Functional characterization of HMGN5 in chromatin architecture and gene expression

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

vorgelegt von

Ingrid Carolina Araya Fuenzalida

aus Santiago de Chile

Im Jahr 2018

Das Promotionsgesuch wurde eingereicht am:
02.03.2018

Die Arbeit wurde angeleitet von:
Prof. Dr. Gernot Längst

Regensburg, 02.03.2018

_____
 Ingrid Carolina Araya Fuenzalida

To Oskar Nahuel, without whom I would have finished this thesis one year before, and to Roman, without whom I would have probably never finished.

"La rebeldía tiene la potencialidad de crear, de imaginar y de proyectar, porque es una energía infinita y transformadora que piensa antes que en un cuarto propio en un cuerpo propio. La rebeldía es el comienzo de la libertad"

(Margarita Pisano Fisher)

Table of contents

Content

Content

List of figures

List of figures

List of figures

List of tables

Abbreviations

List of abbreviations

°C	Degree Celsius
Δ	Deletion
6-FAM	6-carboxyfluorescein
Å	Ångström
aa	Amino acid
Ab	Antibody
ADP	Adenosine diphosphate
Amp	Ampicillin
ANOVA	Analysis of variance
APS	Ammonium persulfate
AS	Alternative splicing
ATP	Adenosine triphosphate
BANF1	Barrier to autointegration factor 1 / BAF
BLAST	Basic Local Alignment Search Tool
bp	basepairs
BSA	Bovine serum albumin
caRNA	Chromatin-associated RNA
cDNA	Complementary DNA
CDS	Coding sequence
ChIP	Chromatin immunoprecipitation
CLIP	UV-crosslinking RNA immunoprecipitation
cm	Centimeters
CMV	Cytomegalovirus
Co-IP	Co-Immunoprecipitation
CpG	Cytosine-phosphate-guanine
CPM	Counts per million
Cryo-EM	Cryogenic electron microscopy
CTCF	CCCTC-binding factor
CTs	Chromosome territories

List of Abbreviations

Cy3	Cyanine 3
Cy5	Cyanine 5
DAPI	4',6-diamidino-2-phenylindole
Df31	Decondensation factor 31
DGE	Differential gene expression
DHSs	DNase I Hypersensitive Sites
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DNAme	DNA methylation
DNMTs	DNA methyltransferase
dNTP	Deoxynucleotide triphosphate
DTT	Dithiothreitol
E.coli	Escherichia coli
EDTA	Ethylenediaminetetraacetic acid
EGR1	Early Growth Response 1
EGTA	Ethylene glycol-bis(β-aminoethyl ether)-N,N,N',N'-tetraacetic acid
EMSA	Electrophoretic mobility shift assay
ENCODE	Encyclopedia of DNA Elements
ER	Estrogen receptor
EtBr	Ethidium bromide
EZH2	Enhancer of Zeste homolog 2
FBS	Fetal bovine serum
FC	Fold change
FISH	Fluorescent in situ hybridization
FRT	Flp recombination target site
Fw	Forward
g	Relative centrifugal force (RCF)
GAPDH	Glyceraldehyde-3-Phosphate Dehydrogenase
GFP	Green fluorescent protein
GO	Gene ontology
GOI	Gene of interest
GST	Glutathione S-Transferase
H1	Linker histone H1

List of Abbreviations

H3K27ac	Acetylation of lysine 27 on the histone H3
H3K4me3	Trimethylation of lysine 4 on the histone H3
HAS	High-affinity sites
HATs	Histone acetyltransferases
HDACs	Histone deacetylases
HF	High Fidelity
His	Histidine
HMG	High mobility group
HMGA	High Mobility Group AT-Hook
HMGB	High mobility group Box
HMGN	High Mobility Group Nucleosome Binding
HMGN5	High Mobility Group Nucleosome Binding Domain 5
HMTs	Histone methyltransferases
HRP	Horseradish peroxidase
HS	High sensitivity
iBAQ	Intensity-based absolute quantification
IDR	Intrinsically disordered region
IF	Immunofluorescence
IGF2/H19	Insulin Like Growth Factor 2/ H19, Imprinted Maternally Expressed Transcript
IP	Immunoprecipitation
IPTG	Isopropyl β-D-1-thiogalactopyranoside
Kb	Kilobase
Kd	Knockdown
kDa	KiloDalton
LacI	*lac* repressor
LacO	l*ac* operon
LADs	Lamina-associated domains
LAP	Lamina-associated protein
LAP2α	Lamina-associated polypeptide 2-alpha
Lap2β	Lamina-associated polypeptide 2, isoforms beta/gamma
LAS1L	LAS1 Like, Ribosome Biogenesis Factor
LB	Luria-Bertani
LCMS/MS	Liquid chromatography-tandem mass spectrometry

List of Abbreviations

LCS	Low-complexity sequences
lncRNA	Long non-coding RNA
LOCK	Large organized chromatin K modification
M	Molar
mA	Milliampere
MBD	Methyl-CpG Binding Domain Protein
MeCP2	Methyl-CpG Binding Protein 2
MCP	MS2 coat protein
min	Minute
miRNA	microRNA
mJ	Millijoules
MM	Master mix
MNase	Micrococcal nuclease
mRNA	Messenger RNA
MSL	Male-specific lethal
MST	Microscale thermophoresis
MW	Molecular weight
MWCO	Molecular weight cut-off
NBD	Nucleosomal binding domain
NBP-45	Nucleosomal binding protein 45
ncRNA	Non-coding RNA
NE	Nuclear envelope
Ni-NTA	Nickel-nitrilotriacetic acid
NL	Nuclear lamina
NLS	Nuclear localization signal
nm	Nanometer
NOL9	Nucleolar Protein 9
NOR	Nucleolar organizer region
NSBP1	Nucleosomal Binding Protein 1
nt	Nucleotides
OD600	Optical density measured at 600 nm
OE	Overexpression
ON	Overnight

List of Abbreviations

ORF	Open reading frame
PAA	Polyacrylamid
PAGE	Polyacrylamide gel electrophoresis
PARD3	Par-3 Family Cell Polarity Regulator
PARP1	Poly [ADP-ribose] polymerase 1
PBS	Phosphate-buffered saline
Pc	Polycomb
PCR	Polymerase chain reaction
PELP1	Proline, Glutamate and Leucine Rich Protein 1
PFA	Paraformaldehyde
PI	Protease inhibitor mix
piRNA	PIWI-interacting RNA
PML	Promyelocytic leukaemia protein
PMSF	Phenylmethylsulfonyl fluoride
PolII	RNA polymerase II
PRC2	Polycomb repressive complex 2
PTMS	Parathymosin
PTMs	Posttranslational modifications
PVDF	Polyvinylidene fluoride
qPCR	Quantitative Real-Time PCR
Rb	Rabbit
RBPs	RNA-binding proteins
RD	Regulatory domain
rDNA	Ribosomal DNA
RNA	Ribonucleic acid
RNP	Ribonucleoprotein
roX	RNA on the X
RPL30	Ribosomal Protein L30
rpm	Revolutions per minute
rRNA	Ribosomal RNA
RT	Room temperature
Rv	Reverse
SAXS	small-angle X-ray scattering

List of Abbreviations

SDM	Site-directed mutagenesis
SDS	Sodium dodecyl sulfate
sec	Seconds
SENP3	SUMO1/Sentrin/SMT3 Specific Peptidase 3
Seq	Sequencing
siRNA	Small interfering RNA
Snf2H	Snf2 homolog protein
snoRNA	Small nucleolar RNA
snoRNP	Small nucleolar ribonucleoprotein
SOC	Super Optimal broth with Catabolite repression
ssDNA	Single-stranded DNA
ssRNA	Single-stranded RNA
SWI/SNF	SWItch/Sucrose Non-Fermentable
TADs	Topologically associated domains
Taq	Thermus aquaticus
TARDBP	TAR DNA Binding Protein
TBE	Tris/Borate/EDTA
TCEP	Tris(2-carboxyethyl)phosphine
TEMED	Tetramethylethylendiamine
Tet	Tetracycline
TEV	Tobacco Etch Virus
TEX10	Testis-Expressed Protein 10
TF	Transcription factor
TRIS	Tris(hydroxymethyl)aminomethane
TSS	Transcription start site
TTF-I	Transcription Termination Factor I
U	Units
UTR	Untranslated region
UV	Ultraviolet
V	Volts
VGF	VGF Nerve Growth Factor Inducible
WB	Western blot
WCE	Whole cell extract

WDR18	WD Repeat Domain 18
wt	Wildtype
XCI	X chromosome inactivation
XIST	X-inactive specific transcript
ZCCHC12	Zinc Finger CCHC-Type Containing 12

1 Summary

The modulation of higher order structure of chromatin has profound implications in the regulation of nuclear processes, like transcription, replication, recombination or DNA repair. Those processes require accessible DNA to recruit large protein complexes to function.

In humans, the architectural proteins of the "high mobility group nucleosomal binding domain" (HMGN) family participate in the opening of chromatin structure and the regulation of gene expression. HMGN proteins bind the nucleosome particle through a conserved nucleosomal binding domain (NBD) and compete with H1 for the binding to chromatin.

Of all HMGN family members, HMGN5 has the biggest effect on transcriptional regulation in mouse and human cells. Moreover, HMGN5 is also able to induce large-scale chromatin decondensation *in vivo.*

In the present work we study the functional role of HMGN5 in the opening of higher-order structure of chromatin and gene expression by biochemical and genome-wide methods.

We identified a novel and specific RNA binding domain overlapping with the NBD of HMGN5. Moreover, by *in vitro* competition assays we demonstrated that HMGN5 exhibits exclusive binding to nucleosomes or to RNA. Furthermore, we showed that the RNA binding activity is a feature of other HMGN members as well, highlighting a novel function for those proteins.

The overexpression and knockdown of HMGN5 in human cell lines affect the expression of about 3000 genes respectively, with 1287 overlapping target genes. ChIP-seq analysis of HMGN5 revealed that HMGN5 mainly associates with active regulatory genomic regions, like promoters and CpG islands, and it localizes to DNase I hypersensitive sites (DHSs). Moreover, we found that the actively regulated target genes belong to the group of genes involved in RNA metabolic processes. HMGN5 binding overlaps with RNA polymerase II binding sites. CLIP-seq analysis of HMGN5-bound RNAs revealed that HMGN5 is able to bind nascent transcripts. In the light of the biochemical results, we propose that HMGN5 participates in the regulation of RNA metabolism by a dual mechanism that enables HMGN5 binding either

Summary

chromatin or RNA since the HMGN5-bound RNAs have no functional relationship with the chromatin function of HMGN5.

Interestingly, by using quantitative mass spectrometry we identified CTCF, BANF1 and seven proteins associated with the pre-ribosomal RNA processing.

Strikingly, our ChIP-seq data revealed that HMGN5 co-localizes with HMGN5. As CTCF constitutes the major organizer of chromatin architecture, the results presented here suggest a cooperative role of both proteins in the organization of higher-order structure of chromatin.

Further functional characterization of the potential HMGN5-CTCF complex, may shed light on the regulation of higher-order chromatin organization.

2 Introduction

2.1 Chromatin as functional organizer of DNA in the nucleus

In eukaryotic cells, the DNA molecule needs to be tightly packed to fit into the nucleus which has a diameter of 10µm, and yet be accessible to allow regulation of processes like transcription, replication or DNA repair. This is possible trough the highly specialized and dynamic packaging of DNA in the chromatin fiber. Chromatin was first discovered and named by Walther Flemming in the 19th century as a stainable fibrous structure within the nucleus of cells. Later, in 1928 the German botanist Emil Heitz made the first categorization of chromatin by discriminating the chromosomes in euchromatin (genetically active) and heterochromatin (genetically inactive) based on its staining properties under the light microscope (Passarge, 1979).

Besides the general categorization in euchromatin and heterochromatin, the chromatin is organized in specialized, spatiotemporal changing compartments that are required for the proper function of all DNA-dependent processes.

Several regulatory mechanisms like ATP-dependent chromatin remodelers, posttranslational modification of histone, DNA modification or regulatory RNAs have been described –globally termed epigenetic mechanisms-, that control the fate of gene expression, by modulating locally (at nucleosomal level) or globally (by long distance chromatin interactions) chromatin accessibility (Bartkuhn & Renkawitz, 2008). This highlights the crucial role of the coordinated and dynamic regulation of chromatin organization on the transcriptional regulation of genes.

2.2 DNA packaging and higher-order chromatin organization

The nucleosome corresponds to the basic level of chromatin organization, and is composed of a nucleosome core particle (NCP) a linker DNA. The nucleosome core particle is composed of 147 bp of two tight 1.65 left handed superhelical turns of DNA wrapped around a histone octamer (Figure 1) which is composed of two copies of the histones H2A, H2B, H3 and H4 (Luger, 2003). In the NCP, the core histones are assembled into four heterodimers, two H2A/H2B and two H3–H4 dimers (Luger, Mäder, Richmond, Sargent, & Richmond, 1997; A. L. Olins & Olins, 1974; Woodcock, Safer, & Stanchfield, 1976). The core histones are evolutionary conserved basic small proteins (ranging between 11-16kDa), that have 2 characteristic functional domains; a histone fold domain mediating histone-DNA and histone-histone interactions required to form the nucleosome particle, and the N-terminal histone tails (and C-terminal tail in H2A and H2B) that are disordered, flexible and accessible structures protruding from the nucleosome core particle (Downs, Nussenzweig, & Nussenzweig, 2007). The histone tails are hotspots of regulatory posttranslational modifications.
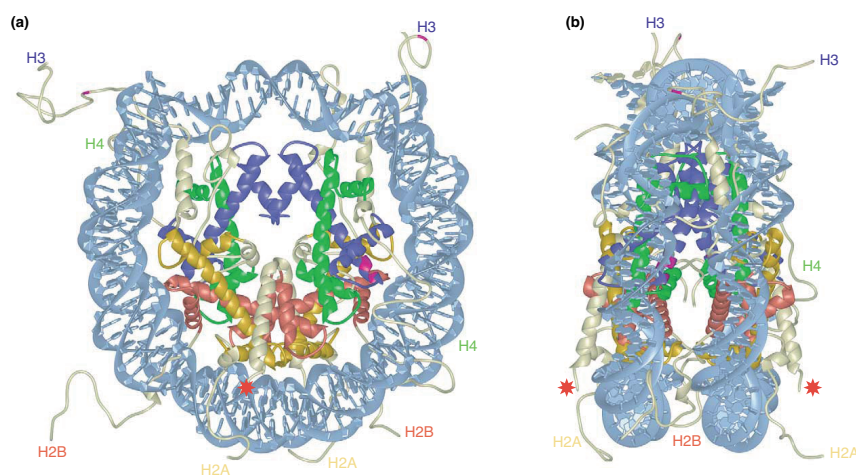
Figure 1. Structure of nucleosome core particle.

Scheme of the nucleosomal structure at the 2.8 Å resolution level obtained from X-ray crystal structure. A) Front view of the nucleosome particle, formed by 147bp of double-stranded DNA (in light blue) wrapped around a histone octamer composed of the histone H2A (yellow), H2B (red), H3 (blue) and H4 (green). The respective histone tails extensions are shown. Red star indicates site of ubiquitination in yeast. B) Side view of the nucleosome with a 90° rotation in the vertical axis. Image obtained after (Luger, 2003).

Histones interact with DNA through different mechanisms including salt bridges, hydrogen bonds with DNA, non-polar contacts with the deoxyribose, electrostatic interactions between the positively charged N-terminal tails with the DNA phosphates and base-specific contacts (Widom, 1998).

The nucleosome core are separated by a flexible DNA linker–with a variable length between 20-80bp depending on the species and the cell type (Felsenfeld & Groudine, 2003)-. This array of nucleosomes forms the so called "10 nm fiber" that resembles a "beads-on-a-string" structure under electron microscopy (Figure 2) at low salt conditions (A. L. Olins & Olins, 1974; Woodcock et al., 1976). A diploid human cell that contains about 6×10^9 nucleotide pairs (with a total average length of 2m), DNA is packaged on average into 30 millions nucleosome cores.

50 nm
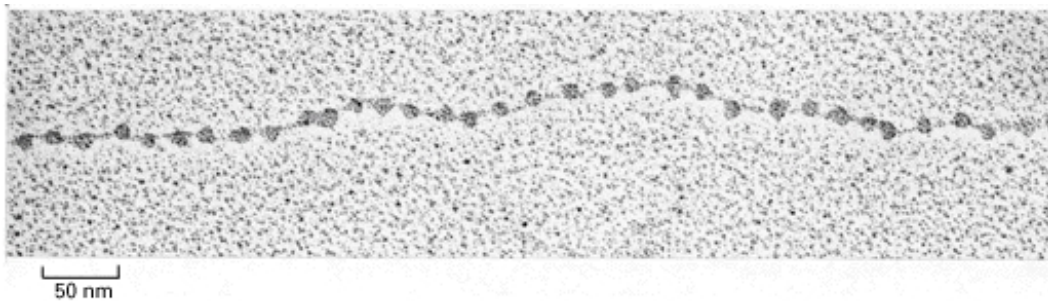
Figure 2. Electron micrograph of the chromatin "beads-on-a-string" structure.

Decondensed "beads-on-a-string" form of purified chromatin visualized by electron microscopy. Picture adapted from Molecular Biology of the Cell. 3rd edition. (Alberts et al., 1994)

2.3 Higher-order chromatin organization

Inside the cell nucleus, chromatin is not linear in structure, but it is rather organized into 3-dimentional higher-order structures. This higher-oder organzation plays critical roles for the regulation of nuclear functions.

At a first level, the binding of the linker histone (H1 and H5) helps the packaging of nucleosomes into the chromatosome core particle (Harshman, Young, Parthun, & Freitas, 2013; Simpson, 1978). Linker histones are composed of a tripartite structure, consisting of a flexible and short N-terminal tail, a conserved central globular domain, and a long (100 amino acids) intrinsically disordered and basic C-terminal domain (Allan, Hartman, Crane-Robinson, & Aviles, 1980).

By using cryo-microscopy (cryo-EM) and crystal structures it was recently shown that the C-terminal domain of H1 associates primarily with a single linker in the nucleosome, while the globular domain interacts with both DNA linkers and the nucleosome dyad, resulting in a reduction of the flexibility of linker DNA and a more compact nucleosome conformation (Bednar et al., 2017). It has been shown that depletion of H1 is lethal in mice (Fan et al., 2003) and *Drosophila* (Lu et al., 2009). Moreover, H1 depletion alters the proper folding of chromosmes during mitosis (Maresca, Freedman, & Heald, 2005) highliting its role in the higher-order chromatin conformation.

2.3.1 Classical view of chromatin higher-order organization

It has been described that chromatin is highly packaged into a hierarchical higher-order structure. In the classcial text-book view of chromatin organization (Figure 3), the nucleosomal array is folded into thought to fold into intermediate fibers of increasing diameter of 30, 120 300 and 700 nm diameter, to finally form the mitotic chromosome, eye-visible under the light microscope.

Two competing models have been proposed to explain the formation of the 30 nm fiber. The solenoid model, in which consecutive nucleosomes are located adjacent to one another in the fiber, and fold into a simple "one-start helix" (Finch & Klug, 1976), and the zig-zag model, which assumes an arrangement of the nucleosomes in a zigzag manner in which two rows of nucleosomes form a "two-start helix" so that

alternate nucleosomes interacts with the neighbor nucleosome of the other row (Dorigo et al., 2004; Horowitz et al., 1994; Worcel et al., 1981). Both models assume a selective internucleosomal interaction of close neigbors nucleosomes on the DNA strand (Maeshima, Imai, Tamura, & Nozaki, 2014).

Despite that for many years the hierarchical folding of chromatin has been widely accepted, the chromatin folding *in vivo* is still a controversial topic.
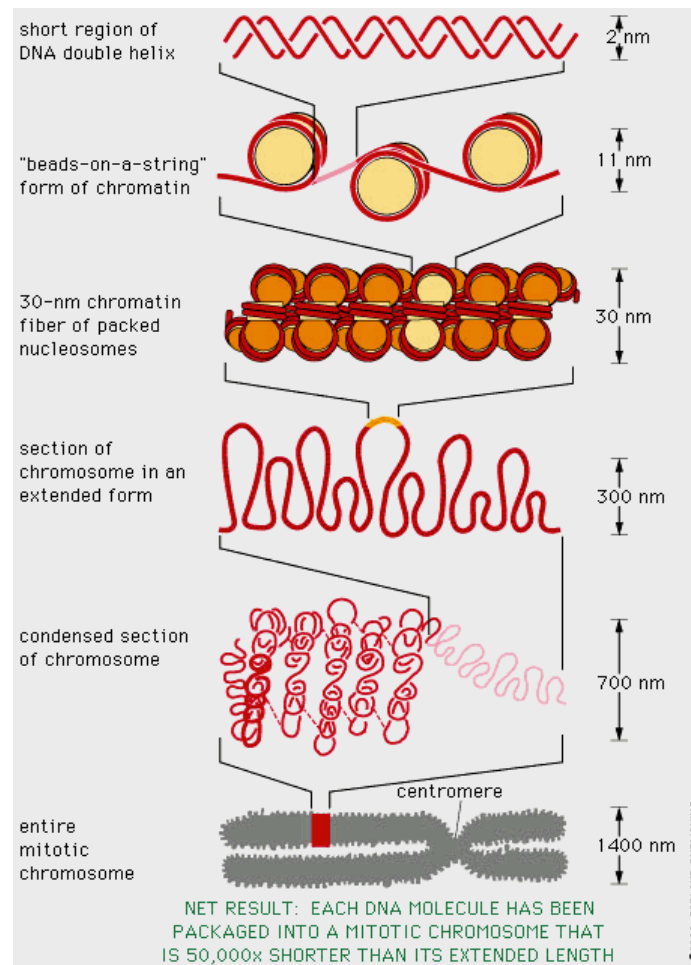
Figure 3. Hierarchical higher-order compaction of chromatin.

DNA is wrapped around histone octamers forming nucleosomes. Individual nucleosomes are separated by free linker DNA and associate with histone H1. The 10 nm fiber twists into a large coil, generating the condensed, supercoiled 30 nm fiber. The coils form loops (300-nm fiber) and the loops coil further, producing the metaphase chromosome as highest condensation level of DNA in eukaryotes (image and description taken from Alberts, 3rd edition.)

2.3.2 A new concept of higher-order chromatin structure

Over years, the formation of the 30 nm fiber has been supported by several biophisical studies, including X-ray crystallography, or small-angle X-ray scattering (SAXS), and the 30 nm chromatin conformation has been the reference for many studies in the chromatin field. However, despite all the efforts made to unravel the specific conformation of chromatin *in vivo* it is still not clear if the 30 nm fiber really exist *in vivo* (Fussner et al., 2011; Luger, Dechassa, & Tremethick, 2012; Maeshima, Hihara, & Eltsov, 2010; Nozaki et al., 2014). This is because the classical methods have technical limitations, as they are mainly based on *in vitro* structure of reconstituted chromatin from DNA and histones (P. J. J. Robinson., 2006; Schalch et al., 2005), or based on chromatin purified from permeabilized cells (Belmont & Bruce, 1994; Horowitz et al., 1994; Worcel et al., 1981), lacking many components of the physiological chromatin context.

By the use of cryo-electron microscopy (cryo-EM) analysis of vitrified chromosomes it was suggested, already three decades ago, that chromatin was irregularly folded *in vivo* (McDowall et al., 1986).

Furthermore, the latest evidence in the field strongly argues against the existence of the 30 nm fiber *in vivo* (Bouchet-Marquis, et al., 2006; Eltsov et al., 2008; Nishino et al., 2012; Ou et al., 2017).

By using a sofisticated method that combines electron microscopy tomography and labeling methods (ChromEMT), the 3D chromatin ultrastructure in the nucleus of living human cells was analyzed for the first time. In the study performed by Ou and colleagues it was revealed that chromatin is organized as irregular polymers with diameters of 5 to 24 nm in interphase and mitotic chromosomes (Ou et al., 2017), instead of the classically described higher-order structures.

Those results are in agreement with the recently proposed "polymer melt model" (Figure 4), in which chromatin is organized as flexible and disordered dynamic folded 10 nm fiber similar to a "polymer melt" (Maeshima et al., 2010). In this model, at low nucleosome concentration, nucleosome fibers can form 30 nm fibers mediated by intra nucleosome interactions, which can explain the observed *in vitro* formation of the 30 nm fibers. However, as the nucleosome content increases nucleosomes form inter fiber associations (due to increased cation concentration or molecular crowding effect) leading to the formation of a polymer-like structure (Nishino et al., 2012). The

"polymer melt" model represents several advantages in determining accessibility of the DNA to the recruiting of regulatory machineries and for the formation of functional chromatin domains through long-distance chromatin interactions.
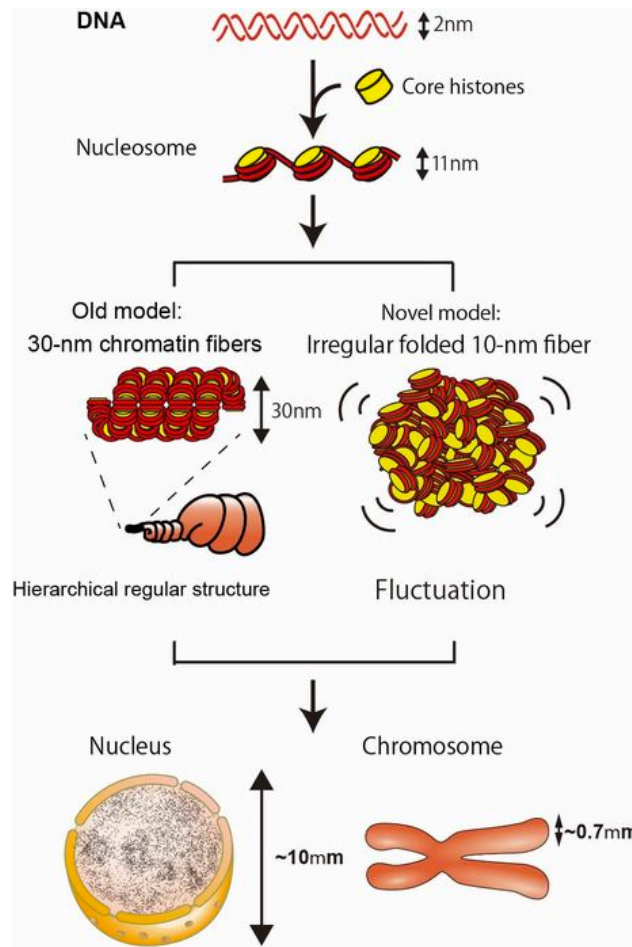
Figure 4. Classical and new model of higher order chromatin folding.

Comparison between the classical view of a hierarchical chromatin folding (left), and the new model of irregularly folded nucleosome fibers ("polymer melt"). In the classical view, nucleosomes are organized in regular fibers of 30 nm, which are subsequently packaged to get a highly compacted mitotic chromosome inside the nucleus. In the novel model, chromatin is organized as irregular folded 10 nm fiber which implies a flexible and less constrained chromatin organization.
Picture taken from:
 https://www.nig.ac.jp/labs/MacroMol/e-more_detailed_description.html (Maeshima et al., 2014)

2.4 Mechanisms regulating chromatin structure

Epigenetic regulatory mechanisms involve posttranslational modification of histones, canonical histone replacement by histone variants, DNA methylation, non-coding RNA and ATP dependent chromatin remodeling. Each mechanism has been extensively studied, revealing that they are highly coordinated, functionally interacting and influencing each other (Armstrong, 2013). The actors in the epigenetic marking system, have been classified as "writers", "erasers", and "readers" due to their ability to add, remove or recognize, respectively, histones or DNA chemical modifications (marks) to establish the gene expression program of the cell (Torres & Fujimori, 2015).

2.4.1 Histone posttranslational modification

Histones are marked by a large number of posttranslational modifications (PTMs), including acetylation, methylation, phosphorylation, ubiquitination, citrullination, sumoylation, ADP-ribosylation and proline isomerization (Rothbart & Strahl, 2014). Most of the modifications are located in the unstructured N-terminal and C-terminal tails that protrude from the nucleosome particle. However, some modifications have been found in the histone-fold domain, required for nucleosome unwrapping and disassembly (Simon et al., 2011).

PTMs have been associated with specific functions, but they are mainly found in combinations, suggesting a "histone code" that finally define a specific chromatin state (Strahl & Allis, 2000) by the recruitment of modifying enzymes, such as histone acetyltransferases (HATs), histone methyltransferases (HMTs,) or histone deacetylases (HDACs), among others, that alter the structure of the surrounding chromatin environment.

Of the PTMs, the mostly described histone modification is lysine methylation. Lysine can be mono-, di- or tri-methylated, and these modifications can be associated with either gene activation or silencing depending on possition and the cross-talk with other PTMs (Torres & Fujimori, 2015). As an example, mono- and trimethylation of the lysine 4 from histone H3 (H3K4) as well the H3 lysine 36 trimethylation

(H3K36me3) are associated with transcription activity, whereas H3K27me3 and H3K9me2 have been associated with transcriptional repression (Woo, Ha, Lee, Buratowski, & Kim, 2017). Specific combinations of histone modifications are hallmarks of the active/inactive regulatory genomic elements like promoters, enhancers, or intron/exon boundaries (V. W. Zhou, Goren, & Bernstein, 2010). A summary of the different histone marks and their associated functions is depicted in (Figure 5).

.
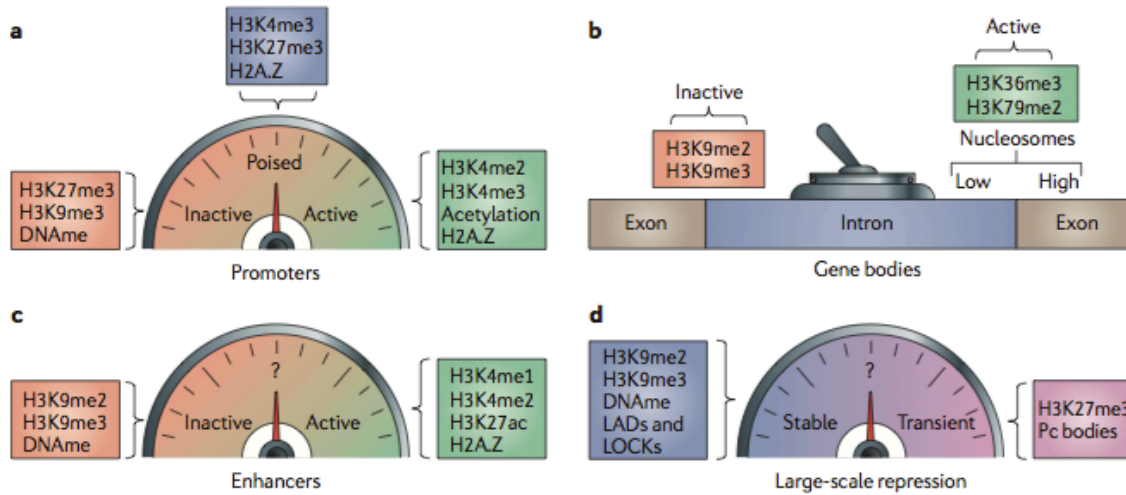
Figure 5. PTMs associated with different chromatin states.

A) At promoters, PTMs contribute to fine-tune gene activity from active to poised to inactive states. B) At gene bodies, they discriminate between active and inactive conformations. C) At distal sites, histone marks correlate with levels of enhancer activity. D) On a global scale they may confer repression of varying stabilities and be associated with different genomic features. For example, lamina-associated domains (LADs) in the case of stable repression and Polycomb (Pc) bodies in the case of context-specific repression. DNAme, DNA methylation; LOCK, large organized chromatin K modification. Figure and description legend obtained after Zhou et al., 2011.

2.4.2 Histone variants

Additionally to the histone PTMs, chromatin can be regulated by the replacement of canonical histones by histone variants.

Histone variants are non-allelic isoforms of the canonical histones that have a specific expression, localization, specie specific distribution pattern (Kamakaka & Biggins, 2005) and play determinant roles in regulating chromatin structure.

Different than canonical histones, that are produced during the DNA synthesis (S) phase of the cell cycle, histone variants are expressed throughout the cell cycle, and are incorporated to chromatin in a replication-independent manner by specific histone chaperones (Biterge & Schneider, 2014).

The genes coding for the (known) histone variants contain intronic sequences, the pre-mRNAs are polyadenylated and can undergo alternative splicing (Biterge & Schneider, 2014).

It is believed that the deposition of histone variants alters the stability of the nucleosome, affecting the interactions between histones and with the DNA, thus helping to modify the chromatin conformation (Gautier et al., 2004; Park, Dyer, Tremethick, & Luger, 2004)

Histone variants can differ from their canonical counterparts by minor modifications (as one amino acid change in the H3 variants H3.1 and H3.2) to drastic structural modifications as in the H2A variant macroH2A (E. Bernstein & Hake, 2006).

To datel, all histones, with the exception of H4 have described histone variants. The most studied histone variants belong to the H2A family and they are involved in diverse cellular processes, associated with activation or repression of transcription, as DNA damage response and centromeres formation (E. Bernstein & Hake, 2006). Among them, H2a.Bbd (associated with active transcription), H2A.X (involved in DNA repair and genome integrity), H2A.Z (activation and repression of transcription and chromosome segregation) and macroH2A which have been associated with X-chromosome inactivation and transcriptional repression (Biterge & Schneider, 2014; Sarma & Reinberg, 2005).

The number of histone H3 variants differs among species. In mammals there are four main isoforms, H3.1, H3.2, H3.3 and CENPA (found at centromeric chromatin) (Filipescu, Mueller, & Almouzni, 2014). However, recent studies have revealed an extended group of H3 variants that possess a tissue-specific and specie-specific

expression pattern. Those variants include the testis-specific histones H3.4, and H3.5, and the histone H3.Y which is conserved among primates (Filipescu et al., 2014).

For many of the histone variants, posttranslational modifications have been described, contributing to the modulation of gene expression (Biterge & Schneider, 2014).

2.4.3 DNA methylation

In vertebrates, DNA methylation is key for regulation of different processes, like gene expression, genomic imprinting, silencing of transposable elements or X-chromosome inactivation. This modification occurs extensively in CpG dinucleotides and is mediated by a group of enzymes called DNA methyltransferases (DNMTs) that catalyze the methylation at the carbon-5 position of cytosine residue at the CpG dinucleotide to form 5-methylcytosine (5-mC) (Prokhortchouk & Defossez, 2008). Methylation of DNA is essential for development and viability in mammals (Jackson-Grusby et al., 2001; Okano, Bell, Haber, & Li, 1999). In humans the CpG methylation patterns can be categorized in two groups; the first group, covering the vast majority of the genome (98%), possesses low CpG frequency (1 each 100bp) but highly methylated. Those CpGs are generally associated with transcriptionally repressed chromatin. In the second group (covering 2% of the genome) the CpGs are highly concentrated in the so-called CpG islands (CGIs), found in a ratio of 1 CpG every 10bp in DNA stretched of about 1000bp (Illingworth et al., 2008). The human genome contains about ~30000 CGIs and most of them (around 21000) are found close to transcription start sites (TSSs) of genes and remain unmethylated which correlates with transcriptional activity. However, about 9000 CGIs are found at gene body, and the methylation at those sites have been associated with enhanced gene expression levels (Ball et al., 2009; Illingworth et al., 2008; Krinner et al., 2014) indicating that DNA methylation is a versatile mark with functions that depend on the genomic context. It has been described that the silencing mediated by DNA methylation can occur by two different mechanisms; DNA methylation can act by masking the genome and therefore preventing the binding of transcription factors, or the

methylation readout can be performed by methyl-binding proteins ("readers") like Methyl-CpG Binding Domain Proteins (MBD1, MBD2 and MBD4), and Methyl-CpG Binding Protein 2 (MeCP2) that can recruit chromatin remodeling complexes, DNA methyltransferases or histone deacetylases, that lead to transcriptional repression (Baubec, Ivanek, Lienert, & Schuebeler, 2013).

It has been lately shown that besides CpG, cytosines followed by adenine, thymine or another cytosine can also be methylated. The methylation of those cytosines is known as Non-CpG methylation and is prevalent in human embryonic stem cells (Ramsahoye et al., 2000) or brain development (Guo et al., 2014; Lister et al., 2013), however, the specific functions of those modifications are still unclear.

2.4.4 ATP dependent remodeling complexes

To allow the binding of regulatory machineries to specific genomic sites, the chromatin needs to be dynamically changed. This active process is mediated by chromatin remodeling complexes, which use ATP-hydrolysis to move, destabilize, eject or restructure nucleosomes (Clapier & Cairns, 2009; Erdel, Krug, Längst, & Rippe, 2011), thereby regulating the accessibility of DNA regulatory factors.

The chromatin remodeling complexes have ATPases, that belong to the SF2 helicase superfamily, and accessory regulatory subunits which are required for targeting and regulation of nucleosome remodelling (Erdel et al., 2011; Längst & Manelyte, 2015).

The remodelers can be classified into four families, the SWI/SNF (SWItch/Sucrose Non-Fermentable), CHD (chromodomain, helicase, DNA binding), ISWI (imitation switch) and INO80 (inositol requiring 80) family (Längst & Manelyte, 2015).

The activity of chromatin remodeling complexes and their targeting to specific genomic regions can be regulated by the interplay with different chromatin signals like DNA sequence, DNA structure, methylation, histone modifications, histone variants, and they can interact with chromatin associated proteins and complexes, as transcription factors (reviewed in Erdel et al., 2011), structural proteins that regulate chromatin compaction as H1 (Hill & Imbalzano, 2000) or HMG chromatin binding proteins (Bonaldi, et al., 2002; Heppet al., 2014; Rattner, Yusufzai, & Kadonaga, 2009). Chromatin remodeling complexes can also interact with regulatory RNAs

(Längst & Manelyte, 2015). Those interactions will help to determine the accessibility of chromatin required for the regulation of gene expression.

2.4.5 Regulatory RNAs

In the last years there have been cumulative evidence showing that non-coding RNAs (ncRNAs) are essential for regulation of chromatin architecture and gene expression. ncRNA have been classified according to their size, biogenesis and function. According to size, they are commonly (and loosely) classified into two sub-categories, small ncRNA, with an average length of less than 200nt, and long non-coding RNAs (lncRNA) with a size of >200nt (Patil, Zhou, & Rana, 2013).

Non-coding RNAs play a role in the regulation of different processes, including gene regulation, translation, splicing, cell cycle control, genome defense and chromosome structure (Brown, Mitchell, & Neill, 2011).

The mechanisms by which ncRNA participate in gene regulation include interactions with different epigenetic regulatory machineries. As an example, they can guide and modulate the activity of transcription factors, direct DNA methylation, recruit histone modifying enzymes to either activate or silence gene expression (main mechanisms described in Mattick, Amaral, Dinger, Mercer, & Mehler, 2009) or they can directly interact with DNA to modulate recruitment of transcriptional regulatory machineries (Schmitz, Mayer, Postepska, & Grummt, 2010).

It was shown already four decades ago that RNA is stably associated with chromatin in different species (Holoubek, et al., 1983; R. C. Huang & Huang, 1969; Huang & Bonner, 1965).

Furthermore it has been shown that RNA also plays a structural role contributing to the higher-order chromatin organization (Caudron-Herger et al., 2011; Rodriguez-Campos & Azorin, 2007; Schubert et al., 2012).

The group of Karsten Rippe demonstrated the existence of RNA transcripts associated with chromatin -which they called "chromatin-interlinking" RNAs or ciRNAs- that were responsible for maintaining chromatin in a decondensed and active state during interphase in human and mouse cell lines (Caudron-Herger et al., 2011).

Moreover, in our laboratory it was shown that the small nucleolar RNAs (snoRNAs) constitute the mayor fraction of chromatin-associated RNAs in *Drosophila* and human cells, and it was demonstrated that they are required to keep the higher-order structure of chromatin in an open state in *Drosophila* (Schubert et al., 2012), indicating that RNA is a key player in the global reorganization of chromatin architecture.

A well-studied example of ncRNAs involved in modulating chromatin conformation is the dosage compensation in mammas and in fly. In female mammals, the lncRNA *Xist* RNA controls the developmentally regulated chromatin-mediated X chromosome inactivation (XCI) by recruiting silencing complexes in *cis*, thus generating a compact heterocromatinized global chromatin architecture in one of the X-chromosomes (Pandya-Jones & Plath, 2016). In *Drosophila melanogaster* on the other hand, the roX RNAs (RNA on the X) are required for balancing dosage by increasing the transcription levels in the single male X-chromosome. This is mainly done by the recruitment of the male-specific lethal (MSL) complex that regulates nucleosome positioning at High-Affinity Sites (HAS) from topological associated domains (TADs), from which the complex is spread leading to the global activation of the male X-chromosome (Ramírez et al., 2015).

To date, a huge number of non-coding RNAs have been discovered –including short ncRNAs and lncRNAs- and many of them have been shown to be required for different nuclear processes -reviewed elsewhere-, however, for most of them, the mechanisms underlying their function still need to be elucidated.

Indeed the main concept that defines the functional borders between coding and non-coding RNA needs to be re-defined, as it has been shown that some non-coding RNAs have coding potential (Andrews & Rothnagel, 2014; Ruiz-Orera, Messeguer, Subirana, & Alba, 2014), and also some classically defined messenger RNAs (mRNAs) have been shown to possess non-coding functions (Kumari & Sampath, 2015; Poliseno et al., 2010; Sampath & Ephrussi, 2016) for which a dual coding/non-coding function have been proposed (Nam, Choi, & You, 2016).

2.5 Nuclear architecture and gene regulation

In the interphase nuclei the chromatin is dynamically compartmentalized. Beyond the global separation between euchromatin and heterochromatin, it has been shown that each chromosome occupies non-random territory (Bodnar & Spector, 2013) called "chromosome territories" (CTs) (Figure 6A and B). By the use of dedicated microscopy methods, together with chromosome conformation capture techniques (from 3C to Hi-C), it was demonstrated that the chromosome territories are further organized into sub-domains called Topologically Associated Domains (TADs) that are formed by long-distance regulatory interactions between the promoters and their regulatory enhancers which are often located hundred of kilobases up to megabases away (Kaiser & Semple, 2017). TADs are delimited by sharp boundaries that are enriched in housekeeping genes and insulator sites that are bound by architectural proteins as the CCCTC-binding factor (CTCF) or cohesins that have been shown to be required for the establishment of TADs through promoter-enhancers interactions through the formation of intrachromosomic loops and in less extent interchromosomal interactions (Dixon et al., 2012; Gonzalez-Sandoval & Gasser, 2016).

This compartmentalization allows the functional clustering of gene-rich (associated with euchromatin) or gene-poor chromatin regions (heterochromatin), keeping them apart from each other, and thereby contributing to the coordinated regulation of transcription of specific set of genes (Meaburn & Misteli, 2007). The gene-rich chromatin is generally located in the nuclear interior while the heterochromatin (like centromeric and pericentromeric chromatin) is clustered in the nuclear periphery generally associated to the nuclear lamina (NL), forming the Lamina-Associated Domains (LADs). LADs comprise large chromatin domains, ranging from 100kb to 10Mb in length and covering around 40% of the genome, that are in close contact with the nuclear lamina in the inner membrane of the nucleus (Guelen et al., 2008; Pickersgill et al., 2006).

TADs are highly conserved between species, and invariant between cell types. It has been shown that the disruption of the boundaries in engineered mouse models recapitulates human developmental disorders (Kaiser & Semple, 2017). Moreover in different limb genetic malformation in Human and Mouse, TADs appear disrupted with altered promoter-enhancer interactions and misexpression (Lupiáñez et al.,

2015) which highlights that gene positioning trough higher-order chromatin organization is crucial for the control of transcriptional fate of the cell.

Besides chromosomal territories, functional nuclear substructures have been characterized, including Cajal bodies, nuclear speckles, promyeolocytic leukemia nuclear bodies (PML NBs) and the nucleolus (Figure 6B); those self-organizing structures are formed by dynamic nucleoprotein complexes that play vital roles in the regulation of the nuclear homeostasis. Cajal bodies, for example, play a role in posttranscriptional modification of spliceosomal components, as the maturation of snRNPs (small nuclear ribonucleoproteins) and snRNA (small nuclear RNAs). Nuclear speckles have been described as sites for splicing factor storage and modification (reviewed in Wood, Garza-Gongora, & Kosak, 2014).

A

B
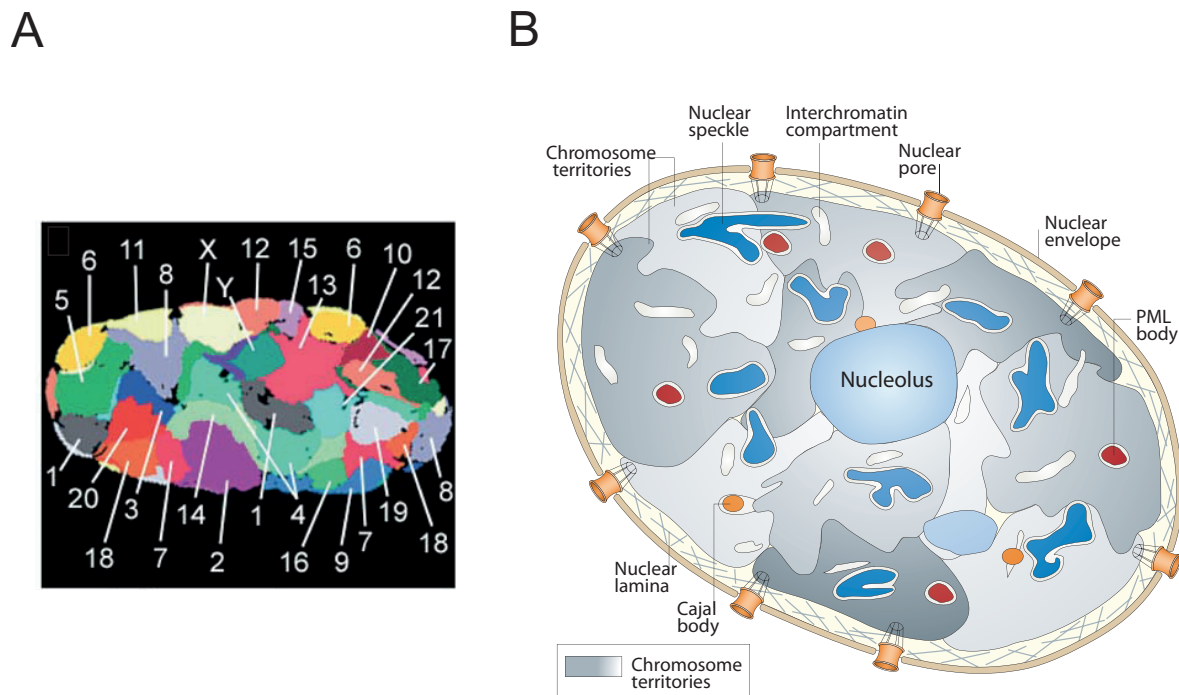
Figure 6. Spatiotemporal organization of nuclear architecture.

A) 24-Color 3D FISH representation and classification of chromosomes in a human G0 fibroblst nucleus. All visible chromosomes in the section are represented with false colors after classification with the program goldFISH. Image modified after Bolzer et al., 2005. B) Diagram depicting the compartmentalization of fuctional nuclear components. Chromosome territories (CTs), nucleolus, nuclear speckles, nuclear pores, Cajal bodies, nuclear lamina, PML bodies and Nuclear envelope are indicated in the figure. Image after Lanctôt et al., 2007.

Of all functional nuclear compartments, the best-characterized example is the nucleolus, where the rRNA transcription, processing and assembly of ribosomal particles take place. The nucleolus is built around the rRNA genes at specific chromosomal loci called "nucleolar organizer regions" (NOR) on the short arms of the acrocentric chromosomes (Nemeth & Längst, 2011). Inside the nucleolus there are specific chromatin conformations associated to the transcriptional activity of the rRNA genes. On the active rRNA genes there have been described specific long-range interactions between promoter and terminator sequences. It has been shown that those interactions are mediated by TTF-I (Transcription Termination Factor), subdividing the rRNA transcription unit into functional chromatin domains (Németh et al., 2008).

Nuclear transcription is carried out by specialized "transcription factories" in which the transcription machinery –like transcription factors, remodeling complexes, coregulators or RNA polymerases- is concentrated and fixed while the DNA template moves relative to the polymerization site (Papantonis & Cook, 2013). In the nucleolus, as an example, the RNA is transcribed by "factories" containing up to 30 active RNA polymerases each (Jackson, Iborra, Manders, & Cook, 1998).

Although in the last years many efforts have been made to understand the mechanisms that regulate chromatin dynamics, it is still not known how the functional organization of the genome into higher-order structures can orchestrate the regulation of transcription.

2.6 HMGN5 regulates higher-order structure of chromatin

The "High Mobility Group" (HMG) proteins constitute another group of players that participate in the regulation of chromatin structure. The HMG super family is divided into three non-related subfamilies, the HMGA (characterized by the AT-hook motif), HMGB (characterized by the HMG-box motif) and HMGN (High mobility group nucleosome-binding) which are non-histone architectural proteins that were shown play roles in the regulation of transcription, replication and DNA repair (Reeves, 2010).

The HMGN family -described so far only in vertebrates- participate in the modulation of higher-order structures of chromatin by direct interaction with the nucleosome through a highly conserved "nucleosomal binding domain" (NBD) (Postnikov & Bustin, 2010).

The HMGN family consists of five members, HMGN1, HMGN2, HMGN3, HMGN4 and HMGN5 (Figure 7), all of them sharing the conserved N-terminal domain containing the Nuclear Localization Signal (NLS) and the NBD, including a conserved octapeptide "RSARLSA". This motif is required for the recruitment of HMGN to the core nucleosome particle, to decompact chromatin architecture, locally and globally, by competing with the binding of linker histone H1 to chromatin (Kugler, Deng, & Bustin, 2012).
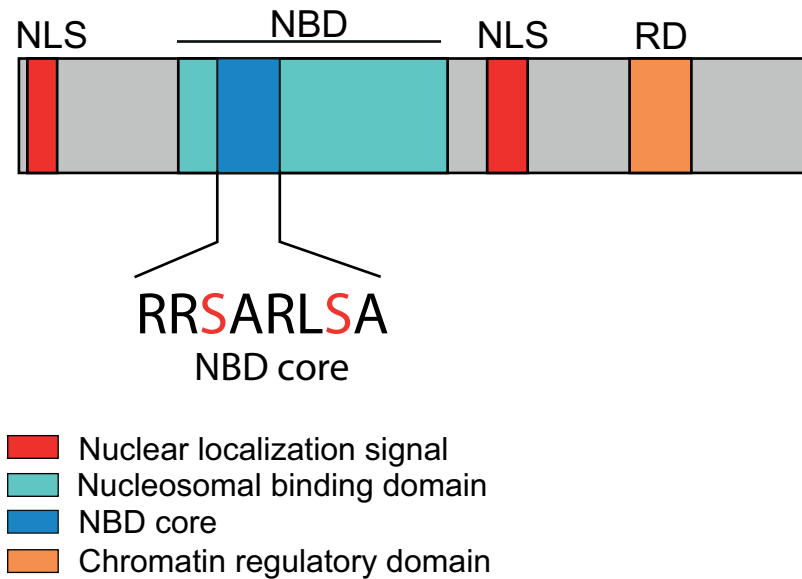
Figure 7. Schematic diagram of HMGN family.

Cartoon depicting the HMGN functional domains; the bipartite nuclear localization signal (NLS) is depicted in red; the conserved Nucleosome Binding Domain (NBD) is indicated in Cyan; the NBD core octapeptide is indicated in blue and the C-terminal chromatin regulatory domain (RD) is indicated in orange. The regulatory serines from the NBD core are highlighted in red.

HMGN5, formerly known as NSBP1 or NBP-45, is the last discovered member of the HMGN family (Shirakawa, Landsman, Postnikov, & Bustin, 2000). It is a highly abundant protein and shares the functional domain NBD with the other members of the family, as well as the ability of counteracting the binding of H1 to chromatin, by a mechanism involving the interaction of the C-terminal tails of both proteins (Rochman et al., 2009). But differentIn contrast to the other HMGNs, HMGN5 is characterized by the presence of a very long negatively charged and unstructured C-terminal domain, of about 300 amino acids in mouse (Shirakawa et al., 2000) and 200 amino acids in humans (King & Francomano, 2001; Rochman et al., 2009). The C-terminal tail of both, mouse and human proteins is enriched in aspartic and glutamic acid and organized in repeated motifs. The acidic motif EDGKE is repeated 11 times in mouse HMGN5 and 4 times in the human protein, but its functional relevance is unknown so far. It was reported that the C-terminal tail is responsible for the specific chromatin location of HMGN5. In mouse, HMGN5 is tethered to euchromatin (Rochman et al., 2009) and in human cells it was shown associated to both, euchromatin and

heterochromatin (Malicetet al., 2011). Moreover, deregulation of HMGN5 lead to changes in the expression of a large set of genes in cells and mouse models (Kugler et al., 2013; Malicet et al., 2011; Rochman et al., 2009; 2011) being the HMGN protein with the strongest effect on transcriptional regulation.

It has been shown that HMGN5 induces large-scale chromatin decondensation *in vivo* -(Rochman et al., 2009) and our own work presented in this dissertation-, and it was suggested that the global chromatin reorganization observed *in vivo* is mediated by the unstructured C-terminal tail. These results highlight the relevance of this protein in the regulation of higher-order structure of chromatin and the cellular transcriptional identity.

Interestingly, several studies have reported an oncogenic role of HMGN5 in various types of cancer (Chen et al., 2012; Yang et al., 2014). Moreover, HMGN5 has an important function in embryonal development, demonstrating that transcriptional changes during differentiation were a result of the induced changes in the global chromatin architecture mediated by HMGN5 (Shirakawa et al., 2009).

It is known that HMGN5, as all HMGNs, is a target of posttranslational modifications. Phosphorylation and acetylation of the NBD control the cell-cycle dependent binding of the protein to chromatin *in vivo* (Moretti et al., 2015; Pogna, Clayton, & Mahadevan, 2010). Furthermore, the recent evidences suggest a role of HMGN5 in the tethering of chromatin to the Nuclear Lamina (NL) (Zhang et al., 2013). However, the mechanism by which HMGN5 can regulate the opening of chromatin structure is far from been understood.

Interestingly, the functional homolog of HMGN5 in the *Drosophila* system, the decondensation factor 31 (Df31) which is also required for opening higher order-structures of chromatin, is tethered to chromatin in an RNA-dependent manner (Schubert et al., 2012; Schubert & Längst, 2013), hinting the possibility of a conserved RNA-dependent mechanism involved in the regulation of higher-order structures of chromatin in the human system.

