and genome integrity. *Trends Cell Biol*.

Sommermeyer V, Beneut C, Chaplais E, Serrentino ME, Borde V. 2013. Spp1, a member of the Set1 Complex, promotes meiotic DSB formation in promoters by tethering histone H3K4 methylation sites to chromosome axes. *MolCell* **49**: 43–54.

Stork CT, Bocek M, Crossley MP, Sollier J, Sanz LA, Chédin F, Swigut T, Cimprich KA. 2016. Co-transcriptional R-loops are the main cause of estrogen-induced DNA damage. *Elife* **5**.

Sun X, Huang L, Markowitz TE, Blitzblau HG, Chen D, Klein F, Hochwagen A. 2015. Transcription dynamically patterns the meiotic chromosome-axis interface. *Elife*.

Taberlay PC, Statham AL, Kelly TK, Clark SJ, Jones PA. 2014. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. *Genome Res* **24**: 1421–1432.

Thomas M, White RL, Davis RW. 1976. Hybridization of RNA to double-stranded DNA: formation of R-loops. *Proc Natl Acad Sci* **73**: 2294–2298. http://www.pnas.org/cgi/doi/10.1073/pnas.73.7.2294.

Umate P, Tuteja N, Tuteja R. 2011. Genome-wide comprehensive analysis of human helicases. *Commun Integr Biol* **4**: 1–20.

Wahba L, Costantino L, Tan FJ, Zimmer A, Koshland D. 2016. S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation. *Genes Dev* **30**: 1327–38.

Wang IX, Grunseich C, Fox J, Burdick J, Zhu Z, Ravazian N, Hafner M, Cheung VG. 2018. Human proteins that interact with RNA/DNA hybrids. *Genome Res*. http://www.ncbi.nlm.nih.gov/pubmed/30108179 (Accessed August 27, 2018).

WATSON JD, CRICK FH. 1953. The structure of DNA. *Cold Spring Harb Symp Quant Biol* **18**: 123–131.

Wickham H. 2009. Ggplot. *Media* **35**: 211.

Wongsurawat T, Jenjaroenpun P, Kwoh CK, Kuznetsov V. 2012. Quantitative model of R-loop forming structures reveals a novel level of RNA-DNA interactome complexity. *Nucleic Acids Res*.

Wright ES. 2016. Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. *R J* **8**: 352–359.

Wu Y, Brosh RM. 2010. G-quadruplex nucleic acids and human disease. *FEBS J* **277**: 3470–88. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2923685&tool=pmcentrez&rendertype=abstract.

Wu Y, Lu J, Kang T. 2016. Human single-stranded DNA binding proteins: Guardians of genome stability. *Acta Biochim Biophys Sin (Shanghai)* **48**: 671–677.

Xu W, Xu H, Li K, Fan Y, Liu Y, Yang X, Sun Q. 2017. The R-loop is a common chromatin feature of the Arabidopsis genome. *Nat Plants* **3**: 704–714.

Yang Y, McBride KM, Hensley S, Lu Y, Chedin F, Bedford MT. 2014. Arginine Methylation Facilitates the Recruitment of TOP3B to Chromatin to Prevent R Loop Accumulation. *Mol Cell* **53**: 484–497.

Yano M, Kato Y. 2014. Using hidden Markov models to investigate G-quadruplex motifs in genomic sequences. *BMC Genomics* **15**.

Yin T, Cook D, Lawrence M. 2012. ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol* **13**: R77.

