SHORT 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

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

by **László Halász, MSc**

Supervisor: Dr. Lóránt Székvölgyi, PhD

Doctoral School of Molecular Cell and Immune Biology, University of Debrecen

Examination Committee:

Head:	Prof. Dr. László Fésüs, MD, PhD, DSc, MHSC
Members:	Dr. György Vámosi, PhD
	Dr. Tibor Pankotai, PhD

The Examination took place at the Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Debrecen

Debrecen; at 11 AM; 28th of April 2017

Reviewers of the thesis:

Prof. Dr. Margit Balázs, PhD, DSc
Dr. László Bodai, PhD

Defense Committee:

Head:	Prof. Dr. László Fésüs, MD, PhD, DSc, MHSC
Members:	Prof. Dr. Margit Balázs, PhD, DSc
	Dr. László Bodai, PhD
	Dr. György Vámosi, PhD
	Dr. Tibor Pankotai, PhD

The PhD Defense takes place at the Lecture Hall of Bldg. A,
Department of Internal Medicine, Faculty of Medicine, University of Debrecen

Debrecen; at 2:15 PM; 15th of February 2019

1. Introduction

1.1. 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 have evolved molecular mechanisms allowing their DNA content to be packed at many scales during interphase (from chromosome territories to interacting chromatin loops).

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. Nucleosomes consist of DNA, structural non-coding RNAs and associated proteins. The nucleosome is composed of a core particle, with DNA wrapped every ~147 base pairs around specialized proteins called histones. The central histone octamer consists of two H2A, H2B heterodimers and two H3/H4 core heterodimers, and a H1 linker. Nucleosomes units are connected to the adjacent nucleosome by short DNA sequences known as the "linker" DNA, creating a nucleosome chain. 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. 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. 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. 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.

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. 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 can make it possible. 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 nuclear compartment and even inferring chromosome trajectories from one compartment to another.

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 chromatin organization became a hot topic in the field of genomics.

1.2. 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. 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. 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.

1.3. 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. It is important to note, that none of these modification changes the charge of the amino acid and thus the structure of the nucleosome, but it serves as a docking site for other effector proteins. Unlike acetylation, which have a half-life of several minutes, methylation is considered 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 its 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 is 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.

1.4. 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). 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. 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.

1.3. 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). 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. Despite of the intense research that discovered many aspects of meiotic DSB formation, the nuclear dynamics of Set1C subunits is still unknown.

1.4. RNA-DNA hybrids (R-loops)

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.

1.5. R-loop formation and functions

R-loops are abundant epigenetic features in mammalian systems. It is estimated, that ~5% of the human genome can form R-loops. Recently, several accepted models exist for R-loop formation. The first models proposed by Lieber and Roy are the 'thread back' and 'extended hybrid' model. 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.

R-loop formation is not a random process. There are several basic determinants in the genome that facilitate or prevent R-loop formation. 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. 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. Studies investigating DNA supercoiling showed, that negative topological stress is tightly linked to R-loop formation. Other observations have indicated that promoter R-loops tend to form over DNA sequences where elevated RNA Polymerase II pausing happens. Open and active chromatin regions (marked by H3K36me3, H3K4me3/me2 and H3K9Ac) and high transcription rates also positively correlate with R-loop formation. 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. 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.

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.

1.6. Detection approaches of the RNA-DNA hybrids

After the initial discovery of R-loops using electron microscopy, 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.

The most commonly used methodology is DNA-RNA immunoprecipitation followed by quantitative PCR (DRIP-qPCR) or next-generation sequencing (DRIP-seq). 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.

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

S1-DRIP-seq 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 and DRIPc-seq 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). 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) and R-ChIP. 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. 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.