Due to the role of HMGN5 in the organization of chromatin architecture and its impact on transcriptional regulation, studying the dynamics and function of HMGN5 may help to understand the contribution of chromatin architecture to the maintenance of cell homeostasis.

3 Objectives

Here we proposed to use the human HMGN5 protein as a model protein to study the dynamics of opening of higher-order structure of chromatin and its contribution to the regulation of transcription, by using integrative genome-wide analysis.

For that purpose, we will analyze the genomic distribution of HMGN5, and study the effects of HMGN5 deregulation on the global transcription pattern. The integration of this data will shed light on the influence of the binding of HMGN5 at specific genomic loci on the transcriptional levels of the associated genes. We will try to determine the functional role of HMGN5 by analyzing its genome-wide association with regulatory elements and characterized chromatin hallmarks, like the histone PTMs H3K27ac and H3K4me3, or RNA polymerase II distribution, and others.

To test the hypothesis of the involvement of RNA in the regulatory mechanism mediated by HMGN5, we will characterize the *in vitro* RNA binding properties of the protein, and we will try to identify the potential RNA interacting partners *in vivo*. The data obtained will be integrated with the genomic distribution of HMGN5 and the transcriptome profile to get more insight into a possible functional mechanism.

In parallel, we will perform quantitative mass spectrometry to analyze the global HMGN5-protein interactome with the aim to reveal potential physiological binding partners that may help to delineate the role of HMGN5 in the regulation of higher-order structure of chromatin.

4 Results

4.1 HMGN5 decompacts chromatin

It was already shown that HMGN5 induces large-scale chromatin decompaction (Rochman et al., 2009). As proof of these results, we tested the global chromatin decompaction *in vivo* by using the LacI/LacO tethering system (Figure 8). This method allows the targeting of a protein to an array of a tandem repeat (256 copies) of the Lac operon (LacO) sequence which is inserted in a highly compacted telomeric region in U2OS cells (Jegou et al., 2009). The HMGN5 protein is fused with a GFP-LacI construct, in which the LacI repressor allows the recruitment to the LacO sequence, and GFP fluorescence is used for *in vivo* visualization. The hHMGN5-GFP-LacI construct (vector psV2_HMGN5-GFP-LacI) was used for transient transfection of U2OS cells containing the LacO array, and GFP fluorescence was monitored by confocal microscopy. When HMGN5 was tethered to the LacO array, large-scale chromatin decompaction induced by HMGN5 was observed, as the LacO array appears decondensed compared with the GFP control, in which a highly condensed chromatin structure is observed (Figure 9). We also tested the human Parathymosin (PTMS) protein (construct pSV2-PTMS-GFP-LacI), which, like HMGN5, is an acidic and unstructured protein with a nuclear and nucleolar localization and that was described to counteract the binding of H1 to chromatin (Martic, 2005) and to decompacts chromatin. As shown in Figure 10, the recruitment of PTMS to the LacO array does not decondense chromatin in our system, indicating the functional specificity of HMGN5 in the global chromatin reorganization.
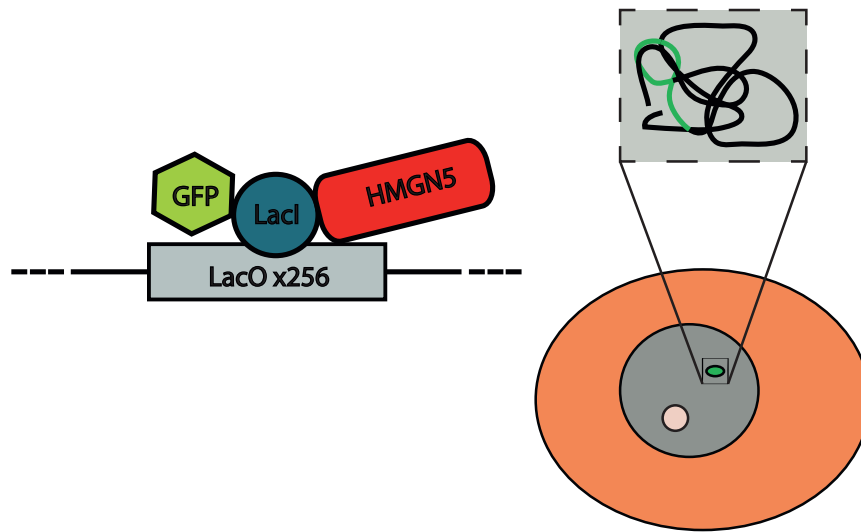
Figure 8. Representation of HMGN5 tethering to LacO array

The diagram illustrates the LacI/LacO tethering system. The fusion protein containing HMGN5-GFP-LacI is tethered through the LacI repressor (depicted in blue in the left diagram) to a tandem array (256 copies) of the LacO sequence, which is stable integrated in a telomeric region is the U2OS cell line. Using confocal microscopy the effect of HMGN5 in LacO array decompaction is visualized by the distribution of GFP signal from the fusion protein. In the right drawing a cartoon of a cell containing the stable integration of the LacO array is indicated.
 Image adapted from:
 https://malone.bioquant.uni-heidelberg.de/methods/single_cell/single_cell.html.
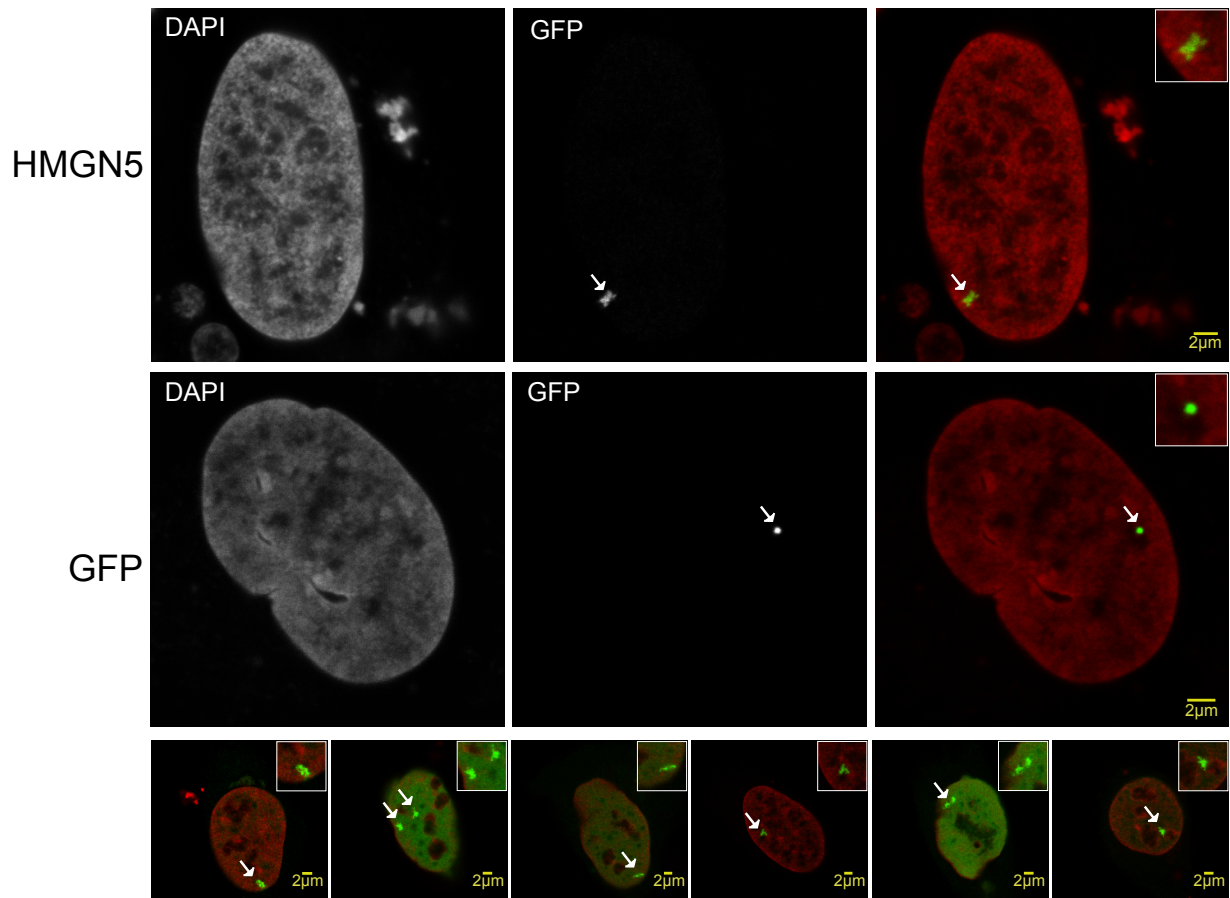
Figure 9. HMGN5-mediated chromatin decondensation.

U2OS cells containing a stable integration of the LacO array were transfected with the construct pSV2_GFP_LacI_HMGN5 or with the control plasmid pSV2_GFP_LacI for transient expression of HMGN5_GFP_LacI or GFP_LacI respectively, to analyze the tethering to the LacO array. 24 h post transfection the cells were fixed with 4% PFA and the DNA was counterstained with DAPI. Cells were visualized by confocal microscopy using a Leica SP8 device. The arrows indicate the tethered protein (seen by GFP fluorescence). The lower panel shows selected example cells transfected with HMGN5. Insets show magnified areas indicated by the arrows. In the merged images, DNA is shown in red and the GFP signal in green. Scale bars represent 2µm.
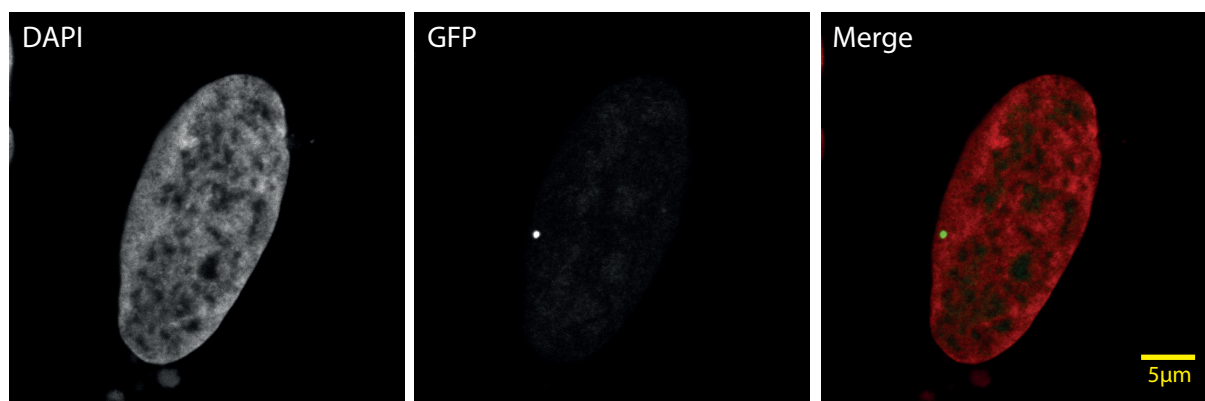
Figure 10. PTMS tethering to the LacO array.

A construct containing PTMS-GFP-LacI (vector pSV2_GFP_LacI_PTMS) was transiently transfected to U2OS cells containing the LacO array. 24 h after transfection the cells were fixed with 4% PFA and the DNA was counterstained with DAPI. Cells were analyzed by confocal microscopy using a Leica SP8 device. The picture shows a representative cell transfected with PTMS. In the merged image the DNA is shown in red and the GFP signal in green. Scale bar represents 5µm.

4.2 HMGN5 is a specific RNA binding protein

It was previously demonstrated in our laboratory that the *Drosophila* protein Df31 (decondensation factor 31), which is a functional homolog of HMGN5, specifically interacts with RNA forming a ribonucleoprotein complex that maintain accessible higher-order structures of chromatin (Schubert et al., 2012).

To explore the potential role of HMGN5 in RNA binding, we tested the *in vitro* interaction of the full-length recombinant human HMGN5, expressed as GST tagged protein (Figure 11), with various fluorescently labeled short single-stranded nucleic acid sequences (Table 1). The interaction was analyzed by electrophoretic mobility shift assays (EMSA) and microscale thermophoresis analysis (MST).

A serial dilution of the protein was performed in a 2:1 ratio and incubated with 50nM of the respective nucleic acid (details in section 7.2.4.2). We used for binding two ssRNA with predicted secondary structure, the snoRNA2T2 (38nt) and snoRNA2T1 (33nt), two RNA without predicted secondary structure, the U-rich En3_TFO_RNA

46

(29nt) and its complementary A-rich RNA sequence En3_RNA_rev (29nt). As control of specificity we used two ssDNA, the En3_TFO_DNA, with the same sequence composition than the En3_TFO_RNA and its complementary DNA En3_DNA_rev, both 29nt long.

Interestingly, HMGN5 is able to specifically interact with RNA as observed by the formation of specific high molecular weight complexes in EMSA with the structured RNA samples snoRNA2T2, snoRNA2T1, and with the unstructured En3_RNA_rev and En3_TFO_RNA (Figure 12A).

We analyzed the binding affinities of the protein to the nucleic acids using Microscale Thermophoresis (MST). Due to technical limitations with fluorescence detection, the analysis could only be performed for the nucleic acids En3_TFO_DNA, EN3_TFO_RNA, snoRNA2T1 and snoRNA2T2. As it was observed in EMSA, the MST analysis showed a higher affinity of HMGN5 for ssRNA over ssDNA (Figure 12B).

When comparing the binding affinity of HMGN5 to the nucleic acids En3_TFO_DNA and En3_TFO_RNA -both having the same sequence- HMGN5 shows a clear preference for binding the RNA over the DNA, as seen by EMSA and after MST analysis (Figure 12A and B) with an EC50 value of 100nM for the binding to En3_TFO_RNA, and 741nM for En3_TFO_DNA. This difference is even more dramatic when comparing the binding to the nucleic acids En3_RNA_rev and En3_DNA_rev, which correspond to the complementary sequences of the En3_TFO_RNA and En3_TFO_DNA respectively, in which the protein showed clear binding for the RNA molecule in EMSA but it does not bind the DNA molecule at the given protein concentrations. It is important to notice that there is no apparent influence of the secondary structure of the RNA in the *in vitro* binding affinity of HMGN5, as judged by the similar binding pattern of the three ssRNA analyzed. Those results suggest a biological function of the RNA binding of HMGN5, and a possible involvement of RNA in the HMGN5 mediated chromatin decondensation.

A

B

Figure 11. Purification of recombinant HMGN5

A) Schematic representation of the full-length human HMGN5 protein. The red box represents the nuclear localization signal (NLS); the nucleosomal binding domain (NBD) containing the core sequence RRSARLSA is represented in cyan, and the green boxes represent the repetition sequence EDGKE (the serines at position 20 and 24 from this sequence are marked in red). B) The human recombinant HMGN5 was expressed in *E.coli* as a GST-tagged construct. 1µg of purified protein was analyzed by denaturing polyacrylamide gel electrophoresis (SDS-PAGE) containing 12% PAA. PageRuler Prestained Protein Ladder Plus was used as a reference for molecular weight. The 70kDa band is indicated.

Table 1. Single-stranded nucleic acids used in EMSA and MST.

Nucleic acid	Sequence	Structure	Label	Length (nt)
snoRNA2T2	GAGUUUAUUACUAAUCUUUCG GGUAUGAAAUUC		Cy5	33
snoRNA2T1	GCUAGCGUGAUAAUGAUUUCA ACUUCACUGCUGACCAG		Fam	38
En3 TFO RNA	CCUCCUUUUUUCUUUUUUUUU UUUUUUCU	Not predicted	Fam	29
En3 TFO DNA	CCTCCTTTTTTCTTTTTTTTTTTT TTTCT	Not predicted	Cy5	29
En3 DNA rev	GGAGGAAAAAAGAAAAAAAAAA AAAAAGA	Not predicted	Cy3	29
En3 RNA rev	GGAGGAAAAAAGAAAAAAAAAA AAAAAGA	Not predicted	Fam	29

Four single-stranded RNA and two single-stranded DNA molecules were used for EMSA and MST experiments. The sequences, structure predictions, fluorescent label and length are shown for each nucleic acid. The secondary structure prediction was performed with the RNAfold web tool.

A

B
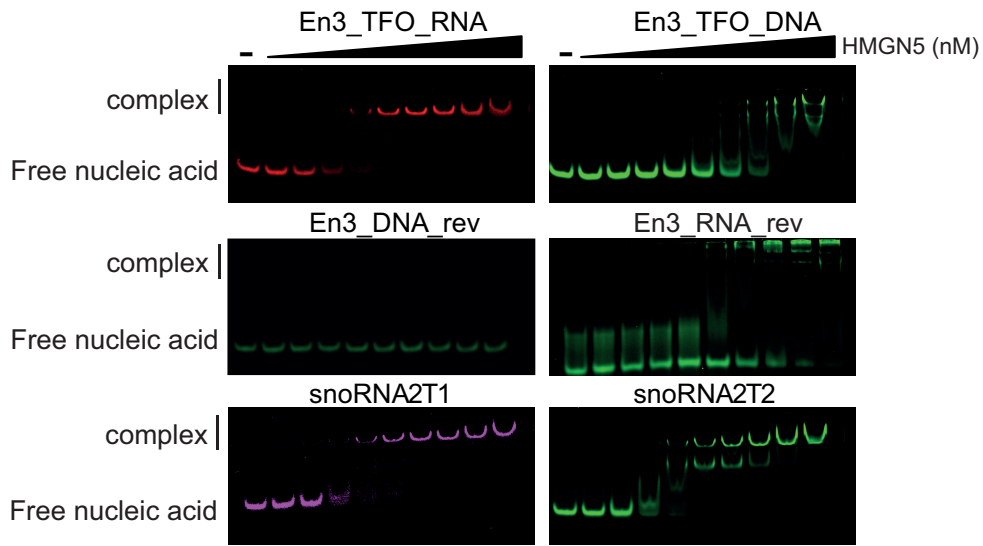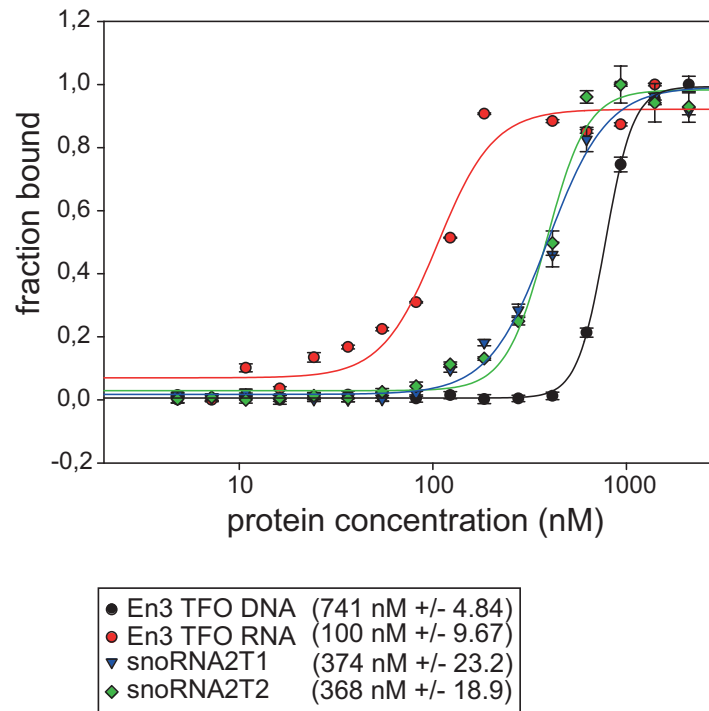
Figure 12. HMGN5 interaction with nucleic acids.

A) HMGN5-nucleic acid interaction analyzed by EMSA. Each fluorescently labeled nucleic acid was kept at a constant concentration of 50nM and a 2:1 dilution of the protein was used starting with 2µM. B) Binding interaction determined by MST. The protein-nucleic acid interactions were quantified by Microscale thermophoresis using the same conditions described for EMSAs. The fluorescence was normalized to fraction bound after adjusting the binding curve with the Hill equation and the EC50 value for each interaction was determined (in Lowe box EC50 +/- standard deviation). The binding interaction by MST was measured for the ssRNAs En3_TFO_RNA, snoRNA2T1, snoRNA2T2 and the ssDNA En3_TFO_DNA.

4.3 The nucleosome-binding domain is required but not sufficient for RNA binding

Since HMGN5 is a highly acidic and unstructured protein, there is no available data regarding the presence of any common RNA binding motif, therefore we focus in the structural characterization according to the nucleosomal binding function of the protein. In HMGN5, like in the other members of the HMGN family, the canonical nucleosomal binding domain (NBD) is crucial for the binding of the protein to nucleosomes. Specifically, NBD contains two serines (S20 and S24 in human HMGN5), that are phosphorylated during mitosis; the phoshorilation was shown to decrease the association of the protein to chromatin (Prymakowska-Bosak et al., 2001; Rochman et al., 2009; Malicet et al., 2011). In order to verify which part of the protein is responsible for the RNA binding ability and test the contribution of the NBD to the RNA binding, we created a set of deletion mutants (as GST-tagged constructs) and three phosphomimetic mutants (Figure 13A and B) -to mimic phosphorylation of HMGN5 during mitosis-, and tested the binding to the Cy5 labeled RNA snoRNA2T2 with which the full length protein showed a specific interaction.

The analysis of interaction reveals that the protein mutant lacking the first 19 amino acids (∆N19) still forms the RNA-protein complex (Figure 13C), but even at the highest protein concentration (2µM) only less than 50% of RNA is bound to HMGN5, indicating that those amino acids contribute to the high binding affinity of the protein. When the first 20 amino acids are deleted (mutant ∆N20) the RNA binding activity is almost completely lost. Furthermore, when deleting the first 24 amino acids (∆N24), containing parts of the NBD including the two serines responsible for the nucleosome binding activity, the RNA interaction is completely abolished. These results indicate that the N-terminal region containing the first 24 amino acids, are required for RNA binding. Interestingly, the deletion mutant ∆C27 that contains only the first 26 amino acids, including the NBD, has also a reduced RNA affinity compared with the wild type and the formation of a defined complex is not observed, but rater a smear at the three higher protein concentration. Moreover, the NBD mutant that contains only the amino acids from the nucleosome-binding domain completely loses the RNA binding ability, indicating that this region is required but not sufficient by itself for RNA binding. This result is confirmed by the RNA binding of the phosphomimetic mutants

(Figure 13D). Here, the serines at position 20 and 24 are replaced by glutamic acid. The mutant S24E shows a similar binding pattern as the ΔN19, while the mutant S20E shows a more pronounced decrease in RNA binding, but still, in both cases, a stable protein-RNA complex is observed. The replacement of both serines by glutamic acid (S20,24E) produces a binding pattern similar to ΔC27, with smeary bands, indicating the formation of intermediate binding species or instable binding. This results show that the serines 20 and 24 have a clear contribution to the RNA binding properties of HMGN5, and it suggests that the C-terminal part of the protein - judged by the binding of the ΔC27 mutant- contribute to stabilize the RNA-protein complexes.
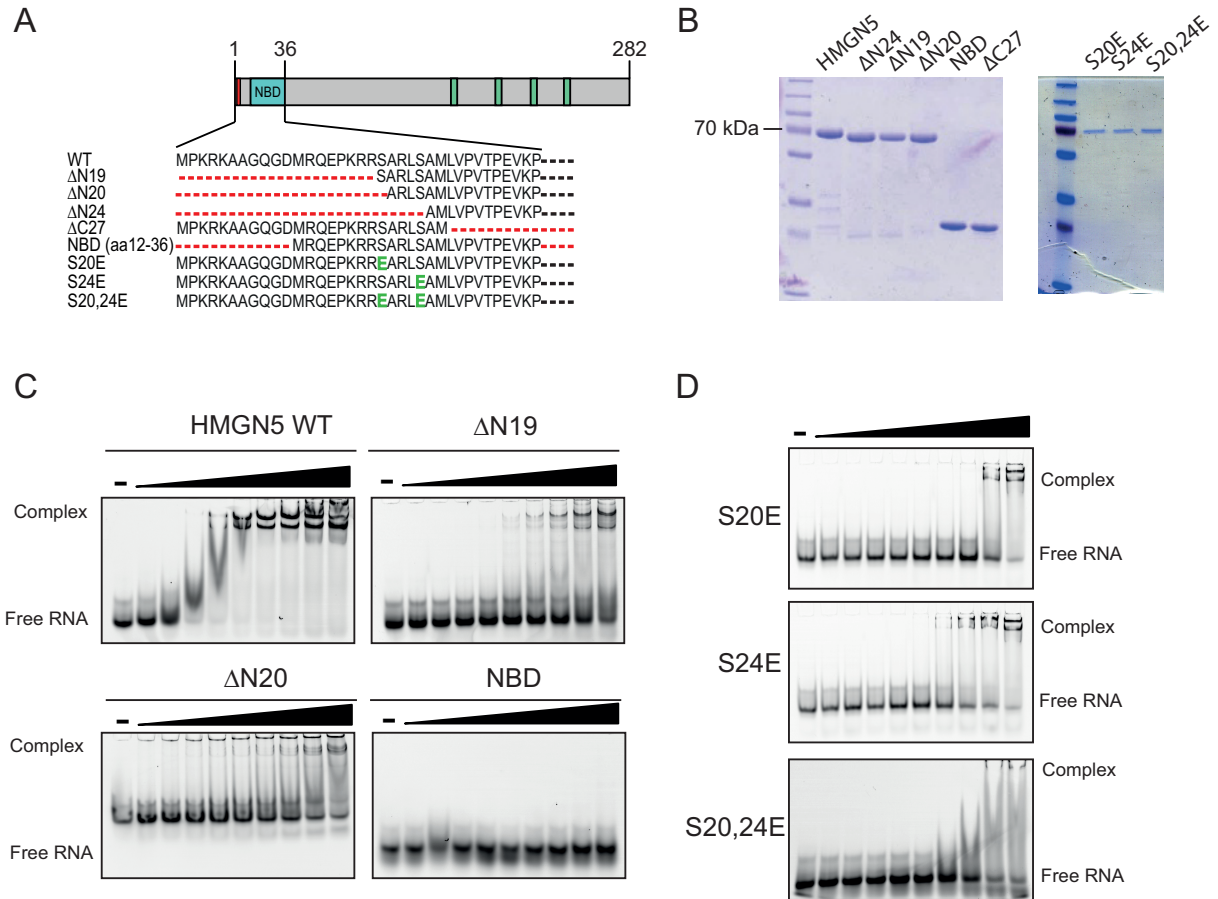
Figure 13. Interaction of deletion and phosphomimetic mutants of HMGN5 with RNA.

A) Schematic representation of HMGN5 deletion and phosphomimetic mutants. The amino acids 1-36 including the nucleosomal binding domain are shown. Red lines represent the deleted amino acids; black lines correspond to non-shown amino acids. The phosphomimetic point mutations S20E, S24E and S20,24E are colored in green. B) Coomassie stained polyacrylamide gel of the recombinant GST-tagged mutants. 1µg of each purified protein was loaded on a 12% denaturing polyacrylamide gel electrophoresis (SDS-PAGE). The 70kDa and 25 kDa bands of the pre-stained protein marker (PageRuler Prestained Protein Ladder Plus) are indicated as molecular weight reference. C) Interaction of HMGN5 deletion mutants and D) HMGN5 phosphomimetic mutants with RNA. The interactions were done using 50nM of the Cy5 labeled snoRNA2T2 and the proteins were serial diluted in a ratio 2:1 starting at 2µM and measured by EMSA assays. Free RNA and the formation of the complex are indicated. A control without protein was loaded in each gel. The interactions were loaded on a 6% native polyacrylamide electrophoresis with 0.4X TBE and visualized and documented using the Typhoon FLA 9500 (GE Healthcare).

4.3.1 Stabilization of RNA-complexes by intramolecular interaction of HMGN5

It has ben shown that several intrinsically disordered proteins (IDPs) fold upon binding of their interacting partners to form a defined structure (Staby et al., 2017) (Mitrea & Kriwacki, 2013). As our previous data suggests that the N-terminal part of HMGN5 is responsible for the RNA-binding and the C-terminal part of the protein may help stabilizing the complexes, we wanted to test if the contribution of the C-terminal domain is determined by the "folding upon binding" hypothesis. In order to do that, we mixed the deletion mutant ΔC27 -containing the N-terminal domain of HMGN5 (which does not show a major specific protein-RNA complex) with equimolar concentrations of the deletion mutant ΔN24 (non-binding mutant) to complement the full-length sequence of HMGN5 and analyze the interaction with RNA. For the interactions, the snoRNA2T2 was kept at a constant concentration of 50nM, while both proteins were serial diluted in a 2:1 ratio starting with 2µM as the highest concentration (See methods 7.2.4.2). Interestingly, the RNA binding activity is recovered when both proteins are mixed as judged by the formation of the specific protein-RNA complexes (Figure 14) compared with the binding of the independent proteins with RNA (ΔC27 and ΔN24 EMSA pictures taken from Figure 13), indicating that there is a specific interaction between both protein domains that allow the formation of a stable RNA-protein complex. Altogether those results suggest that the RNA binding is mediated by the N-terminal domain of HMGN5, with special influence of the serines 20 and 24 from the NBD, and stabilized by intramolecular interactions with the C-terminal domain upon binding of the RNA.
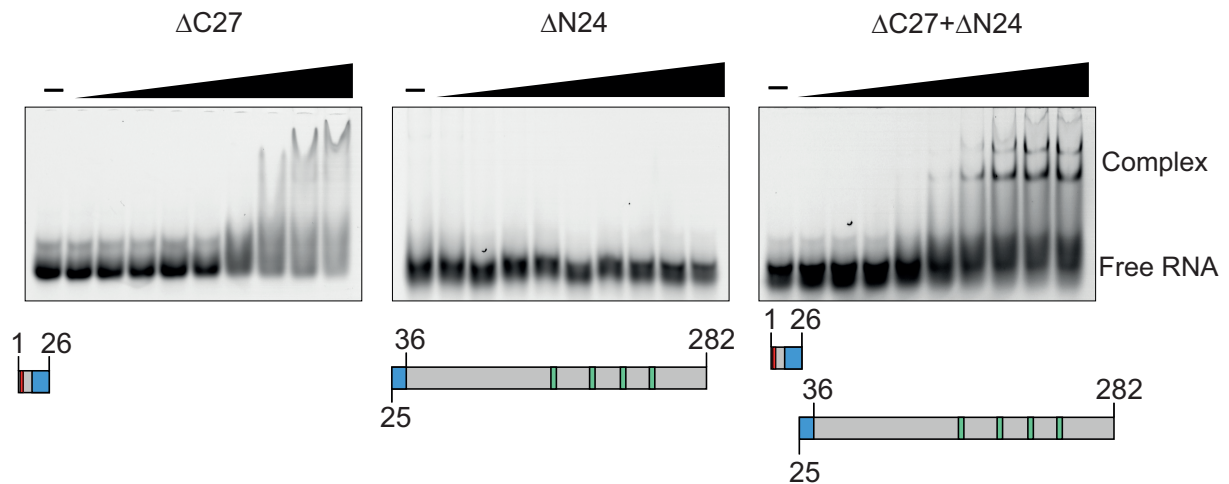
Figure 14. The RNA binding is stabilized by protein intramolecular interactions.

The Cy5-labeled ssRNA snoRNA2T2 was used at a constant concentration of 50nM. Proteins were mixed in equimolar concentration and serial diluted in a 2:1 ratio starting with 2µM. EMSAs corresponding to interaction of mutant ΔC27 and ΔN24 were taken from figure 13. Under each EMSA experiment a schematic representation of the deletion mutants used is shown. The first and last amino acid of each protein is indicated in the schemes. For each experiment a negative binding control without protein was performed. The interactions were loaded in a 6% native PAGE and visualized and documented with a Typhoon FLA 9500 (GE Healthcare).

4.4 The ability to bind RNA is a characteristic of the HMGN family

The N-terminal domain containing the NBD in HMGN5 is a conserved region among all the members of the HMGN family. As this domain is responsible for nucleosome and RNA binding activity in HMGN5, we wanted to test whether this is a general feature of HMGN proteins. Therefore, we expressed, purified and tested the protein HMGN2 and compared the putative RNA binding ability with HMGN5. HMGN2 is composed of 90 amino acids in humans, three times shorter than the human HMGN5 but the minimum binding sequence RRSARLSA from the NBD is conserved (Figure 15).
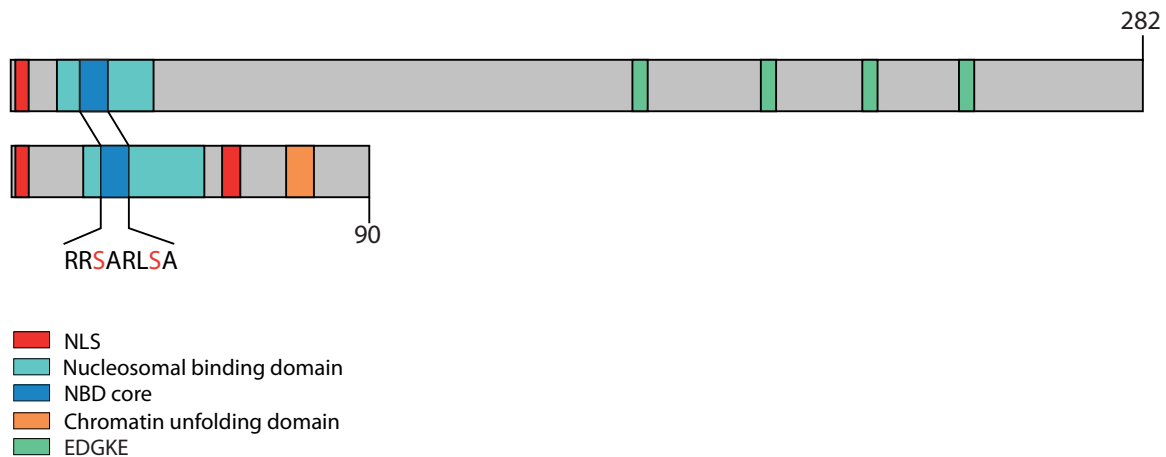
Figure 15. Comparison between HMGN5 and HMGN2 protein features.

Cartoon depicting the HMGN5 and HMGN2 features. The nuclear localization signal (NLS) is indicated in a red box (one NLS in HMGN5 and a bipartite NLS in HMGN2). The nucleosomal binding domain (NBD) and the NBD core containing the conserved octapeptide RRSARLSA are represented in cyan and blue boxes respectively. The two serines of this sequence (responsible for nucleosome binding) are highlighted in red. The orange box in HMGN2 represents the chromatin-unfolding domain, and the green boxes represent the repetitive sequence EDGKE in HMGN5. The last amino acid of HMGN5 (amino acid 282) and HMGN2 (amino acid 90) are indicated in the cartoon.

We created a recombinant human HMGN2 as GST tagged construct (Figure 16A). We prepared a serial dilution of the protein in a 2:1 ratio (Methods 7.2.4.2) and performed the interaction with 50nM of the RNA snoRNA2T2. The interaction was analyzed by EMSA and MST measurement. The result from Figure 16B shows that HMGN2 is also able to specifically interact with RNA, as a clear high molecular weight complex is formed (lanes 6-10). The MST analysis (Figure 16C) shows that HMGN2 binds the RNA with an affinity of 844±7.7nM (EC50 value), while HMGN5 shows, nevertheless, a better binding to the snoRNA2T2, with an affinity of 368nM. This is also visible in the EMSA experiments, in which HMGN5 produces a complete shift of the RNA (as already seen in Figure 12A and Figure 13C), while for HMGN2, even at the highest concentration there is RNA in a free state (Figure 16B).

Additionally, we compared the RNA binding of HMGN5 with HMGB1, member of the HMGB family, characterized by a conserved HMG-box motif. Like HMGN proteins, HMGB1 is able to interact with chromatin and to regulate nuclear processes like transcription, replication or DNA repair (Bustin, 1999). HMGB1 recognizes specifically pre-bent or distorted DNA, and it is able to interact with branched RNA

substrates (A. J. Bell, Chauhan, Woodson, & Kallenbach, 2008). To compare the specificity of HMGN5 binding to the HMGB1-RNA interaction, we measured the RNA binding ability of a recombinant HMGB1 from *rattus norvegicus*, which shares 99% sequence identity with the human HMGB1. The protein was kindly provided by Dr. Klaus Grasser from the University of Regensburg.

The recombinant protein (Figure 16D) was used in the interaction assay in the same way than in HMGN5 and HMGN2 interactions (Materials and methods 7.2.4.2), and we used snoRNA2T2 for binding assays at 50nM concentration. Interestingly, we can observe that at the same concentrations used for HMGN-RNA interactions, HMGB1 does not show RNA binding in the EMSA assay (Figure 16E), since no RNA-protein complex is formed. This result reinforces the specificity of the RNA binding of the HMGN proteins.

A

B

C

D

E
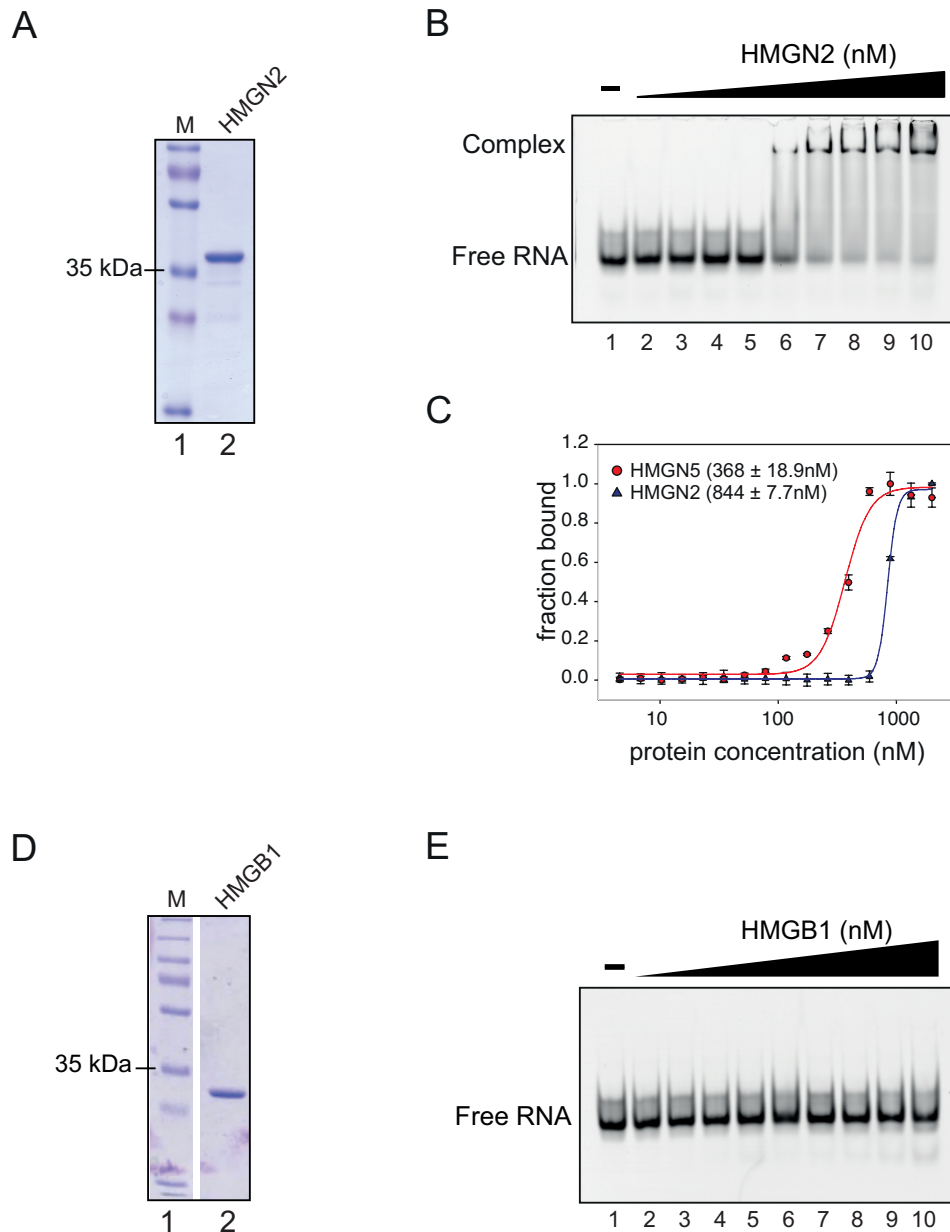
Figure 16. HMGN2 and HMGB1 interactions with RNA.

A) Purification of the recombinant GST-tagged human HMGN2 protein. 1µg of purified protein was loaded on a 12% denaturing polyacrylamide gel electrophoresis (SDS-PAGE) and stained with Coomassie. The 35kDa band of the protein marker is indicated as reference of molecular weight (Lane 1). B) Interaction of HMGN2 with the Cy5-labeled snoRNA2T2 analyzed by EMSA. HMGN5 was serial diluted in a ratio of 2:1 starting with 2µm as highest concentration (Lane 10) and the RNA used in a concentration of 50nM. C) Analysis of HMGN2-RNA interaction by MST. The binding curve was determined according to the Hill equation and the EC50 value was calculated (indicated in the nM range ± standard deviation). The binding was compared to the HMGN5-snoRNA2T2 interaction. D) Recombinant HMGB1 from *rattus norvegicus*. 1µg of recombinant protein was loaded on a 12% SDS-PAGE and stained with Coomassie (Lane 2). Lane 1 indicates the molecular marker (PageRuler Plus prestained protein Ladder) E) Interaction of HMGB1 with snoRNA2T2. The protein-RNA interaction was prepared with the conditions indicated in (B). Lane 1 in (B) and (E) correspond to a negative control without protein.

4.5 Establishment of inducible stable HMGN5 cell line

The *in vitro* analyses revealed the specific binding of HMGN5 to RNA. In order to unravel the biological significance of these interactions and to identify physiological RNA substrates it is necessary to study the interactions in an *in vivo* context. One possibility to achieve that is to immunoprecipitate the endogenous protein from cells using very specific antibodies to isolate the interacting partners and analyze the function of the protein *in vivo*. We tested several commercial antibodies from different companies to test the endogenous protein, and performed immunoprecipitation from HeLa cells. Unluckily none of them was suitable for protein immunoprecipitation (data not shown). Therefore, we created stable cell lines using the Flp-In system from Invitrogen (based on the T-REx™-293 cell line) to generate a stable inducible expression of HMGN5, as a GFP tagged construct (Section 7.2.5.2) for immunoprecipitation with anti-GFP antibodies.

The advantage of producing stable cell lines with the Flp-In system, is the generation of isogenic cell population, with low inter-assay variation and homogeneous expression, due to the tetracycline-regulated protein expression (Savage, Wootten, Christopoulos, Sexton, & Furness, 2013).

Figure 17 shows the successful establishment of the HMGN5-FlpIn cell line. After 24 hours of induction, HMGN5 localizes in the nucleus, correlating with the localization of the endogenous HMGN5 (see Dox+ images in the Figure 17). It can be observed an homogeneus population of cells expressing the protein. In the non-induced cells (Dox-) we do not observe leaky protein expression.
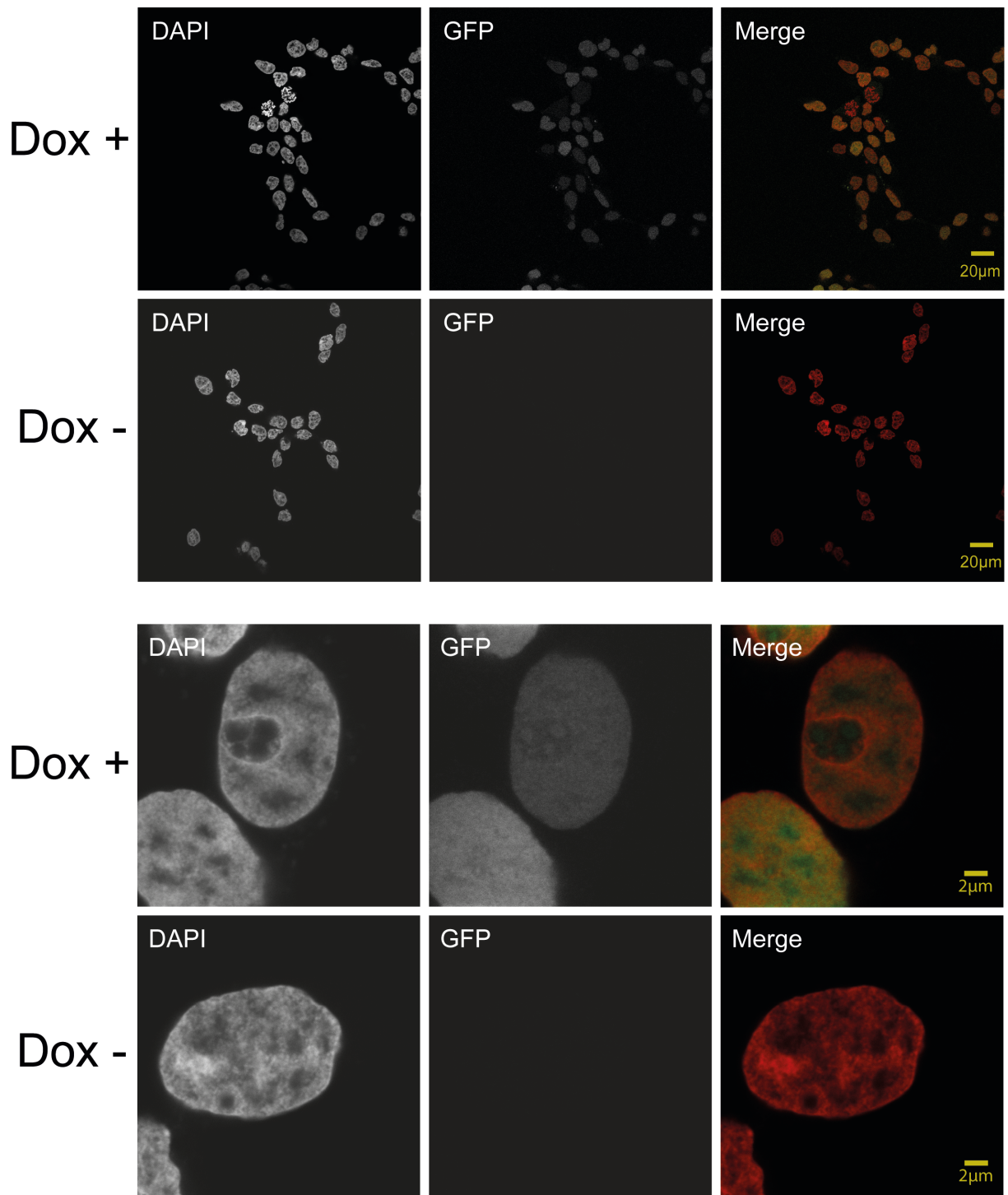
Figure 17. Establishment of HMGN5 FlpIn inducible cell line.

Selected stable clones were seeded in 6-well plates. After 24 h of induction with Doxycycline (Dox+) or without induction (Dox-), cells were fixed with 4% PFA and analyzed with confocal microscopy on a Leica SP8 device. DNA was counterstained with DAPI. HMGN5 expression was visualized by GFP fluorescence. In the merged images the DNA is shown in red and the HMGN5 protein expression in green. Lower images show 10 times magnified captures (2µm scale bar).

To estimate kinetics and levels of protein expression, a time course of HMGN5 induction was performed. Figure 18 shows a western blot after the induction of HMGN5 with Doxycycline, from 0 to 24 h. Between 6-9 hours there are detectable protein levels, without an excess of protein degradation, therefore those time points were selected for the downstream experiments.
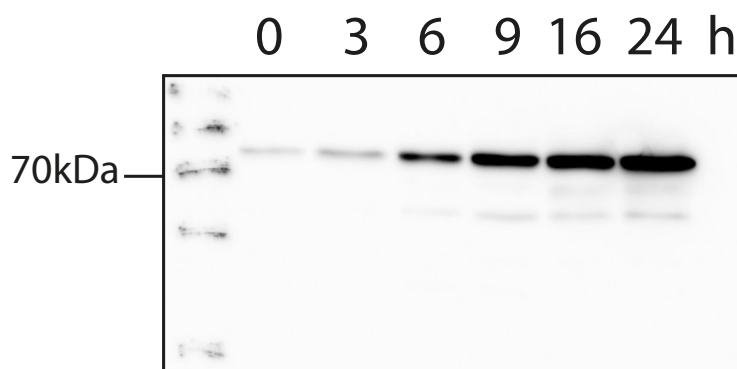
Figure 18. Time course of HMGN5 expression.

 Western blot showing the expression of the construct HMGN5-GFP after inducing the cells with Doxycycline. Total protein extracts were prepared after inducing the HMGN5 FlpIn cells with 1µg/ml Doxyciclyne for 0, 3, 6, 9, 16 and 24 h. 20µg of protein extracts were loaded per lane on a 12% SDS-PAGE and transferred to a PDVF membrane for western blot. HMGN5 was detected using the anti-HMGN5 antibody HPA000511 (Sigma). The 70kDa band of the marker is indicated as reference of molecular weight.

4.6 Effect of HMGN5 deregulation in the global transcriptome pattern

It has been shown that the loss or overexpression of HMGN5 affects the global transcriptome profile of the cells, in a tissue-specific manner (Kugler et al., 2013; Rochman et al., 2009). To analyze the role of the human HMGN5 protein in the transcriptional regulation in our system, we performed an RNA-seq profile and did a comparative expression analysis of genes after 24 h overexpression or knockdown of the protein in the stable cell line (HMGN5_FlpIn). For the knockdown HMGN5_FlpIn cells without selection antibiotics were seeded in 10cm plates per triplicate and transfected with smartpool siRNA complementary to the HMGN5 RNA for 24 h (see Figure 42 in the appendix for knockdown standardization). Two of the three biological replicates were selected for RNA sequencing per condition. Two replicates of un-induced cells were used as control for baseline RNA expression. After treatment, the RNA was extracted (section 7.2.2.11), the libraries were prepared and RNA sequenced (7.2.7.1). Following quality analysis, the raw data was mapped to the hg19 human genome using the STAR aligner. A differential gene expression was performed to estimate the effects of HMGN5 overexpression and knockdown.

The analysis highlights that the up- and down regulation of HMGN5 leads to transcriptional changes in 2999 and 3049 genes respectively (Figure 19A). A large set of genes (1287), corresponding to one third of the total transcripts affected is overlapped between the two conditions. Interestingly, only 90 genes from this set are reciprocally regulated. Detailed comparison of the two groups of genes affected indicates similar number of genes up or down regulated in both conditions, and for most of the genes, Log2 fold change were smaller than 2 (Figure 19B). Principal component analysis of the HMGN5 expression exhibits a correlation of the expression in the biological replicates analyzed for each condition: overexpression, knockdown and control cells (Figure 19C), indicating a good reproducibility of the data.

A

B

C
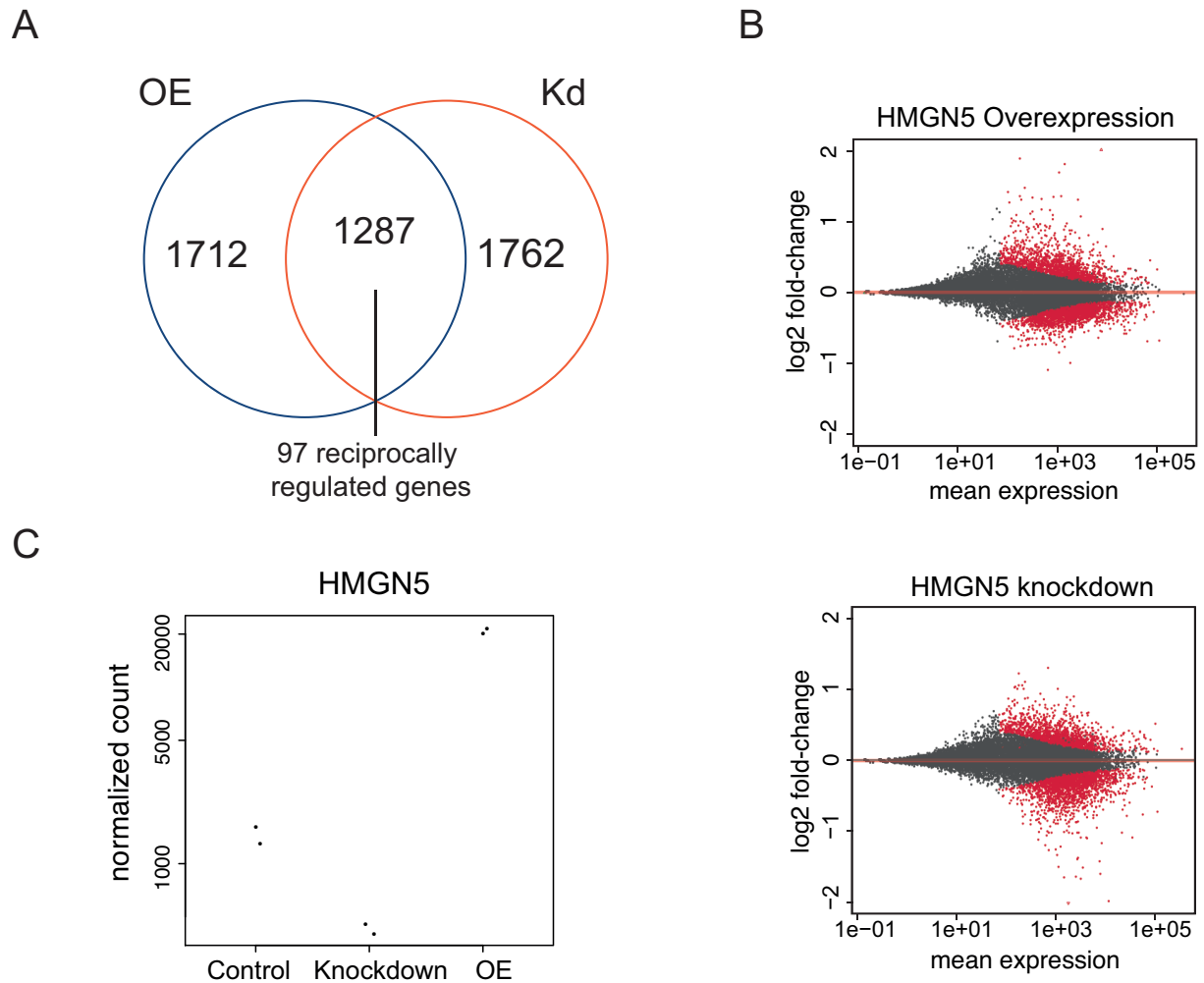
Figure 19. Modulation of global transcriptome profile by HMGN5 misregulation.