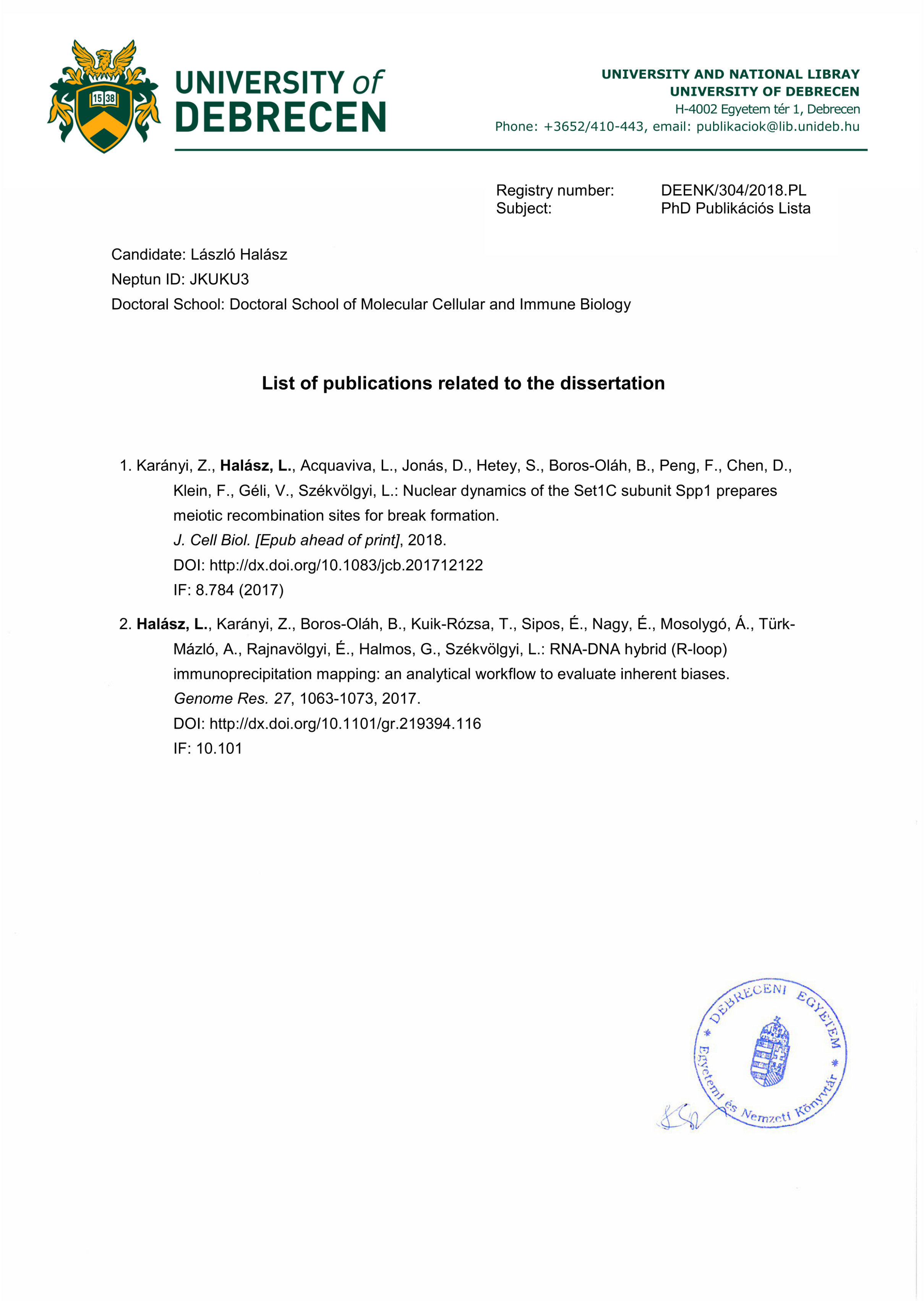
Yuan CC, Matthews AGW, Jin Y, Chen CF, Chapman BA, Ohsumi TK, Glass KC, Kutateladze TG, Borowsky ML, Struhl K, et al. 2012. Histone H3R2 Symmetric Dimethylation and Histone H3K4 Trimethylation Are Tightly Correlated in Eukaryotic Genomes. *Cell Rep*.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2017. Ensembl 2018. *Nucleic Acids Res*. http://academic.oup.com/nar/article/doi/10.1093/nar/gkx1098/4634002.

Zhang Y, Liu T, Meyer C a, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zheng KW, He Y De, Liu HH, Li XM, Hao YH, Tan Z. 2017. Superhelicity Constrains a Localized and R-Loop-Dependent Formation of G-Quadruplexes at the Upstream Region of Transcription. *ACS Chem Biol* **12**: 2609–2618.