2. Specific aims

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

- 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

2. Functional analysis of Spp1 chromatin binding during meiosis.

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

3. Materials and methods

3.1. DNA-RNA immunoprecipitation (DRIP)

Crosslinking of cells was done with 1% paraformaldehyde for 10 minutes, then quenched with 2.5 M glycine for 5 minutes at room temperature. Cells were lysed in 1 ml lysis buffer composed of 500 µl 2x lysis plus 500 µl TE. Cell lysis was performed at two temperatures: either at 65 °C for 7 hours, or at 37 °C, overnight. Total nucleic acid was isolated by a NucleoSpin Tissue Kit and eluted in 100 µl of elution buffer. The purified nucleic acid prep was fragmented by sonication in 300 µl of Tris-HCl pH 8.5 for 2 x 5 min (30 sec ON, 30 sec OFF, LOW) to yield an average DNA fragment size of ~500 bp. Fragment analysis was done by using 1 % agarose gelelectophoreses. If it was necessary, further sonication was applied. The sonicated DNA sample was purified by a NucleoSpin Gel and PCR Clean-up Kit and eluted in 100 µl of elution buffer. 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 in a total volume of 80 µl at 37 °C, overnight. Dynabeads Protein A magnetic beads 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 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), 1ml lysis buffer (high salt), 1ml wash buffer and 1ml TE at 4°C, two times. Elution was performed in 100 µl of elution buffer for 15 min at 65°C. After purification by NucleoSpin Gel and PCR Clean-up Kit, nucleic acids were eluted in 55 µl of elution buffer. The recovered DNA was then analyzed by quantitative real-time PCR. qPCR was performed with LightCycler 480 SYBR Green I Master and analyzed on QuantStudio 12K Flex Real-Time PCR System. The data were analyzed using the comparative C_T method. The RNA-DNA hybrid enrichment was calculated based on the IP/Input ratio.

3.2. 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. Sequenced reads were aligned to the Human reference genome using default parameters of BWA MEM algorithm. Low mapping quality, supplementary alignments, reads mapped to blacklisted regions and redundant reads were omitted from downstream analysis. Replicate experiments were merged and then MACS2 was used to identify enriched regions of the genome normalized to input datasets. Processed and merged alignments were subjected to bamCoverage to generate signal files. RPKM values were calculated for 20 bp bins for each sample and smoothed using a 60 bp sliding window. The generated signal files were visualized in R, using the ggplot2 and ggbio packages.

3.3. Genomic Annotation of RNA-DNA hybrids

We used GenomicRanges 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.

3.4. *In silico* restriction enzyme digestion

To calculate the expected fragment length distribution generated by a combination of restriction enzymes, we cut the human and yeast genomes *in silico* with the DECIPHER R package. 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.

3.5. Spp1 ChIP experiments

50 ml of meiotic cells 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. Cells were resuspended in 500 µl of lysis buffer 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. To obtain whole-cell extract, a 50 µl pre-immunoprecipitation sample was removed and centrifuged at full speed for 10 s to pellet the cell debris. 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, corresponding to 50 µl or 2×10^7 beads, which were preincubated with the 9E11 (monoclonal mouse anti-myc, ab56) or anti-GFP (polyclonal rabbit, ab290) 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, 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 for 3h at 65°C. Nucleic acids were PCR clean up kit purified, and RNA digestion was performed for 1 h at 37°C. The DNA was finally resuspended in 50 µl nuclease-free dH2O.

3.6. 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 and sequenced. Raw reads were aligned to the *S. cerevisiae* reference genome using the default parameters of BWA algorithm and 38–67% of the sequenced reads were retained after removing low mapping quality and PCR duplicate reads.

3.7. 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.

3.8. Statistical analysis

All statistical analyses were performed in R. Group comparisons were performed by ANOVA. Groups were compared with Tukey's post-hoc test. If the data did not fit the normal distribution, we used Kruskal-Wallis's ANOVA and the Mann-Whitney *U* test. Probability values of $P \leq 0.001$ were considered as statistically significant. Significance marks: not significant (ns). $P > 0.05$; *, $P \leq 0.05$; **, $P \leq 0.01$; ***, $P \leq 0.001$; ****, $P \leq 0.0001$.

4. Results

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

4.1.1. 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 forty DRIP experimental schemes so that we assess how they rank different test loci according to their known RNA-DNA hybrid status. The classifiers 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 fixation, *ii.* the method of nucleic acid isolation, *iii.* removal of free RNA, *iv.* the mode of nucleic acid fragmentation and *v.* cell lysis temperature.

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

To derive the parameters of the DRIP classifiers, known positive and negative examples 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 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. 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. 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), and the identified R-loop binding sites significantly overlapped within each group, but sharply stood apart between the two groups. We attribute these differences to the extensive variation of R-loop lengths and heterogeneities of the studied cell types. With the observed variances in mind, our consensus R-loop set was regarded as an amenable reference to benchmark the DRIP classifiers.

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

Positive and negative test regions were selected from the identified R-loop set and were systematically probed for RNA-DNA hybrid enrichment across the DRIP classifiers. Five test regions were frequently used as positive and negative controls in various published DRIP studies (*SNRPN, ZNF554, MYADM, FMR1, APOE*; while the remaining sites were picked at random from the consensus R-loop set (*PRR5L, LOC440704, NOP58, VIM, ING3*).

DRIP-qPCR yields were measured in control and RNase H-treated samples for forty DRIP classifiers, at ten test regions, in five independent experiments. The resulting 4000 DRIP enrichment scores were then readily used as an input parameter of receiver operator characteristics (ROC) calculation.