A) Venn diagram depicting the overlapped transcriptional changes upon overexpression (OE) and knockdown (Kd) of HMGN5 in the HMGN5_FlpIn cell line (24 h of induction). A threshold of adjusted p-value <0.05 was used. B) MA plot of differential expression analysis. Points representing RNAs with significant differential expression compared with the control samples are marked in red. Gene expression changes were normalized to Log2 fold-change ratios. (C) PlotCount graph showing the normalized count of the HMGN5 transcript from the RNA-seq in the non-treated cells (Control), knockdown and overexpression of HMGN5 (OE). Each dot in the groups represents individual biological replicates.

We selected four candidate genes with affected transcription levels in the RNA-seq analysis for validation in qPCR. One of them ZCCHC12 (Zinc Finger CCHC-Type Containing 12) was 2 fold down-regulated after knockdown of HMGN5, and 0.6 fold up-regulated after HMGN5 overexpression. VGF (VGF Nerve Growth Factor Inducible) showed increased expression in both conditions (4-fold and 1.5-fold after knockdown and overexpression respectively). The genes coding for RPL30

(Ribosomal Protein L30) and EGR1 (Early Growth Response 1) changed their expression levels only after one treatment. RPL30 had a Log2FC of -2 after HMGN5 Kd and EGR1 is up-regulated with Log2FC of 1.8 after HMGN5 OE. qPCR validation for all four candidates was performed on three independent biological replicates. The results obtained from the validation were normalized to log2FC and a student t-test was applied for statistical significance. As shown in Figure 20 there is a strict correlation between the log2FC obtained from qPCR for each candidate gene and the one obtained from the RNA-seq analysis, which further validate the RNA-seq analysis.
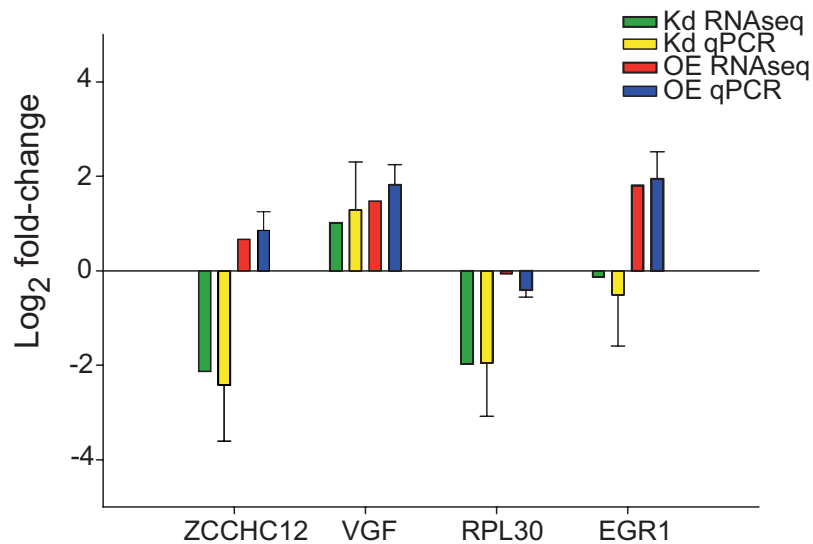
Figure 20. qPCR validation of 4 candidate genes.

Log2 fold-change comparisons between qPCR and RNA-seq analysis for four selected candidates that showed differential expression between the treatments. Log2FC in qPCR estimates relative RNA enrichment in the treated cells (overexpression and knockdown of HMGN5) over non-treated control cells. Standard deviation from 3 independent biological replicates applies only to qPCR results. GAPDH was used for normalization of expression. Two-way ANOVA and Dunett´s post-test were used to define statistical differences of mRNA levels between groups. Kd: knockdown; OE: overexpression of HMGN5. ZCCHC12: Zinc Finger CCHC-Type Containing 12; VGF: Nerve Growth Factor Inducible; RPL30: Ribosomal Protein L30; EGR1: Early Growth Response 1.

We performed gene ontology (GO) enrichment analysis of genes that change their transcription levels after up- and down-regulation of HMGN5. Gene sets were functionally clustered by molecular function and biological process using the ClueGo plug-in (Bindea et al., 2009) of the Cytoscape Network (Cline et al., 2007). Statistical enrichment analysis reveals that mainly genes associated with cell cycle regulation and telomere elongation are down regulated in both conditions (Figure 21B and Figure 22B). HMGN5 overexpression leads to up-regulation of genes classified as "negative regulation of gene expression" (Figure 21A), in which several transcription factors are clustered. We also find upregulated genes associated with the apoptotic process and "cellular protein metabolic process". The results showing that genes associated with cell cycle progression are affected, correlates with the role of HMGN5 in cell cycle regulation and cancer cell growth, previously shown in the literature (reviewed in Shi, Tang, Wu, & Sun, 2015).

Additionally, genes associated to "RNA biosynthetic process", "translation" and "negative regulation of gene expression" are up-regulated after a loss of HMGN5 (Figure 22A), indicating that HMGN5 is important for controlling gene expression.

A
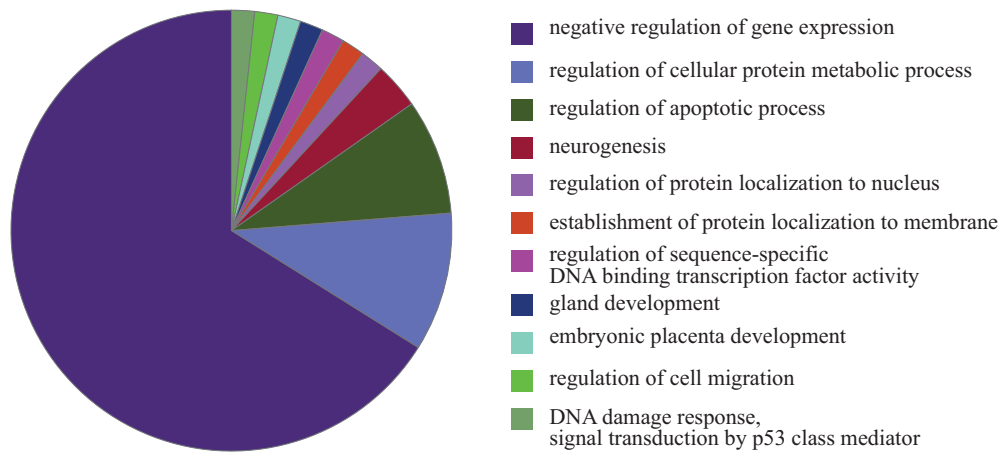
B
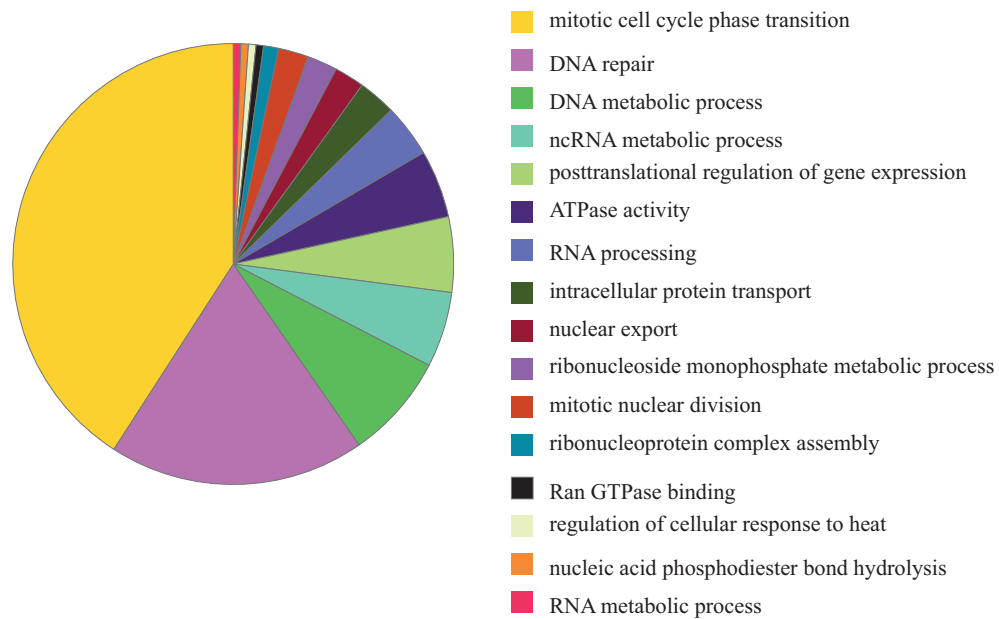
Figure 21. Gene set enrichment analysis after HMGN5 overexpression.

Detailed GO term enriched in genes affected by HMGN5 overexpression is represented as a pie chart. Genes with increased expression levels (A) and with decreased expression levels (B) are analyzed. The groups are organized from higher (up) to lower (bottom) number of genes per group in the color legend, and displayed consecutively in the chart. The analysis was performed using the ClueGo software. Significance of enrichment was estimated with a two-sided hypergeometric test and the p-value correction was performed appliying the Bonferroni step-down method. The leading term correspond to the most significant term in the group. Only term with PValue <0.005 are shown.

A

- RNA biosynthetic process
- translation
- negative regulation of gene expression
- regulation of cyclin-dependent protein serine/threonine kinase activity
- regulation of cellular protein metabolic process
- regulation of apoptotic process
- positive regulation of transcription from RNA polII promoter in response to stress
- cell cycle arrest
- translational elongation

B

- mitotic cell cycle phase transition
- regulation of telomere maintenance
- cellular protein catabolic process
- RNA processing
- positive regulation of protein catabolic process
- neutrophil activation
- regulation of mRNA stability
- intracellular protein transport
- mitotic nuclear division
- purine ribonucleotide binding
- peptide transport
- nucleoside-triphosphatase activity
- protein modification by small protein conjugation or removal
- posttranscriptional regulation of gene expression
- RNA metabolic process
- DNA geometric change
- regulation of cellular response to heat
- regulation of protein complex assembly
- ribonucleoprotein complex assembly

Figure 22. Gene set enrichment analysis after HMGN5 knockdown.

Detailed GO term enriched in genes affected by HMGN5 knockdown is represented as a pie chart. Genes with increased expression levels (A) and with decreased expression levels (B) are analyzed. The groups are organized from higher (up) to lower (bottom) number of genes per group in the color legend, and displayed consecutively in the chart. The analysis was performed using the ClueGo software. Significance of enrichment was estimated with a two-sided hypergeometric test and the p-value correction was performed appliying the Bonferroni step-down method. The leading term correspond to the most significant term in the group. Only term with PValue <0.005 are shown.

When analyzing the 1287 genes overlapped between up- and down-regulation of HMGN5, a similar group of GO terms are overrepresented (Figure 23) in comparison with the separate datasets. Here, again, genes related to cell cycle regulation are mostly affected. It can be observed that genes related to RNA metabolism and processing, as gene associated with protein transport to Cajal bodies are also enriched. Since the subnuclear compartments Cajal bodies are associated with biogenesis, maturation and recycling of small RNAs (Nizami, Deryusheva, & Gall, 2010), those results further confirm that HMGN5 plays a role in the regulation of RNA metabolism.
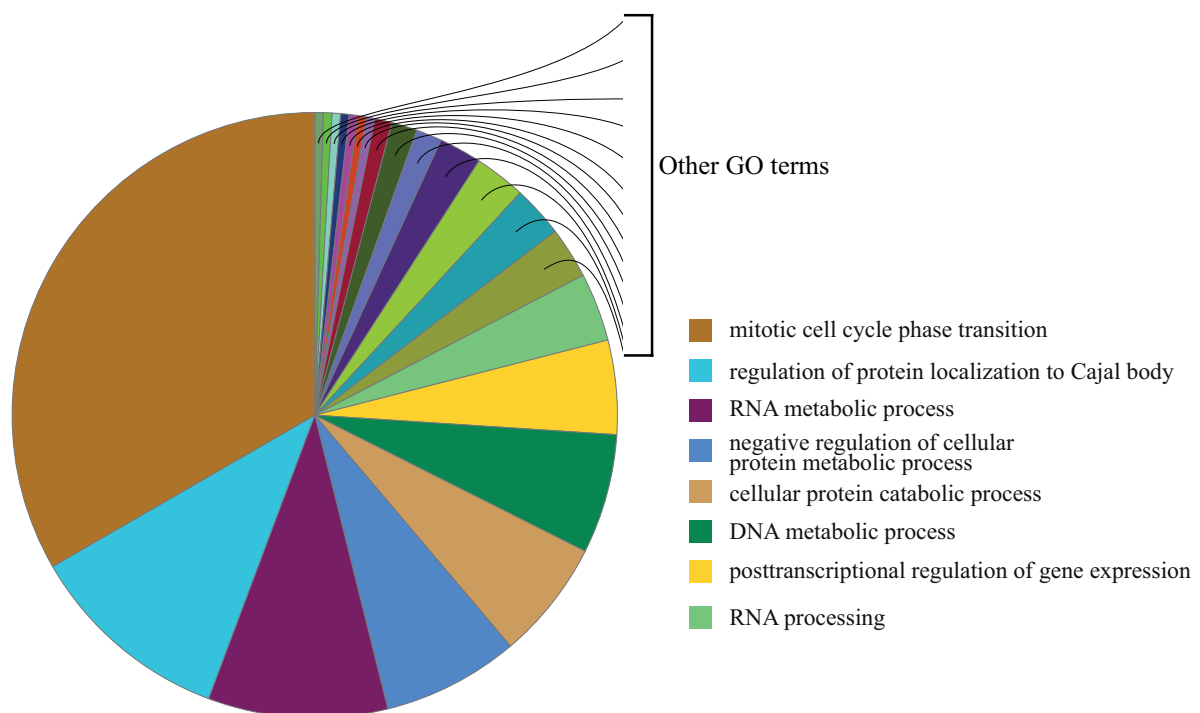

Figure 23. Gene set enrichment analysis of the overlapped genes between overexpression and knockdown of HMGN5.

Detailed GO term enriched in the genes affected by overexpression and knockdown of HMGN5 is represented as pie chart. The groups are organized from higher (up) to lower (bottom) number of genes per group in the color legend, and displayed consecutively in the chart. The analysis was done with the ClueGo software; significance of enrichment was estimated with a two-sided hypergeometric test and the p-value correction with the Bonferroni step-down method. The leading term correspond to the most significant term in the group. Only term with p-value <0.005 are shown.

4.7 HMGN5 UV-crosslinking immunoprecipitation

To identify potential *in vivo* RNA targets of HMGN5, we performed an RNA immunoprecipitation after UV crosslinking (CLIP), using the HMGN5_FlpIn cell line. As control we performed the immunoprecipitation using a GFP_FlpIn cell line available at our laboratory. Protein immunoprecipitation after UV crosslinking is generally used to identify genome-wide RNA-protein interactions *in vivo*. The method (described in Ule, Jensen, Mele, & Darnell, 2005) has the advantage of generating good signal:noise ratio compared with the classical RNA immunoprecipitation method (RIP), and allows the identification of binding sites in the target RNAs (Darnell, 2010). First, cells are treated with UV at 254nm to create covalent bonds between proteins and their interacting RNAs *in vivo*. Afterwards, the complexes are immunoprecipitated from cell lysates while a partial RNase A/T1 digestion on beads is performed. The protein-bound RNAs are subsequently radiolabeled and transferred to a nitrocellulose membrane for isolation. Then, the protein is removed with proteinase K, and RNA linkers added to the isolated RNA to generate the libraries for high throughput sequencing.

 Since HMGN5 is a protein that strongly binds chromatin, the original protocol of CLIP, applied for soluble proteins cannot be directly used. Therefore, we modified the protocol by introducing a sonication step to fragment genomic DNA, before immunoprecipitation (workflow in Figure 24A). First we analyzed the efficiency of UV crosslinking in our samples. Cells from the HMGN5_FlpIn cell line and the control GFP_FlpIn, were treated for 8 h with (Dox+) or without (Dox-) Doxycycline to induce protein expression. Cells were treated with or without crosslinking with 150mJ/cm2 of UV at 254nm followed by protein co-immunoprecipitation. Figure 24B shows the efficiency of HMGN5 crosslinking and immunoprecipitation. For both HMGN5 and GFP, the immunoprecipitation is highly efficient, as judged by the high yield of the specific proteins. HMGN5 corresponds to the prominent band identified at 80-90kDa (Figure 24B, lanes 3-4) and GFP corresponds to the 26kDa band in lanes 6-7 in the figure 24B. Due to its ability of binding nucleosomes, the efficiency of crosslinking between HMGN5 and RNA can be followed by the co-immunoprecipitation of histones. In both induced HMGN5 samples (Lanes 3-4) the core histones are visible,

at a molecular weight between 10 and 15kDa, but when UV crosslinking is applied the amount of co-purified histones increases more than twice, with an observed approximate ratio 1:1 to HMGN5 (Figure 24B, lane 3). Linker histone H1 is also visible after UV crosslinking of HMGN5 samples. When comparing with the GFP control after induction and UV crosslinking, none of those bands are observed (figure 24B lanes 6-7), indicating that the co-purification of histones is not background but rather mediated by specific interaction of HMGN5 with nucleosomes which is stabilized by UV crosslinking.
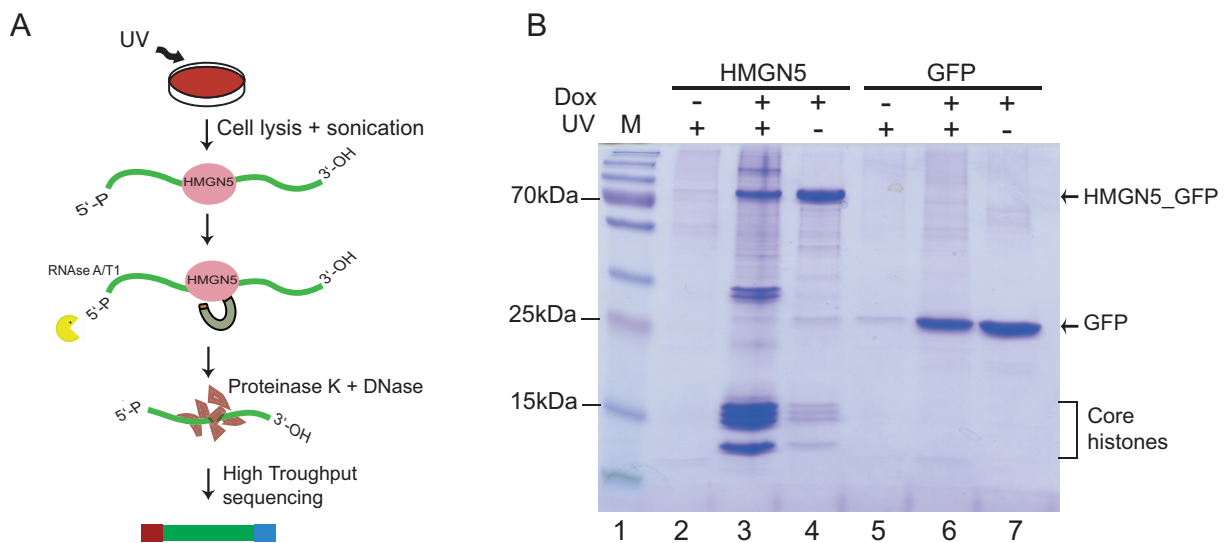

Figure 24. HMGN5 CLIP standardization.

A) Workflow of CLIP protocol to immunoprecipitate HMGN5-bound RNAs. After cells lysis a sonication step is performed for chromatin fragmentation. After immunoprecipitation with GFP_Trap magnetic beads, an on-bead partial RNA digestion is performed with RNAse A/T1. To isolate interacting RNAs, a DNAse and proteinase K treatment is performed to remove DNA and proteins respectively. Purified RNAs are used to prepare libraries for high throughput sequencing. B) Efficiency test of UV crosslinking. Cells grown in 15cm plate (from HMGN5-FlpIn and GFP-FlpIn), are treated with or without induction (Dox+ /Dox-), and subjected to UV crosslinking by applying 150mJ/cm2 UV at 254nm (UV+), or without UV treatment (UV-). Samples were used for protein Co-IP using the GFP_Trap agarose beads. 10% of the immunoprecipitated samples were loaded on a 12% SDS-PAGE and visualized after Coomassie staining. PageRuler Prestained Protein Ladder Plus was used as a reference for molecular weight. The 70, 25 and 15 kDa bands are indicated in the figure.

As already mentioned, in our protocol we applied sonication to shear the chromatin and reduce unspecific background RNAs that were associated to chromatin. We tested the effect of sonication on the RNA quality by performing a sonication time course in the cells. After 20 min of intense sonication (sufficient to shear the chromatin) the total RNA content is mainly intact (Figure 25A, lane 5), as judged by the integrity of the 28S and 18S ribosomal RNA.

Figure 25B shows the RNA partial digestion in lysed cells using the RNase A/T1 cocktail (Thermo Scientific). For downstream experiments we selected the dilution 1:50 (lane 3) in which RNA is digested to fragments up to 600nt approximately.
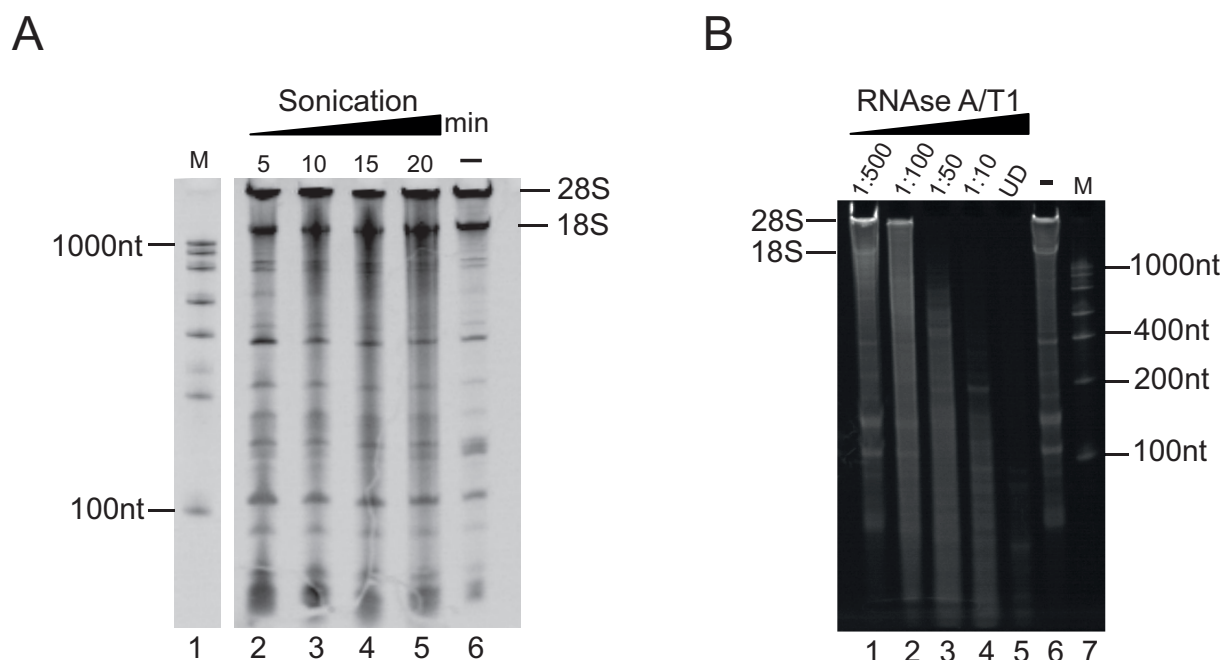

Figure 25. Sonication and RNase treatment for CLIP.

A) Effect of sonication on total RNA quality. Cells grown in 15 cm plate were lysed and sonicated at 0 (-) 5, 10, 15 and 20 min and total RNA was extracted and analyzed on a 4% denaturing urea gel. RNA was visualized by Ethidium bromide (EtBr) staining. The 28S and 18S ribosomal RNA bands are indicated. The RiboRuler Low Range RNA Ladder was used as molecular weight reference (M: Marker). The 1000nt and 100nt bands are indicated in the figure. B) Partial RNA digestion with RNase A/T1. Lysed cells from 15 cm plates were treated with a cocktail of RNase A/T1 at the following dilutions: 1:500, 1:100, 1:50, 1:10, undiluted (UD), and no treatment (-). Per sample 1µg of isolated RNA were loaded on a 4% denaturing urea gel and visualized after EtBr staining. The 28S and 18S ribosomal RNA bands are indicated. The RiboRuler Low Range RNA Ladder was used as molecular weight reference.

4.8 HMGN5 binds RNA *in vivo*

HMGN5-associated RNAs were isolated using the modified CLIP-seq protocol (results 4.7 and methods7.2.6.2). After purifying the RNAs, a stranded library was prepared for sequencing using the Ovation universal RNA-seq kit (Nugen). The quality of the libraries for sequencing is described in Figure 44 of the appendix. The higher throughput sequencing of the barcoded libraries was performed at the KFB Regensburg in an Illumina HiSeq device, using the specific barcodes from the adapters to combine 6 samples in one lane, three HMGN5 biological replicates and three GFP control samples. The sequencing gave a total of 152117704 million reads yielding an average of 25 million per sample. The reads detail of the demultiplexed data is listed in Table 2.

Table 2. Demultiplexed reads from CLIP libraries sequencing.

Sample	Name of raw data sample	Index	#Reads
GFP1	GFP1_GCACTA_L002_R1_001	GCACTA	26.530.237
GFP2	GFP2_ACCTCA_L002_R1_001	ACCTCA	23.277.975
GFP3	GFP3_GTGCTT_L002_R1_001	GTGCTT	23.586.035
HMGN51	HMGN51_AACCAG_L002_R1_001	AACCAG	20.685.221
HMGN52	HMGN52_TGGTGA_L002_R1_001	TGGTGA	33.310.255
HMGN53	HMGN53_AGTGAG_L002_R1_001	AGTGAG	24.727.981

For each sample, the name of the raw data, specific barcode (Index) used from the Ovation Universal-RNA seq kit and the reads obtained after sequencing are shown.

The single-end reads were mapped to the hg38 reference genome using Bowtie (methods section 7.2.7.3). To get the enriched RNA on the HMGN5 samples a differential gene expression analysis (DGE) was done using the edgeR package from R. Due to technical problems with the quality of reads from GFP1 sample, the analysis with the control GFP was performed using only GFP2 and GFP3 samples. For DGE a combination of all alignments from all three HMGN5 samples was used, and as background a combined alignment of GFP2 and GFP3 was used. The reads were normalized to counts per million (CPM) to control for differences in library size between samples, and filtered using a 5 CPM threshold (each gene should have a minimum count of 5-10 CPM in a library to be considered expressed). The reads

count was performed in two different ways, by counting exons or introns obtained from genes, or obtained from transcripts. When performing DGE using exons coming from genes, we obtained a list of 23 RNAs enriched in the HMGN5 samples (FDR cutoff of 0.05) (Table 3).

Table 3. HMGN5-associated transcripts.

Gene ID	Gene name	logFC	logCPM	p-value	FDR
79366	HMGN5	-4.49449498	9.637634792	5.54E-19	5.67E-15
1053	CEBPE	3.958698057	6.628945597	1.65E-16	8.43E-13
57505	AARS2	-6.639625238	4.672186956	4.71E-10	1.61E-06
3508	IGHMBP2	-10.42058477	4.212379836	1.81E-08	4.63E-05
9851	KIAA0753	-10.19893051	3.99528672	4.08E-07	0.000834356
10079	ATP9A	-10.08259598	3.881664093	1.59E-06	0.002707243
8398	PLA2G6	-10.08461392	3.883657698	2.52E-06	0.003683113
79877	DCAKD	-10.04290033	3.842892724	3.38E-06	0.004092192
58496	LY6G5B	-10.02638367	3.826942295	3.60E-06	0.004092192
8452	CUL3	-10.01797036	3.818571999	4.30E-06	0.004399785
41	ASIC1	-9.998895224	3.800061671	4.73E-06	0.004399785
51603	METTL13	-9.961141771	3.763299862	7.63E-06	0.006510971
8318	CDC45	-9.940726861	3.743386566	8.66E-06	0.006580009
84498	FAM120B	-6.83765804	4.009313143	9.28E-06	0.006580009
157313	CDCA2	-9.932440179	3.735407464	9.64E-06	0.006580009
52	ACP1	-9.918149108	3.721273131	1.36E-05	0.008676717
26205	GMEB2	-9.836435056	3.641983308	2.47E-05	0.014699186
80742	PRR3	-9.861485796	3.666656761	2.59E-05	0.014699186
49855	SCAPER	-9.801153022	3.60811899	4.58E-05	0.024650247
25825	BACE2	-7.471101614	3.696949114	5.26E-05	0.026898461
23216	TBC1D1	-7.472213128	3.696984403	6.50E-05	0.031676287
3707	ITPKB	-9.753532537	3.561352804	6.99E-05	0.032528457
4337	MOCS1	-9.751324205	3.559154194	7.59E-05	0.033016902
84447	SYVN1	-9.700511834	3.510182665	7.74E-05	0.033016902
8854	ALDH1A2	-9.710588785	3.52029618	8.11E-05	0.033196156
201158	TVP23C	-9.673225334	3.483877758	9.77E-05	0.038449166
56980	PRDM10	-9.648199646	3.459755756	0.000117533	0.044553581
84326	METTL26	-9.642116562	3.453527431	0.000124256	0.045420042
85313	PPIL4	-9.655885047	3.467047841	0.000137691	0.048595546
85476	GFM1	-9.687933405	3.49804815	0.000143435	0.048935262

List of RNAs generated by counting exons obtained from genes. The Gene ID, name of gene, log fold change (LogFC), normalized Log counts per million (LogCPM), p-value and adjusted false discovery rate (FDR) are shown in the list. FDR cutoff: 0.05. The negative LogFC values correspond to enrichment of HMGN5 over GFP.

Some examples from the list are shown in Figure 26. Peaks are distributed along the RNA body in the transcripts from the genes AARS2 (Alanyl-TRNA Synthetase 2, Mitochondrial), KIAA0753 (OFD1 And FOPNL Interacting Protein) and ASIC1 (Acid Sensing Ion Channel Subunit 1), and enriched in comparison with the GFP samples. The peaks are observed widely distributed in exons and introns.
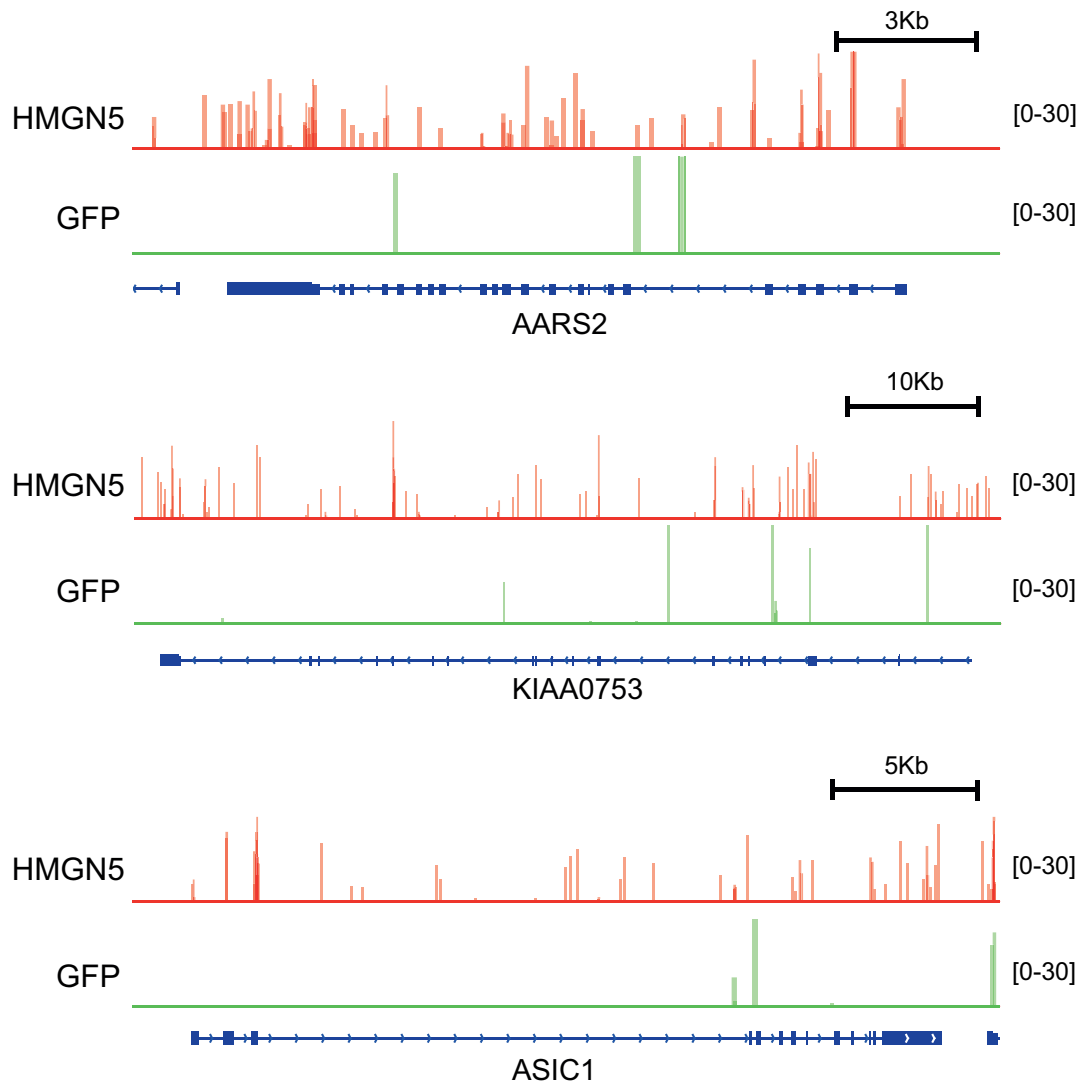

Figure 26. Examples of HMGN5-bound RNAs.

IGV screenshot depicting representative HMGN5-interacting RNAs. HMGN5 peak track (red) and GFP control (green) are indicated. Peak high was normalized to 30 for all samples. Each track represents the merged alignment from the biological replicates (three HMGN5 and two GFP). AARS2: Alanyl-TRNA Synthetase 2, Mitochondrial; KIAA0753: OFD1 And FOPNL Interacting Protein; ASIC1: Acid Sensing Ion Channel Subunit 1. The respective genes are represented under the GFP track for each example. Exons are represented as blue boxes. Lines represent introns, and the arrows in the introns represent direction of the transcript.

When DGE analysis was performed by counting reads of exons coming from transcripts, we were able to identify 150 bound exonic fragments, corresponding to 81 transcripts (Supplementary Table 9). This indicates that HMGN5 does not bind to the complete transcript, but rather to specific motifs in the target RNAs. An example is shown in Figure 27, in which the protein is enriched in exons 16, 17 and 18 in the transcript of SEC31B (SEC31 Homolog B, COPII Coat Complex Component) but not in Exon 14.
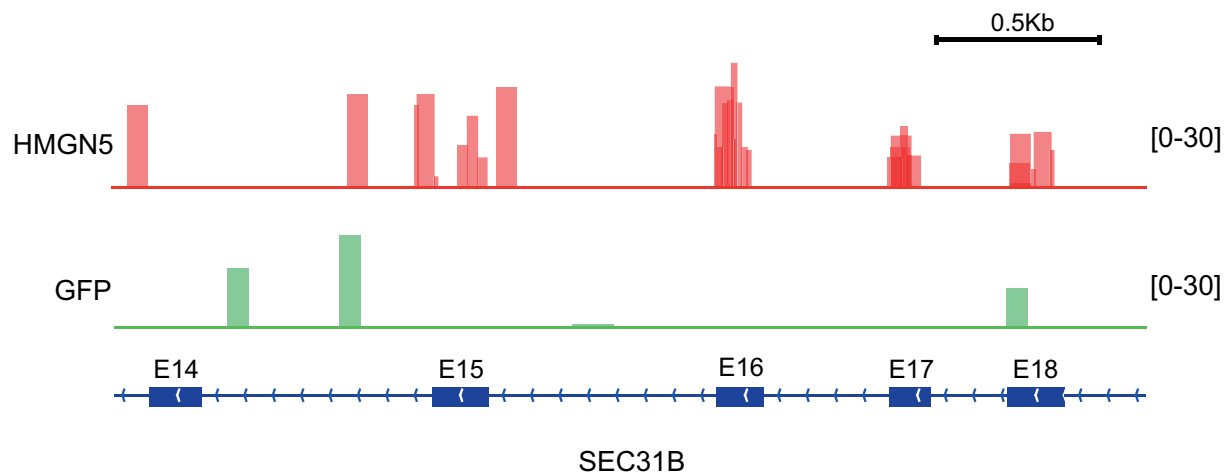

Figure 27. Distribution of HMGN5 peaks in the SEC31B transcript.

IGV snapshot of HMGN5 CLIP-seq peaks in the SEC31B RNA. Peaks from the merged HMGN5 samples (red) and GFP samples (green) are shown. Peaks coming from biological replicates are merged in the figure. The respective gene is represented under the GFP track. Exons are represented as blue boxes and enumerated accordingly to the annotation in the hg38 reference genome. A fragment of the gene from Exon 14 (E14) to Exon18 (E18) is shown. Lines represent introns, and the arrows represent the direction of the transcript. SEC31B: SEC31 Homolog B, COPII Coat Complex Component.

We analyzed the distribution of HMGN5 in the intronic sequences of transcripts, and we found a set of 251 intronic regions associated with the protein. The top 30 intronic transcript regions associated with HMGN5 are listed in Table 4. An example of the distribution is presented in Figure 28 where the HMGN5 peaks are widely distributed in the intronic regions of the gene NFIA (Nuclear Factor I A), with a 7-fold enrichment in comparison to the GFP samples. This result suggests that HMGN5 could be binding to the newly transcribed RNAs.

Table 4. Top 30 HMGN5-associated intronic regions from RNAs

Intron ID	Gene name	logFC	logCPM	p-value	FDR
16397	EPHA6	2.36536253	5.079924622	3.86E-12	9.68E-08
62341	CEBPE	3.969194718	3.276891285	6.87E-11	8.62E-07
28963	CSNK2B	-10.57284844	2.602688125	2.00E-10	1.67E-06
28962	CSNK2B	-10.32288718	2.359067287	4.61E-09	2.89E-05
16398	EPHA6	1.899411654	5.485641471	1.29E-08	6.47E-05
62340	Unidentified	7.136547677	1.936530155	1.58E-08	6.59E-05
70418	TOX3	-7.618927237	2.39595837	4.01E-08	0.00014371
4830	LRRC47	1.815717525	6.181099348	5.59E-08	0.000162901
51553	MOB2	-10.12201589	2.163675689	5.84E-08	0.000162901
28960	LY6G5B	-10.06003812	2.104242312	8.85E-08	0.000209141
25675	G3BP1	-5.1776663	2.731225356	1.08E-07	0.000209141
25674	G3BP1	-5.177663497	2.731225356	1.08E-07	0.000209141
25673	G3BP1	-5.177660694	2.731225356	1.08E-07	0.000209141
67488	WDR90	-10.06870594	2.111911219	1.38E-07	0.000246887
28961	LY6G5B	-9.932080302	1.980545806	1.56E-07	0.000246887
4914	ENO1	-4.060905817	2.989744799	1.57E-07	0.000246887
42278	AKAP2	-3.618873362	3.102655176	2.34E-07	0.000317207
42277	AKAP2	-3.61887332	3.102655176	2.34E-07	0.000317207
56763	GRIN2B	3.885392665	2.701879921	2.40E-07	0.000317207
18061	ZNF385D	2.187543531	4.07306081	3.24E-07	0.000406409
1577	NFIA	-7.238771966	1.916972301	5.72E-07	0.00065174
1578	NFIA	-7.238785294	1.916972301	5.74E-07	0.00065174
43457	TLN1	-9.589848183	1.651615196	5.97E-07	0.00065174
20638	TAPT1-AS1	-3.222021858	3.680936008	6.59E-07	0.000662033
20639	TAPT1-AS1	-3.222020606	3.680936008	6.59E-07	0.000662033
16428	ST3GAL6	-4.203696894	2.717764913	7.83E-07	0.000755889
17381	NAALADL2	1.913529548	4.185663806	8.35E-07	0.00077614
4915	ENO1	-3.991224239	2.866444377	8.73E-07	0.000782529
17380	NAALADL2	1.773638979	4.441169751	9.22E-07	0.000797771

List of RNAs generated by counting introns obtained from transcripts. The Intron ID, name of associated gene, log fold change (LogFC), Log normalized counts per million (LogCPM), p-value and adjusted false discovery rate (FDR) are shown in the list. FDR cutoff: 0.05. The negative LogFC values correspond to enrichment of HMGN5 over GFP. Genes for DGE analysis were extracted from TxDb.Hsapiens.UCSC.hg38.knownGene database of hg38 human reference genome.
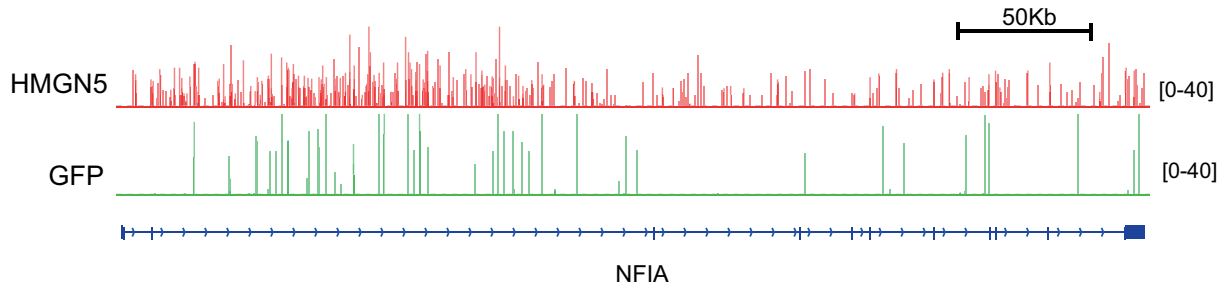
Figure 28. Distribution of HMGN5 peaks in an intronic region of gene NFIA.

The peaks from the merged HMGN5 samples in upper track (red) and GFP samples (lower track in green) in the NFIA transcript are shown. The track height was set at 40 for each sample. The tracks correspond to the overlapped peaks from the independent biological replicates. The NFIA gene is represented under the GFP track. Exons are represented as blue boxes. Lines represent introns and the arrows represent the direction of the transcript. NFIA: Nuclear Factor I A.

We performed genome ontology analysis to evaluate the distribution of HMGN5 on annotated genomic features associated with transcripts. Table 5 summarizes the HMGN5 distribution associated with basic genomic features. When we analyze the overlapped peaks in bp (strict overlap), HMGN5 is mostly found associated with introns (46671bp), intergenic regions (31153bp) exons (20020bp) and protein coding (18374bp). However, when the distribution is normalized according to the coverage of each feature in the genome (LogRE= Log ratio enrichment) we can observe an enrichment in 5' UTR, CpG islands, coding genes, exons, protein-coding and non-coding RNA. Figure 29 shows the enrichment of HMGN5 in some basic genomic features, according to the Log ratio enrichment.

We combined the list of associated transcripts obtained with the different strategies mentioned before (Table 3, Table 9 and Table 10) and we performed GO enrichment analysis. We found that the genes cannot be statistically grouped, indicating that the HMGN5-bound RNAs are not related with each other.

Table 5. Global distribution of HMGN5 in transcript-associated genomic features

Feature	Coverage feature (bp)	Overlap (#peaks)	Overlap (bp)	Expected Overlap (bp)	Log Ratio Enrichment	p-value	Log p-value
exons	84831605	73	20020	4103	1.59	5.14E-26	-58.23
protein-coding	73804459	66	18374	3570	1.64	2.83E-25	-56.53
coding	52471711	67	15453	2538	1.81	5.27E-24	-53.6
Cpg Island	23610399	29	7775	1142	1.92	1.86E-13	-29.31
utr5	7089273	16	2839	342	2.12	2.63E-06	-12.85
ncRNA	8518253	6	1972	412	1.57	2.61E-03	-5.95
introns	1302962824	186	46671	63031	-0.3	5.49E-10	21.32
gaps	159950299	0	0	7737	-8.95	3.34E-12	26.42
intergenic	1781725049	107	31153	86191	-1.02	1.68E-127	291.91

Genome ontology analysis of HMGN5 CLIP-seq. Expected overlap was estimated based on a genome size of 2.00e+09 bp. p-value cutoff 0.05. Negative log p-value= significantly associated; positive log p-value= significantly NOT associated. LogRE: Log ratio enrichment.
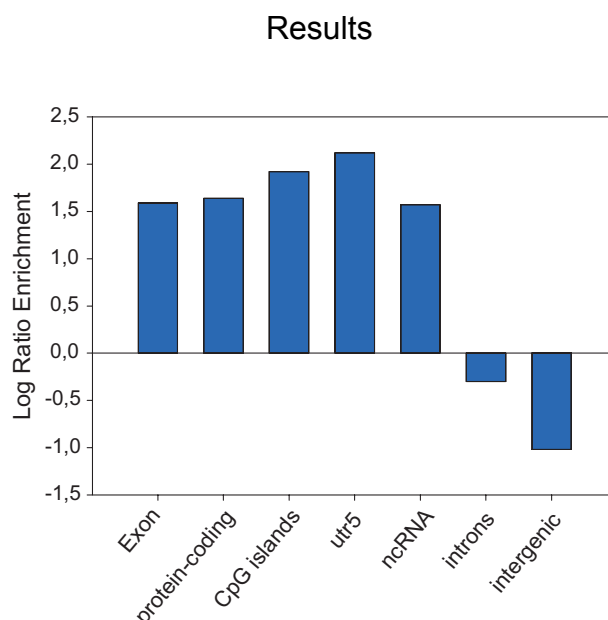
Figure 29. HMGN5 global CLIP peaks distribution.

Genome ontology analysis of the HMGN5-enriched peaks from CLIP-seq analysis. The graph shows the Log ratio enrichment of peaks in the selected genomic features (utr5: 5' untranslated; ncRNA: non coding RNA).

We performed a motif finding against the hg38 genome using 'findMotifsGenome.pl' from Homer to search for specific motif sequences recognized by HMGN5 within the RNAs.

We found a list of eight enriched *de novo* motifs (Figure 30). Interestingly, two of those motifs correspond to poly A and poly U tract, with an enrichment of 30.28% and 24.61% in the target sequences respectively, compared to 12.81% and 9.8% abundance in the background sequences. Remarkably, those two motifs are highly similar to the A-rich (En3_RNA_rev) and U-rich (En3_TFO_RNA) RNAs (Table 1) to which HMGN5 shows a specific and high affinity binding *in vitro* (Figure 12), thus validating the *in vivo* HMGN5-bound RNAs. Moreover, the preference of HMGN5 binding for A rich sequences together with the fact that we mainly identified mRNA with the CLIP-seq suggest that HMGN5 could be binding the polyA tail of mRNA thus participating in the regulation of mRNA metabolism by direct targeting the RNA.

Motif	P- value

$1e^{-27}$

$1e^{-19}$

$1e^{-17}$

$1e^{-15}$

$1e^{-14}$

$1e^{-13}$

$1e^{-13}$

$1e^{-12}$

Figure 30. Significantly enriched HMGN5 *de novo* motifs.

Top eight enriched motifs are shown ranked by p-value. Motif finding was performed with the Homer software using the hg38 reference genome.

4.9 HMGN5 forms distinct complexes either with nucleosomes or with RNA

To compare the binding ability of HMGN5 toward RNA and nucleosomes, and to get insight into the binding mechanism, we performed a competition assay of HMGN5 with RNA and nucleosome by using EMSA assays. First we assembled mononucleosomes *in vitro*, using a Cy3 labeled DNA sequence that contains the strong positioning sequence 601 (Lowary & Widom, 1998) as described in 7.2.4.1. The protein was serial diluted in a 2:1 ratio, starting from 2µM and the concentration of the nucleosome and RNA were kept constant (50ng and 50nM respectively). The result of Figure 31A reveals that HMGN5 interacts either with nucleosome or with the RNA since two distinct complexes are formed, and a super-shift representing HMGN5-RNA-nucleosome complex was not observed. Moreover, the protein binds both molecules with similar affinity, and a complete shift of RNA and nucleosome is observed at a protein concentration of 2µM (Fig 31A, lane 10). Interestingly, the presence of the RNA induces a decrease in the mobility of the HMGN5-nuclesome, as judged by comparison with the HMGN5-nucleosome sample (lane 2). This retardation in the mobility is more pronounced when the HMGN5-nucleosome interaction is prepared in presence of an excess of fragmented total RNA. Total RNA extracted from HeLa cells was subjected to $MgCl_2$-dependent fragmentation (AbouHaidar & Ivanov, 1999), as described in section 7.2.2.14. Figure 31B shows the RNA competition, in which the protein-nucleosome complex was kept constant (lanes 6-9), and increasing concentrations of non-labeled fragmented total RNA were added ranging from 300ng to 2µg. It can be observed that when RNA concentration increases, the mobility of the protein-nucleosome complex is retarded in a RNA concentration-dependent manner. This effect is not visible when only RNA and nucleosome are mixed (lane 10), indicating that the RNA has a specific effect in the HMGN5-nucleosome interaction that tends to displace the binding with the protein. Taken together, those results suggest that HMGN5 might have a dual function, as seen by the specific interaction with RNA and nucleosome.
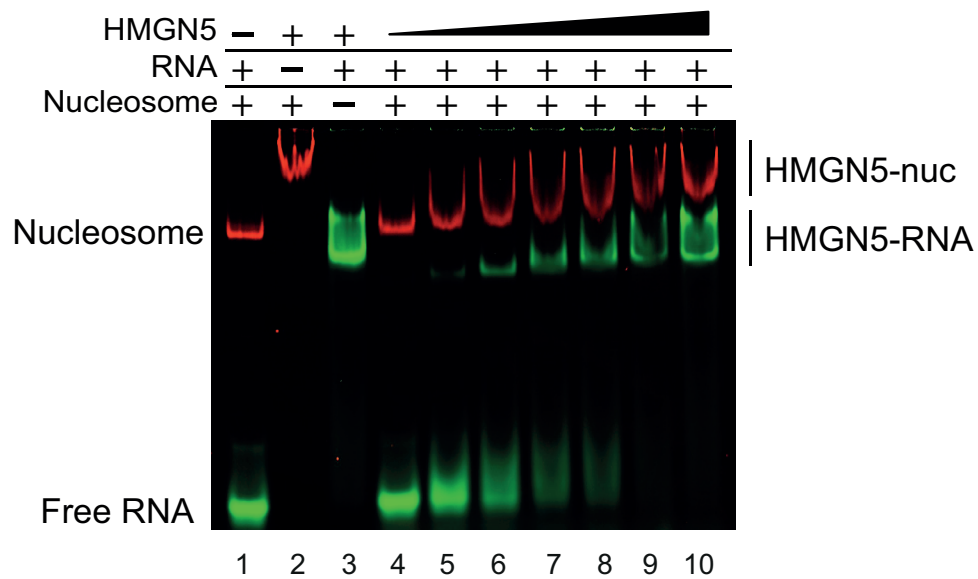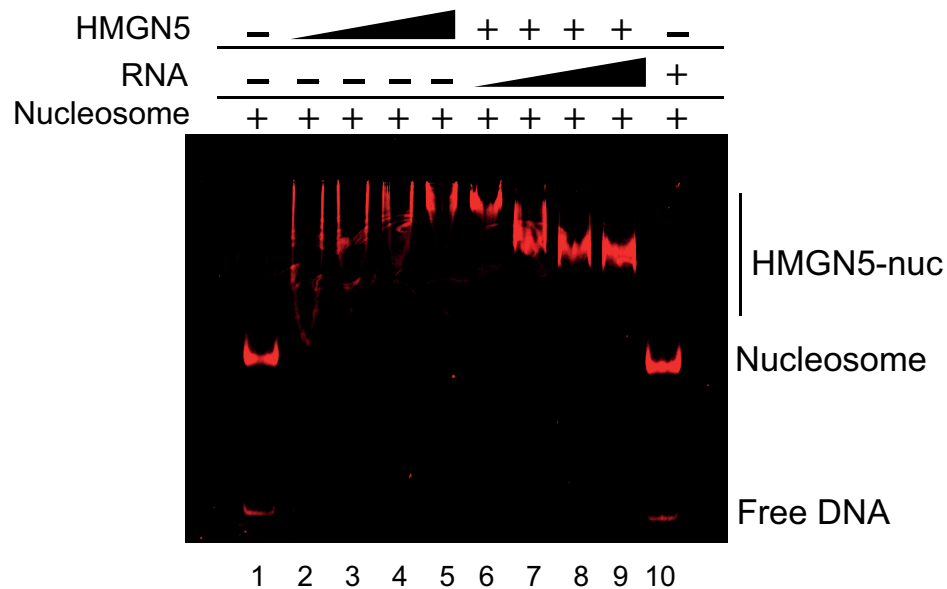
Figure 31. RNA-nucleosome competition assay.

A) Competitive EMSA to compare specificity of HMGN5 to nucleosome and RNA. 50 nM of Cy5 labeled snoRNA2T2 (green) and 50ng of Cy3 labeled mononucleosomes (red) were incubated with increasing concentrations of recombinant HMGN5 up to 2µM (titration in a 2:1 ratio). The interaction was monitored in a 6% Native PAGE and visualized in a Typhoon 9500 device. Controls of only RNA+nucleosome (Lane 1), Protein+nucleosome (lane 2) and protein+RNA (lane3) are shown. Nuc: nucleosome. B) The HMGN5-nucleosome interaction was prepared in presence of an excess of frragmented non-labeled total RNA extracted from HeLa cells. Lane 1, 50ng of Cy3 labeled mononucleosome; lanes 2-5 HMGN5-nucleosome interaction. Lanes 6-9 competition with 300ng, 700ng, 1µg and 2µg of fragmented RNA. Lane 10, 50ng of mononucleosome mixed with 2µg of total RNA. The interactions were loaded in 6% Native PAGE and visualized in a Typhoon 9500 device.

4.10 HMGN5 bind preferentially to regulatory regions and regulates RNA metabolic genes

Based on immunofluorescence analysis, it was previously shown that the human HMGN5 is mainly localized in heterochromatic regions (Malicet et al., 2011), unlike the mouse protein that localizes to euchromatin (Rochman et al., 2009). However the detailed identification of its binding across the genome is not known and is critical to understand its role in the global gene regulation and RNA metabolism.

To analyze the genome-wide distribution of HMGN5, we performed ChIP-seq analysis on the HMGN5-FlpIn cell line with and without induction with Doxycycline and on induced GFP-FlpIn cell as control (7.2.6.1).