8.2. Publication list prepared by the Kenézy Life Sciences Library



UNIVERSITY of DEBRECEN

UNIVERSITY AND NATIONAL LIBRAY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK/304/2018.PL
Subject: PhD Publikációs Lista

Candidate: László Halász
Neptun ID: JKUKU3
Doctoral School: Doctoral School of Molecular Cellular and Immune Biology

List of publications related to the dissertation

1. Karányi, Z., **Halász, L.**, Acquaviva, L., Jonás, D., Hetey, S., Boros-Oláh, B., Peng, F., Chen, D., Klein, F., Géli, V., Székvölgyi, L.: Nuclear dynamics of the Set1C subunit Spp1 prepares meiotic recombination sites for break formation.
   J. Cell Biol. [Epub ahead of print], 2018.
   DOI: http://dx.doi.org/10.1083/jcb.201712122
   IF: 8.784 (2017)
2. **Halász, L.**, Karányi, Z., Boros-Oláh, B., Kuik-Rózsa, T., Sipos, É., Nagy, É., Mosolygó, Á., Türk-Mázló, A., Rajnavölgyi, É., Halmos, G., Székvölgyi, L.: RNA-DNA hybrid (R-loop) immunoprecipitation mapping: an analytical workflow to evaluate inherent biases.
   Genome Res. 27, 1063-1073, 2017.
   DOI: http://dx.doi.org/10.1101/gr.219394.116
   IF: 10.101

List of other publications

3. Hegedűs, É., Kókai, E., Nánási, P. P., Imre, L., **Halász, L.**, Jossé, R., Antunovics, Z., Webb, M. R., El Hage, A., Pommier, Y., Székvölgyi, L., Dombrádi, V., Szabó, G.: Endogenous single-strand DNA breaks at RNA polymerase II promoters in Saccharomyces cerevisiae.
   Nucleic Acids Res. [Epub ahead of print], 2018.
   DOI: http://dx.doi.org/10.1093/nar/gky743
   IF: 11.561 (2017)

4. Roszik, J., Fenyőfalvi, G., **Halász, L.**, Karányi, Z., Székvölgyi, L.: In Silico Restriction Enzyme Digests To Minimize Mapping Bias In Genomic Sequencing.
   Mol. Ther. Methods. Clin. Dev. 6, 66-67, 2017.
   DOI: http://dx.doi.org/10.1016/j.omtm.2017.06.003
   IF: 3.681

Total IF of journals (all publications): 34,127
Total IF of journals (publications related to the dissertation): 18,885

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of Web of Science, Scopus and Journal Citation Report (Impact Factor) databases.

12 September, 2018

9. Keywords

Genomics, bioinformatics, DRIP-seq, fragmentation bias, RNA-DNA hybrids, chromatin structure, H3K4me3, DSB, Set1C, Spp1, Mer2

Tárgyszavak

Genomika, bioinformatika, DRIP-seq, fragmentációs torzítás, RNS-DNS hibridek, kromatin szerkezet, H3K4me3, DSB, Set1C, Spp1, Mer2

10. Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor, Dr. Lóránt Székvölgyi for his guidance, support and patience throughout the years. I greatly appreciate that he has given me the opportunity to work on projects I was interested in.

My sincere thanks also go to Zsolt Karányi for the stimulating discussions. I am especially grateful that he introduced and helped me to learn R programming.

I would like to acknowledge Prof. László Fésüs and Prof. József Tőzsér, the former and recent heads of the Department of Biochemistry and Molecular Biology for the opportunity to work in a well-equipped environment.

I am also thankful to my advisory committee: Prof. Éva Rajnavölgyi and Dr. István Balogh.

Thank you to my co-authors for their contributions to the work presented in this thesis!

Special thanks to Dr. Éva Sipos, Dr. Ágnes Mosolygó-Lukács, Szabolcs Hetey, Éva Nagy, Tímea-Kuik Rózsa, and Ibolya Fürtös who performed the "wet" part of the experiments related to this study.

I also wish to express my deepest gratitude to Krisztina and my family for their support and patience.