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

We quantitated the relative trade-offs between true positive hits and experimental errors by performing ROC analysis on the DRIP-qPCR screen characterizing the classifiers. 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. 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 with a sensitivity of 68.5-75 % and specificity of 68-79 %. Similar ROC parameters were obtained in a repeated experiment using a B lymphoblastoid cell line, demonstrating the reliability of the tested DRIP protocols in other cell types.

Pairwise comparison of the main experimental variables 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 did not change the specificity and sensitivity of the DRIP assay. 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. We explain the adverse effect of RNase A by its reported DNA binding activity that selectively eliminates a vast amount of melted DNA regions upon nucleic acid purification. We confirmed the strong DNA binding of RNase A as migration defects on DNA gels, when a plasmid DNA was incubated with the enzyme. 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 we found a statistically significant difference ($p=0.0002$) in the ROC parameters, suggesting that sonication in more efficient in discriminating true positive signals from false positives, at least within the tested conditions.

4.1.5. 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.

Biased genome sampling, related to the non-random distribution of restriction enzyme recognition sequences, was even more pronounced over exons especially over the first exons. 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 which was significantly reduced over gene coding regions. 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. Precise genomic position of R-loops could be resolved by sonication.

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

4.2.1. 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 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, indicating a group of Spp1 molecules associated with Set1C during meiosis.

Overall, Spp1 & Bre2 peaks and Bre2-only peaks show strong enrichment on ribosomal protein genes, snoRNA/ncRNA genes and transcription start sites, but they are absent from Mer2/Red1 axial sites. 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. Although Spp1 binding sites appear to be more dynamic than common sites, the latter peaks show much higher ChIP signal compared to Spp1-only or Bre2-only sites.

To gain more mechanistic insights into the dynamics of Spp1, we performed unsupervised clustering analysis 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: dynamic sites, which gradually appeared or disappeared as meiosis progressed, and static sites 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.

Functional annotation revealed that i) appearing Spp1 peaks are strongly enriched at chromosome axial sites ii) disappearing Spp1 sites are enriched at RPG and snoRNA genes and iii) constant Spp1 peaks show strong association with ncRNAs.

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.

4.2.2. 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. Both modifications are expected to phenocopy the meiotic phenotype of the Spp1PHDΔ mutation.

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, all four mutations eliminate about 50% of Spp1 binding sites during the meiotic time-course identified in the wild type strain.

We next performed multidimensional scaling analysis on the identified binding sites to highlight temporal and cell type-specific differences in Spp1 chromosomal localization. Wild

type cells and Spp1 PHD- and CxxC-domain mutants behave very differently at the beginning of sporulation. Then, in the first two hours, there will be 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.

In the histone mutant backgrounds, Spp1 binding sites are more like the wild type at the beginning of sporulation. Subsequently, fast and dynamic changes occur in the first few hours such that both mutants quickly move away from the wild type. By the end of the process all three cell types are characterized by a different Spp1 state.

We next analysed the overlap of Spp1 binding sites with annotated functional genomic elements in each mutant. 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 and abrogate the association of Mer2 with the dynamic clusters of Spp1 peaks. 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. 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.

We also showed 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 about 75% of appearing Spp1 peaks detected in wild type cells, however, about half of the remaining Spp1PHDΔ sites still exhibit significant Mer2 enrichment. This contrasts 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. Above the appearing Bre2 binding sites a clearly low Mer2 signal was detected.

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).

5. Discussion

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

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 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 were some DRIP workflows performed unreliably and generated random answers: 2, 10, 11 and 16.

By testing the main parameters of the DRIP experimental scheme - involving 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 overrepresentation of large DRIP fragments, over coding regions, which is especially over first exons. This

phenomenon severely compromised mapping resolution and, therefore, the assignment of clear biological function to a fraction of R-loops. 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.

Based on the above experiences, 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, followed by organic extraction and immunoprecipitation with the S9.6 antibody. 3. If restriction enzyme fragmentation needs to be applied, 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 or lambda exonuclease digestion, 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.

5.2. Functional analysis of Spp1 chromatin binding during meiosis.

By capturing chromosomal binding sites of Spp1 while tracking Set1C with Bre2, we revealed that a specific subpopulation of Spp1 behaves 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 enables the Spo11 to generate DSBs. Disappearing Spp1 sites were 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.

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 in the framework of meiotic chromatin structure.

6. Summary

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

- 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 region. 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 over first exons. This biased genome sampling compromising the mapping resolution and effects the precise annotation of a subset of RNA-DNA hybrids. If used, we advise to check the fragment size distribution both *in silico* and *in vitro*.

6.2. Functional analysis of Spp1 chromatin binding during meiosis.

- 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).

7. Publications Related to Dissertation

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

8. List of Other Publications

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

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