After sequencing and quality check, a trimming step was performed using Homer tools to remove corresponding Illumina adapter sequences (methods 7.2.7.2 and appendix section 9.3.1). The reads were mapped to the hg19 human reference genome using Bowtie2. Peak finding and motif analysis were performed with Homer (v4.7). After inspection of the alignments, peaks of varying sizes were found, spread over larger distances, then a 1000bp size peak was assumed for motif finding (Histone style parameter). The enrichment of peaks over control was performed using each GFP sample as background separately for each individual HMGN5 sample, and the resulting peaks were merged. Only peaks present on both HMGN5 biological replicates with literal overlap were kept in the analysis.

The result generated after ChIP-seq anaylsis reveal 6277 specific HMGN5 peaks. The Figure 32 shows the genomic distribution of HMGN5 at three different example locus. It can be seen an enrichment of HMGN5 in comparison with the GFP samples and the background non-induced HMGN5 cell line (Samples H-1 and H-2). A high correlation is observed between HMGN5 peaks and the genomic enrichment of H3K27ac at the given positions (bottom track in figure 32).
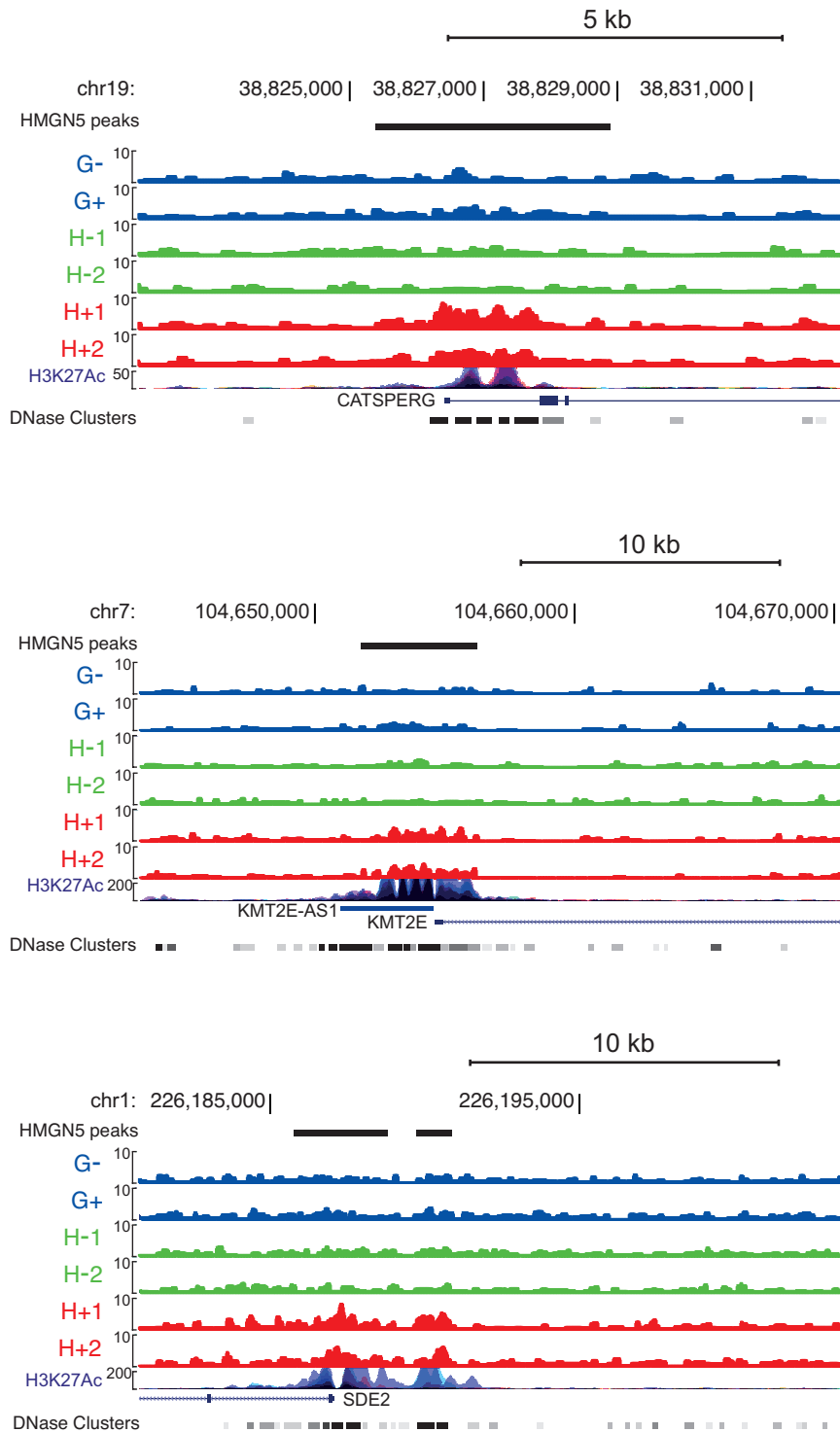
Figure 32. UCSC genome browser tracks depicting the HMGN5 distribution at three different example locus.

Blue tracks represent the GFP control cell line (G- for non-induced GFP expression, and G+ for overexpressed GFP); green tracks represent two independent biological replicates of non-induced HMGN5 control (H-1 and H-2), and in red the overexpressed HMGN5 samples (H+1 and H+2). The genome tracks of the histone mark H3K27Ac and the DNase I hypersensitive sites are also shown. Locus shown: CATSPERG (Cation Channel Sperm Associated Auxiliary Subunit Gamma); KMT2E (Lysine Methyltransferase 2E); SDE2 (SDE2 Telomere Maintenance Homolog).

Of the HMGN5 peaks annotated in the genome, almost 30% are located in promoters and 25% in intronic regions (Figure 33A). We performed Genome Ontology analysis on HMGN5 ChIP-seq to analyze the enrichment of HMGN5 in genomic element by normalizing the data according to the coverage (in bp) of feature in the genome. Strikingly, we found that the protein is highly enriched at promoters close to transcription start sites (TSSs) (Figure 33B, Table 6), corresponding to a Log enrichment of 2.88 in this feature (Log p-value -3863.96) compared with the expected random distribution. Similarly, we observe a strong enrichment in CpG islands and 5'UTR (3.1 and 3.25 Log ratio enrichment respectively), all of them representing regulatory genomic regions. Even though a great number of peaks were found at introns and intergenic regions, as they are highly represented in the genome, the enrichment analysis indicates that HMGN5 is mainly depleted from those genomic attributes.

Table 6. Genome ontology of HMGN5 ChIP-seq peaks

Annotation	Feature coverage (bp)	Overlap (bp)	Expected Overlap (bp)	Log Ratio Enrichment	Log p-value
Promoters-TSS	32837752	2340668	131932	2.88	-3863.96
CpG islands	21842742	1946823	87757	3.1	-3512.42
5'UTR	5995417	622375	24087	3.25	-1159.19
Exons	77105916	1231945	309789	1.38	-660.95
Protein-coding	68356105	1099718	274635	1.39	-590.79
ncRNA	6791910	135252	27287	1.6	-87.98
TTS	30830670	291625	123868	0.86	-68.15
Introns	1230652329	3812530	4944412	-0.26	260.01
Intergenic	1799344491	1867135	7229256	-1.35	7899.65

Ranked list of genomic features associated with HMGN5. Feature coverage: correspond to the total size in base pairs of the respective feature in the hg19 reference genome. Overlap: literal overlap of HMGN5 peaks in base pair in the respective genomic feature. Expected overlap: overlap expected by chance based on the proportion of each attribute in the genome. The Log ratio enrichment is obtained from the HMGN5 peaks overlap (bp) over the expected overlap (bp). The representative selected features are ranked according to the Log p-value. Negative Log P-value corresponds to significant association, and positive p-value corresponds to significantly NOT associated regions.

A
B
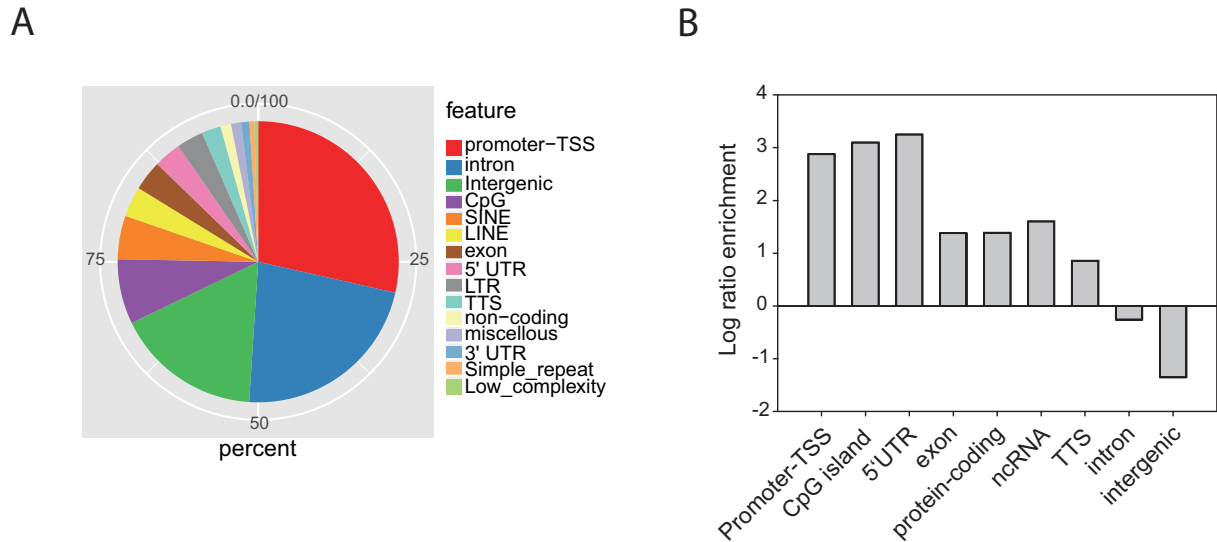
Figure 33. Genome-wide HMGN5 occupancy.

(A) Distribution of total number of identified HMGN5 peaks at annotated genomic regions. Reference percentages are indicated in the pie chart. The name of the respective features is indicated in the color legend. B) Enrichment of HMGN5 in the selected genomic features. The overlap of HMGN5 (in bp) was normalized according to the expected random enrichment at each genomic attribute. The enrichment is represented as Log values (Log ratio enrichment).

As it was shown in the previous result, HMGN5 binds preferentially to regulatory regions. Here we compared the distribution of the protein with the histone marks H3K27ac and H3K4me3, respectively enriched at promoters and enhancer of actively transcribed genes (discussed in Shlyueva, Stampfel, & Stark, 2014). Likewise, we analyzed the correlation of HMGN5 binding with DNase I hypersensitive sites (HS), with RNA polymerase II and CTCF distribution. We obtained the data from public ChIP-seq in the background cell line HEK293T available at the ENCODE project (methods section 7.1.7).

We observe a correlation in the distribution pattern of the activation histone marks H3K27ac (Figure 34A) and H3K4me3 (Figure 34B) with HMGN5 occupancy in the genome, in which the marks are spread close to the center of HMGN5 binding site with a peak flanking at less than 1000bp from the peak center.

Moreover, HMGN5 highly correlates with DNaseI hypersensitive (HS) sites (Figure 34C), hallmark of active chromatin (Boyle et al., 2008; Gross & Garrard, 1988). Furthermore we also find a correlation of HMGN5 peaks with the RNA polymerase II

occupancy (Figure 34D), which marks transcription initiation, altogether demonstrating a role of HMGN5 in transcriptional activation.

Interestingly, we found a strong correlation of HMGN5 binding with the CTCF genomic distribution (Figure 34E). CTCF distribution is observed as a narrow peak close to the center of HMGN5 binding site.

CTCF is a well-characterized transcription factor that can act as transcriptional activator, repressor or insulator and helps to the 3D organization of chromatin by mediating long-range chromatin interactions (Ghirlando & Felsenfeld, 2016). CTCF has been found distributed at domain boundaries acting as barrier that demarcates active from repressive domains (Cuddapah et al., 2008). Taking into account that HMGN5 is a protein that globally decompacts chromatin, this data suggest the either both proteins recognizes the same sequences genome-wide, and then can co-localize, or (and) they have a cooperative function, being part of a complex regulating chromatin architecture so far not described.

A

H3K27ac

B

H3K4me3

C

DNase I HS

D

RNA Pol II

E
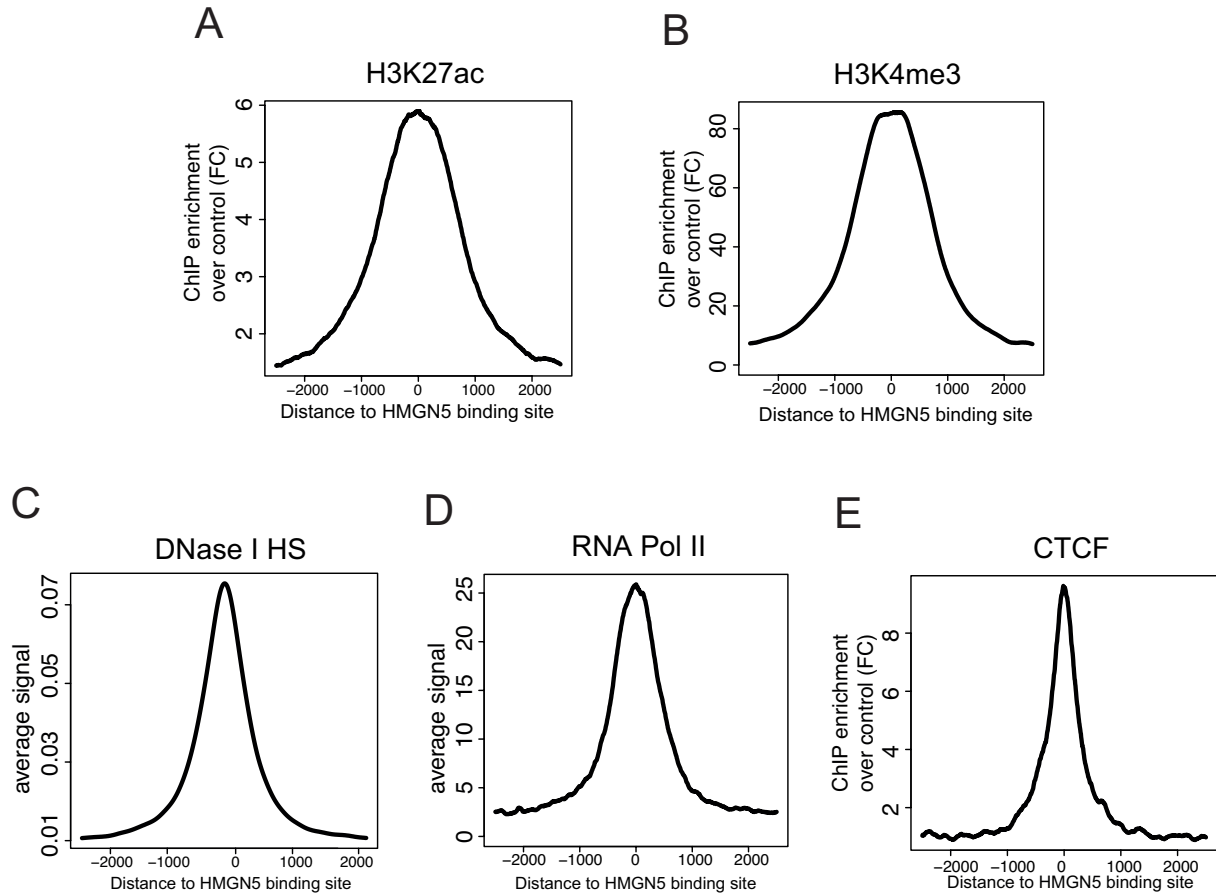
CTCF

Figure 34. Motif distribution around HMGN5 ChIP-seq peaks.

Correlation of HMGN5 occupancy with the distribution of H3K27Ac (A) H3K4me3 (B), DNase I hypersensitive sites (C), RNA Polymerase II (D), and CTCF binding (E). The average genomic distribution of the corresponding signal from the marks was plotted relative to the center of the HMGN5 binding sites (from -2000 to 2000 bp). Datasets were obtained from public available database in the human HEK293T cell line at the ENCODE project.

4.11 HMGN5-dependent transcriptional changes at the DNA-binding sites

We further analyze the HMGN5-dependent transcriptional changes at the DNA-binding sites to test the effect of HMGN5 binding at regulatory elements on the global transcriptional regulation. Figure 35A depicts the overlap between the HMGN5-dependent transcriptional changes and the HMGN5 genomic binding sites at regulatory regions, including all peaks found at promoters, CpG islands and 5'UTR. From the 2394 peaks found at regulatory region, in 882 a transcriptional change is observed in the associated gene, 326 of them only when HMGN5 is overexpressed and 306 exclusively after HGMN5 knockdown. The remaining 250 genes are affected in both conditions. It can be observed a slight tendency to a down-regulation of the transcripts of HMGN5-bound genes compared with non-bound genes in both overexpression and knockdown of HMGN5 (Figure 35B).

A

B
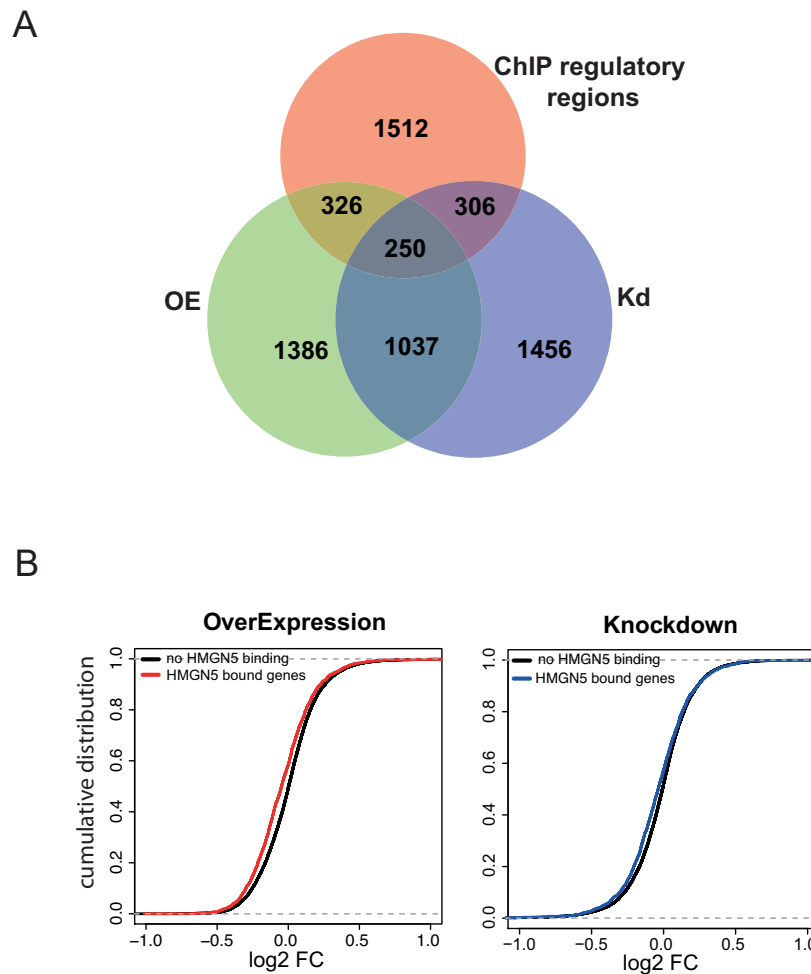
Figure 35. Correlation of HMGN5 dependent transcriptional changes and DNA-binding sites.

(A) Venn diagram depicting the overlapped set of genes affected after overexpression (OE) and knockdown (Kd) of HMGN5, with the identified HMGN5 binding sites at regulatory regions. The Venn diagram was made with the BioVenn online tool (http://www.biovenn.nl/index.php) (B) Cumulative distribution of HMGN5-dependent transcriptional changes at the HMGN5-bound genes. The x-axis shows the log2 fold change (Log2 FC) values of the transcripts and the y-axis shows the cumulative distribution.

Interestingly, when performing Gene Ontology analysis (GO) of the genes with an enrichment of HMGN5 at regulatory regions, we can observe that they are mainly associated with RNA metabolic processes (Figure 36A), being "RNA metabolic process", "translation" and "RNA processing" the most enriched terms in the analyzed dataset. Likewise, when analyzing the subset of HMGN5-bound genes at promoters, with changes in transcription level in both overexpression and knockdown

of HMGN5, the genes are functionally clustered as "posttranscriptional regulation of gene expression", "sequence-specific mRNA binding", and rRNA metabolic process" (Figure 36B and C).

This result strictly correlates with our findings showing that the transcription of genes involved in RNA metabolism is mainly affected by a loss or overexpression of HMGN5, demonstrating a direct role of HMGN5 in the regulation of the RNA metabolism. For a great number of genes the effect could be explained by a direct influence of HMGN5 binding at the associated regulatory elements. Nevertheless, from this evidence is not possible to conclude that the effect is due to a direct change in the chromatin architecture mediated by HMGN5 binding, or is due to the interaction with unknown factors.

A

Other fuctions

- RNA metabolic process
- translation
- RNA processing
- mitotic celly cycle phase transition
- posttranscriptional regulation of gene expression
- Intracellular protein transport

B

sequence-specific
mRNA binding

rRNA metabolic
process

posttranscriptional
regulation
of gene expression

C

posttranscriptional
regulation of gene
expression

rRNA metabolic
process

sequence-specific
mRNA binding

Figure 36. GO analysis of HMGN5-bound genes at regulatory regions.

A) Statistical overrepresentation analysis of gene ontology terms shown as pie chart, from the 2394 genes with HMGN5 enriched at regulatory regions (including promoters, CpG and 5'UTR). The groups are organized from higher (up) to lower (bottom) number of genes per group in the color legend, and displayed consecutively in the pie chart. B) GO analysis of the subset of genes with HMGN5 associated with promoters and that show transcriptional changes after HMGN5 overexpression and knockdown, represented as a pie chart and as a network in (C). The analysis was performed with the ClueGo software, and the significance of the enrichment was performed with a hypergeometric test. p-value correction was estimated sing the Bonferroni step-down method. The leading term correspond to the most significant term in the group. Only term with p-value <0.005 are shown.

4.12 HMGN5 preferentially binds CTCF recognition motif genome-wide

It has been widely accepted that HMGN5, as the other HMGN proteins binds the nucleosome without specificity for the underlying sequence (Bustin, 2001). As the ChIP-seq analysis shows that HMGN5 is enriched at regulatory active regions, we hypothesized that the protein indeed has a preference for binding specific DNA motifs. Using the mapped reads from ChIP-seq we performed *de novo* motif search using the 'findMotifsGenome.pl' command from Homer on hg19 reference genome. We did the analysis assuming a 1000bp region for motif finding, since in the alignment of reads to the reference genome HMGN5 was distributed in broad regions, containing peaks of varying size (200/250/300/350/400/450/500bp).

The motif analysis demonstrates that, different than previously thought, HMGN5 binds specific DNA consensus sequences. In the Figure 37 a list with the five most enriched *de novo motifs* is shown, ranked according to the p-value.
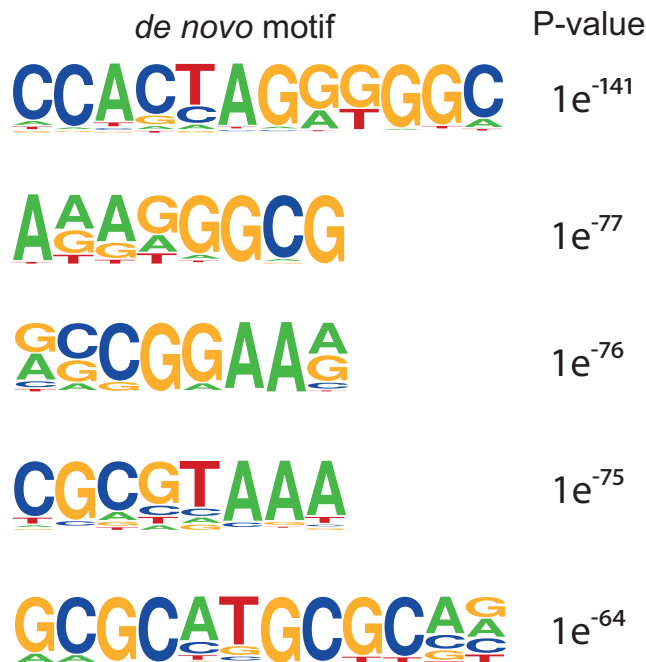
Figure 37. De novo motif analysis in HMGN5 ChIP-seq peaks.

Top five most enriched motifs in the HMGN5 ChIP-seq dataset are ranked according to the p-value. The motifs were searched using the Homer software assuming a peak size distribution of 1000bp.

Remarkably, the most enriched motif in the list, corresponds to the recognition sequence of CTCF (Figure 38) –and the paralog CTCFL- conserved in different human cell lines (Cuddapah et al., 2008; Essien et al., 2009). The motif was found in 17% of the HMGN5 peaks analyzed. It can be noticed that the 12bp HMGN5 motif, is identical to the core sequence of CTCF canonical motif, indicating that both proteins could be co-localizing in the genome. However an interaction between both proteins has not been described in the literature.

Canonical CTCF core motif
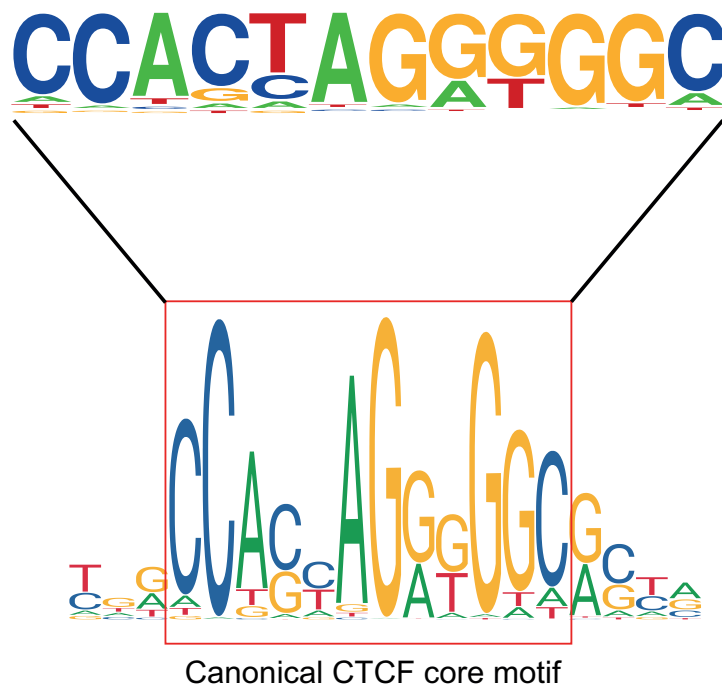
Figure 38. CTCF binding site is the most enriched motif in HMGN5 ChIP-seq data.

Most enriched HMGN5 de novo motif (p-value $1e^{-141}$) compared with the canonical CTCF motif obtained from Jaspar database. In the red box the canonical core of the CTCF motif is indicated.

For comparison, we repeated the analysis using a 200bp peak size for motif finding, which correspond to the pre-defined setting from the Homer software, standard for transcription factor peaks, and is based on the assumption that most of the motifs are found +/- 50-75 bp from the peak center (see description of "findMotifsGenome.pl" command from the Homer software). The analysis shows that the motif for CTCF is also the most significantly enriched motif for HMGN5 using this parameter, with a p-value of $1e^{-96}$ thus demonstrating the robustness of the analysis.

4.13 Identification of HMGN5-interacting partners

As already known in the literature, HMGN5 can destabilize chromatin compaction by directly interacting with H1 and H5 and thus counteracting their binding to chromatin (Rochman et al., 2009). It was also recently shown that the human HMGN5 is able to interact with the lamina-associated protein Lap2α (Lamina-associated polypeptide 2, isoform alpha) (Zhang et al., 2013). However, a global HMGN5-protein interactome that reveals potential physiological binding partners has not been described so far and is critical to better understand the function of HMGN5 and its role in the regulation of RNA metabolism.

Taking advantage of our established HMGN5_FlpIn cell line, we performed HMGN5 Co-immunoprecipitation (Co-IP) to identify potential HMGN5-interacting partners by quantitative mass spectrometry.

Since HMGN5 binds tightly to chromatin, the identification of interacting partners that binds through direct physical contact requires the fractionation of the associated DNA. Therefore, to perform Co-IP we first tested two different conditions to digest the DNA from the chromatin (standardized protocol in section 7.2.6.5). We performed a Micrococcal nuclease (MNase) digestion to generate mono-nucleosomes and a treatment with MNase plus Benzonase. The last treatment should completely degrade the DNA present in the lysed samples since the genetically engineered endonuclease Benzonase degrades the nucleic acids to <8bp fragments (according to the manufacturer). Figure 39A shows a silver-stained SDS-PAGE of immunoprecipitated HMGN5, and the control GFP, using both treatments, MNase (M) and MNase plus Benzonase (M+B) on two biological replicates each. We can see that for both proteins, the immunoprecipitation is highly efficient and the respective treatment does not affect the affinity of the antibody (GFP-Trap) for GFP, as the proteins are efficiently immunoprecipitated (indicated with a red star in the gel, at about 100kDa for HMGN5 and about 30kDa for GFP). When comparing both treatments in each sample, it can be observed that the treatment with MNase plus Benzonase produced cleaner results, since the contaminant background associated to chromatin is reduced. We repeated the Co-IP after both treatments for two new biological replicates, including two technical replicates per condition. The western

blot of the Figure 39B shows that the method is reproducible, since similar levels of HMGN5-GFP are detected with the antibody against HMGN5 after immunoprecipitation in all HMGN5 samples.
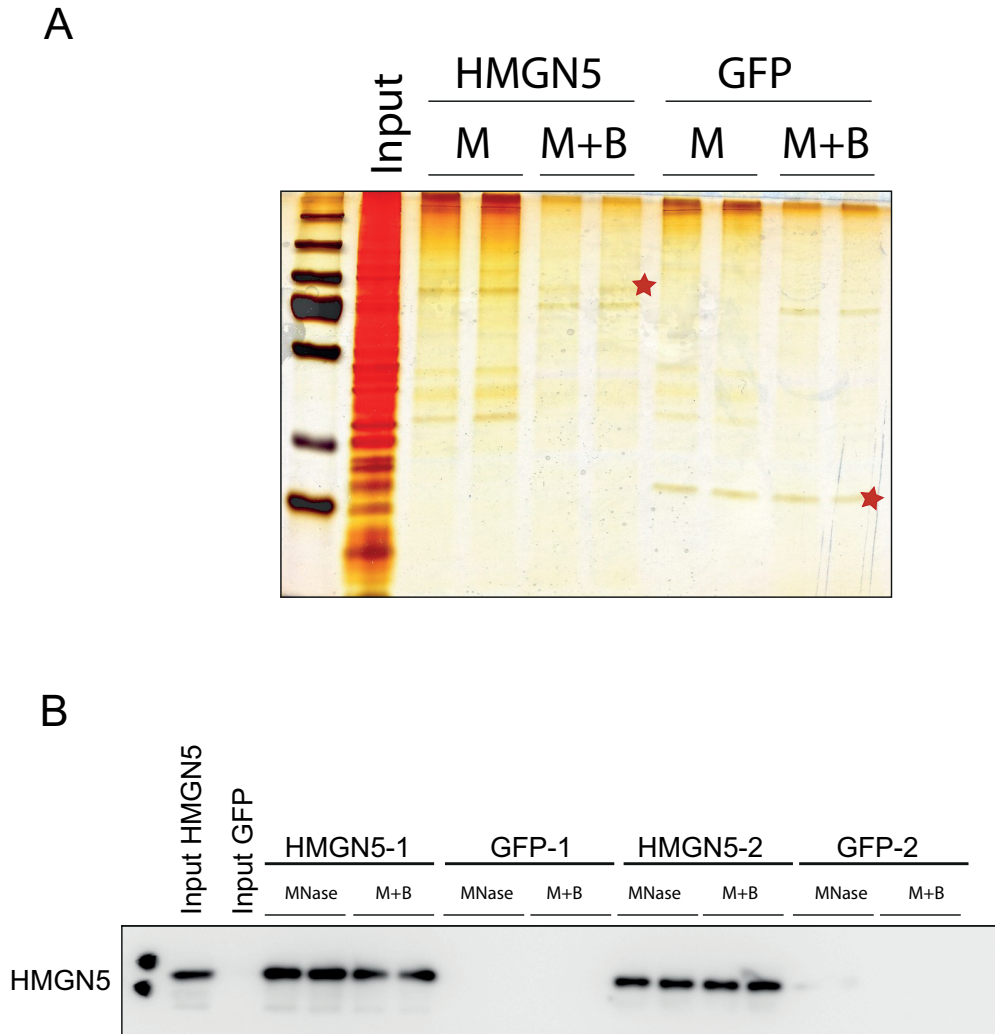

Figure 39. HMGN5 co-IP-standardization for mass spectrometry.

A) Silverstaining of test Co-IP samples. The Co-IP was done with 15µl of GFP_Trap Magnetic beads using different treatments. M corresponds to MNase digested samples and M+B corresponds to a treatment with MNase and Benzonase. 2 independent replicates were used per condition. 10% of the beads used for IP were loaded in each lane. The red star indicates the overexpressed proteins, HMGN5 at around 100kDa, and GFP at around 30 kDa. HMGN5 input was loaded as reference. B) HMGN5 western blot after CoIP. After immunoprecipitation the samples were subjected to western blot using an antibody against HMGN5 (HPA000511). Each lane corresponds to 10% of the beads from the CoIP. Two independent biological replicates were used per cell line (HMGN5-1 and HMGN5-2; GFP-1 and GFP-2), with two technical replicate per treatment.

4.14 HMGN5 binds CTCF *in vivo* and proteins regulating the pre-rRNA processing

After the Co-IP was standardized, three biological replicates per condition (from HMGN5_FlpIn and GFP_FlpIn cells as background control) were prepared and subjected to label-free quantitative proteomics (section 7.2.6.5.1) by liquid chromatography-tandem mass spectrometry (LCMS/MS), in collaboration with the laboratory of Axel Imhof, at the protein analysis unit of the Biomedical Center from the Ludwig Maximilian University. To identify the HMGN5-binding partners, the intensities of the found proteins were normalized with the 'intensity-based absolute quantification' (iBAQ) algorithm, which allows measurement of protein abundance and direct comparison between proteins. The enrichment of the protein in the three HMGN5 samples over the GFP control was calculated (log2 FC IBAQ HMGN5-GFP) and the significance tested with t-test and limma test. Figure 40 shows a volcano plot of logFC vs–log10 p-value, and the significant differentially bound proteins with more than 2 fold change (logFC=2) after p-value correction are represented as red dots (positive FC values for HMGN5 enriched over control, and negative FC the ones depleted compared with the background).
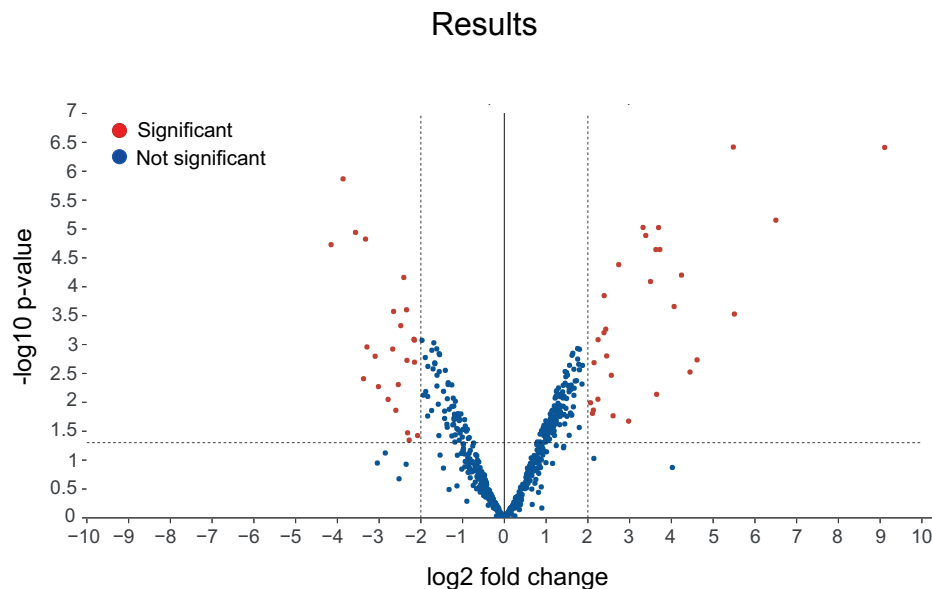
Figure 40. Significant enriched proteins in HMGN5 Co-IP.

Graphical representation of quantitative HMGN5 Co-IP mass spec data illustrated as a volcano plot, according to statistical significance -log10 p-value (y-axis) and log2 fold change HMGN5 over GFP. Significant enriched proteins in the three biological replicates with a log2FC≥2 (x-axis) are shown as red dots.

The mass spectrometry analysis led to the identification of 21 proteins that interact with HMGN5 (Table 7). As prove of the method, the identification of HMGN5 itself is shown in the table (LogFC=9.1). Strikingly, we demonstrated that HMGN5 directly interacts with CTCF. This finding is in agreement with our ChIP-seq results, which show a CTCF distribution close to the HMGN5 peaks (Figure 34E) and with the motif enrichment showing that HMGN5 has the same DNA recognition motif as CTCF (Figure 38). Together, those results suggest a new chromatin regulatory mechanism mediated by HMGN5-CTCF complexes. This hypothesis, however, needs to be validated with mechanistic approaches to better understand the potential function of such a complex.

Remarkably, from the list of HMGN5 associated proteins, we found seven proteins that play a direct role in ribosomal biogenesis. Among them, we found RPL10 (Ribosomal protein L10), which is one of the ribosomal proteins from the 60S large ribosomal particle (Lo et al., 2010). Interestingly, the other six proteins, LAS1L (Ribosomal biogenesis protein LAS1-Like), PELP1 (Proline-, glutamic acid- and leucine-rich protein 1), TEX10 (Testis-expressed sequence 10 protein), WDR18 (WD repeat-containing protein 18) and SENP3 (Sentrin-specific protease 3) NOL9

(Nucleolar protein 9), are all components of a protein complex involved in the pre-rRNA processing and synthesis of the 60S ribosomal subunit (Castle, Cassimere, & Denicourt, 2012; Finkbeiner, Haindl, & Müller, 2011). This complex has been also been linked to (de)sumoylation process in transcriptional regulation (Fanis et al., 2012). The interaction of HMGN5 with this complex suggests a direct role of HMGN5 in the regulation of ribosomal biogenesis.

Furthermore, we also found enriched the BANF1 protein (Barrier to autointegration factor), with a log2 FC ratio of 6.5, the highest ratio from all the identified partners. BANF1 (alternatively named as BAF) is a key component of the nuclear lamina. It has been described that BANF1 interacts with DNA and different groups of proteins, including the LEM domain containing proteins of the nuclear envelope -like LAP2β and LAP2α- histones and lamina proteins (reviewed in Jamin & Wiebe, 2015). BANF1 is essential for nuclear assembly and chromatin organization, and it is involved in transcription regulation, DNA damage response and immunity against foreign DNA (Jamin & Wiebe, 2015). This results support the hypothesis of a potential role of HMGN5 in linking the nuclear lamina to chromatin organization, as it was previously proposed (Naetar, Ferraioli, & Foisner, 2017; Zhang et al., 2013).

Table 7. List of identified HMGN5-binding partners

Symbol	Gene name	log FC IBAQ (HMGN5-GFP)	p.(HMGN5-GFP) limma	NegLogPval limma
HMGN5	High mobility group nucleosome-binding domain-containing protein 5	9,107876181	3,87313E-07	6,4119382743
BANF1	Barrier-to-autointegration factor	6,5017024453	7,03747E-06	5,1525833126
SFXN3	Sideroflexin-3	5,5084969494	0,000296229	3,5283727645
SSR3	Translocon-associated protein subunit gamma	5,4832879267	3.81926E-07	6,4180206783
RPL10	60S ribosomal protein L10	4,6149484045	0.001840618	2,7350364286
HIST1H2BL	Histone H2B type 1-L	4,4458019511	0.002993114	2,5238766954
RAI1	Retinoic acid-induced protein 1	4,2418793154	6.28072E-05	4,201990359
MDC1	Mediator of DNA damage checkpoint protein 1	4,0649654271	0.000219704	3,6581621747
LAS1L	Ribosomal biogenesis protein LAS1L	3,7235194994	2.26486E-05	4,6449587916
TEX10	Testis-expressed sequence 10 protein	3,6921512227	9.4361E-06	5,0252076059
PELP1	Proline-, glutamic acid- and leucine-rich protein 1	3,6289528062	2.26823E-05	4,6443133664
MRPS34	28S ribosomal protein S34, mitochondrial	3,4994632048	8.11635E-05	4,0906393883
WDR18	WD repeat-containing protein 18	3,3864191207	1.29558E-05	4,8875341001
SENP3	Sentrin-specific protease 3	3,3209894845	9.40407E-06	5,0266840162
NOL9	Polynucleotide 5'-hydroxyl-kinase NOL9	2,7380299096	4.13849E-05	4,3831585583
RFC2	Replication factor C subunit 2	2,5615257116	0.003402562	2,468194003
BIRC5	Baculoviral IAP repeat-containing protein 5	2,4494346305	0.001573481	2,8031385063
SLC25A13	Calcium-binding mitochondrial carrier protein Aralar2	2,4282198421	0.000541728	3,2662189037
PBRM1	Protein polybromo-1	2,3899131448	0.0001424	3,8464906259
CTCF	Transcriptional repressor CTCF	2,3859255012	0.000624106	3,2047416097
SUZ12	Polycomb protein SUZ12	2,2442595021	0.000823379	3,084400468
NDUFA10	NADH dehydrogenase [ubiquinone] 1 alpha sub-complex subunit 10, mitochondria	2,1477477616	0.002059266	2,6862876094

Proteins identified by LC-MS/MS are shown and ranked according to the log fold change of HMGN5 over GFP after IBAQ normalization (logFC IBAQ HMGN5-GFP). Threshold LogFC ≥2. Statistical significance after limma test is shown (p(HMGN5-GFP)limma and NegLogPval limma).

5 Discussion

In the past years, many efforts have been made to understand regulation and the dynamics of chromatin structure. This is of particular interest because chromatin structure plays an essential role in transcriptional regulation. However, its structural organization remains unclear even though a hierarchical, highly organized structure is shown in text books (Fussner et al., 2011; Maeshima et al., 2010). Indeed, the 3D chromatin ultrastructure in intact cells was recently analyzed, and it was shown that the chromatin is organized as flexible and disordered granular chain with diameters between 5- to 24 nm (Ou et al., 2017), not representing the classical view of the chromatin folding classically described.

It has been shown that HMGN5 has a direct influence on the global chromatin structure with a high impact on transcriptional regulation. The ability of HMGN5 to induce chromatin decondensation -as seen by others (Rochman et al., 2009) and our own work (Figure 9)- makes the protein a good candidate to study the mechanism of decondensation of higher order structure of chromatin.

In this work we use the human HMGN5 as a model protein to study the opening of higher order structure of chromatin, using genome-wide integrative analysis and *in vitro* techniques, with the aim of helping elucidating the mechanism by which the dynamics of chromatin opening is regulated.

5.1 HMGN5 has a novel RNA binding activity

In the past years it was demonstrated in our laboratory that a fraction of snoRNA that is associated with chromatin plays a critical role in the reversible opening of higher order structures of chromatin in *Drosophila* (Schubert et al., 2012). Furthermore, it was demonstrated that the *Drosophila* decondensation factor 31 (Df31) -first characterized by its ability of decondensing sperm DNA (Gilles Crevel, 2000)- binds chromatin in an RNA-dependent manner (Schubert et al., 2012).

In humans, HMGN5 shares common functional and structural features with Df31, such as the ability of chromatin decompaction and its intrinsically disordered structure and negatively charged aminoacids (Malicet et al., 2011).

Here we explored the potential role of HMGN5 in RNA binding by using *in vitro* techniques like EMSA and MST, and by *in vivo* RNA immunoprecipitation using CLIP-seq analysis.

The recombinant human HMGN5 was used for an interaction screening using different fluorescently labeled single-stranded nucleic acids. As shown in Figure 12, HMGN5 exhibits a novel and specific RNA binding activity, not described so far in the literature. The protein has a strong preference for binding ssRNA over ssDNA containing the same sequence, implying a potential function *in vivo*. It is important to mention that -from the *in vitro* data- HMGN5 does not have an apparent preference for molecules with secondary structures over non-structured ones, as very similar affinities were determined for the en3_RNA_rev, en3_TFO_RNA, without predicted secondary structure, in comparison with the snoRNA2T1 and snoRNA2T2 which form a high number of dsRNA bonds. Interestingly, when analyzing the EMSA interaction of HMGN5 with the snoRNA2T2 (Figure 12A, bottom right EMSA) two distinct protein-RNA complexes are formed, as revealed by the EMSA bands. With increasing molar ratio of HMGN5 to RNA increases (up to 2μM protein concentration) only one major complex remains, corresponding to the higher molecular weight complex. This effect could be mediated by cooperative binding, as it has been demonstrated for other nucleic acid binding proteins (Alfano et al., 2004; Schumacher et al., 2002). Moreover, this result corresponds to the binding mode also described for the HMGN proteins with nucleosomes particles *in vitro*, in which it has been reported the formation of complexes with nucleosome particle *in vitro,* in which it has been reported the formation of complexes as homodimers under physiological

conditions (Postnikov, Trieschmann, Rickers, & Bustin, 1995; Shirakawa et al., 2000).

5.1.1 Intrinsic disorder of HMGN5 in RNA binding

Two initial studies revealed the global protein-mRNA interactome in human cell lines, by using large-scale proteomic analysis (Baltz et al., 2012; Castello et al., 2012), showing more than expected RNA binding proteins. Further studies identified RNA-interacting proteins *in vivo* on a global scale (reviewed in Ryder, 2016), showing that the mRNA interactome is enriched on proteins containing repetitive motifs of low-complexity sequences (LCS) which are intrinsically disordered in their native state. Moreover, recent evidence suggests that the intrinsically disordered regions may contribute to RNA binding (Järvelin, Noerenberg, Davis, & Castello, 2016).

When analyzing HMGN5, we can observe that HMGN5 is generally disordered (Figure 41). An effect attributed mainly to its negative net charge (isoelectric point of 4.2.)
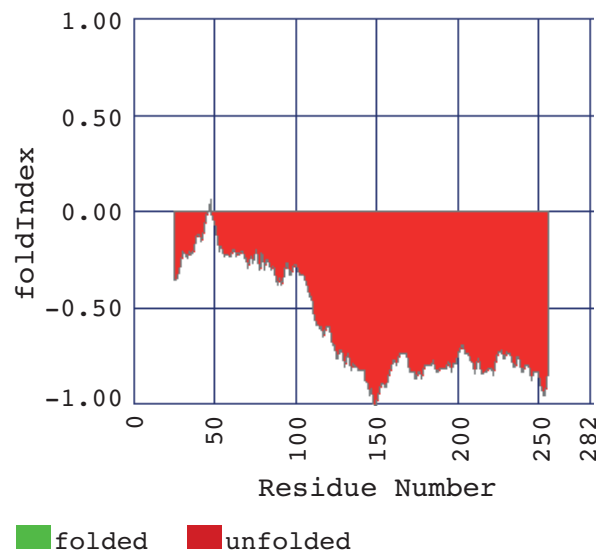
Figure 41. Intrinsic disorder prediction of HMGN5.

The folding prediction of the protein was estimated using the online software FoldIndex. The residue positions in the protein (x-axis) are plotted against the folding tendency (y-axis) in a scale between -1 and +1. Negative values correspond to unfolded residues (red) and positive values represent folded residues (green).

Discussion

A prediction of protein domains involved in RNA binding using the available prediction methods is not feasible. Here we aimed to analyze the contribution of the different regions of the protein to its RNA binding ability, based on the current characterized nucleosomal binding function of HMGN5. As already mentioned in the introduction (section 2.6), the HMGN family is characterized by a conserved N-terminal nucleosomal binding domain, responsible for its binding to chromatin *in vivo* (Bustin, 2001).

The mutational analysis (Figure 13) shows that the nucleosomal binding domain is required for the RNA binding, but is not sufficient by itself for this activity. Indeed, we demonstrated that the N-terminal and C-terminal domains of the protein synergize in trans to bind RNA (Figure 14), in which the negatively charged C-terminal tail of the protein stabilizes the RNA binding that is mediated by the N-terminal domain containing the NBD.

As the NBD has the major contribution to the RNA binding activity of HMGN5, and as this region is highly conserved in the HMGN family, we hypothesized that the RNA binding is conserved in the HMGN proteins.

Interestingly, we demonstrated here that HMGN2, a 90 amino acids long human HMGN protein (one third of human HMGN5), is also able to specifically interact with RNA (Figure 16B and C). Nevertheless, despite the fact that HMGN2 specifically binds RNA molecules *in vitro*, HMGN5 shows higher affinity for the RNA binding compared with HMGN2.

As the N-terminal domain between both proteins is conserved, we hypothesize that the higher binding affinity of HMGN5 is due to the long negatively charged C-terminal domain. We base this assumption on the observed *in vitro* cooperative binding between the N-terminal and C-terminal part of the protein.

Remarkably, the protein HMGB1, for which an RNA binding activity has been previously shown with *in vitro* methods (A. J. Bell et al., 2008), does not bind the RNA molecules at the given concentrations. This indicates the specificity of HMGN binding to ssRNA.

The intrinsically disordered nature of HMGN5 should provide the flexibility to interact with multiple binding partners, a phenomenon known as promiscuity in binding (discussed in Mollica et al., 2016). Indeed, it was already shown that the C-terminal tail confers HMGN5 the ability to counteract the H1 binding to nucleosomes, thereby destabilizing the chromatin structure. Likewise, it was proposed that this region is

responsible for the specific chromatin location and nucleosomes interaction *in vivo* (Rochman et al., 2009). However, the negative charge and the unstructured nature of HMGN5 cannot explain the specific RNA binding and the chromatin decompaction activity, since the human protein Parathymosin (PTMS) which is also highly acidic (isoelectric point of 3.2) and unstructured, does not bind RNA *in vitro* (appendix Figure 47) neither decompacts chromatin in the LacI/LacO tethering system (Figure 10). Likewise, it was already shown that the sequence, rather than the charge of the C-terminal tail of HMGN5 was important for the chromatin localization, judged by immunofluorescence experiments (Rochman et al., 2009).

It is important to mention that PTMS was shown to interact with H1 and to globally decompacts chromatin (Martic, 2005). However, there was only one study showing this -published more than 20 years ago- and using different methods to analyze decompaction. The fact that we could not observe a decompaction of the LacO array indicates that PTMS does not have this function in the U2OS system, but may have a cell-type specific function, or the chromatin reorganization ability is not as dramatic as it is observed for HMGN5.

Recently a large-scale identification of the RNA-interacting sites of native RNA-binding proteins (RBPs), termed RBDmap, was published (Castello et al., 2016). In this study, about half of the RNA-binding sites identified mapped to intrinsically disordered regions (IDR) of the proteins. Among them, an RNA binding motif formed by Lysine (K) combined with negatively charged residues, and G, P, or Q was identified. This motif was repeated in low complexity regions of the AHNAK protein (neuroblast differentiation-associated protein).

Interestingly, HMGN5 possesses a similar motif repeated across the C-terminal tail. The motif EDGKE is repeated 11 times in the mouse HMGN5 and 4 times in the human protein, but the functional relevance is still unknown. We suggest that the repetitive motif EDGKE contributes to the stabilization of RNA-binding mediated by the C-terminal tail of HMGN5. However, this hypothesis has to be tested.

5.2 HMGN5 binds RNA *in vivo*

As the *in vitro* binding analysis strongly suggested the binding of HMGN5 to RNA, we attempted to elucidate the binding partners of HMGN5 in cells. With the use of our established HMGN5_FlpIn cell lines, we performed UV-crosslinking RNA immunoprecipitation using an optimized protocol that includes a chromatin fractionation step prior to RNA immunoprecipitation. Moreover, since the overexpressed HMGN5 protein was fused to GFP, we used a cell line overexpressing GFP (GFP_FlpIn) to control for background binding of RNA, and we also skipped the radiolabeling and transfer step to a nitrocellulose membrane for RNA isolation. The success of our method can be seen, even by the difference observed between the libraries of HMGN5 bound RNAs compared with the GFP, as the biological HMGN5 replicates precipitated high levels of RNA in contrast to GFP (Figure 44).

Interestingly, when analyzing the global peak distribution by genome ontology analysis (with feature size normalization), the protein binds predominantly protein-coding RNAs and is enriched at 5' untranslated region (results section 4.8). Nevertheless, this has to be interpreted with caution, as the different transcripts are not transcribed equally well, therefore the enrichment also depends on the transcriptional levels of the RNAs.

Instead, the analysis of the overlap of HMGN5 with specific genomic features without normalization by feature size is more informative, and reveals that the major HMGN5 peaks are found at intronic regions (46671bp overlap), compared with exons, in which there is less coverage (20020bp overlap). This suggests that the protein is binding co-transcriptionally to the nascent RNA, indicating a regulatory role of HMGN5 in RNA metabolism.

The identification of specific consensus motifs in the HMGN5-bound RNA highlights that binding is sequence-dependent. Remarkably, one of those motifs corresponds to a poly-A tract, highly suggestive of a potential binding of HMGN5 to the poly-A tails of mRNAs –which are the most abundant HMGN5-bound RNAs discovered- thus participating in the regulation of mRNA metabolism by direct targeting the RNA. Unfortunately, as the poly-A tails correspond to repetitive sequences that do not map uniquely, they are filtered out from the mapping step of the analysis. Therefore, with

the pipeline we applied here we cannot show a direct binding of the protein to those sequences.

It has been reported in the literature (and also in this study, discussed in detail in section 5.3) that the deregulation of HMGN5 affects a large set of genes in a cell-specific manner in mouse and human models (Kugler et al., 2013; Malicet et al., 2011; Rochman et al., 2009; 2011), but this is the first time that a direct interaction of the protein with RNA is shown.

As the main described function of HMGN5 is the opening of chromatin structure through direct interaction with the nucleosome particle, we tried to explore the potential mechanism of RNA and nucleosome binding (results section 4.9). To our surprise, HMGN5 is able to interact with either nucleosome or with RNA, suggesting that HMGN5 may have a bimodal state or constitute a switch that enables HMGN5 binding either to chromatin or to RNA. This hypothesis is reinforced by the fact that the interacting RNAs do not have a function related to the regulation of chromatin structure. The RNAs are not clustered in any specific pathway, supporting the dual regulatory function suggested by the *in vitro* data. Still and all, we cannot exclude the possibility of existence of RNA molecules (not detected in the present study) that may cooperate with the chromatin reorganization mediated by HMGN5.