Appendix 1 – Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
S9.6	S9.6 ATCC ® HB-8730™ Mus musculus (B cell)	HB-8730
Goat anti-mouse IgG marked with HRP	Santa Cruz Biotechnology	sc-2005
Rabbit anti-mouse IgG Alexa647	Thermo Fisher Scientific	A16168
Chemicals, Peptides, and Recombinant Proteins		
UltraPure Paraformaldehyde	Sigma-Aldrich	P6148-500G
RNase A solution, 10 mg/ml	UD-GenoMed Ltd.	UDV0322
RNase H	New England Biolabs	M0297L
Proteinase K	Thermo Fisher Scientific	EO0492
HindIII, EcoRI, BsrGI, XbaI and SspI	New England Biolabs	*https://www.neb.com/*
Dynabeads Protein A, 5 ml	Life Technologies	10002D
Penicillin-Streptomycin Solution, 100 ml	Sigma-Aldrich	P4333-100ML
ROX solution 50 µM 1ml	Thermo Scientific	34094
Nitrocellulose Blotting Membrane	GE Healthcare	10600020
Phenol-chloroform-isoamyl alcohol mixture	Sigma-Aldrich	77617-100ml
RPMI-1640	Sigma-Aldrich	R8758-500ML

Critical Commercial Assays		
NucleoSpin Tissue Kit	Macherey-Nagel	740952.50
LightCycler 480 SYBR Green I Master, 2x qPCR master mix	Roche	4887352001
Naive CD4+ T-cell Isolation Kit	Miltenyi Biotec	130-094-131
NucleoSpin Gel and PCR Clean-up (250 preps)	Macherey-Nagel	740609.250
SuperSignal West Femto Trial Kit	Thermo Scientific	34094
Deposited Data		
Raw and analyzed data	(Halász et al. 2017)	SRP095885
Human reference genome NCBI build 37, GRCh37	Genome Reference Consortium	*http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/*
Human GRCh37 blacklisted regions	(Consortium 2013)	ENCFF419RSJ
NTERA2 Chromatin State Data	(Consortium 2013)	ENCSR403MYH
MCF7 Chromatin State Data	(Taberlay et al. 2014)	GEO: GSE57498
K562, IMR90, HEK and Primary Fibroblast Core 15-State Models	NIH Roadmap Epigenomics Mapping Consortium	E123, E017, E086 and E055
NTERA2 DRIP-sequencing data	(Ginno et al. 2012)	GEO: GSE45530
NTERA2 DRIP-sequencing data	(Sanz et al. 2016b)	GEO: GSE70189
K562 DRIP-sequencing data	(Sanz et al. 2016b)	GEO: GSE70189
IMR90 DRIP-sequencing data	(Nadel et al. 2015)	GEO: GSE68953

HEK DRIP-sequencing data	(Nadel et al. 2015)	GEO: GSE68953
Primary Fibroblast DRIP-sequencing data	(Lim et al. 2015)	GEO: GSE57353
MCF7 DRIP-sequencing data	(Stork et al. 2016)	GEO: GSE81851
Human Protein coding genes, exons and introns	Ensembl	*http://www.ensembl.org/*
List of Repetitive Elements	UCSC; RepeatMasker	*https://genome.ucsc.edu/*

Experimental Models: Cell Lines		
Human Jurkat T-lymphoblastoid cell line	Sigma-Aldrich	*https://www.sigmaaldrich.com/*
Human Naive CD4+ T-cells	(Halász et al. 2017)	N/A
Human GM12878 B-lymphoblastoid cell line	Coriell Institute	*https://www.coriell.org/*
Sequence-Based Reagents		
TruSeq ChIP Sample Preparation Kit	Illumina	*www.illumina.com*
Primers for DRIP-qPCR	(Halász et al. 2017)	N/A
Software and Algorithms		
R programming language	R core team	*https://www.r-project.org/*
ggplot2	(Wickham 2009)	*https://cran.r-project.org/*
ggbio	(Yin et al. 2012)	*https://bioconductor.org/*
GenomicRanges	(Lawrence et al. 2013)	*https://bioconductor.org/*
DECIPHER	(Wright 2016)	*https://bioconductor.org/*
Circos	(Krzywinski et al. 2009)	*http://www.circos.ca/*
BWA	(Li and Durbin 2009)	*http://bio-bwa.sourceforge.net/*
SAMtools	(Li et al. 2009)	*http://samtools.sourceforge.net/*

Picard tools	N/A	*https://broadinstitute.github.io/picard/*
BEDTools	(Quinlan and Hall 2010)	*http://bedtools.readthedocs.io/en/latest/*
MACS2	(Zhang et al. 2008)	*https://github.com/taoliu/MACS*
deepTools	(Ramírez et al. 2014)	*https://github.com/fidelram/deepTools*
pROC	(Robin et al. 2011)	*https://cran.r-project.org/*