5.3 HMGN5 is coupling global chromatin architecture and gene expression

It has become clear that the spatial organization of chromatin plays a central role in regulating gene expression. Several attempts have been made to determine the relationship chromatin structure-function in diverse nuclear processes, specifically the ones that involve transcriptional reprograming, like differentiation (Harada, Ohkawa, & Imbalzano, 2017), embryogenesis (Xu & Xie, 2017), or cell cycle regulation (Liu et al., 2017). Nevertheless, there is still a gap of knowledge on how the opening of higher order structure of chromatin can control gene expression. In an effort to try to understand the contribution of HMGN5 to this process, we made an integrative analysis of the genome-wide HMGN5 distribution with the transcriptional changes observed after deregulation of HMGN5.

Discussion

As already mentioned, several studies have revealed that HMGN5 regulates the transcription fidelity of the cells. Using Affimetrix expression arrays it was demonstrated that an overexpression of HMGN5 altered the expression of more than 2000 genes in mouse AtT20 cells (Rochman et al., 2009), and more than 3000 genes in mouse embryonic fibroblast cells (MEF) (Rochman et al., 2011). The human HMGN5 was shown to alter the expression of over 350 genes by approximately two fold in a human breast cancer cell line (Malicet et al., 2011). Moreover, the transcriptional regulation mediated by HMGN5 was shown to be tissue-specific (Kugler et al., 2013). Here we analyzed the effect of HMGN5 deregulation on the transcription profile using the HMGN5_FlpIn cell line. Our results are in line with the previous findings, as we observe a large set of genes (up to 3000 per condition) being affected due to overexpression or knockdown of HMGN5 (see results section 4.6). Interestingly, a large set of genes (1287) is overlapping between HMGN5 up- and down regulation, showing the same type of change. In contrast to the expectation, most of these genes are, for example, repressed in the absence of HMGN5, and also when HMGN5 is overexpressed. We suggest that HMGN5 is part of a multifactor complex (or complexes) and the depletion or overexpression of HMGN5 may produce an imbalance in stoichiometry (Veitia & Potier, 2015), leading to the same phenotype observed in both conditions. A similar effect is described in (P. Wang et al., 2014b).

Interestingly, the GO enrichment analysis indicates that the missregulation of HMGN5 levels by overexpression or knockdown affects genes grouped in different nuclear processes. As an example, both, up- and down regulation of the protein result in the down regulation of genes associated with cell cycle regulation. These results are consistent with previous findings showing that HMGN5 plays a critical role in cell cycle progression and that is directly involved in tumorigenesis. Indeed, it was consistently shown by different studies that the knockdown of HMGN5 by RNA interference reduces cell proliferation, and enhances the apoptotic process, making HMGN5 a promising candidate for therapeutic targeting in different cancer types (reviewed in Shi et al., 2015).

So far only one published study analyzed the binding of HMGN5 in the genome by high throughput sequencing (Zhang et al., 2013). However this study was focused on the context of the reciprocal influence of HMGN5 and the lamina-associated protein LAP2α in their genomic distribution, and did not analyzed an exhaustive genome

occupancy profile of HMGN5. Unfortunately the raw data of their analysis is not available for comparison with our results (they state the analysis as preliminary ChIP-seq results).

When analyzing the strict overlapped distribution of HMGN5 with different genomic features in our data, we can observe that around 30% of the total peaks were found associated with promoters close to the transcription start site (TSS), and 23% were found associated with intronic regions (section 4.10). However, when normalizing the data with respect to the size distribution of each feature in the genome, the results reveal that HMGN5 is highly enriched at regulatory regions, including promoter-TSS, CpG islands and 5'UTR and is, indeed, depleted from introns.

Moreover, the comparison with high throughput sequencing dataset in the HEK293T (background cell line of the stable HMGN5_FlpIn) available at the ENCODE project, reveals that HMGN5 highly correlates with the process of transcriptional activation, as revealed by the co-localization with RNA polymerase II distribution (Cho et al., 2016)-, and the histone marks H3K4me3 and H3K27ac, respectively being associated with promoters of active genes and active enhancers (Shlyueva et al., 2014). Although the sole presence of those markers does not constitute an indication of transcriptional activity -as the histone modification H3K4me3 has been lately shown to be associated with 75% of all protein-coding genes in human embryonic stem (ES) cells (Guenther, Levine, Boyer, Jaenisch, & Young, 2007)-, still the high correlation of the genome-wide HMGN5 distribution with the DNase I Hypersensitive Sites (DHSs), a hallmark of open chromatin regions -marking diverse classes of cis-regulatory elements (Boyle et al., 2008; He et al., 2014)- together with the histone marks, RNA polymerase II and CpG island binding, reveal that HMGN5 is associated with transcriptional activity.

Interestingly, from the total peaks found at regulatory regions, including promoter-TSS, CpG islands and 5'UTR, almost 40 percent (882 peaks) are found associated with genes that show transcriptional changes after deregulation of HMGN5. Of those peaks, 250 were found in regulatory regions of the genes that change transcription after both up-and down regulation of HMGN5. This result reinforces the suggested direct role of HMGN5 in gene regulation.

Strikingly, more than 50 percent of those genes can be classified as genes involved in RNA metabolism (Figure 36A). Moreover, when analyzing only the genes with HMGN5 enrichment at promoter regions close to the transcription start site (TSS)

found overlapping with the RNA-seq profile of HMGN5 missregulation by both, HMGN5 overexpression and knockdown, the group is specifically categorized in RNA related metabolic processes, like post-transcriptional regulation of gene expression, rRNA metabolic process and sequence-specific mRNA binding. This demonstrates a direct role of HMGN5 in the regulation of RNA metabolism, explained by a direct influence of HMGN5 targeting at the associated regulatory regions.

Nevertheless, it is important to mention that for several of the genes with altered transcription and with RNA metabolism-related functions in the RNA-seq profile, HMGN5 was not found binding regulatory regions, thus we believe that the transcriptional changes observed in those cases correspond to an indirect effect of HMGN5, like subsequent effects of altered gene expression or due to the binding with interacting factors.

Remarkably, the fact that HMGN5 binds directly to RNAs in our CLIP-seq data, and its apparent preferential binding to introns, make us speculate that the transcriptional regulation is modulated at different levels. 1) pre-transcriptional regulation given by the recruitment of HMGN5 to regulatory genomic regions and thus reorganizing the chromatin, and 2) a co-transcriptional regulation given by the interaction of the protein with the nascent transcript. This apparent dual function has been recently seen for other chromatin binding proteins. As an example, EZH2 (Enhancer of Zeste homolog 2), the catalytic subunit of the Polycomb repressive complex 2 (PRC2) - histone methyltransferase required for maintaining epigenetic silencing during development and cancer (Davidovich, Zheng, Goodrich, & Cech, 2013)- occupies active promoters while contacting the 5' region of nascent transcripts. This effect was correlated with decreased levels of H3K27me3. This model of binding was suggested to be a mechanism by which PRC2 senses the activation states of promoters through contacts with nascent RNAs (Kaneko, Son, Shen, Reinberg, & Bonasio, 2013). As additional example, it has been shown that CTCF plays a direct role on splicing regulation by promoting the pausing of RNA polymerase II (Shukla et al., 2011).

Moreover, in a recent study, using a new high-throughput technology called "Systematic Parallel Analysis of Endogenous RNA Regulation Coupled to Barcode Sequencing" (SPAR-seq) it was revealed that hundreds of proteins, including transcription factors, DNA-binding and chromatin binding proteins, affect the Alternative Splicing (AS) (Han et al., 2017). In various cases by a dual mechanism by

directly binding the RNA adjacent to target exons and affecting in parallel the expression of splicing factors that control the AS (Han et al., 2017).

Remarkably, we do observe that for the 126 identified HMGN5-bound RNA molecules, 56 exhibit transcriptional changes when HMGN5 is deregulated. From those, 23 specifically change when HMGN5 is overexpressed, and 16 change only after HMGN5 knockdown. The remaining 17 RNAs exhibit transcriptional changes after up- and down regulation of HMGN5. This result emphasizes the dual role of HMGN5 in coupling chromatin architecture with transcriptional regulation, probably acting as a sensor of the transcriptional state of the cells and translating the information by modifying the chromatin structure at the associated regulatory regions. This effect could be mediated by interaction with regulatory factors, as proteins (see section 5.4) and possibly with regulatory RNAs that were not uncovered by our method.

It is important to notice that for all the identified interacting RNAs the transcriptional changes are mild, with log2FC less than 1. However, we are confident with the fact that those are biologically significant changes, since in the method the counts are normalized by transcript abundance (read count per gene), passing multiple-test adjustment to determine the statistical significance. Moreover, the qPCR validation in three independent biological replicates of the gene ZCCHC12 (Zinc Finger CCHC-Type Containing 12) which has a Log2FC of 0.6 in the RNA-seq, showed a similar transcriptional change in the samples tested by qPCR (Figure 20), indicating the consistency in the expression changes.

5.4 A Regulatory HMGN5-CTCF network

Strikingly, the analysis of our ChIP-seq data reveals that there is a strict correlation of the genome-wide HMGN5 distribution with the CTCF occupancy (Figure 34E). Our *de novo* motif analysis shows that the most enriched motif in the HMGN5 ChIP-seq corresponds to the CTCF (and the paralog BORIS/CTCFL) binding site (Figure 38), a conserved sequence in different human cell lines (Cuddapah et al., 2008; Essien et al., 2009). CTCF motif was found in 17 percent of total HMGN5 peaks in the genome, indicating a high co-occupancy of both proteins.

First described as a transcription factor (TF) (Klenova et al., 1993; Lobanenkov et al., 1990), the highly conserved zinc finger protein CTCF (CCCTC-binding factor) has been shown to play a key role in regulating transcription at different levels. It has been identified as a repressor of transcription (Burcin et al., 1997; Klenova et al., 1993), as transcriptional activator (Vostrov & Quitschke, 1997) in heterologous reporter assays, playing a role in gene imprinting, and X-chromosome inactivation (Phillips & Corces, 2009). However, the principal function of CTCF is related to the organization of the 3D genome architecture by mediating long-range inter and intra chromosomal interactions over tens or even hundreds of kilobases creating large loop-domains, thus bringing gene enhancers into proximity with their target promoters (Ong & Corces, 2014). A widely described long-range interaction mediated by CTCF is the insulation activity (A. C. Bell, West, & Felsenfeld, 1999; Ciavatta, Rogers, & Magnuson, 2007). Here, CTCF acts as enhancer-blocking, preventing the inappropriate action of an enhancer on a neighboring locus (A. C. Bell et al., 1999) or buffering the heterochromatin spreading (barrier) (J. Wang, Lawry, Cohen, & Jia, 2014a). CTCF has also been shown as an activator by establishing long-distance contacts to create an active "chromatin hub" in the β-globin locus during the activation of specific globin genes (Tolhuis, Palstra, Splinter, Grosveld, & de Laat, 2002). These many variety of CTCF functions are explaned by its interplay with interacting complexes in a lineage-specific manner (Cuddapah et al., 2008).

The insulation activity, as an example, is influenced by the binding of cohesins (Wendt et al., 2008) and the DEAD-box RNA helicase p68 with its associated ncRNA (Yao et al., 2010).

Interestingly, it has been shown that CTCF has a critical role in establishing the Topologically Associated Domains (TADs) as it is highly enriched at TAD boundaries (Dixon et al., 2012). In addition CTCF-mediated loops were strongly enriched at the border of Lamina-Associated Domains (LADs) (Guelen et al., 2008).

LADs are large domains of chromatin associated with the nuclear lamina (NL) that are required for the spatial organization of the genome in the nuclei. LADs have well defined borders which are enriched for active promoters, CpG islands and CTCF in mammals (Guelen et al., 2008).

The strict correlation that we observe with CTCF binding sites and the enrichment of HMGN5 in regulatory regions like promoters and CpG islands highly suggests that

HMGN5 and CTCF could have similar and/or cooperative functions in the interaction with the LADs.

One of our more striking results in this regard, is that we demonstrate a direct interaction of HMGN5 with CTCF *in vivo*, as identified by quantitative mass spectrometry after Co-IP (see section 4.14). Since HMGN5 is tightly bound to chromatin, we completely hydrolyzed DNA by a combination of MNase and Benzonase to only isolate interacting partners that made physical contact with HMGN5. The result then indicates that both proteins could have a direct protein-protein interaction or that they are part of the same protein complex.

Interestingly, both proteins have been shown to physically interact with components of the nuclear lamina, which let us speculate about a role of an HMGN5-CTCF complex in the regulation of the lamina-chromatin interactions.

Besides the high enrichment observed at the LADs borders, CTCF has been functionally linked to nuclear lamina in two recent reports. Honglei Zhao and colleagues showed that a complex between CTCF and PARP1 (Poly [ADP-ribose] polymerase 1) promotes the recruitment of circadian genes (as IGF2/H19, TARDBP and PARD3) to the lamina resulting in the attenuation circadian genes (Zhao et al., 2015). Another study revealed that the CTCF-nuclear lamina complex plays an active role in the transcriptional control of Estrogen Receptor (ER) target genes in breast cancer cells (Fiorito et al., 2016). It was shown that in this complex CTCF directly associates the lamina components Lamin B and/or Lap2β (Lamina-associated polypeptide 2, isoforms beta/gamma).

It was also shown by the group of Michael Bustin that HMGN5 directly interacts with the lamina-associated polypeptide 2-alpha (Lap2α) and that they reciprocally influence their genomic binding sites (Zhang et al., 2013).

LAP2α is a LAP isoform not anchored to the membrane. This protein is required to maintain the nucleoplasmic lamin A/C in the nuclear interior and has been shown to be bound to hetero- and euchromatin (Gesson et al., 2016).

Additionally, it was demonstrated that the HMGN5 dependent global chromatin decompaction decreases the sturdiness, elasticity and rigidity of the nucleus. Also mice overexpressing HMGN5 exhibit a phenotype characterized by deformed nuclei and disrupted lamina, which is associated with premature death because of cardiac malfunction (Furusawa et al., 2015). This study demonstrated a direct role of HMGN5

in regulating the nuclear architecture by linking the higher order organization of chromatin to the nuclear lamina.

Strikingly, amongst the HMGN5-interacting proteins identified in our study, the Barrier to autointegration factor 1 (BANF1 and alternatively named BAF), key component of the lamina, was the most abundant protein. BANF1 is involved in different pathways associated with genome maintenance. Loss of BANF1 in *Drosophila* and *C. elegans* is associated with altered nuclear envelope morphology and aberrant mitotic phenotypes like anaphase chromosome bridges (Jamin & Wiebe, 2015)

BANF1 also directly interacts with Lap2α, LAP2β, histones and lamina proteins (Jamin & Wiebe, 2015) being an essential factor for nuclear architecture and chromatin organization.

Those results support the hypothesis of HMGN5 playing a role in linking the nuclear lamina to chromatin organization as previously proposed (Naetar et al., 2017; Zhang et al., 2013) and it is tantalizing to speculate that the interaction of HMGN5 with CTCF may play a role in this higher order chromatin-lamina interaction. Such a function, though, needs to be experimentally validated.

As it was discussed in section 5.3, HMGN5 could play a dual regulatory role by modifying the global chromatin architecture and directly affecting the RNA metabolism. Due to its intrinsically disordered nature, we hypothesized that it may form distinct interaction complexes that, as it has been described for CTCF, would determine its specific function.

Specific targeting of HMGN5 to the genome could be determine by the binding to regulatory RNAs, as shown for the *Drosophila* Df31 (Schubert et al., 2012) or CTCF protein (Kung et al., 2015).

By using the MS2 tethering assay *in vivo* (Shevtsov & Dundr, 2011), we tried to analyze the role of candidate RNAs (from the HMGN5-RNA CLIP) in tethering and chromatin reorganization mediated by HMGN5. In this assay we tagged candidates RNAs with an MS2 stem loop and tried to tether them to the MS2 coat protein (MCP) that was bound to a LacO array in cells. Unfortunatelly, but we could not immobilize the MCP protein to the LacO (data not shown). This possibility needs to be validated with the use of a different method.

Furthermore, we observed that HMGN5 interacts with seven proteins that are associated with pre-rRNA processing and synthesis of the 60S ribosomal subunit (Castle et al., 2012; Finkbeiner et al., 2011). The identified proteins include a

complete complex associated with the (de)sumoylation process in transcriptional regulation (section 4.14) -composed by the proteins LAS1L, PELP1, TEX10, WDR18, SENP3, and NOL9- suggests a direct role of HMGN5 in the regulation of ribosomal biogenesis.

The specific interaction networks may finally determine the specific regulatory functions of HMGN5.

The results presented here highlight the highly versatile and critical role of HMGN5 in the higher-order structure of chromatin and regulation of gene expression.

6 Conclusion and Perspectives

Here we revealed that HMGN5 possesses a novel RNA binding activity. Furthermore, we showed that the RNA binding ability is also characteristic of other HMGN proteins, indicating a novel function of the HMGN family.

The missregulation of HMGN5 in cells by overexpression and knockdown affected the transcription of a large set of genes (about 3000 respectively), one third of them being overlapped in both conditions.

By ChIP-seq we showed that HMGN5 was preferentially associated with active regulatory regions like promoters and CpG islands. HMGN5 also correlates with RNA polymerase II target sites, and localizes at DNase I HS sites.

The actively regulated target genes possess functions related with RNA metabolism.

By using CLIP-seq we revealed that HMGN5 binds nascent RNA *in vivo*. Moreover, the *in vitro* competitive assays revealed that HMGN5 form distinct complexes with nucleosome and with RNA. We propose that HMGN5 plays a role in regulating RNA metabolism by a dual binding mechanism: 1) by opening chromatin structure at regulatory genomic regions and 2) and by directly contacting nascent RNA.

We revealed that HMGN5 co-localizes and directly interacts with CTCF *in vivo,* suggesting a cooperative role of both proteins in the organization of higher order structure of chromatin.

Moreover, the interaction of HMGN5 with several proteins associated with the pre-rRNA processing -revealed by quantitative mass spectrometry- suggests a role of HMGN5 in the regulation of rRNA metabolism.

Furthermore, amongst the identified interacting proteins we also found enriched the protein BANF1, component of the nuclear lamina.

Since both, CTCF and HMGN5 interact with factors associated with nuclear lamina, we hypothesized that a CTCF-HMGN5 complex could participate in the lamina-chromatin interplay.

To characterize the potential HMGN5-CTCF complex, different *in vivo* and *in vitro* studies need to be performed.

By using pull-down assays and *in vitro* binding measurements (MST or EMSA), the protein domains involved in the interaction can be determined.

The reciprocal influence on the binding targets can be followed by analyzing their mutual genome-wide distribution after overexpression or knockdown of one of both proteins.

By using the LacI/LacO tethering system, the influence of CTCF on the recruitment of HMGN5 to the chromatin can be studied.

It is important to characterize the role of HMGN5 in the lamina-chromatin interactions. In this regard, the association of HMGN5 with BANF1 and LAP2α could be analyzed using standard *in vitro* and *in vivo* methods, to test how they influence the chromatin and RNA binding ability of HMGN5. As an example, the effect of the lamina proteins on the decompaction mediated by HMGN5 can be tested by MNase accessibility assays, the LacI/LacO tethering system and by ChIP-seq analysis.

As we showed that HMGN5 preferentially binds the canonical CTCF binding motif *in vivo*, an *in vitro* characterization by EMSA and MST could help to understand if the sequence motif influences the genomic binding preferences of HMGN5, or if the bnding is determined by the interacting network.

7 Materials and methods

7.1 Materials

Unless otherwise stated, all common chemicals and kits were purchased from GE Healthcare, Fermentas, Invitrogen, Merck, New England Biolab, Promega, Roche, Roth, Serva, Bio-Rad, Stratagene/Agilent, Sigma-Aldrich, BD Biosciences, Macherey-Nagel, Nugen, Pierce and Qiagen. Cell culture reagents were purchased from Gibco, Invitrogen, Clontech and Promega.

7.1.1 Equipment and consumables

7.1.1.1 Consumables

Consumable	Supplier
1.5 ml and 2 ml micro centrifuge tubes	Eppendorf/Sarstedt
15 and 50mL centrifuge tubes	Sarstedt
Centrifugal spin filter column 0.5ml	Merck Millipore
Barrier food wrap	Saran
Cell culture 6 and 24 well-plate	Sarstedt
Cell scraper	Sarstedt
Cover slips, 22x22 mm	Roth
Cryovials	Roth/Nunc
Dialysis membrane	Spectrum Laboratories
Filter paper Whatman 3MM	Whatman
Filter tips	Sarstedt
Gel cassettes (disposable)	Invitrogen
Glassware	Schott
Immobilon-P Membrane	Millipore
Laboratory film	Parafilm®

Consumable	Supplier
Latex gloves	Roth
Micro cuvettes	Sarstedt
Microscope glass slides	Roth
Monolith NT115 standard treated capillaries	NanoTemper Technologies
Pasteur glass pipettes	VWR
PCR-reaction tubes 0.2 ml	Kisker Biotech
Petri dishes and tissue culture plates	Sarstedt
Pipette tips	Sarstedt
qPCR reaction strip tubes	Kisker
Serological pipettes	Sarstedt
Syringes	Braun
Vivaspin 500 Protein Concentrators (Vivaspin)	Sigma

7.1.1.2 Equipment

Equipment	Supplier
- 20 °C freezer	Siemens/Ewald
- 80 °C freezer	Sanyo
2100 Bioanalyzer	Agilent
Agarose gel chambers	University of Regensburg
Agarose gel UV imaging system	GelMax, Intas
Autoclave	Zirbus
Bioruptor Standard Sonicator	Diagenode
Centrifuge 3-16K	Sigma
Centrifuge Avanti J-26 XP	Beckman Coulter
Centrifuge Avanti J-26-S XP	Beckman Coulter
Centrifuge Centrikon T-324	Kontron Instruments
CO_2 incubator (mammalian cells)	SANYO/Binder
Fluorescence Microscope, Axiovert 200M	Carl Zeiss

Materials and methods

Equipment	Supplier
Heatblock neoBlock-Heizer Duo	neoLab
Ice machine	Ziegra
Image Reader LAS-3000	Fuji
Incubator Shaker	Infors
Laminar flow hood	Antair BSK
Leica TCS SP8 confocal microscope	Leica Microsystems
Magnetic plate SPRIplate® 96-Ring	Agencourt Bioscience
Magnetic stirrer MR 3001	Heidolph
Microbiological incubator	MEMMERT
Micropipettes	Gilson
Microwave	Sharp
Monolith NT.115	Nanotemper technologies
Nanodrop® ND-1000 Spectrophometer	peQLab Biotechnologie GmbH
Neubauer counting chamber	Marienfeld
Overhead rotator	Scientific Industries/Labinco
Peristaltic Pump PUMPDRIVE 5001	Heidolph
Peristaltic Pump LKB-P1	GE Healthcare
pH meter	Knick
Pipette controller	Integra/eppendorf
Power Supply EPS301	Ashram Biosciences
PURELAB Ultra	ELGA LabWater
Qubit 2.0 fluorometer	Invitrogen
Reax top vortexer	Heidolph
Real Time Thermocycler Rotor Gene Q	Qiagen
Safe Imager	Invitrogen
Scale	Sartorius
Sonifier 250	Branson
Stratalinker UV crosslinker	Stratagene
Table top centrifuge	Eppendorf
Thermocycler	Peqlab/Applied Biosystems
Thermomixer compact	Eppendorf

Equipment	Supplier
Trans-Blot® SD Semi-Dry Transfer Cell	Bio-Rad
Typhoon FLA 9500	GE Healthcare
Uvikon Spectrophotometer 922	Kontron Instruments
Vacuum Pump FB70155	Fischerbrand
XCell SureLock™ Mini-Cell system	Invitrogen

7.1.1.3 Kits

Kit	Supplier
Gateway BP Clonase II Mix	Invitrogen
Gateway LR Clonase II Mix	Invitrogen
High Sensitivity DNA Kit	Agilent
iScript™ cDNA Synthesis Kit	BioRad
NEBNext ChIP-Seq Library Prep Master Mix Kit	New England Biolabs
NucleoSpin RNA Kit	Macherey-Nagel
Ovation® Universal RNA-Seq System 1-16	Nugen
PureLink® HiPure Plasmid Maxiprep Kit	Invitrogen
QIAquick PCR purification Kit	Qiagen
QIAEX® II Gel Extraction Kit	Qiagen
QIAprep Spin Miniprep Kit	Qiagen
Qiagen Plasmid Purification Midi Kit	Qiagen
Qubit® dsDNA BR Assay Kit	Invitrogen
Qubit® Protein Assay Kit	Invitrogen
Qubit® RNA BR Assay Kit	Invitrogen
RNA 6000 Pico Kit	Agilent
SuperSignal WEST Dura WB Kit	Pierce
SuperScript II Reverse Transcriptase (RT)	Invitrogen
SYBR® Green qPCR Kit	Qiagen

7.1.2 Reagents

7.1.2.1 Common reagents

Product	Supplier
Agencourt AMPure XP beads	Agencourt Bioscience
Acetic acid	Sigma/VVR
Agarose, UltraPure™	Invitrogen
Ammonium acetate	Merck
Ammonium persulfate (APS)	Roth
Ampicillin	Roth
Aprotinin	Roth
Bacto Agar	BD Biosciences
Bacto Tryptone	BD Biosciences
Bacto Yeast Extract	BD Biosciences
Bovine serum albumin (BSA) (albumin fraction V)	Serva electrophoresis
Bovine serum albumin (BSA), 10 mg/ml	Sigma-Aldrich
Bradford protein assay	Bio-Rad
Bromophenol Blue	Roth
Calcium chloride ($CaCl_2$)	Roth
Chloroform	Merck
Coomassie Brilliant Blue G250	Thermo scientific
D-Sucrose	Roth
DAPI	Sigma-Aldrich
DMSO	Merck
DTT	Sigma-Aldrich
dNTP mix	Qiagen
EDTA	Merck
EGTA	Sigma-Aldrich
Elution buffer (EB)	Qiagen
Ethanol p. A.	Merck

Materials and methods

Product	Supplier
Ethidium bromide	Roth
Formaldehyde	Merck/Sigma
Formamide	Roth
GFP-Trap agarose beads	Chromotek
GFP-Trap magnetic particles	Chromotek
Glutathione	Thermo scientific
Glutathione-agarose beads	Jena Bioscience
Glycerol	Merck/Roth
Glycine	Serva electrohporesis
Glycogen	Sigma-Aldrich
Hydrochloric acid	Merck
IGEPAL CA-630	Sigma-Aldrich
Imidazol	Roth
IPTG	Roth
Isopropanol	VVR chamicals
Kanamycin	Roth
Leupeptin	Roth
Lithium chloride (LiCl)	Roth
Magnesium chloride (MgCl$_2$)	Merck
Methanol	Merck
Milk powder	Sucofin
Monopotassium phosphate (KH$_2$PO$_4$)	Merck
Ni-NTA agarose	Qiagen
Nuclease-free water	Ambion
Orange G	Sigma-Aldrich
Paraformaldehyde (PFA)	EMS
Pepstatin	Roth
PMSF	Roth
Potassium chloride (KCl)	Merck
Roti®-Phenol/Chloroform/Isoamyl alcohol	Roth
Rotiphorese®Gel30 (37.5:1) (acrylamide/bis-acrylamide)	Roth
Sodium acetate	Merck

Product	Supplier
Sodium chloride (NaCl)	Sigma/VVR
Sodium dodecyl sulfate (SDS)	Roth
Sodium hydrogen carbonate (NaHCO$_3$)	Merck
Sodium hydroxide (NaOH)	Merck
SYBR® Green	Invitrogen
SYBR® Safe	Invitrogen
TEMED	Roth
TCEP	Thermo Scientific
TRIS	Sigma-Aldrich
Triton X-100	Sigma-Aldrich
Trizol	Invitrogen
Tween-20	Roth
Vectashield mounting medium	Vector Laboratories Inc.
β-Mercaptoethanol	Sigma-Aldrich

7.1.2.2 Ladders

Ladder	Supplier
GeneRuler™1kb Plus DNA Ladder	Fermentas
PageRuler Prestained Protein Ladder Plus	Fermentas
RiboRuler Low Range RNA Ladder	Thermo scientific

7.1.2.3 Cell culture reagents and buffers

Reagent	Supplier
(DMEM), low glucose, GlutaMAX™	Gibco
0.05% Trypsin - EDTA (1X)	Gibco
Blasticidin	Gibco
Doxycycline	Clontech
Fetal Bovine serum	Gibco
Fetal Bovine serum, Tetracycline-Free	Biochrom
FuGENE® HD Transfection Reagent	Promega
Hygromycin	Invitrogen
Lipofectamine® 3000	Invitrogen
Lipofectamine® RNAiMAX	Invitrogen
Opti-MEM® I Reduced Serum Medium	Gibco
Trypan Blue	Gibco
Zeocin	Invitrogen

7.1.2.4 Buffers and medium

Common buffers were prepared according to standard protocols (Sambrook and Russel 2001; LabFAQS, Roche) and additional buffers for specific methods are described in each individual method section. If not stated differently, all buffers are stored at room temperature.

Buffer	Composition
ChIP High Salt Buffer	20 mM Tris-HCl pH 8.0 500 mM NaCl 2 mM EDTA 1% Triton X-100
ChIP LiCl Buffer	10 mM Tris-HCl pH 8.0 250 mM LiCl 1 mM EDTA 1 % IGEPAL CA-630 1 % (w/v) sodium deoxycholate

Buffer	Composition
ChIP Low Salt Buffer	20 mM Tris-HCl pH 8.0 150 mM NaCl 2 mM EDTA 1% Triton X-100
ChIP SDS Lysis Buffer	50 mM Tris-HCl, pH 8.0 10 mM EDTA 1% (w/v) SDS 1X protease inhibitor mix (PI)
Chromatin Assembly High Salt Buffer	10 mM Tris-HCl pH 7.6 2M NaCl 1 mM EDTA 0.05% IGEPAL CA-630 2 mM β-mercaptoethanol
Chromatin Assembly Low Salt Buffer	10 mM Tris-HCl pH 7.6 50 mM NaCl 1 mM EDTA 0.05% IGEPAL CA-630 2 mM β-mercaptoethanol
Chromatin Assembly Low Salt Buffer	10 mM Tris-HCl pH 7.6 50 mM NaCl 1 mM EDTA 0.05% IGEPAL CA-630 2 mM β-mercaptoethanol
CLIP Elution Buffer	10 mM Tris-HCl, pH 7.5 (prepared in nuclease-free water)
CLIP High Salt Buffer	50 mM Tris-HCl pH 7.5 500 mM NaCl 1 mM $MgCl_2$ 0.1% SDS 0.05% IGEPAL CA-630 100units/mL RNAsin 1X protease inhibitor mix (PI)
CLIP Low Salt Buffer	20 mM Tris-HCl pH7.5 250 mM NaCl 1 mM $MgCl_2$ 0.025% SDS 0.05% IGEPAL CA-630
CLIP Lysis Buffer	20 mM Tris-HCl pH7.6 100 mM NaCl 0.1 mM EDTA 0.5% IGEPAL CA-630 1X protease inhibitor mix (PI) 100U/mL Rnasin

Materials and methods

Buffer	Composition
CLIP Proteinase K Buffer	10 mM Tris-HCl pH 7.6 10 mM NaCl 5 mM CaCl$_2$
Coomassie staining solution	0.1% (w/v) Coomassie Brilliant Blue 45% methanol 10% acetic acid
Co-IP Lysis Buffer	20 mM Tris-HCl pH 7.5 150 mM NaCl 0.5% IGEPAL CA-630 3 mM CaCl$_2$ 1X protease inhibitor mix (PI)
Co-IP Wash Buffer 1	20 mM Tris-HCl pH 7.5 150 mM NaCl 0.5% IGEPAL CA-630 0.5 mM EDTA 1X protease inhibitor mix (PI)
Co-IP Wash Buffer 2	50 mM Tris-HCl pH 8
EX-X buffers	20 mM Tris-HCl pH 7.6 1.5 mM MgCl$_2$ 0.5 mM EGTA 10% glycerol X mM KCl
Immunofluorescence Fixation solution	1X PBS 4% (w/v) Paraformaldehyde
GST Lysis and Wash buffer	50 mM Tris-HCl pH 8 300 mM NaCl 1 mM TCEP 0.1% Triton X-100 1X protease inhibitor mix (PI)
GST Elution Buffer	20 mM Tris-HCl pH 8 10 mM Glutathione pH adjusted to 8
His Elution Buffer	EX100 250 mM Imidazol 1X protease inhibitor mix (PI)
His Lysis Buffer	50 mM Tris pH7.5 300 mM NaCl 10 mM imidazol 1X protease inhibitor mix (PI)
His Wash Buffer	50 mM Tris pH 7.5 500 mM NaCl 20 mM Imidazol 1X protease inhibitor mix (PI)

Buffer	Composition
Immunofluorescence Blocking Solution	1X PBS 0.1% Tween 20 4% BSA
Immunofluorescence Permeabilization Buffer 1	1X PBS 0.5% Triton X-100
Immunofluorescence Permeabilization Buffer 2	1X PBS 0.01% Triton X-100
Immunofluorescence Wash buffer	1X PBS 0.1% Tween 20
Lämmli buffer (6X) (protein loading buffer)	350 mM Tris-HCl pH 6.8 10% (w/v) SDS 30% (v/v) Glycerin 5% (v/v) β-Mercaptoethanol 0.2% (w/v) Bromophenol blue
Luria-Bertani (LB) medium	1.0% (w/v) Bacto‒Tryptone 1.0% (w/v) NaCl 0.5% (w/v) Bacto‒Yeast extract
MST/EMSA binding buffer	EX100 buffer 0.05% IGEPAL CA-630 (for MST)
Orange G loading dye (10X)	50% Glycerin 10 mM EDTA 0.03% (w/v) Orange G
PBS-T buffer	1X PBS 0.1% Tween 20
Permeabilization Buffer (genomic DNA extraction)	15 mM Tris pH 7.5 300 mM Sucrose 60 mM KCl 15 mM NaCl 3 mM $CaCl_2$ 1.5 mM $MgCl_2$ 0.5 mM EGTA 0.2% (v/v) IGEPAL CA-630
PBS (Phosphate-Buffered Saline)	137 mM NaCl 2.7 mM KCl 10 mM Na_2HPO_4 1.8 mM KH_2PO_4 pH adjusted to 7.4

Buffer	Composition
1X Protease inhibitor mix (PI)	1µg Aprotinin 1µg/ml Leupeptin 1µg/ml Pepstatin 1 mM PMSF (All final concetrations)
SDS-PAGE Lower Buffer (4X)	1.5 M Tris-HCl pH 8.8 0.4% SDS
SDS-PAGE Running buffer (1X)	25 mM Tris-HCl 192 mM Glycine 0.1% SDS
TBE (1X)	90 mM Tris-HCl pH 8.0 90 mM boric acid 2 mM EDTA
Transfer buffer (Towbin)	25 mM Tris-HCl 192 mM Glycin 20% Methanol 0.05% SDS
Western blot Blocking solution	1x PBS-T 5% dried milk
Whole Cell Lysis Buffer	20 mM Tris-HCl pH 7.5 100 mM NaCl 0.1 mM EDTA 0.5% IGEPAL CA-630

7.1.2.5Enzymes

Enzyme	Supplier
Antarctic phosphatase	New England Biolabs
Benzonase	Sigma Aldrich
MNase	Sigma-Aldrich
MNase (lab made)	Längst's group
Phusion® DNA polymerase	New England Biolabs
Proteinase K (10mg/ml)	Sigma-Aldrich
Q5 DNA Polymerase	New England Biolabs
Restriction endonucleases	New England Biolabs
RNase A	Invitrogen
RNase A/T1	Thermo Scientific
RQ1 DNase	Promega
T4 DNA ligase	New England Biolabs
T4 DNA polymerase	New England Biolabs
Taq DNA polymerase (Lab made)	Längst's group
TURBO DNA-free™ DNase	Ambion

7.1.2.6 Antibodies

Antibody	Reactivity	Biological source	Supplier	Applications
Anti HMGN5 (HPA000511)	Human HMGN5	Rabbit polyclonal	(Sigma)	WB 1:2500 IF 1:500
Anti actin (A2066)	Wide range including human	Rabbit polyclonal	Sigma	WB 1:5000
Anti rabbit HRP conjugated	Ab raised in rabbit	Goat		WB 1:5000
anti-Rb-DL488	Ab raised in Rabitt	Donkey	Jackson	IF 1:200

7.1.3 Cell lines

Cell line	Description	Supplier	Growth conditions
HEK293T	Human embryonic kidney	ATCC, cells were received second hand from Dr. A.Németh	37 °C, 5% CO_2 Low glucose DMEM−GlutaMAX, 10% FBS
U2OS /LacO	Homo sapiens bone osteosarcoma containing a stable integration of an array of the operon Lac (LacO) inserted in a telomeric region	Karsten Rippe's laboratory	37 °C, 5% CO_2 Low glucose DMEM−GlutaMAX, 10% FBS
T-REx™-293 Flp-In	Based in HEK-293T for inducible protein expression	Life Technologies/ Prof. Dr. Wagner	37 °C, 5% CO_2, Low glucose DMEM−GlutaMAX, 10% Tetracycline-Free FBS, 100 µg/ml Hygromycin, 10 µg/ml Blasticidin
GFP_FlpIn	GFP gene stably integrated in T-REx 293 cells	Received from Gunter Meister's group	37 °C, 5% CO_2, Low glucose DMEM−GlutaMAX, 10% Tetracycline-Free FBS, 100 µg/ml Hygromycin, 10 µg/ml Blasticidin
HMGN5_FlpIn	HMGN5 gene stably integrated in T-REx 293 cells	Ingrid Araya	37 °C, 5% CO_2, Low glucose DMEM−GlutaMAX, 10% Tetracycline-Free FBS, 100 µg/ml Hygromycin, 10 µg/ml Blasticidin

7.1.3.1 Bacterial strains

Strain	Application	Remarks
DH5α	General cloning and storage of common plasmids, blue/white screening.	RecA1 and endA1 mutations increase insert stability and plasmid quality in miniprep DNA
XL1blue	General cloning and storage of common plasmids, blue/white screening.	RecA1 and endA1 mutations increase insert stability and plasmid quality in miniprep DNA
BL21 (DE3) pLysS	Protein expression	T7 RNA Polymerase under the Control of the lac UV5 Promoter for IPTG inducible protein expression.
Rosetta 2 (DE) pLysS	Protein expression	Expression from T7 promoter with codon bias correction

7.1.4 Plasmids

7.1.4.1 LacI/LacO tethering plasmids

Vector	Remarks	TAG	Bacterial resistance	Cloning strategy	Obtained from
pSV2_GFP_LacI	Mammalian expression vector. Transient transfection for LacI/LacO system	GFP-LacI	Kanamycin	Not available	Dr. Karsten Rippe's Laboratory
gBlock_pSV2_GFP_LacI	Mammalian expression vector. Transient transfection for LacI/LacO system	GFP-LacI	Kanamycin	Modification of psV2_GFP_LacI to add EcoRI and XhoI restriction sites downstream of LacI sequence	Ingrid Araya
pSV2_GFP_LacI_HMGN5	Mammalian expression vector. Transient transfection of HMGN5 for LacI/LacO system	GFP-LacI	Kanamycin	HMGN5 CDS inserted in gblock_pSV2_GFP_LacI with EcoRI and XhoI restriction sites	Ingrid Araya

Vector	Remarks	TAG	Bacterial resistance	Cloning strategy	Obtained from
pSV2_GFP _LacI_ PTMS	Mammalian expression vector. Transient transfection of PTMS for LacI/LacO system	GFP-LacI	Kanamycin	PTMS CDS was inserted in gblock_pSV2_GF P_LacI with EcoRI and XhoI restriction sites	Ingrid Araya

7.1.4.2 Plasmids for stable cell line Flp-In

Vector	Remarks	TAG	Bacterial resistance	Cloning strategy	Obtained from
pCDNA5 /FRT/TO _GFP	Mammalian Inducible expression vector designed for use with the Flp-In™ T-REx™ System	GFP	Amp	Not available	Dr. Helen Hoffmeister
pCDNA5 /FRT/TO _HMGN 5-GFP	Mammalian Inducible expression vector designed for use with the Flp-In™ T-REx™ System	GFP	Amp	HMGN5 CDS inserted in pCDNA5/FRT/ TO_GFP with HindIII and Xma restriction sites	Ingrid Araya
pOG44	Flp recombinase expression plasmid, For efficient integration into Flp-In™ T-REx™ Cell Line		Amp	Not available	Life Technologies/ Prof. Dr. Wagner

7.1.4.3 Cloning plasmids

Vector	Remarks	TAG	Bacterial resistance	Cloning strategy	Obtained from
pGEX-4T3	Bacterial expression vector	N-GST /thrombin	Amp		Vendor
pGEX-TEV	Bacterial expression vector	N-GST/ PreScission /TEV	Amp	Created from pGEX4T3 replacing the Thrombin site for the PreScission and TEV recognition site	Ingrid Araya
pGEX-TEV_HMGN5_WT	Bacterial expression vector expressing HMGN5 WT	N-GST/ PreScission /TEV	Amp	HMGN5 cloned in pGEX-4T3_TEV with KpnI and XhoI	Ingrid Araya
pGEX-TEV_HMGN5_dn19	Bacterial expression vector expressing HMGN5 ΔN19	N-GST/ PreScission /TEV	Amp	HMGN5 ΔN19 cloned in pGEX-TEV with KpnI and XhoI	Ingrid Araya
pGEX-TEV_HMGN5_dn20	Bacterial expression vector expressing HMGN5 ΔN20	N-GST/ PreScission /TEV	Amp	HMGN5 ΔN20 cloned in pGEX-TEV with KpnI and XhoI	Ingrid Araya
pGEX-TEV_HMGN5_dn24	Bacterial expression vector expressing HMGN5 ΔN24	N-GST/ PreScission /TEV	Amp	HMGN5 ΔN24 cloned in pGEX-TEV with KpnI and XhoI	Ingrid Araya
pGEX4T3_HMGN5_dC27	Bacterial expression vector expressing HMGN5 ΔC27	N-GST/ PreScission /TEV	Amp	HMGN5 ΔC27 cloned in pGEX-4T3 with BamHI and SalI	Ingrid Araya

Materials and methods

Vector	Remarks	TAG	Bacterial resistance	Cloning strategy	Obtained from
pGEX-TEV_NBD	Bacterial expression vector expressing HMGN5 NBD	N-GST/ PreScission /TEV	Amp	HMGN5 NBD cloned in pGEX-4T3 with BamHI and SalI	Ingrid Araya
pGEX-TEV_HMGN5 _S20,24E	Bacterial expression vector expressing HMGN5 S20,24E	N-GST/ PreScission /TEV	Amp	Created with Site directed mutagenesis from pGEX-TEV_HMGN5_ WT	Ingrid Araya
pGEX-TEV_HMGN5 _S20E	Bacterial expression vector expressing HMGN5 S20E	N-GST/ PreScission /TEV	Amp	Created with Site directed mutagenesis from pGEX-TEV_HMGN5_ WT	Ingrid Araya
pGEX-TEV_HMGN5 _S24E	Bacterial expression vector expressing HMGN5 S24E	N-GST/ PreScission /TEV	Amp	Created with Site directed mutagenesis from pGEX-TEV_HMGN5_ WT	Ingrid Araya
pGEX-TEV_ HMGN2	Bacterial expression vector expressing human HMGN2	N-GST/ PreScission /TEV	Amp	HMGN2 CDS cloned into pGEG-TEV with EcoRI and XhoI from a synthesized DNA fragment	Ingrid Araya

7.1.5 Oligonucleotides

Cloning primers were purchased from Sigma-Aldrich and MWG. Fluorescent primers were obtained from Sigma-Aldrich.

7.1.5.1 Cloning primers

Name	Gene/ binding site	Orien-tation	Sequence	Remarks	Description
IAP'#001	gBlock for pSV2_ GFP_ LacI	FW	CGCAATGCGCGCC ATTACCGAG		For modification of pSV2_GFP_LacI
IAP'#002	gBlock for pSV2_GFP_ LacI	RV	TAGATCCGGTGGA TCCCGGGTCAG		For modification of pSV2_GFP_LacI
IAP'#003	Parathymosin	FW	CGGAATTCATGTC GGAGAAAAGCGTG	EcoRI site	For cloning in pSV2_GFP_ LacI
IAP'#004	Parathymosin	RV	CTCTCGAGTCACG CCGATGCCCCA	XhoI site	For cloning in pSV2_GFP_LacI
IAP'#005	pSV2_GFP_ LacI	FW	GAGGGCATCGTTC CCACTGCG		Sequencing primer
IAP'#006	pSV2_GFP_ LacI	RV	TTCAGGGGGAGGT GTGGGAGG		Sequencing primer
IAP'#007	HMGN5	FW	GCTGCAAGCTTAC CATGCCCAAAAGA AAGGCTG	HindIII site	For cloning in pCDNA5/FRT/TO
IAP'#008	HMGN5	RV	CTGGTCCCGGGTA ACAATACTCTGTGG CTCCTC	XmaI site	For cloning in pCDNA5/FRT/TO
IAP'#013	HMGN5	FW	AGCAGTTGCTGAA ACCAAGC		qPCR primer and sequencing
IAP'#014	HMGN5	RV	AGAAGCAGTGGCA GCAGAAG		qPCR primer and sequencing
IAP'#015	HMGN5	FW	CGGAATTCATGCC CAAAAGAAAGGCT GC	EcoRI site	For cloning in pSV2GFPLacI
IAP'#016	HMGN5	RV	CTCTCGAGCTAAAC AATACTCTGTGGC	XhoI site	For cloning in pSV2GFPLacI
IAP'#033	HMGN5	FW	CATCGGATCCATG CCCAAAAGAAAGG CTGCAGG	BamHI site	For cloning in pGEX-4T3
IAP'#034	HMGN5	RV	ATCCGTCGACCATA GCAGACAACCTGG CAGATCTTC	SalI site	For cloning in pGEX-4T3
IAP'#035	HMGN5	RV	ATCTGCTCGAGCTA AACAATACTCTGTG GC	XhoI site	For cloning into pGEX-TEV

Name	Gene/ binding site	Orien-tation	Sequence	Remarks	Description
IAP'#036	HMGN5	FW	CATCGGTACCGCC AGGTTGTCTGCTAT GC	KpnI site	For cloning into pGEX-TEV
IAP'#037	HMGN5	FW	CATCGGTACCGCT ATGCTTGTGCCAGT TAC	KpnI site	For cloning into pGEX-TEV
IAP'#039	IAG_TEV_ PreScission	FW	TTCGAAGATCGTTT ATGTC		For introducing TEV_PreScission in pGEX-4T3
IAP'#040	IAG_TEV_ PreScission	RV	CGAGGCAGATCGT CAGTCAG		for introducing TEV_PreScission in pGEX-4T3
IAP'#041	HMGN5	FW	GTGAGGTACCATG AGGCAGGAGCCA AAG AG	KpnI site	For cloning into pGEX-TEV
IAP'#042	HMGN5	RV	ACGTCTCGAGCTA AGGCTTCACCTCT GGTGTAAC	XhoI site	For cloning into pGEX-TEV
IAP'#043	HMGN5	FW	ATCTGGTACCTCTG CCAGGTTGTCTGC TATG	KpnI site	For cloning into pGEX-TEV
IAP'#044	HMGN5	RV	ATCCGGTACCATG CCCAAAAGAAAGG CTGCAG	KpnI site	For cloning into pGEX-TEV
IAP'#080	HMGN5 (in pGEX-TEV)	FW	AGAGAAGAGAGGC CAGGTTGGAGGCT ATGCTTGTGCCAGT TACACCAGAGGTG AAGC		Site directed mutagenesis HMGN5 S20,24E
IAP'#081	HMGN5 (in pGEX-TEV)	RV	GCATAGCCTCCAA CCTGGCCTCTCTTC TCTTTGGCTCCTGC CTCATATCACCTTG		Site-directed mutagenesis HMGN5 S20,24E
IAP'#082	HMGN5 (in pGEX-TEV)	FW	AGAGAAGAGAGGC CAGGTTGTCTGCTA TGCTTGTG		Site-directed mutagenesis HMGN5 S20E
IAP'#083	HMGN5 (in pGEX-TEV)	RV	CAACCTGGCCTCT CTTCTCTTTGGCTC CTGCCTC		Site-directed mutagenesis HMGN5 S20E
IAP'#084	HMGN5 (in pGEX-TEV)	FW	CCAGGTTGGAGGC TATGCTTGTGCCAG TTACAC		Site-directed mutagenesis HMGN5 S24E
IAP'#085	HMGN5 (in pGEX-TEV)	RV	AGCATAGCCTCCA ACCTGGCAGATCTT CTCTTTG		Site-directed mutagenesis HMGN5 S24E
IAP'#118	HMGN2 in gBlock	FW	TATGAATTCCCCAA GAGAAAGGCTGAA GG	EcoRI	For HMGN2 cloning into pGEX_TEV
IAP'#119	HMGN2 in gBlock	RV	ACTCTCGAGTCACT TGGCATCTCCAGC AC	XhoI	For HMGN2 cloning into pGEX_TEV

Name	Gene/ binding site	Orien-tation	Sequence	Remarks	Description
LP09	T7 promotor/lac operator	FW	AATACGACTCACTA TAGGGGAATTGT		Sequencing primer
LP017	CMV promoter	FW	GCAAATGGGCGGT AGGC		Sequencing primer
LP019	M13 (-49)	RV	GAGCGGATAACAA TTTCACACAGG		Sequencing primer
LP103	GST gene pGEX vector	FW	ATAGCATGGCCTTT GCAG		For pGEX sequencing and PCR
LP104	pGEX rev standard	RV	GAGCTGCATGTGT CAGAGG		For pGEX sequencing and PCR
LP090	EGFP	RV	CGTCGCCGTCCAG CTCGACCAG		Sequencing primer
LP091	EGFP	RV	TCCAGCAGGACCA TGTGATC		Sequencing primer

7.1.5.2 qPCR primers

Name	Gene/ binding site	Orientation	Alternative name	Sequence
IAP'#049	VGF	FW	VGF.fw	TGATCAGCAGAAGGCAGAAG
IAP'#050	VGF	RV	VGF.rv	GACCCTCCTCTCCACCTCTC
IAP'#061	RPL30	FW	RPL30.fw	ATTCTCGCTAACAACTGCCCA
IAP'#062	RPL30	RV	RPL30.rv	ATTTTCCGCATGCTGTGCC
IAP'#073	ZCCHC12	FW	ZCCHC12.fw	GCAGAGGCTGGAAGTGAAAT
IAP'#074	ZCCHC12	RV	ZCCHC12.rv	AGCTTAATCTGCAGGTGGGA
IAP'#075	EGR1	FW	EGR1.fw	CACCTGACCGCAGAGTCTTTT
IAP'#076	EGR1	RV	EGR1.rv	CAGGGAAAAGCGGCCAGTAT

7.1.5.3 Primers for *in vitro* mononucleosomes assembly

Name	Binding site	Orientation	Sequence	Modification
RMP#012	AvaI digested pUC18 12X601	FW	CCCTGGAGAATCC CGG	
RMP#014	AvaI digested pUC18 12X601	RV	CACAGGATGTATA TATCTGACAC	5' Cy3

7.1.5.4 Fluorescent single-stranded nucleic acids

Nucleic acid	Sequence	Modification
snoRNA2T2	GAGUUUAUUACUAAUCUUUCGGGUAUGAAAUUC	Cy5
snoRNA2T1	GCUAGCGUGAUAAUGAUUUCAACUUCACUGCUGACCAG	FAM
En3_TFO_RNA	CCUCCUUUUUUCUUUUUUUUUUUUUUUUCU	FAM
En3_TFO_DNA	CCTCCTTTTTTCTTTTTTTTTTTTTTCT	Cy5
En3_DNA_rev	GGAGGAAAAAGAAAAAAAAAAAAAAGA	Cy3
En3_RNA_rev	GGAGGAAAAAGAAAAAAAAAAAAAAGA	FAM

7.1.6 Software and databases

Name	Description	Supplier	Reference
Adobe Illustrator	Vector graphics editor	Adobe Systems	
Axiovision	Microscope imaging software	Zeiss	
BEDtools (v2.24.0)	Fast and flexible suit of utilities for genomic analysis tasks	http://bedtools.readthedocs.io/en/latest/	(Pham et al., 2010)
BioVenn	Area-proportional Venn diagrams generation	http://www.biovenn.nl/	(Hulsen, de Vlieg, & Alkema, 2008)
BLAST	The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences	http://blast.ncbi.nlm.nih.gov/Blast.cgi	(ALTSCHUL, GISH, MILLER, MYERS, & LIPMAN, 1990)
Bowtie (v2.1.1)	Tool for aligning sequencing reads to long reference sequences	http://bowtie-bio.sourceforge.net/index.shtml	(Langmead, Trapnell, Pop, & Salzberg, 2009)
Bowtie2 (v2.2.4)	Tool for aligning sequencing reads to long reference sequences	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml	(Langmead & Salzberg, 2012)
ClueGo	Analyze interrelations of terms and functional groups in biological networks	http://apps.cytoscape.org/apps/cluego	(Bindea et al., 2009)

Materials and methods

Name	Description	Supplier	Reference
Cytoscape (v3.5.1)	Network Data Integration, Analysis, and Visualization	http://www.cytoscape.org/	(Shannon et al., 2003)
FastQC	Quality control tool for high throughput sequence data	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/	
FoldIndex©	Analysis of protein folding		(Prilusky et al., 2005)
Geneious 9.0.2	Organization and analysis of sequence data	Biomatters Ltd. http://www.geneious.com/	(Kearse et al., 2012)
HOMER (v4.9)	Motif discovery and next generation sequencing analysis	http://homer.ucsd.edu/homer/	(Heinz et al., 2010)
ImageJ	Image processing and analysis	https://imagej.nih.gov/ij/	(C. A. Schneider, Rasband, & Eliceiri, 2012)
Integrative genomics viewer (IGV)	Visualization tool for interactive exploration of large, integrated genomic datasets	http://software.broadinstitute.org/software/igv/	(J. T. Robinson et al., 2011)
Jaspar	Open-access database of curated, non-redundant transcription factor (TF) binding profiles	http://jaspar.genereg.net/	(Khan et al., 2017)
Leica LAS AF Lite	Visualization and analysis of Leica confocal images	https://www.leica-microsystems.com/	
OligoAnalyzer 3.1	Evaluation of physical properties of primers	IDT	
Pubmed	Database of biomedical literature	https://www.ncbi.nlm.nih.gov/pubmed/	PubMed [Internet]. Bethesda (MD): National Library of Medicine (US). [1946] - [cited 2017 12 25]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/
Quartzy	Online laboratory management platform	https://www.quartzy.com/	
R (v3.1.2)	Software environment for statistical computing and graphics	https://www.r-project.org/	R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/

Name	Description	Supplier	Reference
RNAfold	RNA secondary prediction tool	http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi	
Rotorgene 6000 Software		Qiagen (Corbett Research)	
SAMtools (v1.3.1)	Post-processing alignments in the SAM format, such as indexing, variant caller and alignment viewer,	http://samtools.sourceforge.net/	(H. Li et al., 2009)
Sigmaplot V12.5	Graphing and statistical analysis software	Systat https://systatsoftware.com/	
STAR	Ultrafast universal RNA-seq aligner	https://code.google.com/archive/p/rna-star/	(Dobin et al., 2012)
UCSC Genome Browser viewer	Visualization and analysis of genomic data	https://genome.ucsc.edu/	(Kent et al., 2002)
Uniprot	Protein sequence and functional information database	http://www.uniprot.org/	

7.1.7 High throughput datasets

Dataset	Cell type	ENCODE Accession	Producer Laboratory	Reference
H3K27ac ChIP-seq	HEK293	ENCFF631VZK	Peggy Farnham, USC	(C. A. Davis et al., 2018)
H3K4me3 ChIP-seq	HEK293	ENCFF001FJO	John Stamatoyannopoulos, UW	(C. A. Davis et al., 2018)
DNase-seq	HEK293	ENCFF000SPI	Gregory Crawford, Duke	(C. A. Davis et al., 2018)
POLR2A ChIP-seq	HEK293	ENCFF000WYF	Sherman Weissman, Yale	(C. A. Davis et al., 2018)
CTCF ChIP-seq	HEK293	ENCFF924LOC	John Stamatoyannopoulos, UW	(C. A. Davis et al., 2018)

7.2 Methods

7.2.1 Microbiological methods

7.2.1.1 Bacterial transformation

For transformation of chemically competent *Escherichia coli* (*E. coli*), 50µl of the bacterial strain were thawed on ice for 10 min. When transforming plasmids, 50ng were added to the bacteria, and for ligation products the complete ligation reaction was used. The samples were mixed by shortly flicking the tube. A heat-shock was performed at 42 °C for 45 sec in a thermomixer followed by an additional incubation on ice for 2 min. The transformed bacteria were resuspended in 1ml sterile pre-warmed medium (LB or SOC) and incubated for 30 min at 37 °C with shaking at 300 rpm in a thermomixer (for Ampicilin resistant plasmids the incubation was done in 10 min to avoid the degradation of Ampicillin by Beta-lactamases). 150µl of the mixture were plated on agar plates containing the appropriate antibiotics. Plates were incubated at 37 °C overnight (ON) and plates were inspected after 12 – 16 h.

7.2.1.2 Glycerol stock

To create bacterial glycerol stocks for long-term storage, 1ml of an ON culture was mixed with 420µl sterile 50% glycerol, snap frozen in liquid nitrogen and then frozen at -80 °C.

7.2.2 Nucleic acids methods

7.2.2.1 DNA quantification and quality measurement

The concentration and purity of DNA was generally measured by spectrophotometry using the NanoDrop ND1000 spectrophotometer (Peqlab). The concentration is given by the absorbance at 260 nm and the relative purity of the sample analyzed by the ratio of the absorbance at 260/280, being a ratio of ~1.8 accepted as "pure" for DNA. For more accurate measurements prior to high throughput sequencing of ChIP samples and for qPCR measurements, the DNA was quantified with the Qubit® dsDNA BR Assay Kit (Invitrogen) and the quality measured using the High Sensitivity DNA Kit (Agilent) in the 2100 Bioanalyzer device.

7.2.2.2 Agarose gel electrophoresis

Agarose gel electrophoresis was used to separate DNA or total RNA according to size and as quality control of DNA isolations or PCR reactions. The method was performed using gels with agarose concentration between 0.8 -1.5% depending on the predicted size of the nucleic acid fragments to separate, in 1X TBE buffer supplemented with 0.01% SYBR Safe (Invitrogen). Before loading, the samples were mixed with 10x Orange G loading dye. In the case of DNA, the approximate size of the samples was followed by comparison with the 1Kb DNA standard (Fermentas). The electrophoresis was done in 1X TBE buffer at constant voltage of 100-120V. The gels were documented using the Safe Imager blue light transilluminator system (Invitrogen) and analyzed with the ImageJ software.

7.2.2.3 Polymerase chain reaction

Depending on the downstream experiments, PCR reactions were performed using different polymerases, and programs.

The PCR performed for cloning (either restriction cloning, Gibson assembly or site-directed mutagenesis) were done using the Q5 high fidelity DNA polymerase (NEB) or the Phusion DNA polymerase (NEB). The annealing temperature, time and elongation time were adjusted for each reaction. A Positive and a water negative control were always included. A standard reaction is done in 50µl final volume, containing 0.2µM of each primer, 0.2mM dNTP mix, 1X PCR reaction buffer containing 2 mM $MgCl_2$ at final (1X) reaction concentrations, (provided commercially together with the specific DNA polymerases), from 0.4 to 1 U of DNA polymerase, and from 0.5 to 5ng DNA, according to the nature of the template.

7.2.2.4 Colony PCR

To test for the presence of insert in the plasmid constructs, colony PCR was performed. Individual colonies were selected and resuspended in 25µl of sterile H_2O in a 0.2 ml PCR reaction tube. Then 25µl PCR mastermix was added to each sample, and the PCR run was started. As no fidelity is required for the screening, a selfmade Taq polymerase was used. The primers were designed to match one in the insert and one in the vector. Afterwards, the PCR reactions were analyzed on an agarose gel for presence of the amplicon.

Basic thermocycler program for colony PCR

Cycle step	Temperature (°C)	Time	Number of cycles
Initial denaturation	95	5 min	1
Denaturation	95	30 sec	
Annealing	55-60 (depending on the template)	30 sec	35
Extension	72	1 min per Kb	
Final extension	72	5 min	1

7.2.2.5 Quantitative Real-Time PCR (qPCR)

Real‑Time quantitative PCR was used to accurately validate candidates from the RNA-seq dataset obtained in this study using a relative quantification. The primers for mRNA amplification were designed to flank exon-intron boundaries by using the software Primer-BLAST from the NCBI website. The list of primers used for qPCR for the candidate mRNAs ZCCHC12, VGF, RPL30 and EGR1 can be found in 7.1.5.2. An efficiency test for each set of primers was performed. GAPDH was used as reference gene for quantification. Quantification was carried out using the algorithm proposed by (Pfaffl, 2001). To define statistical differences between mRNA levels from knockdown, overexpression and control groups Two-way ANOVA and Dunett´s post-test were used (Dunnett, 1955).

The pipetting was done under a hood using dedicated pipettes for qPCR to avoid contamination. A master mix (MM) that contains the buffer, $MgCl_2$ and dNTP (purchased from Qiagen) was prepared, aliquoted and kept at -20 °C. The reaction is described in the following table:

Master Mix [MM]	µl/rxn
10X PCR-Buffer Qiagen	2
25 mM $MgCl_2$ Qiagen	0.8
25 mM dNTPs	0.16
H_2O	6.71
Total volume	**9.67**

The individual qPCR reactions were performed in technical triplicates in 0.1 mL PCR tubes, in a final volume of 20µl. A typical reaction is shown in the table below:

Component	Volume (µl)
MM	9.67
HotStar Taq Qiagen 5u/µl	0.08
[1:400000] SyberGreen Stock solution	0.25
Primer Fw [10µM]	0.40
Primer Rv [10µM]	0.40
H_2O	5.2
DNA	4
Total volume	20

qPCR was performed using the Rotor-Gene Q system from Qiagen and analyzed using the Rotorgene 6000 Software. After statistical analysis the qPCR validation were plotted using the Sigmaplot software (v12.5).

7.2.2.6 DNA precipitation

To precipitate DNA obtained from large-scale PCR or DNA purified from cells, the samples were mixed with 2 volumes ice-cold 100% ethanol (Merck) and 0.5 volumes 7.5 M ammonium acetate and they were incubated for 10-30 min at -20 °C. The samples were centrifuged at >13000*g* at 4 °C and the supernatant was carefully removed. The samples were washed twice with 700µl of 70% ethanol, followed by centrifugation steps of 10 min at >13000*g*. After the final wash, the ethanol was removed carefully and the pellet was air-dried for approximately 10 min. Finally, DNA was dissolved in the appropriate amount of Elution Buffer (Qiagen) and stored at -20 °C.

7.2.2.7 Gel extraction and PCR purification

When purifying PCR products for cloning, or when purifying large-scale PCR, the QIAquick PCR Purification Kit was used. From the large-scale PCR, DNA was first precipitated according to 7.2.2.6 and then the concentrated DNA was purified with the QIAquick PCR Purification Kit following the instructions of the manufacturer. To purify DNA fragments after enzymatic restriction digested for downstream applications (PCR and or cloning), the digestion was loaded on a 1-1.5% agarose gel, separated by electrophoresis (7.2.2.2) and the band extracted from the gel by cutting it with a clean scalpel using the blue-light screen and an orange filter. The DNA purification from the extracted band was carried out with the Qiaex II® Gel Extraction Kit according to the manufacturer's instructions.

7.2.2.8 Isolation of plasmid DNA

Depending on the amount of plasmid DNA required, a small-scale (minipep) or large-scale plasmid DNA purification (maxiprep) was performed. For that, a culture of bacteria, taken from a glycerol stock or fresh colony from plate, was performed ON in LB medium with the appropriate antibiotics. The minipreps were prepared from 5ml ON cultures using the QIAprep Spin Miniprep Kit from Qiagen, and the maxipreps were generated from 200ml ON culture using the PureLink® HiPure Plasmid Maxiprep Kit (Invitrogen) according to the instructions of the manufacturer. The isolated DNA was quantified as described in 7.2.2.1.

7.2.2.9 Isolation of genomic DNA from human cells

Starting from a 10 cm plate grown to 60-90% confluence, the medium was soaked from the plate, and the cells rinsed once with 1X PBS at RT. The cells were collected with a cell scraper in 1ml 1X PBS and transferred to a 1.5ml microcentrifuge tube. After a short spin down the PBS was removed and the cells lysed in 500µl Permeabilization Buffer. Then the cellular RNA was hydrolyzed by the addition of 100µg of RNase A. After 2 h of digestion at 37 °C, 250µg proteinase K were added

and the samples incubated at 55 °C ON. The next day the samples were precipitated with the addition of 2 volumes ice-cold 100% ethanol and 0.5 volumes 7.5 M ammonium acetate. The samples were incubated 10 min at -20 °C and then centrifuged for 30 min at 4 °C at 13000g. The ethanol was removed and the samples washed twice with 700µl ice-cold 70% ethanol, with centrifugation steps of 10 min at 4 °C. After carefull removal of all traces of ethanol by pipetting, the genomic DNA was air dried for 10 min, and then resuspended in 200µl Elution Buffer (Qiagen). Since the genomic DNA is viscous and difficult to get dissolved, the samples were incubated ON at 50 °C in a thermoblock to ensure that the DNA is completely dissolved in the buffer. The DNA concentration was determined and about 1µg DNA loaded on a 1.3% agarose gel for visualization. The DNA was stored at 4 °C.

7.2.2.10 Cloning

7.2.2.10.1 Restriction cloning

This method was used to insert the cDNA of a gene of interest (GOI) amplified by PCR into a vector, by using restriction enzyme digestion and ligation to the linearized vector with compatible ends. The *in silico* design of the cloning strategy was done with Geneious® 9.0.2 (www.geneious.com). The specific restriction endonucleases were obtained from New England Biolabs (NEB) and used according to the recommendation of the manufacturer. When possible, High Fidelity (HF) enzymes or enzymes with compatible buffer were used, to simplify the digestion protocol. For analytical digest e.g. identification of positive clones, 500ng DNA were incubated with 5 units of the respective restriction endonuclease in a total volume of 20µl, for 1 h at 37 °C (when both enzymes were HF enzymes, the incubation time was reduced to 20 min). The large-scale digestion for subsequent ligation reaction was done with 5µg of plasmid DNA and 2.5µg of insert. In both cases 20 units of each restriction endonuclease in a total volume of 30-40µl were used. To verify the completion of the reaction, DNA was electrophoretically separated using 0.8-1.5% TBE agarose gels supplemented with SYBR Safe (Invitrogen).

7.2.2.10.1.1 Ligation

Ligation of DNA fragments obtained from restriction digest with cohesive ends was performed with T4 DNA ligase (NEB). The molar ratio insert:vector was 3:1 using the following formula

$$ng\ vector \times desired\ molar\ ratio \times \frac{fragment\ lenght\ (bp)}{vector\ lenght\ (bp)} = ng\ fragment$$

The ligation was normally done in 20-30µl volume, using 1µl (400 U) of T4 DNA ligase, in 1X T4 DNA Ligase Reaction Buffer, with incubation at RT for 30 min. The reaction mixture was immediately used for bacterial transformation as described in 7.2.1.1.

7.2.2.10.2   Gibson assembly

When cloning large insert fragments, or in case no compatible restriction sites are present in the fragment or the vector, the Gibson assembly method was used(Gibson et al., 2009)

Primers with 30-40 bp overlap were designed (Geneious® 9.0.2), and the vector and DNA fragment (cDNA or genomic DNA) were amplified with Q5 High Fidelity DNA polymerase. The PCR products were analyzed on a 0.8% agarose gel, and if the desired amplification products were pure, the raw PCR mix was used for the assembly reaction, otherwise the fragments were gel purified and then assembled.

The fragments were assembled using the Gibson Assembly kit (NEB) following the instructions of the manufacturer. Briefly, a molar ratio of 3:1 insert/vector was used, with 50ng of vector, mixed with the Gibson Assembly Master Mix in a final volume of 20µl and incubated at 50 °C for 1 h. NEB 5-alpha competent *E. coli* cells were transformed with 2µl of the master mix/fragment mixture using the transformation protocol described in 7.2.1.1, using SOC medium. The next day, a few colonies were selected and grown ON in 5ml LB medium supplemented with antibiotics, and the plasmid was purified for sequencing according to the protocol described in 7.2.2.8.

7.2.2.10.3   Site-directed mutagenesis

To create the phosphomimetic mutants of HMGN5 (S20E, S24E and S20,24E), the site-directed mutagenesis strategy (SMD) was used. The starting DNA clone mutated was pGEX_TEV_HMGN5, containing a N-terminal GST tag. For the mutation of 2 amino acids at the same time (HMGN5 S20,24E) the primers were designed following the protocol described in (Zheng, 2004), using partially overlapping primers of about 50nt, containing the mutated sequences (designed with Geneious® 9.0.2). 50ng of the HMGN5 plasmid was fully amplified with the Q5 HF DNA polymerase (NEB), using 16 cycles of amplification and 1 min of extension/Kilobase at 72 °C, in a final volume of 50μl. 4μl of the PCR product were loaded on a 0.8% agarose gel to check the amplified plasmid. The product was purified with the QIAquick PCR Purification Kit and then subjected to DpnI treatment to hydrolyze the methylated parental plasmid. For this treatment 10 U of DpnI were used to treat 100-150ng of DNA, at 37 °C for 1 h. After incubation, the samples were heat-inactivated at 80 °C for 20 min. Then 1μl of the product was used to transform DH5α cells using the protocol described in 7.2.1.1. The selected clones were then sequenced and used to transform BL21 cells for protein expression and purification.

7.2.2.11   RNA purification

Total RNA was purified with Trizol to prepare cDNAs, or to isolate RNA from CLIP assays. The Trizol method is an improved adaptation of the "Acid guanidinium thiocyanate-phenol-chloroform extraction" (Chomczynski & Sacchi, 1987). In all cases, the RNA extraction was performed in a fume hood, always wearing a lab coat, gloves and safety glasses. A dedicated set of pipettes, and "Safe-lock" microcentrifuge tubes were used. In general, RNA purification was performed using 1 confluent 10 cm culture dish as starting material. RNA purification for high throughput sequencing was performed with the NucleoSpin RNA Kit (MACHEREY-NAGEL). The purified RNA was dissolved in nuclease-free H_2O and remaining genomic DNA was eliminated with TURBO DNA-free™ DNase (Ambion) according with the manufacturer's instructions.

7.2.2.12   Measurement of quality and quantity of total RNA

The RNA quantity was assessed using the Qubit RNA HS Assay Kit (Invitrogen), which is an accurate method for measuring low RNA concentrations. A working solution master mix was prepared by diluting the Qubit® RNA HS Reagent 1:200 in Qubit® RNA HS Buffer. 1µl of RNA (total RNA or RNA coming from CLIP samples) was added to 199µl of working solution. After vortexing and incubation at RT for 2 min, the RNA concentration was determined based on a standard calibration curve. The quality of the total RNA, used for routine cDNA synthesis, was determined by loading 200-500ng of the quantified sample on a native 1% agarose gel electrophoresis. The gels were stained with Ethidium bromide (EtBr) and visualized with the Agarose gel UV imaging system (Intas). The quality of total RNA purified for high throughput sequencing, or the RNA purified from CLIP experiments was measured using the Agilent RNA 6000 Pico Kit on a Bioanalyzer device.

7.2.2.13   cDNA synthesis

500 ng of total RNA (section 7.2.2.11) were used to prepare cDNA using the SuperScript II reverse transcriptase (Invitrogen) and random hexamers (Invitrogen) to prime the reactions. The protocol was performed following the recommendations of the manufacturer. cDNA was then used for qPCR validation (7.2.2.5) or quality control of RNA-seq. When corresponding, cDNAwas used as template for cloning.

7.2.2.14   MgCl$_2$-dependent RNA fragmentation

To get fragmented RNA for competition EMSA assays (methods 7.2.4.2.2), total RNA from one 10 cm plate of HeLa cells was extracted with Trizol according to section 7.2.2.11 and dissolved in buffer Ex0 (containing 1.5 mM MgCl$_2$) at a concentration of 500ng/µl in a volume of 50µl. The RNA was incubated at 95 °C for increasing time points (up to 10 min) in a thermocycler (standardization in Figure 46 in Appendix). After incubation, the fragmented RNA was loaded on a 1% agarose gel

electrophoresis as described in 7.2.2.2 and stained with EtBr. Gels were documented using the Agarose gel UV imaging system (Intas).

7.2.3 Proteins

Unless otherwise indicated, proteins were always manipulated on ice and for long-term storage they were aliquoted, snap frozen with liquid nitrogen and kept at -80 °C. Depending on the experiment setup the buffers were pre-chilled at 4 °C and supplemented with protease inhibitors mix (PI) when necessary.

7.2.3.1 Protein quantification

7.2.3.1.1 Qubit quantification

The purified recombinant proteins were quantified using the Qubit® Protein Assay Kit (Invitrogen). For this purpose, a working solution was prepared by diluting the Qubit® Protein Reagent 1:200 in Qubit® Protein Buffer. 1-5µl of the stock proteins were mixed with the working solution in a final volume of 200µl, vortexed for a few seconds (avoiding the formation of bubbles) and incubated for 15 min at RT. The protein concentration was determined based on a standard calibration curve of protein standards provided with the kit.

7.2.3.1.2 Bradford assay

Whole cell extracts (WCE) or nuclear extracts (NE), prepared for western blot analysis were quantified by using a modified protocol of the Bradford Coomassie brilliant blue assay (Bradford, 1976). 6 standards of BSA dilutions were prepared in water (0, 0.1, 0.25, 0.5, 1.5, and 5 µg/µl). 10µl of each standard, 1-10µl protein sample and a background control (water or buffer) were added to 1ml of a 1:5 dilution

of the Bradford reagent, respectively, in a disposable micro cuvette. The samples were mixed and the absorbance at 590 and 450 nm was measured. To get the linearization of the concentration curve over the protein concentration range, the absorbance was corrected using the ratio of absorbance at 590 nm over 450 nm by a mathematical equation described in (Ernst & Zor, 2010).

7.2.3.2 Denaturing Polyacrylamide Gel Electrophoresis (SDS-PAGE)

To visualize and determine the quality/purity of proteins, and when performing western blots, the proteins were analyzed using denaturing Polyacrylamide Gel Electrophoresis (SDS-PAGE). The stacking and resolving gels were prepared according to standard protocols, using a ready to use 30% Acrylamide/Bisacrylamide solution (Rotiphorese® NF-Acrylamide/Bis-solution 30%). Depending on the size of proteins to be analyzed, the separation gels were prepared with concentrations of polyacrylamide ranging between 10 and 17%, 1X SDS-PAGE Lower Buffer, 0.1% APS, and 0.1% TEMED. The gels were prepared in the cassettes of the XCell SureLock™ Mini-Cell system, filling the separation gel up to 2 cm under the rim of the cassette. The gels were then covered with a layer of isopropanol until the polymerization was completed. A 5% stacking gel was prepared with the Rotiphorese solution, 1X SDS-PAGE Upper Buffer, 0.1% APS and 0.1% TEMED. After complete removal of the isopropanol, the freshly prepared solution was poured covering the resolving gels, and a 12-15 wells comb was inserted. The samples to be analyzed were mixed with 1X Lämmli Buffer (Protein Loading Buffer), boiled for 5 min at 95 °C and loaded on the gels. The proteins were separated at 40mA in SDS-PAGE running buffer for 50-55 min. The electrophoretic mobility of the proteins and the size was followed by comparison with the Prestained protein ladder PageRuler™ Plus. For visualization, the gels were stained with Coomassie Brilliant Blue solution for 5-10 min and then destained by boiling in water using a microwave, until the bands were visible. The gels were documented using a common office scanner.

7.2.3.3 Native Polyacrylamide gel electrophoresis (PAGE)

Native polyacrylamide gel electrophoresis was mainly used for EMSA experiments to analyze protein-RNA, protein-DNA or protein-nucleosome interactions, or to reveal the assembly of nucleosomes *in vitro*. The gels were prepared with a ready to use 30% Acrylamide/Bisacrylamide solution (Rotiphorese® NF-Acrylamide/Bis-solution 30%) to a final concentration between 6 to 8 %, 0.4X TBE, 80µl APS per 10ml solution and 8µl TEMED per 10ml solution. The XCell SureLock™ cassettes were fully filled with the freshly prepared PAA solution and 10 well combs were inserted. Before loading the samples, the gels were rinsed with a syringe and pre run at 80V for 30 min at room temperature to remove unpolymerized acrylamide. The electrophoretic mobility was done at room temperature at 100V between 30 and 60 min depending on the application. The gels were visualized and documented using the Typhoon FLA 9500 (GE Healthcare).

7.2.3.4 Whole cell extract preparation

Tissue culture from 6-well plates or 10 cm culture dishes were washed twice with ice-cold PBS, and then harvested with a cell scraper in 1ml ice-cold PBS. The cells were transferred to a 1.5ml microcentrifuge tube and sedimented by centrifugation at 500-1000g for 2 min at 4 °C. The supernatant was aspirated and the cells resuspended in 100-500µl Whole Cell Lysis Buffer supplemented with protease inhibitor mix (PI). After incubation for 15 min on ice with vigorously vortexing every 5 minutes, the extracts were sonicated with a Bioruptor device at low intensity, during 1 min (15 sec ON/ 15 sec OFF) to help disrupting the membranes. The cell lysate was centrifuged at 13000g for 15 min at 4 °C. Subsequently, the supernatant was saved and quantified (7.2.3.1.2) or stored at -80 °C for later downstream applications.

7.2.3.5 Nuclei preparation

The nuclei were isolated using the "rapid, efficient, and practical method" (REAP) (Nabbi & Riabowol, 2015). Tissue cultures were grown to 70-90% confluence in 10 cm culture dishes. The culture dishes were placed on ice, the medium was aspirated and the cells were washed twice in ice-cold PBS before scraping them of the culture dish with 1ml ice-cold PBS. The cells were transferred to a 1.5ml microcentrifuge tube and centrifuged at 500g for 2 min at 4 °C. The supernatant was removed and the cell pellet resuspended in 500µl ice-cold PBS containing 0.1% IGEPAL CA-630. The cell suspension was resuspended five times on ice with a P1000 micropipette and then centrifuged for 5–10 sec at 10.000 rpm. Resuspension step and centrifugation were repeated 2 more times in 500µl ice-cold PBS/0.1% IGEPAL CA-630. Finally the supernatant was aspirated, and the nuclei (white pellet) were resuspended in 100-200µl ice-cold PBS/0.1% IGEPAL CA-630 supplemented with protease inhibitor mix (PI) and sonicated for 1 min using a Bioruptor device at high intensity (15 sec ON/15 sec OFF). The samples were stored at -80 °C or immediately used in the downstream experiments.

7.2.3.6 Semi dry Western blot

Proteins to be analyzed by western blot were first subjected to a SDS-PAGE (7.2.3.2). Depending on the quality of the antibodies and the relative abundance of the proteins to be detected, 20-100µg of total protein extract was used. To mount the transfer to a PVDF membrane, four Whatman filter papers were cut to the size of the gel (6X8 cm), the membrane was activated in 100% methanol for 1-2 min and then soaked in Transfer Buffer (Towbin) together with the Whatman papers and the gel for 5-10 min. The "sandwich" was build up in the Bio-Rad 'Trans-Blot SD Apparatus' from anode to cathode, starting with two soaked Whatman papers, followed by the activated membrane, then the gel and finally two soaked Whatman papers, making sure that no air bubbles were trapped in the blot. Depending on the gel percentage, the transfer took place for 40 min to 1 h at 1.5-2.5mA/cm^2.

After transfer, the membranes were incubated for 30 min in blocking solution (1X PBS-T, containing 5% dried milk) at room temperature while gently shaking. The

membranes were then incubated with an appropriate dilution of primary antibody prepared in blocking solution, for 1 h at room temperature or ON at 4 °C with gentle shaking. The membranes were washed four times for 2-5 min with PBS-T, and then incubated for 45 min with the secondary antibody (generally HRP conjugated). The membranes were then washed 4 times for 2-5 min with PBS-T. The antigen-antibody complexes were detected using Supersignal West Dura (Pierce) by preparing a 1:1 dilution of the reagents given in the kit (300-500µl each depending on the size of the membrane), and mixed on a piece of parafilm (bigger than the membrane). The mixture was spread over the parafilm in form of drops, and the membrane was placed on top. The membranes were incubated for 2-3 min protected from light. For signal detection the membranes were placed between two layers of a transparent disposal bag, and the protein ladder was marked with a fluorescent marker. Western blot signals were detected with a LAS-3000 Imager.

7.2.3.7 Expression and purification of proteins

7.2.3.7.1 Small-scale bacterial protein production

Prior to large-scale protein expression, His-tagged or GST-tagged proteins were expressed in a small-scale to check for the proper expression of the proteins. In general, BL21 cells, or modified BL21 strains were used. 2 or 3 colonies were selected after transformation and grown ON in 2-3ml of LB medium supplemented with the specific antibiotics. The next day 1ml culture was taken to prepare a glycerol stock (7.2.1.2); the remaining culture was used to inoculate 15-20ml of LB containing antibiotics and grown until the OD600 reached 0.3-0.6 and 1ml of the culture was taken as "non-induced" sample. Protein expression was induced with 1 mM IPTG, and 1ml of samples were taken after 1, 2, and 4 hours, or until the OD600 reached 1.5. All the samples were pelleted for 30 sec at >2000g, the supernatant was discarded, and the samples were resuspended in 1X Lämmli buffer. Lämmli buffer was added according to the formula: 1 OD=150µl Lämmli buffer. 15-20µl bacterial

whole cell preparation were loaded on a 10-15% SDS-PAGE (see 7.2.3.2), depending on the size of the protein.

7.2.3.7.2 GST-tagged protein purification

7.2.3.7.2.1 Large-scale protein expression

For every 50ml expression volume, a 5ml pre–culture was inoculated and incubated at 37 °C with shaking (180 rpm) overnight. The pre-culture was diluted to an OD600 of about 0.05-0.1 with fresh LB medium supplemented with antibiotics. The samples were incubated at 37 °C on a shaker until the OD600 reached 0.3-0.6 and then protein expression was induced by the addition of 1 mM IPTG and incubated for 3-4 h. Afterwards, the samples were centrifuged for 10 min at >4000g at 4 °C, the supernatant was removed and the pellet were frozen in liquid nitrogen and stored at -80 °C until purification.

7.2.3.7.2.2 Lysis and protein purification

For the purification of GST-tagged proteins, pellets corresponding to 50ml bacterial culture were resuspended with 10ml ice-cold GST Lysis Buffer supplemented with protease inhibitor mix (PI) on ice, in a 15ml falcon tube. The cell lysates were prepared by sonication using a Branson sonifier 250D (50% amplitude, 50% Duty cycle, 4X 20 sec ON/20 sec OFF) keeping the samples on ice. The lysates were centrifuged at >4000g, 4 °C for 30 min. The supernatant was transferred to a new tube for purification and the pellets were discarded.

For every sample to be purified, 150µl of GST–agarose beads were washed two times in 1ml 1X PBS by resuspending and spinning down in a tabletop centrifuge at 500g for 1 min. The bacterial lysates were then mixed with the washed beads and incubated in an overhead rotator for 0.5-1 h at room temperature (RT). The supernatant was discarded and the beads were washed four times with 10ml GST–

wash buffer at RT for 5 min with gentle agitation. The proteins were eluted with 150µl GST Elution Buffer (pH adjusted to 8.0 with NaOH directly prior to use, adding 2-3µl of 2 M NaOH per ml of buffer and monitored with a pH indicator) for 20 min in an overhead rotator. The beads were spun down and the supernatant was transferred into a concentrator column (3kDa or 10kDa, depending on the protein size) for concentration and buffer exchange. The samples were quantified, aliquoted and stored at -80 °C.

7.2.3.7.3 His-tagged protein purification

7.2.3.7.3.1 Large-scale expression

For the large-scale protein purification of His-tagged proteins an overnight culture of 20-30ml was added to 500ml of LB supplemented with antibiotics and incubated on a shaker at 37 °C until the OD600 reached 0.3-0.6. Protein expression was induced with 1mM IPTG and incubated on a shaker at 37°C until the OD was 1.5 (about 4 h). The cultures were centrifuged at 10000g for 10 min and the pellets were frozen at -80 °C, or immediately used for protein purification.

7.2.3.7.3.2 Preparation of the bacterial cell extract

The cell pellets obtained in 7.2.3.7.3.1 were resuspended in 45ml ice-cold His Lysis buffer + PI on ice, and the cell suspension was sonicated 6X 20s on ice, with a Branson sonifier (50% duty cycle, intensity level 5, 20 sec ON/ 20 sec OFF).
 The insoluble cell fragments were spun down at >4000g for 30 min at 4 °C in a 50ml centrifuge tube and the supernatant was transferred into a fresh 50ml tube for further purification.

7.2.3.7.3.3 Purification of the His-tagged Proteins

The supernatant (7.2.3.7.3.2) was mixed with 1-1.5ml of 50% Ni-NTA beads slurry (pre-washed with 10ml of His Lysis Buffer). The reaction was incubated at 4 °C for 1 h with gentle agitation. The beads were spun down at 2000*g* for 5 min at 4 °C and the supernatant was discarded. The beads were transferred into a 15ml falcon tube and washed 4 times with 10ml His Wash Buffer for 5 min at 4 °C on a rotating wheel with centrifugation steps of 5 min at 2000*g*.

The proteins were eluted with 3ml His Elution Buffer for 60 min at 4 °C on a rotating wheel. The beads were then centrifuged at 2000*g* for 5min and the supernatant transferred to a concentrator column (Vivaspin 500) for concentration and buffer exchange. Samples were quantified using the Qubit system (described in 7.2.3.1.1), aliquoted, snap frozen and stored at -80 °C for the downstream experiments.

7.2.4 *In vitro* interactions

7.2.4.1 *In vitro* reconstitution of nucleosomes

Mononucleosomes were assembled using the salt gradient dialysis method (Rhodes & Laskey, 1989) using histones purified from chicken blood. For that purpose the plasmid pUC18 12x601 that contains an array of the strong nucleosome positioning sequence 601 was hydrolyzed with the enzyme AvaI to obtain a 200 bp fragment containing the 601 sequence. DNA fragments were purified from an agarose gel (7.2.2.7) and used as templates for a large-scale PCR reaction to generate a 149 bp long Cy3 labeled DNA (Primers RMP#012 and RMP#014). The assembly reactions were performed in dialysis chamber prepared on a lid of siliconized 1.5ml tubes. A hole was melted in the middle of the lid (so called O-ring), the cap was smoothened with a scalpel and the bottom of the tube was cut off. A piece of dialysis membrane with a MWCO of 6–8 kDa (Spectrapor) was pre-incubated for 5 min in Chromatin Assembly High Salt Buffer and then placed between the tube and the lid to create the dialysis chamber. The chambers were placed in a Styrofoam floater and transferred

to a 3 liter beaker filled with 300ml Chromatin Assembly High Salt Buffer, containing a magnetic stirrer. The air bubbles below each membrane were removed by using a bent Pasteur pipet, and the chambers were checked for leakage. In general, a small-scale assembly was performed to determine the appropriate ratio of histones to DNA. The small-scale reactions were prepared in a final volume of 50µl of High Salt Buffer, and BSA in a final concentration of 200ng/µl, containing 1µg of the purified fluorescently labeled DNA, and histones to DNA concentration ratios between 0.4:1 to 1.5:1. The large-scale assembly was performed in a final volume of 250µl High Salt Buffer, containing 5µg of DNA, 200ng/µl BSA and the previously determined histones concentration (supplementary Figure 45).

The assembly reactions were loaded into the dialysis chambers and dialysis was started by pumping 3 liter of Chromatin Assembly Low Salt Buffer into the beaker at a flow rate of about 150–300 ml/h at RT. The beaker containing the reactions was covered with aluminum foil to protect the samples from light. After the dialysis was complete, the volume of the samples was determined to estimate the concentration of the nucleosomes.

The quality of the assembly reaction was analyzed on a 6% native PAA gel in 0.4X TBE (section 7.2.3.3) and run for 1 h at RT. The assembled nucleosomes were stored at 4 °C protected from light.

7.2.4.2 Protein-nucleic acids interactions

7.2.4.2.1 EMSA

To analyze the interaction of proteins with RNA, DNA or nucleosomes, electrophoretic mobility shift assay (EMSA) were performed. Protein concentration was titrated over the potential interaction partner. Therefore, proteins were used in a dilution ratio of 2:1 keeping constant the fluorescently labeled partner, being 50nM for DNA and RNA, and 50ng of nucleosome, respectively. The reactions had a final volume of 10µl in EX100 buffer and were incubated for 10 min at RT. 4µl of each reaction (a total of 9 serial dilutions and one negative control without protein) were loaded on a 6% native PAA gel. The EMSAs were run for 30 min at 100V at RT when

analyzing protein-DNA or protein-RNA interactions, or 1 h at 100V for protein-nucleosome interactions. The gels were documented using a Typhoon 9500 device and analyzed using the ImageJ software.

7.2.4.2.2 Competitive binding assay

To perform the competitive binding assays between RNA and nucleosome, the protein-nucleosome complexes were kept at constant concentration (2µM protein and 50ng of nucleosome) adding increasing amounts of non-labeled fragmented HeLa RNA (RNA fragmentation decribed in 7.2.2.14) (up to 2µg) for competition. The reactions were loaded on 6% native PAA gel and run 1 h at 100V. The gels were documented using a Typhoon 9500 device and analyzed using the ImageJ software.

7.2.4.2.3 Microscale thermophoresis (MST)

Microscale thermophoresis was used to determine the binding affinity of HMGN5 to different nucleic acids. The technique is a powerful method to determine molecular interactions in solution. MST is based on the motion of molecules along a temperature gradient, which depends on changes in the hydration shell, charge or size (Jerabek-Willemsen, Wienken, Braun, Baaske, & Duhr, 2011).

For the interactions, titration series of the proteins were prepared by diluting the proteins in a 2:1 ratio with EX100 Buffer (as indicated in the plots). The binding reactions were performed in a final volume of 10µl containing 50nM of the fluorescently labeled nucleic acids (Table 1), varying protein concentrations and 0.005% IGEPAL CA-630. 15 serial dilutions and one control without protein were incubated for 10 min at RT and used to fill up MonolithNT™ standard capillaries. The binding was measured using a Monolith NT.015T/NT.115 device (NanoTemper) at RT with laser ON for 40 sec, LED 20 and Laser power 40. The thermophoresis signals were plotted using the Sigmaplot software (v12.5) and fitted according to the Hill equation $f = b + \left[\frac{m-b}{1+(IC_{50}/[C])^n}\right]$ were *f* is the measured raw fluorescence, *m* and *b* correspond to the maximum and minimum fluorescence values of the titration curve, *C* corresponds to the concentration of the unlabeled protein, and *n* is the Hill coefficient. For data evaluation, the thermophoresis signals were normalized to

fraction bound (X) $X = \frac{f-b}{m-b}$ to correct for the maximum (*m*) and minimum (b) values of the binding curve.

7.2.5 Mammalian Cell culture

7.2.5.1 Basic cell culture techniques

7.2.5.1.1 Techniques for mammalian cell manipulation

The manipulation of the mammalian cell cultures was done in accordance with the "Guidance for a Good Cell Culture Practice" (GCCP), in compliance with regulations and ethical principles, using safety-working procedures. Materials were handled with standard precautions under a sterile hood in laminar flow. All the materials and the surface of the working place were wiped with 70% ethanol, and the medium and trypsin were always warmed up to 37 °C directly before use. Medium and trypsin were kept at 4 °C when not in use and PBS was always stored at room temperature. Antibiotics were stored at 4 °C or at -20 °C depending on the recommendations of the manufacturer.

Cell lines used in this thesis were cultured in low glucose DMEM GlutaMAX™ medium supplemented with 10% FBS and grown in humidified incubators at 37 °C (atmosphere 95% air/ 5% CO_2).

7.2.5.1.2 Thawing of mammalian frozen cells

New batches of cells were thawed quickly in a water bath at 37 °C and transferred to a 75cm^2 culture flask containing 25ml of pre-warmed medium. After 20-24 h of incubation at 37 °C the cells were checked for viability under a microscope. The medium was aspirated using a sterile Pasteur pipette, the cells were rinsed once with

10ml PBS to remove non-attached cells and remaining DMSO, and fresh medium was added.

7.2.5.1.3 Culturing adherent cells

For maintenance, the cells were grown in 75cm^2 flasks. Splitting was performed every 2-3 days when 80-90% confluence was reached. For splitting, the medium was aspirated with a Pasteur pipette and the cells were washed once with 10ml sterile PBS, afterwards 5ml of 0.05% Trypsin - EDTA (1X) solution (Gibco) were added and incubated at 37 °C for 3-5 min. Cells were detached using 5ml trypsin. The detachment was monitored under a light microscope. Trypsin was inactivated with 4-5 volumes of fresh medium and the cells were resuspended several times using a 10ml serological pipette until all cell clumps were dissolved. The cells were then transferred into a 75cm^2 culture flask, diluting the cells by a factor of 1:6-1:10 in 25ml fresh medium. When performing large-scale experiments, the cells were seeded in 15 cm culture plates using 20-25ml of fresh medium. Every splitting step increased the passage number by 1 and was documented. After 15 passages cells were discarded.

7.2.5.1.4 Cryopreservation of cells

The cryopreservation of cells was always conducted as early as possible, best at low passage numbers (not more than 5). Before cryopreservation, cells were checked for contamination under a microscope. Cells with 70-80% confluence were detached from the flasks with 0.05% Trypsin - EDTA (1X) for 3-5 min, then resuspended with 5 volumes of fresh medium, centrifuged at 500g for 5 min at RT and the medium was removed. The pellets were resuspended in 10ml fresh medium. 500µl of cells suspension was mixed with Trypan Blue (Gibco) to analyze cellular integrity and to count the cells in a Neubauer chamber. The remaining cells were centrifuged at 500g for 2-3min, the medium was aspirated and the cells were resuspended quickly in

complete medium containing 10% DMSO reaching a cell number of of 1×10^7 cells/ml. 1ml aliquots were transferred into labeled cryovials and quickly placed in a insulator box and stored overnight at -80 °C. The insulator box ensures a freezing rate of 1 °C/min. After 24 h the cryovials were permanently stored at -80 °C.

7.2.5.2 Generation of stable Flp-In cell line

To establish an inducible cell line expressing HMGN5, the T-REx™-293 Flp-In cell line (Life Technologies) was used. The system uses Flp recombinase-mediated integration derived from *Saccharomyces cerevisiae*, that allows Tetracycline-inducible expression of a gene of interest from a specific genomic location(O'Gorman, Fox, & Wahl, 1991). The cell line contains a Flp recombination target site (FRT site) integrated in the genome. The coding sequence of HMGN5 (NM_030763.2) was cloned in frame with a C-terminal GFP in the expression vector pcDNA5/FRT/TO, which contains the recombination site FRT and a Hygromycin resistance cassette for selection. The expression of the protein is controlled by a hybrid human cytomegalovirus (CMV)/TetO2 promoter for high-level, Tetracycline-regulated expression. The plasmid is co-transfected with the pOG44 vector that transiently expresses the Flp recombinase, mediating recombination between the FRT sites on the plasmid and the genome, resulting in a stable single integration of the gene of interest in the genome.

For maintenance, T-REx cells were grown in low glucose DMEM medium as described in section 7.2.5.1.3) plus Tetracycline-Free FBS (Biochrom) supplemented with 100 µg/ml Zeocin and 10 µg/ml Blasticidin freshly before use. For transfection, the cells were seeded in T25 culture flasks at 50-60% confluence in medium without antibiotics. After 24 h of incubation at 37 °C the cells were transfected with the pOG44 and pcDNA5/FRT/TO© constructs in a 9:1ratio. For this, the plasmids were diluted to 10ng/µl in 150µl Opti-MEM® I Reduced Serum Medium. Subsequently the FuGENE® HD Transfection Reagent was added to the mixture in a ratio of 3µl per µg of DNA and incubated for 15 min at RT. The mixture was then transferred to the T25 flask containing the cells in 3ml medium without antibiotics.

After 24 h of incubation at 37 °C, the cells were washed with 5ml PBS and fresh medium was added. 48 h post transfection the medium was aspirated and the cells

were trypsinized with 1ml 0.05% Trypsin - EDTA (1X) for 3-5 min. The cells were resuspended in 5ml fresh medium and 3 aliquots of 0.5, 1 and 3.5ml were taken and transferred to P10 dishes in a final volume of 8ml with medium plus antibiotics for selection (100 µg/ml Hygromycin and 10 µg/ml Blasticidin). The medium with the selective antibiotics was changed every 3 days until foci could be identified by eye (about 10-12 days).

The Hygromycin-resistant foci were picked by scraping them with a pipet tip and resuspended in 500µl medium plus antibiotics for further expansion in 24-well plates. When 60-70% confluence was reached, cells were split in a 1:5 ratio. One part was used for testing protein expression with 1 µg/ml Doxycycline in 3ml of complete medium (without antibiotics). The Doxycycline induction was tested in 6-well plates containing coverslips for immunofluorescence analysis. One part of the splitted cells was taken and cultivated without Doxycycline induction as control. The remaining cells were grown in 6-well plates and incubated with medium plus antibiotics. Once cells reached 70-90% confluence, they were split and expanded in 75cm^2 flasks to prepare cryocultures as described in 7.2.5.1.4. Only clones showing stable protein expression (analyzed by the GFP expression level) after induction and not (or undetectable) in the non-induced control cells were picked for further expansion.

7.2.5.3 Recruitment HMGN5 to a cellular LacO array

The tethering of HMGN5 to a specific genomic locus was performed using the LacO/lacI system in U2OS cells. The method allows the targeting of a protein to an array of tandem repeat (256 copies) of the bacterial Lac operator (LacO) which is inserted in a highly compacted telomeric region in U2OS cells (Jegou et al., 2009), allowing for the analysis of chromatin decompaction.

The coding sequence of the human HMGN5 (NM_030763.2) was cloned downstream of the fusion gene GFP-LacI in the vector pSV2_GFP-LacI, to generate the construct pSV2_GFP-LacI-HMGN5.

U2OS cells (kindly provided by Dr. Karsten Rippe) were maintained in low glucose DMEM medium supplemented with 10% FBS. For transfection, cells were seeded on coverslips in 6-well plates at 70% confluence (150000 cells approx.) in 3ml complete DMEM and incubated for 24 h at 37 °C. Cells were then transfected with the vector

pSV2_GFP-LacI-HMGN5, and control cells were transfected with the vector pSV2_GFP-LacI (to express GFP-LacI). The plasmids were diluted to 20ng/µl of in 153µl OptiMEM. 12µl of Lipofectamine 3000 (Invitrogen) reagent (ratio reagent:DNA 4:1) were added to the mixture and carefully mixed by pipetting to form the complex. After 15 min incubation at RT, 150µl of complex were added to each well and gently mixed. Cells were incubated for 24 hour at 37 °C and analyzed by immunofluorescence microscopy according to 7.2.5.4.

7.2.5.4 Immunofluorescence microscopy

7.2.5.4.1 Fixation and Permeabilization of cells

Adherent cells grown on coverslips were washed twice with 1X PBS at RT for 2-3 min with gentle agitation. Cells were then fixed in 4% paraformaldehyde (PFA)/PBS for 10 min at RT on an overhead rotator. A few drops of Immunofluorescence Permeabilization buffer 1 (0.5% Triton X-100/PBS) were added after 9 minutes. Samples were washed 3 times with Immunofluorescence Permeabilization buffer 2 (0.01% Triton /PBS) for 3 min at RT. An additional wash step using 0.5% Triton X-100/PBS for 5 min at RT was performed. Then the cells were washed twice with 1X PBS and an immunofluorescence step was performed (see next section).

7.2.5.4.2 Immunofluorescence

All steps were performed at RT; the wash steps were done in an overhead rotator. The blocking solution was prepared using Albumin Bovine Fraction V (Serva). Cells were incubated with the primary antibody diluted in 4% BSA/PBS-Tween for 45 min. After three wash steps with 0.1% PBS-T for 5 min, the samples were incubated with the secondary antibody diluted in 4% BSA/PBS-T for 30 min and washed twice in 0.1% PBS-T for 5 min. DNA counterstaining was performed with 50 ng/ml DAPI in PBS-T for 5 min, then rinsed once with 0.1% PBS-T followed by rinsing with 1X PBS.

The coverslips with the cells were mounted on microscopy slides using Vectashield mounting medium (Sigma). The edges of the coverslips were sealed with nail polish and the samples were stored at 4 °C. The next day they were analyzed by microscopy.

7.2.6 Chromatin specific methods

7.2.6.1 Chromatin Immunoprecipitation with GFP-Trap

The HMGN5-FlpIn stable cell line was used to perform the chromatin immunoprecipitation (ChIP) of HMGN5. As control, the GFP-FlpIn cell line was used. Experiments were performed in triplicates, seeding cells in one 15 cm plate for each experiment, using low glucose DMEM/Tet free FBS without antibiotics. When the cells were 70% confluent, protein expression was induced with 1 µg/ml Doxycycline for 24 h at 37 °C. The cells were washed twice with 1X PBS for 5 min at RT with gentle agitation. Formaldehyde was added to a final concentration of 0.8 % in 15ml PBS and incubated 10 min with gentle agitation at RT. The reaction was stopped with 125 mM glycine and incubated for 5 min with gentle shaking at RT. Cells were washed twice with 10ml ice-cold PBS and then harvested in 4ml ice-cold PBS by pipetting (cells can be easily detached from the plates). Cells were then transferred into 15ml falcon tubes and centrifuged at 2000g for 5 min at 4 °C. The pelleted cells were washed with 2ml ice-cold PBS and pelleted for 5 min at 2000g. The pellets were resuspended in 900µl ChIP SDS Lysis Buffer and incubated for 10 min on ice. For sonication the samples were split into three aliquots of 300µl in 1.5ml tubes (maximum volume that can be processed in the Bioruptor device). The Bioruptor was precooled by the addition of crashed ice to the water bath and 3 cycles of 10 min with 30s ON/ 30s OFF at high intensity were applied (fresh ice was added to the water bath between cycles). The cell debris was pelleted at 13000g for 10 min at 4 °C. The supernatant containing the chromatin was transferred to a new microcentrifuge tube. Cross-linked material was either immediately used for immunoprecipitation or stored at -80 °C.

For the immunoprecipitation only siliconized 1.5ml Eppendorf tubes were used. Per sample 10µl GFP-Trap agarose beads were equilibrated with 300µl ChIP Low Salt Buffer. The beads were centrifuged for 2 min at 500*g*, and the supernatant was discarded. The chromatin was added to the equilibrated beads and diluted to 500µl with ChIP Low Salt Buffer. The binding reaction was incubated for 2 h at 4 °C on a rotating wheel. The beads were spun down for 1 min at 500*g* at RT and the supernatant was removed.

Bead washing was performed for 2-3 min at RT on a rotating wheel, followed by centrifugation steps of 1 min at 500*g* and RT. The beads were washed twice with 700µl ChIP Low Salt Buffer, then one time with 700µl ChIP High Salt Buffer followed by a wash with ChIP LiCl Buffer. A final washing step was done with 500µl EB buffer (Qiagen) and the beads were finally resuspended in 500µl EB buffer.

Since the GBP (GFP Binding Protein) of the GFP-Trap beads is covalently linked to the beads, and the interaction with GFP is very strong, the protein-DNA complexes were not eluted but directly decrosslinked on the beads and isolated. 8µl RNase A (20mg/ml) were added to each sample and incubated for 2 h at 37 °C with occasional vortexing. 5µl of 10mg/ml Proteinase K were added to each sample followed by an incubation of 1 h at 55 °C in a thermomixer. To revert the crosslinking, a final ON incubation at 65 °C was performed.

7.2.6.1.1 DNA purification

The samples were transferred to "Safe-lock" tubes, and a phenol/chloroform extraction was performed. Briefly, 1 volume of phenol/chloroform/isoamyl alcohol (Roth) was added to each sample and mixed vigorously by vortexing. A centrifugation step for 5 min at 13000g RT was performed to separate the organic and the aqueous phase. The upper aqueous phase containing the DNA was carefully transferred to a fresh tube. The phenolic extraction was repeated. To remove traces of phenol the samples were mixed with 1 volume of chloroform p.a., mixed by vortexing and centrifuged for 5 min at RT. The aqueous phase was transferred to a fresh tube and the DNA was precipitated with 2 volumes ice-cold 100% EtOH and 0.1 volume 3 M sodium acetate pH 5.2. DNA was precipitated for 20 min at -20 °C. The samples were centrifuged for 15 min at 13000*g* and 4 °C, the supernatant was

aspirated carefully and the DNA washed with 500µl of 70% ice-cold ethanol. The ethanol was removed with a pipette and the pellet was air-dried for approx. 15min. The pellets were dissolved in 60µl of sterile H_2O and stored at -20 °C until analyzed.

7.2.6.2 RNA Immunoprecipitation with GFP-Trap

For RNA immunoprecipitation the HMGN5-FlpIn and the control cell line GFP-FlpIn were used. The cells were seeded in 15 cm culture dishes in low glucose DMEM/ 10% Tet-free FBS without antibiotics and grown at 37 °C until 70-80% confluence was reached. Protein expression was induced with 1µg/ml Doxycycline for 6 h at 37 °C. After induction the cells were washed twice with PBS. All PBS traces were removed and the plates were placed without lid on ice into the stratalinker UV crosslinker (Stratagene). UV crosslinking was performed by applying 150mJ/cm^2 UV light at 254 nm. While on ice, cells were harvested in 1.5 ml ice-cold PBS with a scraper and transferred into a 2 ml tube. Cells were centrifuged at 2000g for 3 min at 4 °C and washed with 2ml ice-cold PBS. For sonication, the samples were resuspended in 400µl CLIP Lysis Buffer and vortexed for a few seconds. The samples were split into 4 aliquots of 100µl in 0.6 ml tubes and sonicated for 15 min with high intensity in a precooled Bioruptor (30 sec ON/ 30 sec OFF). The cell debris was pelleted for 5 min at 8000g and 4 °C and the supernatant was transferred to a fresh tube. For the immunoprecipitation steps only siliconized microcentrifuge tubes were used. Per sample, 10µl of GFP-Trap_M (magnetic) beads were equilibrated with 300µl CLIP Low Salt Buffer and placed in a 96-well magnetic plate (SPRIplate® 96-Ring). The magnetic beads on suspension were concentrated to a ring for 5 min, until the solution became clear, and the supernatant was removed. The equilibrated beads were mixed with 200µl of the sonicated samples and diluted to 500µl with CLIP Medium Salt Buffer + PI + 100 U/ml RNasin. The samples were incubated for 2 hours at 4 °C on a rotating wheel.

After incubation the samples were placed in the magnetic plate for 5 min to remove the supernatant. All the washing steps were performed at RT incubating with buffer for 5 min on a rotating wheel, and the beads were placed in a magnetic plate for 5 min followed by complete removal of the supernatant with a 200µl pipette. Beads were washed 3X with 700µl CLIP Low Salt Buffer, then 2X with 700µl CLIP High Salt Buffer and finally once with 500µl CLIP Elution buffer (EB). RNA was partially

digested on-bead using 500µl EB plus 1µl RNase A/T1 cocktail pre-diluted 1:100 (the dilution was previously standardized (See results 4.7). RNA hydrolysis was performed for 10 min at 37 °C on a rotating wheel. The samples were then placed in a magnetic plate, incubated for 1-2 min and the supernatant was removed. The beads were washed once with 500µl CLIP Elution Buffer and immediately resuspended in 30µl 1X DNase Q1 buffer containing 100U/ml RNasin and 1µl DNase Q1 to remove contaminant DNA. The DNA was hydrolyzed for 1 h at 37 °C in a thermomixer. The samples were diluted to 200µl with Proteinase K Buffer, and proteinase K was added to a concentration of 300µg/ml. The reaction was incubated for 2 h at 37 °C.

Finally, the RNA was purified using the Trizol reagent (using Safe-lock tubes), following the instructions of the manufacturer and being careful in not carrying over magnetic beads. The RNA was eluted in 20µl nuclease-free water, quantified and analyzed in a Bioanalyzer using the Agilent RNA 6000 Pico Kit (Section 7.2.2.12). The RNA was stored at -80 °C and used to prepare libraries (section 7.2.7.3).

7.2.6.3 Knockdown of HMGN5

HMGN5-FlpIn cells seeded in 6-well plates were grown to 70% in 2.5ml low glucose DMEM medium without antibiotics. Tho perform the transfection, 100pmol of a SmartPool siRNA against HMGN5 or a non-targeting siRNA pool (synthesized at Gunter Maister's laboratory) were diluted in 250µl Opti-MEM. 5µl of Lipofectamine RNAimax reagent (Invitrogen) were diluted in 250µl Opti-MEM and were combined with the diluted siRNA. To form the complexes RNAi-Lipofectamine, the mixture was incubated 20 min at RT, and then added by drops to the cells. 24 h post transfection the cells were harvested with a cell scraper and used for the downstream experiments. To check the efficient knockdown of HMGN5, whole cell protein extract were prepared (7.2.3.4) and HMGN5 was detected by western blot (methods 7.2.3.6) using the antibody anti-HMGN5 HPA000511 (Sigma).

7.2.6.4 RNA preparation and quality control for RNA-seq

To perform RNA-seq, total RNA was prurified from HMGN5-FlpIn cells after HMGN5 overexpression, knockdown and from non-treated control cells. For HMGN5 overexpression, cells were grown in 10 cm culture plates in low glucose DMEM medium without antibiotics until 70% confluence was reached. Protein expression was induced with 1µg/ml Doxycycline for 24 h at 37°C. HMGN5 knockdownd was performed as described in 7.2.6.3. RNA isolation was prepared from 3 biological replicates per condition using the NucleoSpin RNA Kit (Macherey-Nagel) as described in 7.2.2.11. Isolated RNA was quantified with the Qubit RNA kit (methods 7.2.2.12). Before library preparation, the samples were subjected to quality control. For this purpose 300ng of isolated RNA were loaded on a 1% TAE-agarose gel electrophoresis and stained with Ethidium bromide (EtBr). RNA quality was estimated by the integrity of 28S and 18S ribosomal RNAs (supplementary Figure 43A). As an additional quality control, cDNA was prepared (7.2.2.13) and 0.5µl cDNA were used for PCR amplification of a 611bp fragment of human beta actin (Supplementary Figure 43B). In general, purified RNA was stored at -80 °C and the cDNA at -20 °C. After quality control, total RNA was processed for high throughput sequencing as indicated in 7.2.7.1.

7.2.6.5 HMGN5 Co-Immunoprecipitation (Co-IP)

Cells were grown to 70% confluence in 15 cm plates in low glucose DMEM without antibiotics. Protein expression was induced with 1µg/ml Doxycycline for 8 h (3 plates for HMGN5 and 3 for GFP control). Cells were washed once with 1X PBS and harvested in 2ml ice-cold PBS using a cell scraper. Subsequently, the cells were lysed using 1ml Co-IP Lysis Buffer containing ***MNase/benzonase** on ice (to hydrolyze genomic DNA) and incubated for 10 min at 37 °C. The reaction was stopped with EDTA in a final concentration of 20mM. Co-IP was performed with 500µl of lysate per each biological replicate. The samples were mixed with 500µl Co-IP Wash Buffer 1 and transferred to a 1.5ml microcentrifuge tube. To each sample 15µl of GFP_Trap magnetic beads (pre-washed with 300µl Co-IP Lysis Buffer) were added and incubated for 1 h at 4 °C with gentle rotation. Samples were placed on a

magnetic rack for 5 min and the supernatant was removed. Beads were washed 3X with 700μl ice-cold Co-IP Wash buffer 1, and 3X with 100μl ice-cold Co-IP Wash Buffer 2. To remove wash solution, the beads were placed on a magnetic rack, incubated for 2-3 min and the supernatant was carefully removed by pipetting. Beads were stored dried at -20 °C until used for mass spectrometry (7.2.6.5.1).

*MNaseBenzonase treatment: 2μl Benzonase were previously added to 6ml Co-IP Lysis Buffer and 1ml buffer per sample was used for lysis. Then 1μl of MNase (Sigma) was added per lysed sample.

7.2.6.5.1 Mass spectrometry of HMGN5 Co-IP

After HMGN5 and GFP control co-IP (methods 7.2.6.5), the independent biological replicates were digested on-bead and subjected to mass spectrometry. A label-free quantitative proteomics was performed by liquid chromatography tandem mass spectrometry (LC-MS/MS) at the Protein Analysis Unit of the Biomedical Center from the Ludwig Maximilian University, in collaboration with the laboratory of Dr. Axel Imhof. The data obtained was processed with the Perseus software to calculate Intensity-based absolute quantification (iBAQ) as a measure of abundance of the identified proteins. The UniProt database was used for protein identification. Protein enrichment in the HMGN5 samples compared to the GFP control was calculated by Log2 fold change (Log2 FC) normalization over the GFP control (log2 FC IBAQ HMGN5-GFP). The significance was evaluated using t-test and limma test, both adjusted for multiple comparisons. Proteins were considered enriched in the HMGN5 samples over GFP only when Log2 fold change was ≥2 and with statistical significance (p-value <0.05) in both, t-test and limma test.

7.2.7 High throughput sequencing

The processing steps of high throughput sequencing datasets from RNA-seq, ChIP-seq and CLIP-seq were performed by Julia Wimmer from the University clinic of the University of Regensburg (UKR). Comparison of HMGN5 ChIP-seq with ENCODE datasets was performed by Uwe Schwartz.

7.2.7.1 RNA-seq

At least 3µg of RNA and 2 biological replicates per condition -HMGN5 knockdown, overexpression and control cells (RNA isolation described in 7.2.2.11)- were sent to the GeneCore Facility (EMBL Heidelberg) for library preparation and high throughput sequencing. The sequencing was performed on an Illumina Hiseq2000 platform with a read length of 50bp paired-end. Raw data was delivered as compressed FASTQ files.

Prior to mapping to the reference genome, the quality control of the reads was performed using the software FastQC. All pre-processing steps were performed with the software packages SAMtools or BEDtools. The sequencing reads were mapped to the hg19 human reference genome using the STAR aligner with standard settings. The DEseq2 package from R (v3.1.2) was used to perform a differential gene expression analysis of HMGN5 overexpression and knockdown over the non-treated control using standard settings.

The RNA-seq was further validated by qPCR using four genes that were differentially expressed, using the method described in 7.2.2.5.

Gene Ontology Enrichment Analysis was performed using the ClueGo plug-in (v2.3.3) from the Cytoscape software (v.3.5.1). The significance of the enrichments was obtained with a hypergeometric test and the p-value correction with the Bonferroni step-down method (p-value <0.005). The selected leading terms correspond to the most significant term in each group.

When corresponding, plots were performed using the Sigmaplot software (v12.5).

Visualization of grouped datasets was performed with the BioVenn tool (available online at http://www.biovenn.nl/index.php).

7.2.7.2 ChIP-seq

To perform high throughput sequencing of the HMGN5 and GFP ChIP samples (7.2.6.1), the NEBNext® ChIP-Seq Library kit was used to prepare the libraries. Quality of the libraries was analyzed with the High Sensitivity DNA kit (Agilent) on a Bioanalyzer. Deep sequencing was performed at the KFB Regensburg, on an Illumina HiSeq 1000 platform, using 50bp single-end sequencing. Quality control was performed with the FastQC software (v0.10.1). As the library adapters were overrepresented (~10%), a filtering step to remove reads with adapter sequences was done with Homer tools (v4.7). The sequences were aligned to the hg19 reference genome using Bowtie2 (v2.2.4) with standard settings, keeping only uniquely mapped reads. Peak calling was done using the 'findPeaks' tool of the Homer software, and bed files were created for UCSC genome browser analysis using the 'makeUCSCfile' command.

As the ChIP enrichments covered broad regions, peak finding was performed appling the "histone style" parameter, including 7 different peak sizes (200, 250, 300, 350, 400, 450 and 500bp). The commandlines are described in "merge2" strategy (Appendix 9.3.1).

Peak finding was performed twice using either Hneg1 or Hneg2 (non-induced controls) as background for each individual Hpos sample (HMGN5 induced: Hpos1, Hpos2). The resulting peaks were combined in one peak file using the 'mergePeaks' command.

Subsequently, the peaks of Hpos1 and Hpos2 were compared and only the peaks present in both samples were kept (common peaks with literal overlap).

Finally gene (GO) and genome ontology analyses were performed. Peaks were annotated using the 'annotatePeaks.pl' command of Homer, using the hg19 reference genome and default settings. The output file strategy2_merge2.ann.txt was generated.

The motif enrichment analysis was done on the strategy2_merge2.ann.txt peak file with the 'findMotifsGenome.pl' command on hg19 genome assuming a peak size distribution of 1000bp (see above on peak finding description). Motif search was repeated using a 200bp peak size distribution (smallest peak finding size) as control. This parameter is the standard option in the Homer software (transcription factor

peak distribution) based on the assumption that most of the motifs are found +/- 50-75 bp from the peak center (see description of "findMotifsGenome.pl" command from the Homer software).

Gene ontology enrichment analysis was performed using the ClueGo plug-in (v2.3.3) from the Cytoscape software (v.3.5.1). The significance of the enrichments was performed with a hypergeometric test and the p-value correction with the Bonferroni step-down method (p-value <0.005). The selected leading terms correspond to the most significant term in the group (Bindea et al., 2009).

When corresponding, plots were performed using the Sigmaplot software (v12.5). Visualization of grouped datasets was performed with the BioVenn tool.

7.2.7.3 CLIP-seq

After isolation of immunoprecipitated RNA by CLIP (section 7.2.6.2), stranded libraries were prepared using the Ovation Universal RNA-seq kit (Nugen) according to the manufacturer's instructions. Library size distribution was analyzed using a High Sensitivity DNA kit (Agilent) and run in the Bioanalyzer device (See supplementary Figure 44). Each library was prepared with a specific barcode for multiplex sequencing (Table 8) and they were sent to the KFB Regensburg for high throughput sequencing. The samples were quantified by kappa PCR and then combined in one lane, and 50bp single-end sequencing was performed on an Illumina HiSeq 1000 platform.

Table 8. Barcodes for CLIP-seq

Sample	Ligation adaptor mix	Barcode sequence
HMGN5 1	L2V12DR-BC1	AACCAG
HMGN5 2	L2V12DR-BC2	TGGTGA
HMGN5 3	L2V12DR-BC3	AGTGAG
GFP 1	L2V12DR-BC4	GCACTA
GFP 2	L2V12DR-BC5	ACCTCA
GFP 3	L2V12DR-BC6	GTGCTT

For each biological replicate of HMGN5 and GFP samples, the specific name of Adaptor Ligation mix from the Ovation universal RNA-seq kit (Ligation adaptor mix) and the associated barcode sequence are indicated.

The command lines to process the CLIP-seq data are detailed in the Appendix section 9.3.2.

FASTQ files containing the raw data were subjected to quality control analysis using the FastQC software. As there was less than 10% adapter contamination, the trimming step was ommited. The reads were mapped to the hg38 human reference genome using the Bowtie software (v2.1.1).

The pre-build bowtie indexes for hg38 genome (w/o alternate haplotypes) were obtained from ftp://ftp.ccb.jhu.edu/pub/data/bowtie_indexes/GRCh38_no_alt.zip.

The alignments were exported as BAM files for visualization and indexes were created using the BEDtools utilities (v2.24.0).

To identify the enriched regions a peak finding was performed using the 'findPeaks' command of Homer (v4.9), for each of the three HMGN5 samples and additionally a combined alignment of all three HMGN5 samples. As background we used combined alignments of the GFP2 and GFP3 experiments. For the analysis only uniquely mapping reads were selected.

To obtain the enriched RNA bound to HMGN5, a differential gene expression analysis (DGE) was done using the edgeR software package from R (v3.1.2). The transcription database for DGE analysis was obtained from the annotation package "TxDb.Hsapiens.UCSC.hg38.knownGene" from Bioconductor (v3.0). Read count was generated according to the settings described in

http://htseq.readthedocs.io/en/master/count.html. The RNA reads were normalized to counts per million (CPM) to control the differences in library sizes. Three settings for filtering were tested, using 5, 10 and 20 CPM as threshold. The final analysis was done using a threshold of 5 CPM (A gene is considered to be expressed in a library - thus informative- when it has a count of at least 5-10 CPM). Read counting was performed in two different ways, by counting exons/introns obtained from genes, or obtained from transcripts and the resulting data was combined.

Finally we searched for overrepresented motifs in the HMGN5-bound RNAs using the 'findMotifsGenome.pl' command of Homer (v4.9). Motif finding was done with the merged common HMGN5 peaks over GFP2+3 using the hg38 genome. 'mergePeaks' and 'findMotifsGenome.pl' were used to obtain common peaks with literal overlap. As control, the peak search was repeated determining peaks with a distance of 50bp and 100pb away from the peak center. Peaks were annotated using the 'annotatePeaks.pl' command for the hg38, using default setting. Additionally, gene ontology and genome ontology analysis were performed with Homer.

Gene ontology enrichment analysis was analyzed using the ClueGo plug-in (v2.3.3) of the Cytoscape software (v.3.5.1). The significance of the enrichments was performed with a hypergeometric test and the p-value correction with the Bonferroni step-down method (p-value <0.005). The selected leading terms correspond to the most significant term in the group. When corresponding, the plots were performed using the Sigmaplot software (v12.5).

8 References

AbouHaidar, M. G., & Ivanov, I. G. (1999). Non-enzymatic RNA hydrolysis promoted by the combined catalytic activity of buffers and magnesium ions. *Zeitschrift Fur Naturforschung. C, Journal of Biosciences*, *54*(7-8), 542–548.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., & Watson, J. D. (1994). Molecular Biology of the Cell.

Alfano, C., Sanfelice, D., Babon, J., Kelly, G., Jacks, A., Curry, S., & Conte, M. R. (2004). Structural analysis of cooperative RNA binding by the La motif and central RRM domain of human La protein. *Nature Structural & Molecular Biology*, *11*(4), 323–329.

Allan, J., Hartman, P. G., Crane-Robinson, C., & Aviles, F. X. (1980). The structure of histone H1 and its location in chromatin. *Nature*, *288*(5792), 675–679.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W., & LIPMAN, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Andrews, S. J., & Rothnagel, J. A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. *Nature Publishing Group*, *15*(3), 193–204.

Armstrong, L. (2013). Epigenetics. Garland Science.

Ball, M. P., Li, J. B., Gao, Y., Lee, J.-H., LeProust, E. M., Park, I.-H., et al. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature Biotechnology*, *27*(4), 361–368.

Baltz, A. G., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., et al. (2012). The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Molecular Cell*, *46*(5), 674–690

Bartkuhn, M., & Renkawitz, R. (2008). Long range chromatin interactions involved in gene regulation. *Biochimica Et Biophysica Acta*, *1783*(11), 2161–2166.

Baubec, T., Ivanek, R., Lienert, F., & Schuebeler, D. (2013). Methylation-Dependent and -Independent Genomic Targeting Principles of the MBD Protein Family. *Cell*, *153*(2), 480–492.

Bednar, J., Garcia-Saez, I., Boopathi, R., Cutter, A. R., Papai, G., Reymer, A., et al. (2017). Structure and Dynamics of a 197 bp Nucleosome in Complex with Linker Histone H1. *Molecular Cell*, *66*(3), 384–397.e8.

Bell, A. C., West, A. G., & Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, *98*(3), 387–396.

References

Bell, A. J., Chauhan, S., Woodson, S. A., & Kallenbach, N. R. (2008). Biochemical and Biophysical Research Communications. *Biochemical and Biophysical Research Communications*, *377*(1), 262–267.

Belmont, A. S., & Bruce, K. (1994). Visualization of G1 chromosomes: a folded, twisted, supercoiled chromonema model of interphase chromatid structure. *Journal of Cell Biology*, *127*(2), 287–302.

Bernstein, E., & Hake, S. B. (2006). The nucleosome: a little variation goes a long way. *Biochemistry and Cell Biology*, *84*(4), 505–517.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, *25*(8), 1091–1093.

Biterge, B., & Schneider, R. (2014). Histone variants: key players of chromatin. *Cell and Tissue Research*, *356*(3), 457–466.

Bodnar, M. S., & Spector, D. L. (2013). Chromatin meets its organizers. *Cell*, *153*(6), 1187–1189. http://doi.org/10.1016/j.cell.2013.05.030

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., et al. (2005). Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. *PLoS Biology*, *3*(5), e157.

Bonaldi, T., Längst, G., Strohner, R., Becker, P. B., & Bianchi, M. E. (2002). The DNA chaperone HMGB1 facilitates ACF/CHRAC-dependent nucleosome sliding. *EMBO Journal*, *21*(24), 6865–6873.

Bouchet-Marquis, C., Dubochet, J., & Fakan, S. (2006). Cryoelectron microscopy of vitrified sections: a new challenge for the analysis of functional nuclear architecture. *Histochemistry and Cell Biology*, *125*(1-2), 43–51.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., et al. (2008). High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*, *132*(2), 311–322.

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, *72*, 248–254.

Brown, J. D., Mitchell, S. E., & Neill, R. J. O. A. (2011). Making a long story short: noncoding RNAs and chromosome change, *108*(1), 42–49.

Burcin, M., Arnold, R., Lutz, M., Kaiser, B., Runge, D., Lottspeich, F., et al. (1997). Negative protein 1, which is required for function of the chicken lysozyme gene silencer in conjunction with hormone receptors, is identical to the multivalent zinc finger repressor CTCF. *Molecular and Cellular Biology*, *17*(3), 1281–1288.

Bustin, M. (1999). Regulation of DNA-dependent activities by the functional motifs of the high-mobility-group chromosomal proteins. *Molecular and Cellular Biology*, *19*(8), 5237–5246.

References

Bustin, M. (2001). Chromatin unfolding and activation by HMGN(*) chromosomal proteins. *Trends in Biochemical Sciences*, *26*(7), 431–437.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B. M., Strein, C., et al. (2012). Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell*, *149*(6), 1393–1406.

Castello, A., Fischer, B., Frese, C. K., Horos, R., Alleaume, A.-M., Foehr, S., et al. (2016). Comprehensive Identification of RNA-Binding Domains in Human Cells. *Molecular Cell*, *63*(4), 696–710.

Castle, C. D., Cassimere, E. K., & Denicourt, C. (2012). LAS1L interacts with the mammalian Rix1 complex to regulate ribosome biogenesis. *Molecular Biology of the Cell*, *23*(4), 716–728.

Caudron-Herger, M., Müller-Ott, K., Mallm, J.-P., Marth, C., Schmidt, U., Fejes-Tóth, K., & Rippe, K. (2011). Coding RNAs with a non-coding function: Maintenance of open chromatin structure. *Nucleus*, *2*(5), 410–424.

Chen, P., Wang, X.-L., Ma, Z.-S., Xu, Z., Jia, B., Ren, J., et al. (2012). Knockdown of HMGN5 expression by RNA interference induces cell cycle arrest in human lung cancer cells. *Asian Pacific Journal of Cancer Prevention : APJCP*, *13*(7), 3223–3228.

Cho, W.-K., Jayanth, N., English, B. P., Inoue, T., Andrews, J. O., Conway, W., et al. (2016). RNA Polymerase II cluster dynamics predict mRNA output in living cells. *eLife*, *5*.

Chomczynski, P., & Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical Biochemistry*, *162*(1), 156–159.

Ciavatta, D., Rogers, S., & Magnuson, T. (2007). Drosophila CTCF is required for Fab-8 enhancer blocking activity in S2 cells. *Journal of Molecular Biology*, *373*(2), 233–239.

Clapier, C. R., & Cairns, B. R. (2009). The Biology of Chromatin Remodeling Complexes. *Annual Review of Biochemistry*, *78*, 273–304.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., et al. (2007). Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols*, *2*(10), 2366–2382.

Cuddapah, S., Jothi, R., Schones, D. E., Roh, T. Y., Cui, K., & Zhao, K. (2008). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Research*, *19*(1), 24–32.

Darnell, R. B. (2010). HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdisciplinary Reviews - RNA*, *1*(2), 266–286.

Davidovich, C., Zheng, L., Goodrich, K. J., & Cech, T. R. (2013). Promiscuous RNA

binding by Polycomb repressive complex 2. *Nature Structural & Molecular Biology*, *20*(11), 1250–1257.

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*, *46*(D1), D794–D801.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., et al. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21.

Dorigo, B., Schalch, T., Kulangara, A., Duda, S., Schroeder, R. R., & Richmond, T. J. (2004). Nucleosome arrays reveal the two-start organization of the chromatin fiber. *Science*, *306*(5701), 1571–1573.

Downs, J. A., Nussenzweig, M. C., & Nussenzweig, A. (2007). Chromatin dynamics and the preservation of genetic information. *Nature*, *447*(7147), 951–958.

Dunnett, C. W. (1955). A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association*, *50*(272), 1096–1121.

Eltsov M., Maclellan K. M., Maeshima K., Frangakis A. S., Dubochet J. (2008). Analysis of cryo-electron microscopy images does not support the existence of 30-nm chromatin fibers in mitotic chromosomes in situ. Proc. Natl. Acad. Sci. U.S.A. 105, 19732–19737.

Erdel, F., Krug, J., Längst, G., & Rippe, K. (2011). Biochimica et Biophysica Acta. *BBA - Gene Regulatory Mechanisms*, 1–12.

Ernst, O., & Zor, T. (2010). Linearization of the Bradford Protein Assay. *Journal of Visualized Experiments*, (38).

Essien, K., Vigneau, S., Apreleva, S., Singh, L. N., Bartolomei, M. S., & Hannenhalli, S. (2009). CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. *Genome Biology*, *10*(11), R131.

Fan, Y., Nikitina, T., Morin-Kensicki, E. M., Zhao, J., Magnuson, T. R., Woodcock, C. L., & Skoultchi, A. I. (2003). H1 linker histones are essential for mouse development and affect nucleosome spacing in vivo. *Molecular and Cellular Biology*, *23*(13), 4559–4572.

Fanis, P., Gillemans, N., Aghajanirefah, A., Pourfarzad, F., Demmers, J., Esteghamat, F., et al. (2012). Five Friends of Methylated Chromatin Target of Protein-Arginine-Methyltransferase[Prmt]-1 (Chtop), a Complex Linking Arginine Methylation to Desumoylation. *Molecular & Cellular Proteomics : MCP*, *11*(11), 1263–1273.

References

Felsenfeld, G., & Groudine, M. (2003). Controlling the double helix. *Nature*, *421*(6921), 448–453.

Filipescu, D., Mueller, S., & Almouzni, G. (2014). Histone H3 Variants and Their Chaperones During Development and Disease: Contributing to Epigenetic Control. *Annual Review of Cell and Developmental Biology, Vol 30*, *30*, 615–646.

Finch, J. T., & Klug, A. (1976). Solenoidal model for superstructure in chromatin. *Proceedings of the National Academy of Sciences of the United States of America*, *73*(6), 1897–1901.

Finkbeiner, E., Haindl, M., & Müller, S. (2011). The SUMO system controls nucleolar partitioning of a novel mammalian ribosome biogenesis complex. *The EMBO Journal*, *30*(6), 1067–1078.

Fiorito, E., Sharma, Y., Gilfillan, S., Wang, S., Singh, S. K., Satheesh, S. V., et al. (2016). CTCF modulates Estrogen Receptor function through specific chromatin and nuclear matrix interactions. *Nucleic Acids Research*, *44*(22), 10588–10602.

Furusawa, T., Rochman, M., Taher, L., Dimitriadis, E. K., Nagashima, K., Anderson, S., & Bustin, M. (2015). Chromatin decompaction by the nucleosomal binding protein HMGN5 impairs nuclear sturdiness. *Nature Communications*, *6*, 6138.

Fussner, E., Ching, R. W., & Bazett-Jones, D. P. (2011). Living without 30 nm chromatin fibers. *Trends in Biochemical Sciences*, *36*(1), 1–6.

Gautier, T., Abbott, D. W., Molla, A., Verdel, A., Ausió, J., & Dimitrov, S. (2004). Histone variant H2ABbd confers lower stability to the nucleosome. *EMBO Reports*, *5*(7), 715–720.

Gesson, K., Rescheneder, P., Skoruppa, M. P., Haeseler, von, A., Dechat, T., & Foisner, R. (2016). A-type lamins bind both hetero- and euchromatin, the latter being regulated by lamina-associated polypeptide 2 alpha. *Genome Research*, *26*(4), 462–473.

Ghirlando, R., & Felsenfeld, G. (2016). CTCF: making the right connections. *Genes & Development*, *30*(8), 881–891.

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

Gilles Crevel, H. H. A. S. C. (2000). Df31 is a novel nuclear protein involved in chromatin structure in Drosophila melanogaster. *Journal of Cell Science*, *114*(Pt 1), 37–47.

Gonzalez-Sandoval, A., & Gasser, S. M. (2016). On TADs and LADs: Spatial Control Over Gene Expression. *Trends in Genetics*, *32*(8), 485–495.

Gross, D. S., & Garrard, W. T. (1988). Nuclease hypersensitive sites in chromatin. *Annual Review of Biochemistry*, *57*, 159–197.

References

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, *453*(7197), 948–951.

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., & Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, *130*(1), 77–88.

Guo, J. U., Su, Y., Shin, J. H., Shin, J., Li, H., Xie, B., et al. (2014). Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nature Neuroscience*, *17*(2), 215–222.

Han, H., Braunschweig, U., Gonatopoulos-Pournatzis, T., Weatheritt, R. J., Hirsch, C. L., Ha, K. C. H., et al. (2017). Multilayered Control of Alternative Splicing Regulatory Networks by Transcription Factors. *Molecular Cell*, *65*(3), 539–553.e7.

Harada, A., Ohkawa, Y., & Imbalzano, A. N. (2017). Temporal regulation of chromatin during myoblast differentiation. *Seminars in Cell and Developmental Biology*, *72*, 77–86.

Harshman, S. W., Young, N. L., Parthun, M. R., & Freitas, M. A. (2013). H1 histones: current perspectives and challenges. *Nucleic Acids Research*, *41*(21), 9593–9609.

He, Y., Carrillo, J. A., Luo, J., Ding, Y., Tian, F., Davidson, I., & Song, J. (2014). Genome-wide mapping of DNase I hypersensitive sites and association analysis with gene expression in MSB1 cells. *Frontiers in Genetics*, *5*.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, *38*(4), 576–589.

Hepp, M. I., Alarcon, V., Dutta, A., Workman, J. L., & Gutierrez, J. L. (2014). Nucleosome remodeling by the SWI/SNF complex is enhanced by yeast High Mobility Group Box (HMGB) proteins. *BBA - Gene Regulatory Mechanisms*, *1839*(9), 764–772.

Hill, D. A., & Imbalzano, A. N. (2000). Human SWI/SNF nucleosome remodeling activity is partially inhibited by linker histone H1. *Biochemistry (Washington)*, *39*(38), 11649–11656.

HOLOUBEK, V., DEACON, N. J., BUCKLE, D. W., & NAORA, H. (1983). A Small Chromatin-Associated Rna Homologous to Repetitive Dna-Sequences. *European Journal of Biochemistry*, *137*(1-2), 249–256.

HOROWITZ, R. A., AGARD, D. A., SEDAT, J. W., & Woodcock, C. L. (1994). The 3-Dimensional Architecture of Chromatin in-Situ - Electron Tomography Reveals Fibers Composed of a Continuously Variable Zigzag Nucleosomal Ribbon. *Journal of Cell Biology*, *125*(1), 1–10.

References

Huang, R. C., & Huang, P. C. (1969). Effect of protein-bound RNA associated with chick embryo chromatin on template specificity of the chromatin. *Journal of Molecular Biology*, *39*(2), 365–378.

HUANG, R., & BONNER, J. (1965). Histone-Bound Rna a Component of Native Nucleohistone. *Proceedings of the National Academy of Sciences of the United States of America*, *54*(3), 960–967.

Hulsen, T., de Vlieg, J., & Alkema, W. (2008). BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics*, *9*.

Illingworth, R., Kerr, A., Desousa, D., Jørgensen, H., Ellis, P., Stalker, J., et al. (2008). A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biology*, *6*(1), e22.

Jackson, D. A., Iborra, F. J., Manders, E. M., & Cook, P. R. (1998). Numbers and organization of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei. *Molecular Biology of the Cell*, *9*(6), 1523–1536.

Jackson-Grusby, L., Beard, C., Possemato, R., Tudor, M., Fambrough, D., Csankovszki, G., et al. (2001). Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nature Genetics*, *27*(1), 31–39.

Jamin, A., & Wiebe, M. S. (2015). Barrier to Autointegration Factor (BANF1): interwoven roles in nuclear structure, genome integrity, innate immunity, stress responses and progeria. *Current Opinion in Cell Biology*, *34*, 61–68.

Järvelin, A. I., Noerenberg, M., Davis, I., & Castello, A. (2016). The new (dis)order in RNA regulation. *Cell Communication and Signaling*, 1–23.

Jegou, T., Chung, I., Heuvelman, G., Wachsmuth, M., Görisch, S. M., Greulich-Bode, K. M., et al. (2009). Dynamics of telomeres and promyelocytic leukemia nuclear bodies in a telomerase-negative human cell line. *Molecular Biology of the Cell*, *20*(7), 2070–2082.

Jerabek-Willemsen, M., Wienken, C. J., Braun, D., Baaske, P., & Duhr, S. (2011). Molecular Interaction Studies Using Microscale Thermophoresis. *ASSAY and Drug Development Technologies*, *9*(4), 342–353.

Kaiser, V. B., & Semple, C. A. (2017). When TADs go bad: chromatin structure and nuclear organisation in human disease. *F1000Research*, *6*, 314.

Kamakaka, R. T., & Biggins, S. (2005). Histone variants: deviants? *Genes & Development*, *19*(3), 295–310.

Kaneko, S., Son, J., Shen, S. S., Reinberg, D., & Bonasio, R. (2013). PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nature Structural & Molecular Biology*, *20*(11), 1258–1264.

References

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*(6), 996–1006.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., et al. (2017). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*. 4;46(D1):D260-D266

King, L. M., & Francomano, C. A. (2001). Characterization of a Human Gene Encoding Nucleosomal Binding Protein NSBP1. *Genomics*, *71*(2), 163–173.

Klenova, E. M., Nicolas, R. H., Paterson, H. F., Carne, A. F., Heath, C. M., Goodwin, G. H., et al. (1993). CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Molecular and Cellular Biology*, *13*(12), 7612–7624.

Krinner, S., Heitzer, A. P., Diermeier, S. D., Obermeier, I., Längst, G., & Wagner, R. (2014). CpG domains downstream of TSSs promote high levels of gene expression. *Nucleic Acids Research*, *42*(6), 3551–3564.

Kugler, J. E., Deng, T., & Bustin, M. (2012). The HMGN family of chromatin-binding proteins: Dynamic modulators of epigenetic processes. *Biochimica Et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1819*(7), 652–656.

Kugler, J. E., Horsch, M., Huang, D., Furusawa, T., Rochman, M., Garrett, L., et al. (2013). High Mobility Group N Proteins Modulate the Fidelity of the Cellular Transcriptional Profile in a Tissue- and Variant-specific Manner. *Journal of Biological Chemistry*, *288*(23), 16690–16703.

Kumari, P., & Sampath, K. (2015). cncRNAs: Bi-functional RNAs with protein coding and non-coding functions. *Seminars in Cell and Developmental Biology*, *47-48*, 40–51.

Kung, J. T., Kesner, B., An, J. Y., Ahn, J. Y., Cifuentes-Rojas, C., Colognori, D., et al. (2015). Locus-Specific Targeting to the X Chromosome Revealed by the RNA Interactome of CTCF. *Molecular Cell*, *57*(2), 361–375.

Lanctôt, C., Cheutin, T., Cremer, M., Cavalli, G., & Cremer, T. (2007). Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nature Publishing Group*, *8*(2), 104–115.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2.

<div align="center">References</div>

Nature Methods, *9*(4), 357–U54.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3).

Längst, G., & Manelyte, L. (2015). Chromatin Remodelers: From Function to Dysfunction. *Genes*, *6*(2), 299–324.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D., et al. (2013). Global Epigenomic Reconfiguration During Mammalian Brain Development. Science, 341(6146) :1237905.

Liu, Y., Chen, S., Wang, S., Soares, F., Fischer, M., Meng, F., et al. (2017). Transcriptional landscape of the human cell cycle. *Proceedings of the National Academy of Sciences*, *114*(13), 3473–3478.

Lo, K.-Y., Li, Z., Bussiere, C., Bresson, S., Marcotte, E. M., & Johnson, A. W. (2010). Defining the Pathway of Cytoplasmic Maturation of the 60S Ribosomal Subunit. *Molecular Cell*, *39*(2), 196–208.

Lobanenkov, V. V., Nicolas, R. H., Adler, V. V., Paterson, H., Klenova, E. M., Polotskaja, A. V., & Goodwin, G. H. (1990). A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene*, *5*(12), 1743–1753.

Lowary, P. T., & Widom, J. (1998). New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal of Molecular Biology*, *276*(1), 19–42.

Lu, X., Wontakal, S. N., Emelyanov, A. V., Morcillo, P., Konev, A. Y., Fyodorov, D. V., & Skoultchi, A. I. (2009). Linker histone H1 is essential for Drosophila development, the establishment of pericentric heterochromatin, and a normal polytene chromosome structure. *Genes & Development*, *23*(4), 452–465.

Luger, K. (2003). Structure and dynamic behavior of nucleosomes. *Current Opinion in Genetics & Development*, *13*(2), 127–135.

Luger, K., Dechassa, M. L., & Tremethick, D. J. (2012). New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nature Publishing Group*, *13*(7), 436–447.

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, *389*(6648), 251–260.

References

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., et al. (2015). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, *161*(5), 1012–1025.

Maeshima, K., Hihara, S., & Eltsov, M. (2010). Chromatin structure: does the 30-nm fibre exist in vivo? *Current Opinion in Cell Biology*, *22*(3), 291–297.

Maeshima, K., Imai, R., Tamura, S., & Nozaki, T. (2014). Chromatin as dynamic 10-nm fibers. *Chromosoma*, *123*(3), 225–237.

Malicet, C., Rochman, M., Postnikov, Y., & Bustin, M. (2011). Distinct Properties of Human HMGN5 Reveal a Rapidly Evolving but Functionally Conserved Nucleosome Binding Protein. *Molecular and Cellular Biology*, *31*(13), 2742–2755.

Maresca, T. J., Freedman, B. S., & Heald, R. (2005). Histone H1 is essential for mitotic chromosome architecture and segregation in Xenopus laevis egg extracts. *Journal of Cell Biology*, *169*(6), 859–869.

McDowall AW, Smith JM, Dubochet J (1986) Cryo-electron microscopy of vitrified chromosomes in situ. EMBO J 5:1395– 1402

Martic, G. (2005). Parathymosin Affects the Binding of Linker Histone H1 to Nucleosomes and Remodels Chromatin Structure. *Journal of Biological Chemistry*, *280*(16), 16143–16150.

Mattick, J. S., Amaral, P. P., Dinger, M. E., Mercer, T. R., & Mehler, M. F. (2009). RNA regulation of epigenetic processes. *BioEssays*, *31*(1), 51–59.

Meaburn, K. J., & Misteli, T. (2007). Cell biology - Chromosome territories. *Nature*, *445*(7126), 379–381.

Németh, A. N., & Längst, G. (2011). Genome organization in and around the nucleolus. *Trends in Genetics*, *27*(4), 149–156.

Mitrea, D. M., & Kriwacki, R. W. (2013). Regulated unfolding of proteins in signaling. *FEBS Letters*, *587*(8), 1081–1088.

Mollica, L., Bessa, L. M., Hanoulle, X., Jensen, M. R., Blackledge, M., & Schneider, R. (2016). Binding Mechanisms of Intrinsically Disordered Proteins: Theory, Simulation, and Experiment. *Frontiers in Molecular Biosciences*, *3*, 52.

Moretti, F., Rolando, C., Winker, M., Ivanek, R., Rodriguez, J., Kriegsheim, Von, A., et al. (2015). Growth Cone Localization of the mRNA Encoding the Chromatin Regulator HMGN5 Modulates Neurite Outgrowth. *Molecular and Cellular Biology*, *35*(11), 2035–2050.

Nabbi, A., & Riabowol, K. (2015). Rapid Isolation of Nuclei from Cells In Vitro. *Cold Spring Harbor Protocols*, *2015*(8), pdb.prot083733.

Naetar, N., Ferraioli, S., & Foisner, R. (2017). Lamins in the nuclear interior - life

outside the lamina. *Journal of Cell Science*, *130*(13), 2087–2096.

Nam, J.-W., Choi, S.-W., & You, B.-H. (2016). Incredible RNA: Dual Functions of Coding and Noncoding. *Molecules and Cells*, *39*(5), 367–374.

Németh, A., Guibert, S., Tiwari, V. K., Ohlsson, R., & Längst, G. (2008). Epigenetic regulation of TTF-I-mediated promoter–terminator interactions of rRNA genes. *The EMBO Journal*, *27*(8), 1255–1265.

Nishino Y, Eltsov M, Joti Y, Ito K, Takata H, Takahashi Y, Hihara S, Frangakis AS, Imamoto N, Ishikawa T, Maeshima K. (2012). Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure. *EMBO J.* 31, 1644–1653

Nizami, Z., Deryusheva, S., & Gall, J. G. (2010). The Cajal Body and Histone Locus Body. *Cold Spring Harbor Perspectives in Biology*, *2*(7), a000653–a000653.

Nozaki, T., Kaizu, K., Pack, C.-G., Tamura, S., Tani, T., Hihara, S., et al. (2014). Flexible and dynamic nucleosome fiber in living mammalian cells. *Nucleus*, *4*(5), 349–356.

O'Gorman, S., Fox, D. T., & Wahl, G. M. (1991). Recombinase-mediated gene activation and site-specific integration in mammalian cells. *Science*, *251*(4999), 1351–1355.

Okano, M., Bell, D. W., Haber, D. A., & Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, *99*(3), 247–257.

Olins, A. L., & Olins, D. E. (1974). Spheroid chromatin units (ν bodies). *Science*, *183*(4122), 330–332.

Ong, C.-T., & Corces, V. G. (2014). CTCF: an architectural protein bridging genome topology and function. *Nature Publishing Group*, *15*(4), 234–246.

Ou, H. D., Phan, S., Deerinck, T. J., Thor, A., Ellisman, M. H., & O'Shea, C. C. (2017). ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science*, *357*(6349), eaag0025.

Pandya-Jones, A., & Plath, K. (2016). The "lnc" between 3D chromatin structure and X chromosome inactivation. *Seminars in Cell and Developmental Biology*, *56*, 35–47.

Papantonis, A., & Cook, P. R. (2013). Transcription factories: genome organization and gene regulation. *Chemical Reviews*, *113*(11), 8683–8705.

Park, Y.-J., Dyer, P. N., Tremethick, D. J., & Luger, K. (2004). A new fluorescence resonance energy transfer approach demonstrates that the histone variant H2AZ stabilizes the histone octamer within the nucleosome. *The Journal of Biological Chemistry*, *279*(23), 24274–24282.

References

Passarge, E. (1979). Emil Heitz and the concept of heterochromatin: longitudinal chromosome differentiation was recognized fifty years ago. *American journal of human genetics* (Vol. 31, pp. 106–115).

Patil, V. S., Zhou, R., & Rana, T. M. (2013). Gene regulation by non-coding RNAs. *Critical Reviews in Biochemistry and Molecular Biology*, *49*(1), 16–32.

Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Research*, *29*(9), e45.

Pham, T.-H., Minderjahn, J., Schmidl, C., Hoffmeister, H., Schmidhofer, S., Chen, W., et al. (2010). Mechanisms of in vivo binding site selection of the hematopoietic master transcription factor PU.1. *Bioinformatics*, *26*(13), 6391–6402.

Phillips, J. E., & Corces, V. G. (2009). CTCF: master weaver of the genome. *Cell*, *137*(7), 1194–1211.

Pickersgill, H., Kalverda, B., de Wit, E., Talhout, W., Fornerod, M., & van Steensel, B. (2006). Characterization of the Drosophila melanogaster genome at the nuclear lamina. *Nature Genetics*, *38*(9), 1005–1014.

Pogna, E. A., Clayton, A. L., & Mahadevan, L. C. (2010). Signalling to chromatin through post-translational modifications of HMGN. *BBA - Gene Regulatory Mechanisms*, *1799*(1-2), 93–100.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W. J., & Pandolfi, P. P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, *465*(7301), 1033–1038.

Postnikov, Y. V., Trieschmann, L., Rickers, A., & Bustin, M. (1995). Homodimers of chromosomal proteins HMG-14 and HMG-17 in nucleosome cores. *Journal of Molecular Biology*, *252*(4), 423–432.

Postnikov, Y., & Bustin, M. (2010). Regulation of chromatin structure and function By HMGN proteins. *Biochimica Et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1799*(1-2), 62–68.

Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., Rydberg, E. H., Man, O., Beckmann, J. S., et al. (2005). FoldIndex((c)): a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, *21*(16), 3435–3438.

Prokhortchouk, E., & Defossez, P.-A. (2008). The cell biology of DNA methylation in mammals. *Biochimica Et Biophysica Acta (BBA) - Molecular Cell Research*, *1783*(11), 2167–2173.

Prymakowska-Bosak, M., Misteli, T., Herrera, J. E., Shirakawa, H., Birger, Y., Garfield, S., & Bustin, M. (2001). Mitotic phosphorylation prevents the binding of HMGN proteins to chromatin. *Molecular and Cellular Biology*, *21*(15), 5169–5178.

Ramírez, F., Lingg, T., Toscano, S., Lam, K. C., Georgiev, P., Chung, H.-R., et al.

(2015). High-Affinity Sites Form an Interaction Network to Facilitate Spreading of the MSL Complex across the X Chromosome in Drosophila. *Molecular Cell*, *60*(1), 146–162.

Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P., & Jaenisch, R. (2000). Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(10), 5237–5242.

Rattner, B. P., Yusufzai, T., & Kadonaga, J. T. (2009). HMGN proteins act in opposition to ATP-dependent chromatin remodeling factors to restrict nucleosome mobility. *Molecular Cell*, *34*(5), 620–626.

Reeves, R. (2010). Biochimica et Biophysica Acta. *BBA - Gene Regulatory Mechanisms*, *1799*(1-2), 3–14.

Rhodes, D., & Laskey, R. A. (1989). Assembly of nucleosomes and chromatin in vitro. *Methods in Enzymology*, *170*, 575–585.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26.

Robinson, P. J. J., Fairall, L., Van A T Huynh, & Rhodes, D. (2006). EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(17), 6506–6511.

Rochman, M., Postnikov, Y., Correll, S., Malicet, C., Wincovitch, S., Karpova, T. S., et al. (2009). The Interaction of NSBP1/HMGN5 with Nucleosomes in Euchromatin Counteracts Linker Histone-Mediated Chromatin Compaction and Modulates Transcription. *Molecular Cell*, *35*(5), 642–656.

Rochman, M., Taher, L., Kurahashi, T., Cherukuri, S., Uversky, V. N., Landsman, D., et al. (2011). Effects of HMGN variants on the cellular transcription profile. *Nucleic Acids Research*, *39*(10), 4076–4087.

Rodriguez-Campos, A., & Azorin, F. (2007). RNA Is an Integral Component of Chromatin that Contributes to Its Structural Organization. *PLoS ONE*, *2*(11).

Rothbart, S. B., & Strahl, B. D. (2014). Interpreting the language of histone and DNA modifications. *Biochimica Et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1839*(8), 627–643.

Ruiz-Orera, J., Messeguer, X., Subirana, J. A., & Alba, M. M. (2014). Long non-coding RNAs as a source of new peptides. *eLife*, *3*, e03523.

Ryder, S. P. (2016). Protein-mRNA interactome capture: cartography of the mRNP landscape. *F1000Research*, *5*, 2627.

Sampath, K., & Ephrussi, A. (2016). CncRNAs: RNAs with both coding and non-

coding roles in development. *Development (Cambridge, England)*, *143*(8), 1234–1241.

Sarma, K., & Reinberg, D. (2005). Histone variants meet their match. *Nature Reviews. Molecular Cell Biology*, *6*(2), 139–149.

Savage, E. E., Wootten, D., Christopoulos, A., Sexton, P. M., & Furness, S. G. B. (2013). A simple method to generate stable cell lines for the analysis of transient protein-protein interactions. *BioTechniques*, *54*(4), 217–221.

Schalch, T., Duda, S., Sargent, D. F., & Richmond, T. J. (2005). X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature*, *436*(7047), 138–141.

Schmitz, K. M., Mayer, C., Postepska, A., & Grummt, I. (2010). Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes & Development*, *24*(20), 2264–2269.

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, *9*(7), 671–675.

Schubert, T., & Längst, G. (2013). Changes in higher order structures of chromatin by RNP complexes. *RNA Biology*, *10*(2), 175–179.

Schubert, T., Pusch, M. C., Diermeier, S., Benes, V., Kremmer, E., Imhof, A., & Längst, G. (2012). Df31 Protein and snoRNAs MaintainAccessible Higher-Order Structures of Chromatin. *Molecular Cell*, *48*(3), 434–444.

Schumacher, M. A., Miller, M. C., Grkovic, S., Brown, M. H., Skurray, R. A., & Brennan, R. G. (2002). Structural basis for cooperative DNA binding by two dimers of the multidrug-binding protein QacR. *EMBO Journal*, *21*(5), 1210–1218.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504.

Shevtsov, S. P., & Dundr, M. (2011). Nucleation of nuclear bodies by RNA. *Nature Cell Biology*, *13*(2), 167–173.

Shi, Z., Tang, R., Wu, D., & Sun, X. (2015). Research advances in HMGN5 and cancer. *Tumour Biology,* 37(2):1531-9,

Shirakawa, H., Landsman, D., Postnikov, Y. V., & Bustin, M. (2000). NBP-45, a novel nucleosomal binding protein with a tissue-specific and developmentally regulated expression. *The Journal of Biological Chemistry*, *275*(9), 6368–6374.

Shirakawa, H., Rochman, M., Furusawa, T., Kuehn, M. R., Horigome, S., Haketa, K., et al. (2009). The nucleosomal binding protein NSBP1 is highly expressed in the placenta and modulates the expression of differentiation markers in placental Rcho-1 cells. *Journal of Cellular Biochemistry*, *106*(4), 651–658.

References

Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: fromproperties to genome-wide predictions. *Nature Publishing Group*, *15*(4), 272–286.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., et al. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, *479*(7371), 74–79.

Simon, M., North, J. A., Shimko, J. C., Forties, R. A., Ferdinand, M. B., Manohar, M., et al. (2011). Histone fold modifications control nucleosome unwrapping and disassembly. *Proceedings of the National Academy of Sciences*, *108*(31), 12711–12716.

Simpson, R. T. (1978). Structure of the chromatosome, a chromatin particle containing 160 base pairs of DNA and all the histones. *Biochemistry (Washington)*, *17*(25), 5524–5531.

Staby, L., O'Shea, C., Willemoës, M., Theisen, F., Kragelund, B. B., & Skriver, K. (2017). Eukaryotic transcription factors: paradigms of protein intrinsic disorder. *The Biochemical Journal*, *474*(15), 2509–2532.

Strahl, B. D., & Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, *403*(6765), 41–45.

Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F., & de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell*, *10*(6), 1453–1465.

Torres, I. O., & Fujimori, D. G. (2015). Functional coupling between writers, erasers and readers of histone and DNA methylation. *Current Opinion in Structural Biology*, *35*, 68–75.

Ule, J., Jensen, K., Mele, A., & Darnell, R. B. (2005). CLIP: A method for identifying protein–RNA interaction sites in living cells. *Methods (San Diego, Calif.)*, *37*(4), 376–386.

Veitia, R. A., & Potier, M. C. (2015). Gene dosage imbalances: action, reaction, and models. *Trends in Biochemical Sciences*, *40*(6), 309–317.

Vostrov, A. A., & Quitschke, W. W. (1997). The zinc finger protein CTCF binds to the APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation. *The Journal of Biological Chemistry*, *272*(52), 33353–33359.

Wang, J., Lawry, S. T., Cohen, A. L., & Jia, S. (2014a). Chromosome boundary elements and regulation of heterochromatin spreading. *Cellular and Molecular Life Sciences : CMLS*, *71*(24), 4841–4852.

Wang, P., Zhou, Z., Hu, A., Ponte de Albuquerque, C., Zhou, Y., Hong, L., et al. (2014b). Both decreased and increased SRPK1 levels promote cancer by interfering with PHLPP-mediated dephosphorylation of Akt. *Molecular Cell*, *54*(3),

References

378–391.

Wendt, K. S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., et al. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*, *451*(7180), 796–801.

Widom, J. (1998). Structure, dynamics, and function of chromatin in vitro. *Annual Review of Biophysics and Biomolecular Structure*, *27*, 285–327.

Woo, H., Ha, S. D., Lee, S. B., Buratowski, S., & Kim, T. (2017). Modulation of gene expression dynamics by co-transcriptional histone methylations, *49*(4), e326–9.

Wood, A. M., Garza-Gongora, A. G., & Kosak, S. T. (2014). Biochimica et Biophysica Acta. *BBA - Gene Regulatory Mechanisms*, *1839*(3), 178–190.

Woodcock, C. L., Safer, J. P., & Stanchfield, J. E. (1976). Structural repeating units in chromatin. I. Evidence for their general occurrence. *Experimental Cell Research*, *97*, 101–110.

WORCEL, A., STROGATZ, S., & RILEY, D. (1981). Structure of Chromatin and the Linking Number of Dna. *Proceedings of the National Academy of Sciences of the United States of America*, *78*(3), 1461–1465.

Xu, Q., & Xie, W. (2017). Epigenome in Early Mammalian Development: Inheritance, Reprogramming and Establishment. *Trends in Cell Biology*.

Yang, C., Gao, R., Wang, J., Yuan, W., Wang, C., & Zhou, X. (2014). High-mobility group nucleosome-binding domain 5 increases drug resistance in osteosarcoma through upregulating autophagy. *Tumour Biology*, *35*(7), 6357–6363.

Yao, H., Brick, K., Evrard, Y., Xiao, T., Camerini-Otero, R. D., & Felsenfeld, G. (2010). Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. *Genes & Development*, *24*(22), 2543–2555.

Zhang, S., Schones, D. E., Malicet, C., Rochman, M., Zhou, M., Foisner, R., & Bustin, M. (2013). High Mobility Group Protein N5 (HMGN5) and Lamina-associated Polypeptide 2 (LAP2) Interact and Reciprocally Affect Their Genome-wide Chromatin Organization. *Journal of Biological Chemistry*, *288*(25), 18104–18109.

Zhao, H., Sifakis, E. G., Sumida, N., Millán-Ariño, L., Scholz, B. A., Svensson, J. P., et al. (2015). PARP1- and CTCF-Mediated Interactions between Active and Repressed Chromatin at the Lamina Promote Oscillating Transcription. *Molecular Cell*, *59*(6), 984–997.

Zheng, L. (2004). An efficient one-step site-directed and site-saturation mutagenesis protocol. *Nucleic Acids Research*, *32*(14), e115–e115.

Zhou, V. W., Goren, A., & Bernstein, B. E. (2010). Charting histone modificationsand the functional organization ofmammalian genomes. *Nature Publishing Group*,

12(1), 7–18.

9 Appendix

9.1 Supplementary Figures
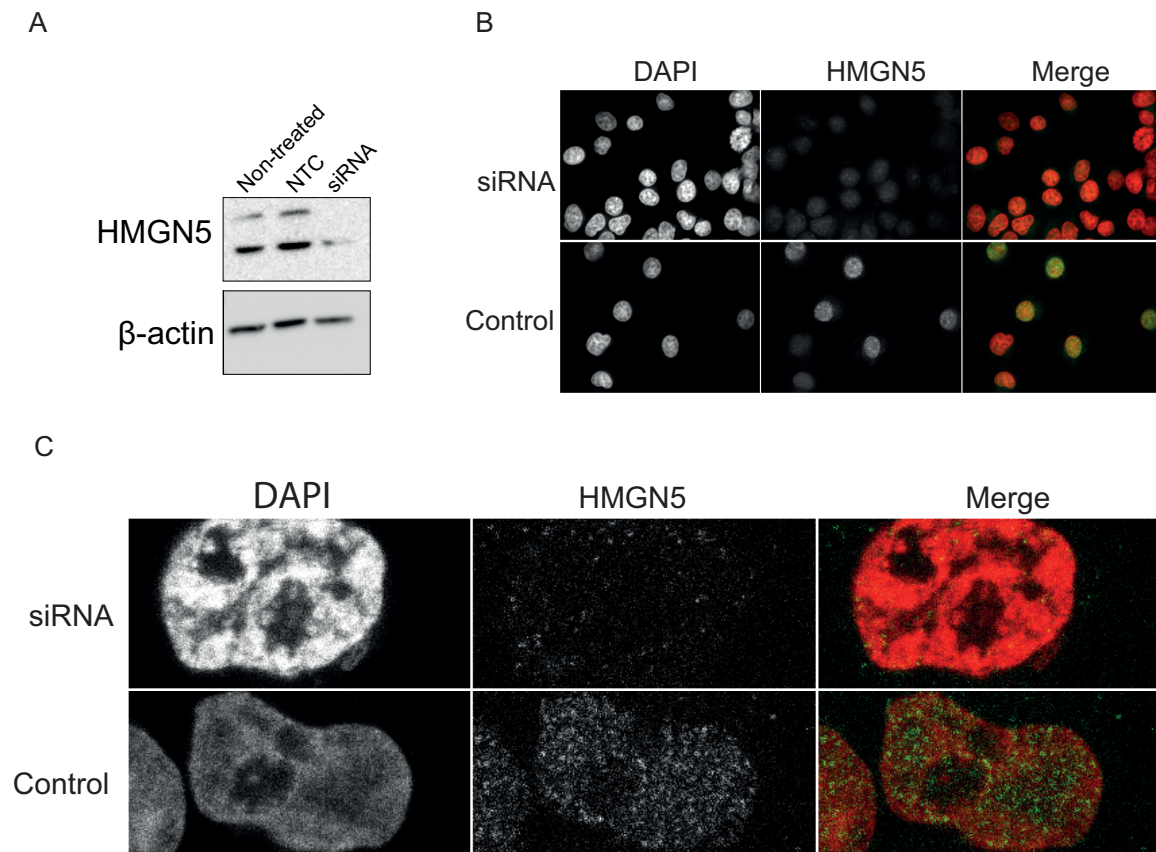
Figure 42. Standardization of HMGN5 knockdown.

HEK293T cells were transfected with 40nM (final concentration) of a smartpool siRNA against hHMGN5 or 40nM of a non-targeting control smartpool siRNA (NTC). A no-treatment control was prepared in parallel. A) Knockdown analyzed by western blot. After 24 h of incubation the cells were lysed and 20µg of total protein extract were loaded on 12% SDS-PAGE and transferred to a PDVF membrane for western blot detection. HMGN5 was detected with the anti-HGMN5 antibody HPA000511 (Sigma) at a dilution of 1:2500. β-actin was used as a loading control, detected with the anti-actin antibody A2066 (Sigma) at a 1:5000 dilution. B) Immunofluorescence detection of HMGN5 in HEK293T cells. After 24 h of knockdown the cells were fixed in 4% PFA and subjected to immunostaining using the anti HMGN5 antibody HPA000511 at a dilution of 1:500. DNA was counterstained with DAPI. In the merged images the DNA is shown in red and the HMGN5 signal in green. The immunofluorescences were visualized with the Fluorescent Microscope Axiovert 200M. C) The same cells shown in (B) were visualized by confocal mocriscopy using a Leica SP8 device.

197

A

B
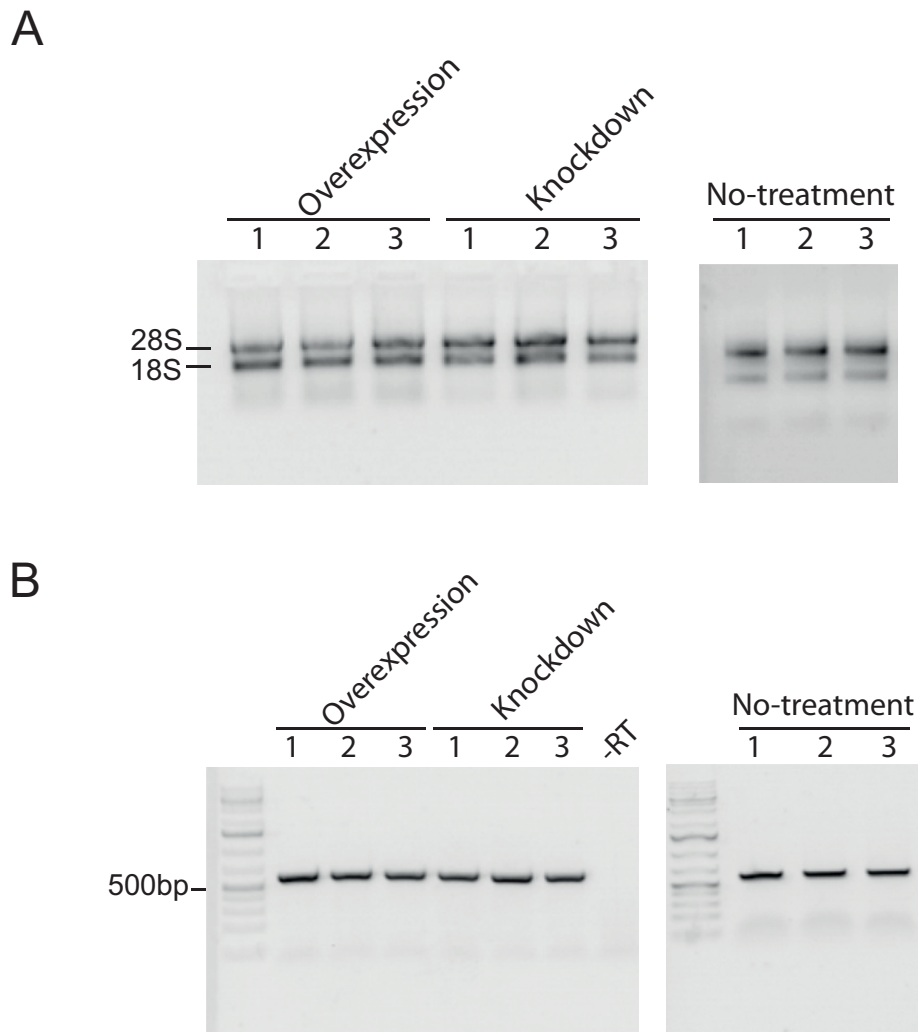
Figure 43. Total RNA quality control.

A) Total RNA extracted from each biological sample after overexpression, knockdown and no-treatment control of HMGN5-FlpIn cell lines. 300ng of extracted RNA was loaded on a 1% Native TAE-agarose gel and separated by electrophoresis. The RNA was stained with BrEt and visualized with UV. The 28S and 18S ribosomal RNA are indicated. B) cDNA quality control. 500ng of each RNA samples were used for cDNA synthesis, and 0.5µl cDNA were used for PCR amplification of a 611bp fragment of human beta actin in every biological sample of the specific treatments. An –RT (only RNA) control was used to estimate genomic DNA contamination. 1Kb DNA ruler was used as molecular marker reference.
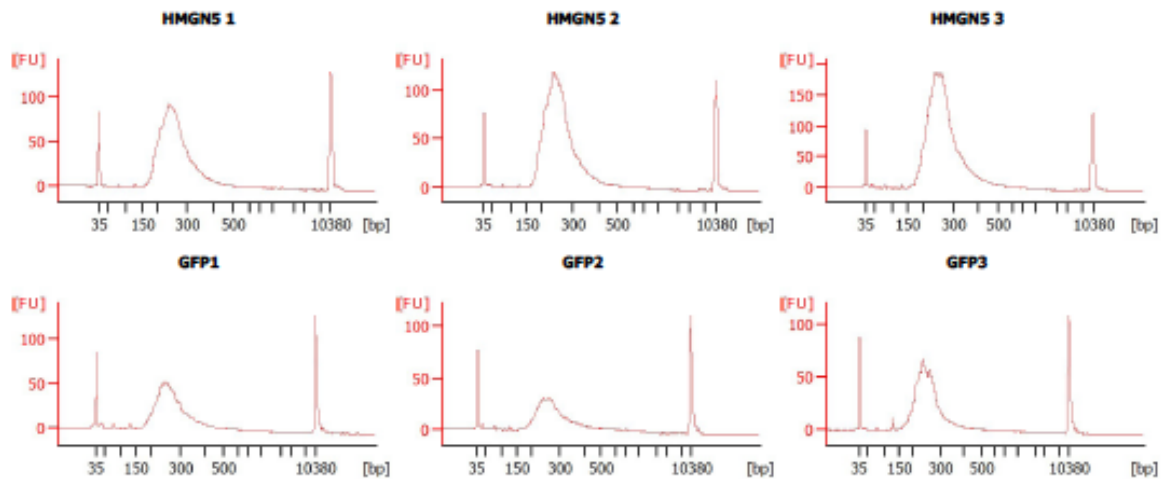
Figure 44. Electopherogram of the libraries for CLIP-seq run with the High sensitivity DNA chip.

Isolated RNAs from each CLIP replicate (3 HMGN5 and 3 GFP samples) were used to prepare stranded-libraries with the Ovation Universal RNA-seq kit from Nugen. 1 µl of each library was loaded on a High sensitivity DNA chip and analyzed on a Bioanalyzer instrument. The peaks at 35 and 10380bp correspond to the lower and upper DNA markers respectively. Each individual library presents a peak size distribution around 250bp.

Figure 45. *In vitro* **mononucleosomes assembly.**

Mononucleosomes were reconstituted using an AvaI digested Cy3 labeled PCR fragment containing the 601 sequence, and histones purified from chicken blood. The assembly was performed using a histones:DNA ratio of 1.4:1 and 2:1. Assembled nucleosomes were loaded on a 6% Native PAGE and visualized with the Thyphoon 9500 device.
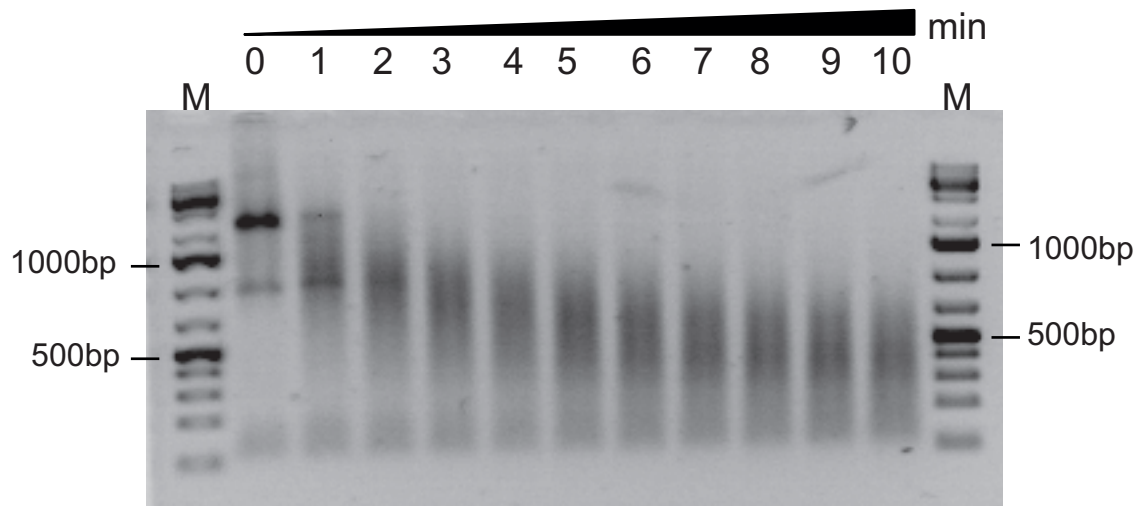
Figure 46. MgCl2-dependent RNA fragmentation.

Total RNA extracted from HeLa cells was dissolved at 500ng/µl in EX0 buffer (containing 1.5 mM $MgCl_2$) and was subjected to a time course fragmentation. Samples were incubated at 95°C. from 0 to 10 min After incubation, per lane 1µg of RNA was loaded on a 1% agarose gel and separated by electrophoresis. The gel was stained with Ethidum bromide and visualized in an UV transiluminator. 1kb GeneRuler™1kb Plus DNA Ladder was used as molecular weight reference (M). The 1000bp and 500bp bands of the marker are indicated.
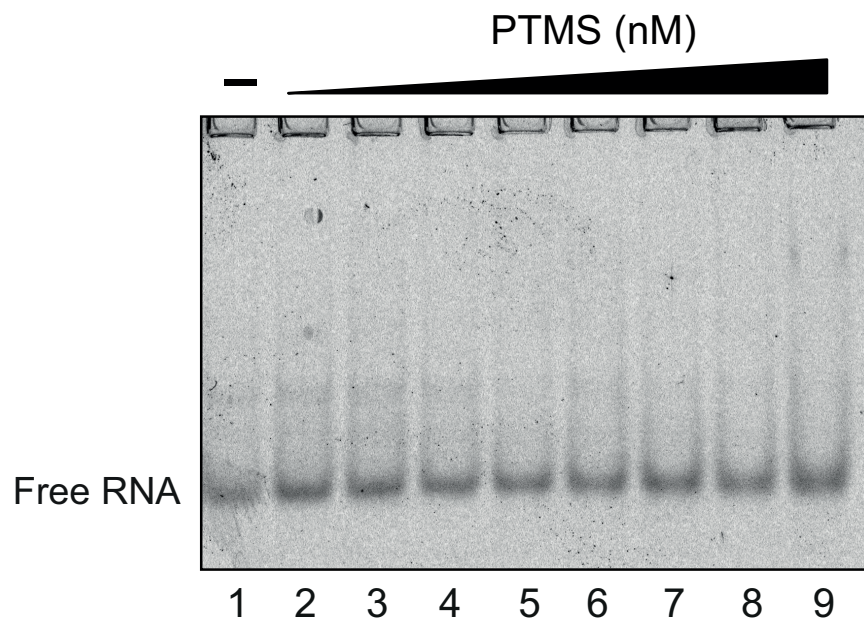
Figure 47. Interaction of PTMS with RNA.

A recombinant his-tagged human parathymosin (PTMS) was used for RNA interactions using EMSA experiments. The fluorescently labeled snoRNA2T2 was kept at a constant concentration of 50nM and a 2:1 serial dilution of the protein was used starting with 2000nM as highest concentration. A negative control without protein was performed (lane 1). The interactions were loaded on a 6% native PAGE and documented with a Typhoon FLA 9500 (GE Healthcare).

9.2 Supplementary tables

Table 9. Identified HMGN5-associated exons

Tx_id	Gene	logFC	logCPM	p-value	FDR
91266	HMGN5	-4.5199134	8.966831944	6.73E-19	1.92E-14
62341	CEBPE	3.93205741	5.935729426	5.45E-16	7.79E-12
31974	AARS2	-6.666727435	4.003751002	3.65E-10	3.48E-06
59363	RRP36	-10.3608153	3.473334418	1.24E-07	0.000707709
59364	RRP36	-10.3608153	3.473334418	1.24E-07	0.000707709
32747	EPB41L2	-10.28473286	3.398826771	2.00E-07	0.000953513
48327	SEC31B	-10.2232589	3.338682149	4.50E-07	0.001657266
50387	IGHMBP2	-10.22493055	3.340293141	4.64E-07	0.001657266
73789	KIAA0753	-10.20720177	3.323085719	6.19E-07	0.001967874
73788	KIAA0753	-10.16746698	3.284294108	1.08E-06	0.003096795
20448	ADD1	-10.25766138	3.37206897	1.27E-06	0.00322348
43459	TLN1	-10.15798845	3.274855612	1.37E-06	0.00322348
59365	ARHGEF7	-10.16218473	3.278974165	1.53E-06	0.00322348
73790	KIAA0753	-10.13290018	3.250561664	1.58E-06	0.00322348
84788	ATP9A	-10.09323531	3.211655277	1.99E-06	0.003557829
84787	ATP9A	-10.09323534	3.211655277	1.99E-06	0.003557829
78607	PDCD5	-4.031511994	4.269416698	2.15E-06	0.003616876
78174	CC2D1A	-7.901146497	3.430437839	3.09E-06	0.004904088
14893	CUL3	-10.02978445	3.149642084	5.16E-06	0.00704886
14892	CUL3	-10.02978458	3.149642084	5.18E-06	0.00704886
7700	DCAF8	-9.998470469	3.119095594	7.14E-06	0.008167912
3449	METTL13	-9.971481171	3.0929969	7.33E-06	0.008167912
3452	METTL13	-9.971481176	3.0929969	7.34E-06	0.008167912
54925	ASIC1	-9.970167116	3.091660389	7.98E-06	0.008167912
54926	ASIC1	-9.970167118	3.091660389	7.98E-06	0.008167912
54924	ASIC1	-9.970167174	3.091660389	8.00E-06	0.008167912
29482	ABCC10	-9.954154388	3.076345219	1.02E-05	0.008846739
38347	CDCA2	-9.941953488	3.064324407	1.02E-05	0.008846739

38346	CDCA2	-9.941953465	3.064324407	1.02E-05	0.008846739
38345	CDCA2	-9.941953443	3.064324407	1.02E-05	0.008846739
54927	ASIC1	-9.926145208	3.048819717	1.21E-05	0.00972174
86359	CDC45	-9.919433798	3.042275815	1.22E-05	0.00972174
64841	PML	-7.692991943	3.222443584	1.59E-05	0.011640079
64840	PML	-7.69299472	3.222443584	1.59E-05	0.011640079
64839	PML	-7.692997395	3.222443584	1.59E-05	0.011640079
3451	METTL13	-9.889495357	3.013287054	1.63E-05	0.011680737
3450	METTL13	-9.86103196	2.985603946	2.13E-05	0.014861697
75115	DCAKD	-9.844252237	2.96906729	2.47E-05	0.01628555
86358	CDC45	-9.83198955	2.957323856	2.60E-05	0.01628555
86357	CDC45	-9.831989554	2.957323856	2.60E-05	0.01628555
78296	TPM4	-9.860248869	2.984476632	2.63E-05	0.01628555
59366	ARHGEF7	-9.825891433	2.951382351	2.84E-05	0.016739451
50388	IGHMBP2	-9.857010535	2.981249972	2.94E-05	0.016797307
7294	SLC39A1	-5.640989368	3.482744067	3.21E-05	0.017971912
7292	SLC39A1	-5.62828656	3.471041421	3.61E-05	0.019328159
7699	DCAF8	-9.804317271	2.930250661	3.65E-05	0.019328159
78428	CRTC1	-9.808523772	2.934275856	4.31E-05	0.020317771
37084	TRIM4	-9.787641912	2.91409594	4.32E-05	0.020317771
66863	SCAPER	-9.808257867	2.934789942	4.33E-05	0.020317771
66864	SCAPER	-9.808257864	2.934789942	4.33E-05	0.020317771
66862	SCAPER	-9.808257711	2.934789942	4.33E-05	0.020317771
75116	DCAKD	-9.78103501	2.907608006	4.48E-05	0.020503133
75973	MCRIP1	-6.178905944	3.257883937	4.58E-05	0.020503133
75972	MCRIP1	-6.178891633	3.257883937	4.59E-05	0.020503133
85406	BACE2	-7.495704088	3.027084977	5.15E-05	0.022608053
71698	SPECC1	-2.716336112	4.323516314	5.30E-05	0.022608053
71697	SPECC1	-2.716335494	4.323516314	5.30E-05	0.022608053
70897	MBTPS1	-9.742503243	2.870393567	5.50E-05	0.022608053
5630	TRAPPC3	-9.737891512	2.866083448	5.57E-05	0.022608053
71694	SPECC1	-2.686800027	4.337597447	5.61E-05	0.022608053
71695	SPECC1	-2.686799425	4.337597447	5.61E-05	0.022608053
18044	SGO1	-9.727819078	2.856381195	6.52E-05	0.025868325
75790	FBF1	-9.732859964	2.860880471	6.60E-05	0.025868325
86360	CDC45	-9.719659768	2.84844862	6.79E-05	0.026227123

66866	SCAPER	-9.749753071	2.878140554	6.90E-05	0.026313421
52879	SYVN1	-9.711015275	2.83998755	7.33E-05	0.026565937
52880	SYVN1	-9.711015277	2.83998755	7.33E-05	0.026565937
63062	RPS6KL1	-9.720383526	2.848773645	7.39E-05	0.026565937
78609	PDCD5	-3.787592467	3.986420955	7.76E-05	0.026565937
7298	SLC39A1	-5.532997894	3.382789591	7.81E-05	0.026565937
7297	SLC39A1	-5.532997067	3.382789591	7.81E-05	0.026565937
7295	SLC39A1	-5.53297912	3.382789591	7.83E-05	0.026565937
66390	ALDH1A2	-9.718240545	2.847474065	7.98E-05	0.026565937
66392	ALDH1A2	-9.718240468	2.847474065	7.99E-05	0.026565937
66391	ALDH1A2	-9.718240438	2.847474065	7.99E-05	0.026565937
85407	BACE2	-7.442804267	2.973952239	8.09E-05	0.026597203
84965	GMEB2	-9.706515658	2.835477464	8.41E-05	0.027337047
66867	SCAPER	-9.721925995	2.851228435	8.79E-05	0.028236376
54928	ASIC1	-9.688534523	2.818194895	8.92E-05	0.028351876
31853	MOCS1	-9.727425018	2.855496723	0.000102811	0.032265093
7296	SLC39A1	-5.496364407	3.348997119	0.000105086	0.032265093
49838	FAM111B	-9.702449886	2.832328217	0.000105787	0.032265093
49837	FAM111B	-9.702449621	2.832328217	0.000106061	0.032265093
85408	BACE2	-7.398819961	2.929892825	0.00011397	0.032522213
58455	SBNO1	-7.399294461	2.93040712	0.000114974	0.032522213
58457	SBNO1	-7.399297916	2.93040712	0.000115039	0.032522213
58456	SBNO1	-7.39930012	2.93040712	0.000115081	0.032522213
8721	ITPKB	-9.686794227	2.816269053	0.000115715	0.032522213
54027	PRDM10	-9.657084008	2.788060667	0.000116724	0.032522213
54028	PRDM10	-9.657083993	2.788060667	0.000116765	0.032522213
75118	DCAKD	-9.673585531	2.803312674	0.000116928	0.032522213
75117	DCAKD	-9.673585683	2.803312674	0.000117142	0.032522213
69493	METTL26	-9.6532322	2.783871434	0.000119957	0.032688532
27300	FBN2	-2.021098282	5.018327303	0.000120027	0.032688532
75114	DCAKD	-9.668736263	2.798617086	0.000121191	0.032694061
84964	GMEB2	-9.636869587	2.768338499	0.000128964	0.03440545
1363	MAST2	-9.646714006	2.777893021	0.000129941	0.03440545
87834	LARGE1	-9.68244887	2.811655497	0.00013647	0.035802696
17245	GFM1	-9.698583161	2.827999453	0.000139747	0.036276104
48319	SEC31B	-2.281418361	4.586867657	0.000140812	0.036276104

31852	MOCS1	-9.653563792	2.78403644	0.000145279	0.036430819
31851	MOCS1	-9.653563827	2.78403644	0.000145357	0.036430819
48318	SEC31B	-2.267283815	4.574554244	0.000146742	0.036430819
58454	SBNO1	-9.618315851	2.750345106	0.000147411	0.036430819
51938	PLEKHA7	-7.374309822	2.90516	0.000147782	0.036430819
77879	CLEC4GP1	-3.967050588	3.839826948	0.0001566	0.037362005
77878	CLEC4GP1	-3.967049405	3.839826948	0.000156665	0.037362005
77877	CLEC4GP1	-3.967049286	3.839826948	0.000156672	0.037362005
77876	CLEC4GP1	-3.967047246	3.839826948	0.000156786	0.037362005
81444	HNRNPL	-9.623904151	2.755221998	0.00016411	0.038784133
18560	IFRD2	-6.500859328	3.12196974	0.000165506	0.038793541
6463	TMED5	-7.343712711	2.874557967	0.000183447	0.041987491
6462	TMED5	-7.343714306	2.874557967	0.000183492	0.041987491
6461	TMED5	-7.343715901	2.874557967	0.000183537	0.041987491
43369	KIF24	-9.623930227	2.755961617	0.000185408	0.042078798
18559	IFRD2	-6.485583918	3.108310158	0.000187891	0.042306498
67487	WDR90	-9.658550229	2.78855843	0.000197222	0.043601746
11717	RETREG2	-9.729158643	2.856432138	0.000197773	0.043601746
11716	RETREG2	-9.729159046	2.856432138	0.000198217	0.043601746
79138	ZNF221	-9.566886813	2.700756563	0.000208392	0.044543376
79137	ZNF221	-9.566886783	2.700756563	0.000208729	0.044543376
52881	SYVN1	-9.573300526	2.706716256	0.000212459	0.04500361
48016	GRID1	-9.587169575	2.720416632	0.000221991	0.046566128
83425	PIGT	-9.608036752	2.740926827	0.000223494	0.046566128
15426	CAPN7	-9.588152196	2.720510229	0.000224721	0.046566128
20746	TBC1D1	-7.330114125	2.859920377	0.00022653	0.046603351
49927	TMEM132A	-7.345403318	3.047752684	0.00023267	0.046660021
49928	TMEM132A	-7.345402447	3.047752684	0.00023269	0.046660021
79507	AP2A1	-6.418692675	2.915843246	0.000233052	0.046660021
79506	AP2A1	-6.418689901	2.915843246	0.000233333	0.046660021
57438	SARNP	-9.541039705	2.675616291	0.000245386	0.048729595
9015	ACP1	-9.552529483	2.686284877	0.000254666	0.049967445
71026	SPATA2L	-5.98890898	2.975698486	0.000257333	0.049967445
54026	PRDM10	-9.539806772	2.674902931	0.000261272	0.049967445
54025	PRDM10	-9.539806701	2.674902931	0.000261566	0.049967445
54024	PRDM10	-9.539806613	2.674902931	0.000261927	0.049967445

54023	PRDM10	-9.53980657	2.674902931	0.000262104	0.049967445
20637	Unidentified	-3.561796171	4.681635546	3.74E-06	0.005632355
30742	Unidentified	-6.860895713	3.338915124	5.56E-06	0.007227569
28961	Unidentified	-9.961374362	3.083402336	9.69E-06	0.008846739
28960	Unidentified	-9.929488654	3.052354046	1.22E-05	0.00972174
28838	Unidentified	-9.86862523	2.993366863	2.67E-05	0.01628555
28839	Unidentified	-9.868624941	2.993366863	2.68E-05	0.01628555
31459	Unidentified	-9.84642236	2.971374302	2.87E-05	0.016739451
72136	Unidentified	-7.518678577	3.202010083	3.31E-05	0.018204906
30739	Unidentified	-6.640802006	3.127765811	4.14E-05	0.020317771
30740	Unidentified	-6.640801998	3.127765811	4.14E-05	0.020317771
30741	Unidentified	-7.417265061	3.108249451	7.34E-05	0.026565937
28959	Unidentified	-9.666892724	2.798393146	0.000204649	0.044391396
28958	Unidentified	-9.666892507	2.798393146	0.000204912	0.044391396

List of RNAs generated by counting exons obtained from transcripts. The Transcript Id (Tx_id), name of associated gene, log fold change (LogFC), Log normalized counts per million (LogCPM), p-value and adjusted false discovery rate (FDR) are shown in the list. FDR cutoff: 0.05. The negative LogFC values correspond to enrichment of HMGN5 over GFP. Genes for DGE analysis were extracted from TxDb.Hsapiens.UCSC.hg38.knownGene database of hg38 human reference genome.

Table 10. Identified HMGN5-associated introns

Tx_id	Gene name	logFC	logCPM	p-value	FDR
28963	CSNK2B	-10.57284844	2.602688125	2.00E-10	1.67E-06
28962	CSNK2B	-10.32288718	2.359067287	4.61E-09	2.89E-05
70418	TOX3	-7.618927237	2.39595837	4.01E-08	0.00014371
51553	MOB2	-10.12201589	2.163675689	5.84E-08	0.000162901
28960	LY6G5B	-10.06003812	2.104242312	8.85E-08	0.000209141
25675	G3BP1	-5.1776663	2.731225356	1.08E-07	0.000209141
25674	G3BP1	-5.177663497	2.731225356	1.08E-07	0.000209141
25673	G3BP1	-5.177660694	2.731225356	1.08E-07	0.000209141
67488	WDR90	-10.06870594	2.111911219	1.38E-07	0.000246887
28961	LY6G5B	-9.932080302	1.980545806	1.56E-07	0.000246887
4914	ENO1	-4.060905817	2.989744799	1.57E-07	0.000246887
42278	AKAP2	-3.618873362	3.102655176	2.34E-07	0.000317207
42277	AKAP2	-3.61887332	3.102655176	2.34E-07	0.000317207
43457	TLN1	-9.589848183	1.651615196	5.97E-07	0.00065174
1578	NFIA	-7.238785294	1.916972301	5.74E-07	0.00065174
1577	NFIA	-7.238771966	1.916972301	5.72E-07	0.00065174
20638	TAPT1-AS1	-3.222021858	3.680936008	6.59E-07	0.000662033
20639	TAPT1-AS1	-3.222020606	3.680936008	6.59E-07	0.000662033
16428	ST3GAL6	-4.203696894	2.717764913	7.83E-07	0.000755889
4915	ENO1	-3.991224239	2.866444377	8.73E-07	0.000782529
16426	ST3GAL6	-4.173234815	2.688983048	1.25E-06	0.000976947
16427	ST3GAL6	-4.173234784	2.688983048	1.24E-06	0.000976947
16429	ST3GAL6	-4.173234753	2.688983048	1.24E-06	0.000976947
68438	FUS	-7.229492411	2.031812665	1.70E-06	0.001211976
85477	PDE9A	-7.06423973	1.747890562	1.74E-06	0.001211976
16431	ST3GAL6	-4.110383791	2.629681577	2.28E-06	0.001506548
71982	RASL10B	-9.409586198	1.48000323	2.55E-06	0.001640172
73388	BAHCC1	-6.333293588	1.969561588	3.47E-06	0.002122117
16430	ST3GAL6	-4.09189762	2.612267693	3.46E-06	0.002122117
2383	ATP1A1	-9.207676185	1.288596489	5.05E-06	0.002869415
42874	GPSM1	-9.19334333	1.274637329	5.21E-06	0.002869415
73434	RAC3	-6.990524563	1.80474533	5.08E-06	0.002869415

72063	Unidentified	-1.999661976	4.024715531	5.42E-06	0.002869415
21458	GSTCD	-4.019514004	2.574695205	7.13E-06	0.003649981
28178	MAPK9	-3.677815912	2.808176011	7.99E-06	0.003884737
28179	MAPK9	-3.677815014	2.808176011	7.99E-06	0.003884737
28177	MAPK9	-3.6778013	2.808176011	8.05E-06	0.003884737
72572	NFE2L1	-6.227636108	1.759044953	1.00E-05	0.004414283
80063	ZNF587	-6.027533835	1.678877426	9.76E-06	0.004414283
80062	ZNF587	-6.027530192	1.678877426	9.78E-06	0.004414283
51552	MOB2	-9.66244054	1.720311	1.03E-05	0.004419196
32918	Unidentified	-2.307741185	4.076507551	1.22E-05	0.004574152
12169	SH3YL1	-6.013069032	1.556948572	1.37E-05	0.004839221
12165	SH3YL1	-6.008227086	1.55238919	1.39E-05	0.004839221
46812	EDRF1	-5.928230959	2.255867817	1.38E-05	0.004839221
18334	CCDC12	-3.827549312	2.573040691	1.52E-05	0.004960297
51899	RRAS2	-3.133201226	3.079162806	1.51E-05	0.004960297
51900	RRAS2	-3.133198654	3.079162806	1.51E-05	0.004960297
51902	RRAS2	-3.133195835	3.079162806	1.51E-05	0.004960297
65638	Unidentified	-4.269157755	2.513311881	1.71E-05	0.005430562
51609	NAP1L4	-5.966074876	1.621452482	1.96E-05	0.006144728
36079	RAPGEF5	-5.837762476	1.499646664	2.00E-05	0.006203235
33882	DBNL	-5.881068939	1.54115342	2.08E-05	0.006350793
17058	SPSB4	-1.860596418	4.098032492	2.39E-05	0.00715133
1139	SMAP2	-5.292999922	1.895757613	2.55E-05	0.007155936
45045	RBM17	-4.972575825	2.079436449	2.79E-05	0.007155936
1138	SMAP2	-5.105723364	2.055875768	3.15E-05	0.007977015
51033	TTC12	-5.845517077	1.507746274	3.22E-05	0.007999808
18335	CCDC12	-3.7651382	2.514353313	3.20E-05	0.007999808
51032	TTC12	-5.815217117	1.479198321	3.38E-05	0.008256094
51031	TTC12	-5.815212986	1.479198321	3.39E-05	0.008256094
13171	LRRTM4	-1.377708427	5.852014679	3.46E-05	0.008293965
13172	LRRTM4	-1.377708383	5.852014679	3.47E-05	0.008293965
80485	CD320	-5.433449976	1.947411152	3.51E-05	0.008308383
46813	EDRF1	-5.767291619	2.109662116	4.44E-05	0.009960837
46811	EDRF1	-5.767276585	2.109662116	4.45E-05	0.009960837
45044	RBM17	-4.463915503	2.229249849	4.38E-05	0.009960837
80486	CD320	-5.330577109	1.849679784	4.81E-05	0.010587271

4037	SYT14	-1.635851737	4.304743494	4.99E-05	0.010611306
4036	SYT14	-1.635851725	4.304743494	4.99E-05	0.010611306
4033	SYT14	-1.635851716	4.304743494	4.99E-05	0.010611306
4034	SYT14	-1.635851714	4.304743494	4.99E-05	0.010611306
72310	TUBG2	-6.758361706	1.596828218	5.05E-05	0.010641458
4035	SYT14	-1.630021533	4.299967331	5.35E-05	0.011190942
393	VPS13D	-2.434268944	3.325006473	5.83E-05	0.012084784
80484	CD320	-5.244417064	1.768045288	6.33E-05	0.013028851
4038	SYT14	-1.612974292	4.286033027	6.39E-05	0.01303249
29530	SLC29A1	-4.870576526	1.763017338	6.81E-05	0.013142134
29528	SLC29A1	-4.870576082	1.763017338	6.82E-05	0.013142134
29529	SLC29A1	-4.870575882	1.763017338	6.82E-05	0.013142134
26672	PDE4D	-1.819552239	4.546007944	6.52E-05	0.013142134
28180	MAPK9	-3.604547129	2.560134371	7.44E-05	0.013622223
46850	GLRX3	-4.730550724	1.783908083	8.01E-05	0.01457424
31808	MTCH1	-5.27636376	1.451268676	9.82E-05	0.016933501
31810	MTCH1	-5.276362244	1.451268676	9.80E-05	0.016933501
46784	Unidentified	-5.210693438	1.819722321	9.85E-05	0.016933501
58189	ATXN2	-1.349895097	5.293590955	0.000107317	0.017719305
58188	ATXN2	-1.349894868	5.293590955	0.000107213	0.017719305
38905	PKIA	-5.95146042	1.484124771	0.000129135	0.020527733
38906	PKIA	-5.951457889	1.484124771	0.000129097	0.020527733
81345	ZNF461	-3.650259427	2.411623679	0.000142552	0.022084051
81344	ZNF461	-3.650256847	2.411623679	0.000142445	0.022084051
81343	ZNF461	-3.650255074	2.411623679	0.000142371	0.022084051
58187	ATXN2	-1.30784954	5.314083908	0.00015518	0.023747319
58186	ATXN2	-1.307849435	5.314083908	0.000155066	0.023747319
89300	EDA	-1.644938825	5.643596671	0.000159824	0.024168013
89299	EDA	-1.644938688	5.643596671	0.000159855	0.024168013
71240	ALOX12P2	-2.557815117	3.318003849	0.000168563	0.024891552
71241	ALOX12P2	-2.557814662	3.318003849	0.000168608	0.024891552
77413	PIGN	-2.222032016	3.333440086	0.000168507	0.024891552
77412	PIGN	-2.222031837	3.333440086	0.000168479	0.024891552
9823	ZNF638	-5.143363823	1.931245397	0.000174723	0.025201351
51034	TTC12	-5.577986361	1.259082348	0.000179625	0.025607137
25534	KIAA0141	-4.52964464	1.728096885	0.000189826	0.025607137

57018	GXYLT1	-4.558711725	1.821714497	0.000198957	0.02641921
83102	CBFA2T2	-2.818637183	2.721158381	0.000216543	0.028305068
83101	CBFA2T2	-2.818636767	2.721158381	0.00021645	0.028305068
31809	MTCH1	-5.195393297	1.374945988	0.00021807	0.028356963
76088	SMCHD1	-1.90360414	4.047435412	0.00022656	0.029036245
76087	SMCHD1	-1.903602411	4.047435412	0.000226764	0.029036245
65435	ASB7	-5.059992576	1.942727428	0.000239388	0.029890198
57017	GXYLT1	-4.48800068	1.755777926	0.000244026	0.030318391
64623	TPM1	-4.312421181	1.646272933	0.000267261	0.032560488
69730	CDIP1	-4.186341168	1.835627765	0.000271001	0.032703069
6064	NDC1	-2.002650579	3.547331115	0.000279506	0.033403678
44484	PSMB7	-3.25853715	2.766563948	0.000282212	0.033567236
23911	CENPU	-4.247762059	1.859132912	0.000302883	0.035520923
23910	CENPU	-4.247757708	1.859132912	0.000303335	0.035520923
59387	ATP11A	-1.937303161	3.662112336	0.000300971	0.035520923
392	VPS13D	-1.845492069	3.659357642	0.000304299	0.035520923
69732	CDIP1	-4.178310332	1.827698887	0.000306711	0.035636684
88133	CENPM	-2.296072168	3.183440912	0.00031904	0.036898406
6065	NDC1	-1.989589108	3.536378794	0.000331349	0.038146172
75873	DNAH17	-1.308332025	4.982111803	0.000348112	0.039892999
22051	WDR17	-3.623857522	2.581048664	0.000355669	0.040051991
69728	CDIP1	-4.120428292	1.773884539	0.000377255	0.041526156
69729	CDIP1	-4.120424641	1.773884539	0.000376915	0.041526156
69731	CDIP1	-4.120420992	1.773884539	0.000376576	0.041526156
61153	ARID4A	-3.473198428	2.360606864	0.000383244	0.042001185
6063	NDC1	-1.975635227	3.524689148	0.000392629	0.042291036
21459	GSTCD	-2.652168151	3.047205409	0.00039894	0.042787168
40959	PTK2	-2.275105225	3.148395097	0.000421621	0.04502732
26191	SDHAP3	-1.466666921	4.25373813	0.000452308	0.047695729
91504	COL4A6	-4.423758122	1.562553024	0.000463071	0.048423708
91505	COL4A6	-4.423756917	1.562553024	0.000462647	0.048423708

List of RNAs generated by counting introns obtained from transcripts. The Transcript Id (Tx_id), name of associated gene, log fold change (LogFC), Log normalized counts per million (LogCPM), p-value and adjusted false discovery rate (FDR) are shown in the list. FDR cutoff: 0.05. The negative LogFC values correspond to enrichment of HMGN5 over GFP. Genes for DGE analysis were extracted from TxDb.Hsapiens.UCSC.hg38.knownGene database of hg38 human reference genome.

9.3 High throughput sequencing command lines

9.3.1 ChIP-seq analysis command lines

The following script was applied to identify the HMGN5-binding sites in the genome using the HMGN5 Flp-In cell line (background HEK293T). The storage of raw and processed dataset is indicated in the script. ChIP-seq analysis was performed by Dr. Julia Wimmer from the University clinic of the University of Regensburg.

```
##Installed Software:

homer-4.7
hg19 package
bowtie2-2.2.4
blatSuite
FastQC
ghostscript-9.14


#bash_profile augmented with new paths, bowtie2 variables set as
described on bowtie2 website.
#Genome: hg19 obtained from UCSC ("chromFa.tar.gz",
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/ , 2015-02-
06) > stored in
/users/admin/NGSanalysis_JuliaW/GenomeFasta/hg19_unmasked >unzip
with % tar -xfvz chromFa.tar.gz


#Genomes collection for hg19 downloaded from bowtie2 website > *.bt2
files copied to /Users/admin/HTS-Software/bowtie2-2.2.4/index


###ChIP command line:
##raw data
50bp, seq
```

##Mapping to the (unmasked) hg19 reference genome (Illuminaís iGenome collection contained bt2 index files> no need to build bowtie2 genome manually > copied to ../HTS-software/bowtie2_2.2.4/index. Allow multi mapper, homer will later take only uniquely mapped reads (if advised) according to SAM flag. Use standard parameters for alignment, scoring, reporting, and effort. Works with compressed fastq.gz files!

#Usage

#bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r>} -S [<hit>]

#Options
#-q Reads are fastq files

(#ó-phred33   phreed33+ quality scores are used by very latest Illumina pipelines, should be default)
#-p <int>   number of threads (here: 24 cores available)
#-x <bt2-idx>   normally path to genome index files, but path was exported to .bash_profile > ìhg19î info should be sufficient
#-U <string>   path to (unzipped?) fastq file
#-S <string>   path to SAM output (with name!)


```
bowtie2 -q -p 20 -x /Users/admin/HTS-Software/bowtie2-
2.2.4/index/hg19 -U
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Hpos1_ACAGTG_L006
_R1_001.fastq.gz -S
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Hpos1;
bowtie2 -q -p 20 -x /Users/admin/HTS-Software/bowtie2-
2.2.4/index/hg19 -U
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Hpos2_GCCAAT_L006
_R1_001.fastq.gz -S
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Hpos2;
bowtie2 -q -p 20 -x /Users/admin/HTS-Software/bowtie2-
2.2.4/index/hg19 -U
```

```
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Hneg1_CGATGT_L006
_R1_001.fastq.gz -S
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Hneg1;
bowtie2 -q -p 20 -x /Users/admin/HTS-Software/bowtie2-
2.2.4/index/hg19 -U
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Hneg2_TGACCA_L006
_R1_001.fastq.gz -S
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Hneg2;
bowtie2 -q -p 20 -x /Users/admin/HTS-Software/bowtie2-
2.2.4/index/hg19 -U
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Gneg1_CAGATC_L006
_R1_001.fastq.gz -S
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Gneg1;
bowtie2 -q -p 20 -x /Users/admin/HTS-Software/bowtie2-
2.2.4/index/hg19 -U
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Gneg2_CTTGTA_L006
_R1_001.fastq.gz -S
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Gneg2;
bowtie2 -q -p 20 -x /Users/admin/HTS-Software/bowtie2-
2.2.4/index/hg19 -U
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Gpos1_ATCACG_L006
_R1_001.fastq.gz -S
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Gpos1;


##Creating tag directories (Homer)

#Usage

% makeTagDirectory <Output Directory Name> [options] <alignment
file1> [alignment file2] Ö

#Options

#-unique   (default) keep only uniquely aligned reads

#-genome <string> -checkGC This analysis will produce several output
files in addition to the basic quality control analysis
```

214

```
makeTagDirectory
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_hg19 -genome
hg19 -checkGC
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Hpos1;
makeTagDirectory
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_hg19 -genome
hg19 -checkGC
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Hpos2;
makeTagDirectory
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_hg19 -genome
hg19 -checkGC
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Hneg1;
makeTagDirectory
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_hg19 -genome
hg19 -checkGC
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Hneg2;
makeTagDirectory
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Gpos1_hg19 -genome
hg19 -checkGC
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Gpos1;
makeTagDirectory
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Gneg1_hg19 -genome
hg19 -checkGC
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Gneg1;
makeTagDirectory
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Gneg2_hg19 -genome
hg19 -checkGC
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Gneg2;


##Making Genome Browser Files (homer)

#Usage

% makeUCSCfile <tag directory> -o auto
#Options
#-o auto   output file is stored in the corresponding tag directory
```

```
cd /Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/

makeUCSCfile HMGN5_Hpos1_hg19 -o auto;
makeUCSCfile HMGN5_Hpos2_hg19 -o auto;
makeUCSCfile HMGN5_Hneg1_hg19 -o auto;
makeUCSCfile HMGN5_Hneg2_hg19 -o auto;
makeUCSCfile HMGN5_Gpos1_hg19 -o auto;
makeUCSCfile HMGN5_Gneg1_hg19 -o auto;
makeUCSCfile HMGN5_Gneg2_hg19 -o auto;
```

>>> Fastqc results failed for uper base sequence content and persequence GC content > each samples has Illumina adapter sequences overrepresented (~10%), which could cause these failures >>> #use homer tools to remove corresponding adapter sequences and run fastqc again.

###Trimming reads with adapter sequences

##Major adapter sequences

Hpos1 13.5% GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGC
      TruSeq Index 5

Hpos2 8.4% GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGC
      TruSeq Index 6
Hneg1 8.7% GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC
      TruSeq Index 2
Hneg2 7.7% GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGC
      TruSeq Index 4
Gpos1 8.6% GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGC
      TruSeq Index 1
Gneg1 11.0% GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGC
      TruSeq Index 7
Gneg2 10.4% GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGC
      TruSeq Index 12

```
##Usage

% homerTools trim [options] <sequence file1>
#Options
-len <#> (trim sequences to this length)
-min <#> (remove sequence that are shorter than this after timming)
-3 <#> (trim this many bp off the 3' end of the sequence)
-5 <#> (trim this many bp off the 5' end of the sequence)
-3 <ACGT> (trim adapter sequence (i.e. "-3 GGAGGATTT") from the 3'
end of the sequence)
-5 <ACGT> (trim adapter sequence (i.e. "-5 GGAGGATTT") from the 5'
end of the sequence)


cd
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Project_S124_11-
12/Sample_Hpos1/;


homerTools trim -3
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGC -min 50
Hpos1_ACAGTG_L006_R1_001.fastq;


#Trimmed output: 22391698 of 25900301 reads



cd
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Project_S124_11-
12/Sample_Hpos2/;


homerTools trim -3
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGC -min 50
Hpos2_GCCAAT_L006_R1_001.fastq;


#Trimmed output: 28692900 of 31326646 reads
```

Appendix

```
cd
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Project_S124_11-
12/Sample_Hneg1/;

homerTools trim -3
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC -min 50
Hneg1_CGATGT_L006_R1_001.fastq;

#Trimmed output: 28389091 of 31100985 reads

cd
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Project_S124_11-
12/Sample_Hneg2/;

homerTools trim -3
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGC -min 50
Hneg2_TGACCA_L006_R1_001.fastq;

#Trimmed output: 27471492 of 29771949 reads

cd
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Project_S124_11-
12/Sample_Gpos1/;

homerTools trim -3
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGC -min 50
Gpos1_ATCACG_L006_R1_001.fastq;

#Trimmed output: 28224082 of 30880894 reads

cd
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Project_S124_11-
12/Sample_Gneg1/;

homerTools trim -3
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGC -min 50
Gneg1_CAGATC_L006_R1_001.fastq;

#Trimmed output: 33851770 of 38068436 reads
```

```
cd
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Project_S124_11-
12/Sample_Gneg2/;
```

```
homerTools trim -3
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGC -min 50
Gneg2_CTTGTA_L006_R1_001.fastq;
```

```
#Trimmed output: 25439957 of 28410531 reads
```

```
>>> change file name *.fastq.trimmed to *.trimmedAdapt.fastq to
enable analysis with fastqc.
```

```
>> all parameters improved. Still, kmer content, per base sequence
content and per sequence Gc content give warnings/failures >>>
remaining adapter contaminations in reads (but too low to be
reported as overrepresented sequence!
```

```
>Repeat mapping with cleaned read files.
```

```
##Mapping with bowtie2
```

```
old alignment files were deleted!
```

```
#Usage
```

```
% bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r>} -S
[<hit>]
```

```
#Options
```

```
#-q  Reads are fastq files
```

```
(#ó-phred33    phreed33+ quality scores are used by very latest
Illumina pipelines, should be default)
```

```
#-p <int>        number of threads (here: 24 cores available)


#-x <bt2-idx>    normally path to genome index files, but path was
exported to .bash_profile > ìhg19î info should be sufficient


#-U <string>     path to (unzipped?) fastq file


#-S <string>     path to SAM output (with name!)
```

```
bowtie2 -q -p 20 -x /Users/admin/HTS-Software/bowtie2-
2.2.4/index/hg19 -U
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Project_S124_11-
12/Sample_Hpos1/Hpos1_ACAGTG_L006_R1_001.trimmed.fastq -S
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Hpos1_trimme
d;


bowtie2 -q -p 20 -x /Users/admin/HTS-Software/bowtie2-
2.2.4/index/hg19 -U
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Project_S124_11-
12/Sample_Hpos2/Hpos2_GCCAAT_L006_R1_001.trimmed.fastq -S
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Hpos2_trimme
d;


bowtie2 -q -p 20 -x /Users/admin/HTS-Software/bowtie2-
2.2.4/index/hg19 -U
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Project_S124_11-
12/Sample_Hneg1/Hneg1_CGATGT_L006_R1_001.trimmed.fastq -S
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Hneg1_trimme
d;


bowtie2 -q -p 20 -x /Users/admin/HTS-Software/bowtie2-
2.2.4/index/hg19 -U
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Project_S124_11-
12/Sample_Hneg2/Hneg2_TGACCA_L006_R1_001.trimmed.fastq -S
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Hneg2_trimme
d;
```

```
bowtie2 -q -p 20 -x /Users/admin/HTS-Software/bowtie2-
2.2.4/index/hg19 -U
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Project_S124_11-
12/Sample_Gneg1/Gneg1_CAGATC_L006_R1_001.trimmed.fastq -S
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Gneg1_trimme
d;


bowtie2 -q -p 20 -x /Users/admin/HTS-Software/bowtie2-
2.2.4/index/hg19 -U
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Project_S124_11-
12/Sample_Gneg2/Gneg2_CTTGTA_L006_R1_001.trimmed.fastq -S
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Gneg2_trimme
d;


bowtie2 -q -p 20 -x /Users/admin/HTS-Software/bowtie2-
2.2.4/index/hg19 -U
/Users/admin/NGSanalysis_JuliaW/RawData/ChIP/HMGN5/Project_S124_11-
12/Sample_Gpos1/Gpos1_ATCACG_L006_R1_001.trimmed.fastq -S
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Gpos1_trimme
d;
```

>>>as expected, the number of uniquely aligned reads (which will be
used to create tag directories) is the same, but the alignment
statistics have improved. In the now smaller, but still existing
number of unmappable reads the remaining adapter contaminations are
contained!

##Creating tag directories (Homer)

old tag dirs from untrimmed mappings are kept for the moment!!

```
#Usage
% makeTagDirectory <Output Directory Name> [options] <alignment
file1> [alignment file2] Ö
#Options
#-unique  (default) keep only uniquely aligned reads
#-genome <string> -checkGC This analysis will produce several output
files in addition to the basic quality control analysis
```

Appendix

```
makeTagDirectory
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
-genome hg19 -checkGC
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Hpos1_trimme
d;

makeTagDirectory
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-genome hg19 -checkGC
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Hpos2_trimme
d;

makeTagDirectory
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
-genome hg19 -checkGC
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Hneg1_trimme
d;

makeTagDirectory
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19
-genome hg19 -checkGC
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Hneg2_trimme
d;

makeTagDirectory
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Gpos1_trimmed_hg19
-genome hg19 -checkGC
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Gpos1_trimme
d;

makeTagDirectory
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Gneg1_trimmed_hg19
-genome hg19 -checkGC
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Gneg1_trimme
d;

makeTagDirectory
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Gneg2_trimmed_hg19
-genome hg19 -checkGC
```

```
/Users/admin/NGSanalysis_JuliaW/ChIP/bowtie2_hg19/HMGN5_Gneg2_trimme
d;
```

Making Genome Browser Files (homer)

```
old ones from untrimmed mappings are kept for the moment!!
#Usage
% makeUCSCfile <tag directory> -o auto
#Options
#-o auto   output file is stored in the corresponding tag directory
```

```
cd /Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/
```

```
makeUCSCfile HMGN5_Hpos1_trimmed_hg19 -o auto;
makeUCSCfile HMGN5_Hpos2_trimmed_hg19 -o auto;
makeUCSCfile HMGN5_Hneg1_trimmed_hg19 -o auto;
makeUCSCfile HMGN5_Hneg2_trimmed_hg19 -o auto;
makeUCSCfile HMGN5_Gpos1_trimmed_hg19 -o auto;
makeUCSCfile HMGN5_Gneg1_trimmed_hg19 -o auto;
makeUCSCfile HMGN5_Gneg2_trimmed_hg19 -o auto;
```

```
###FINDING PEAKS using HOMER's findPeaks tool
#extended regions are covered by reads >use -region option
#'regions' show spikes of varying size > find peaks with varying
initial peak size (before stitching into regions) and combine
generated peak files
#high background and middle enrichment success > use poisson
distribution p-value<0.001 (lowers tag threshold) instead of default
FDR<0.1% to identify initial peaks and lower enrichment over
background to 2-fold
#Use 2 strategies to identify peaks: strategy1 - keep peaks from
both Hpos samples    //   strategy2 - keep only reads found for
both Hpos samples
#Use two different 'peak size assumptions':    merge -
~200bp(auto)+350bp+500bp(histone)    //   merge2 -
~200bp(auto)+250bp+300+350bp+400bp+450bp+500bp(histone)
```

223

*********MERGE*******************************

~~~~~~
~~~~~~
~~~~~~
~~~~~~~~~~~~

~~~~~~~~~~~r_sauto_d1000_F2_p10-
3~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

```
cd /Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/

mkdir region_sauto_D1000_F2_p10-3/

cd region_sauto_D1000_F2_p10-3/

 findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
-region -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_sauto_D1000_F2
_p10-3/HMGN5_Hpos1.1_r_sauto_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
;
```

202bp, 8 tags

        168925 peaks passed threshold
#threshold here: poisson distribution p-value <0.001

        Differential Peaks: 42421 of 168925 (25.11% passed)
    #Enrichment over Input

        Stitching together putative peaks into regions
    #Retained peaks are merged into regions (here: when closer than
1000bp)

```
        Checking regions against input...


        Differential Peaks: 31634 of 31918 (99.11% passed)
     #Enrichment over Input in regions


        Clonal filtering: 31585 of 31634 (99.85% passed)   #Clonal
Filtering


        Total Peaks identified = 31585
findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
-region -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_sauto_D1000_F2
_p10-3/HMGN5_Hpos1.2_r_sauto_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19


202bp, 8 tags


        168925 peaks passed threshold


        Differential Peaks: 51669 of 168925 (30.59% passed)


        Stitching together putative peaks into regions


        Checking regions against input...


        Differential Peaks: 36843 of 37179 (99.10% passed)


        Clonal filtering: 36748 of 36843 (99.74% passed)


        Total Peaks identified = 36748


cd
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_sauto_D1000_F2
_p10-3
```

```
mergePeaks -prefix g HMGN5_Hpos1.1_r_sauto_d1000_F2_p10-3.txt
HMGN5_Hpos1.2_r_sauto_d1000_F2_p10-3.txt;


HMGN5_Hpos1.1_r_sauto_d1000_F2_p10-3.txt
HMGN5_Hpos1.2_r_sauto_d1000_F2_p10-3.txt        Total    Name


       X        21202    HMGN5_Hpos1.2_r_sauto_d1000_F2_p10-3.txt


X                16028    HMGN5_Hpos1.1_r_sauto_d1000_F2_p10-3.txt


X       X        15520    HMGN5_Hpos1.1_r_sauto_d1000_F2_p10-
3.txt|HMGN5_Hpos1.2_r_sauto_d1000_F2_p10-3.txt



mv  g_HMGN5_Hpos1.1_r_sauto_d1000_F2_p10-
3.txt_HMGN5_Hpos1.2_r_sauto_d1000_F2_p10-3.txt
g_Hpos1_r_sauto_d1000_F2_p10-3.txt;



 findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-region -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_sauto_D1000_F2
_p10-3/HMGN5_Hpos2.1_r_sauto_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
;

204bp, 9tags

            276310 peaks passed threshold

        Differential Peaks: 59232 of 276310 (21.44% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 41890 of 42388 (98.83% passed)
```

```
        Clonal filtering: 41821 of 41890 (99.84% passed)


        Total Peaks identified = 41821
```

```
findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-region -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_sauto_D1000_F2
_p10-3/HMGN5_Hpos2.2_r_sauto_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19
;
```

```
204bp, 9tags
```

```
        276310 peaks passed threshold

        Differential Peaks: 73259 of 276310 (26.51% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 49454 of 50018 (98.87% passed)

        Clonal filtering: 49348 of 49454 (99.79% passed)

        Total Peaks identified = 49348
```

```
cd
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_sauto_D1000_F2
_p10-3
```

```
mergePeaks -prefix g HMGN5_Hpos2.1_r_sauto_d1000_F2_p10-3.txt
HMGN5_Hpos2.2_r_sauto_d1000_F2_p10-3.txt;
```

```
HMGN5_Hpos2.1_r_sauto_d1000_F2_p10-3.txt
HMGN5_Hpos2.2_r_sauto_d1000_F2_p10-3.txt
```

```
 Total    Name


       X        30874     HMGN5_Hpos2.2_r_sauto_d1000_F2_p10-3.txt


X                23319     HMGN5_Hpos2.1_r_sauto_d1000_F2_p10-3.txt


X        X       18442     HMGN5_Hpos2.1_r_sauto_d1000_F2_p10-
3.txt|HMGN5_Hpos2.2_r_sauto_d1000_F2_p10-3.txt



mv  g_HMGN5_Hpos2.1_r_sauto_d1000_F2_p10-
3.txt_HMGN5_Hpos2.2_r_sauto_d1000_F2_p10-3.txt
g_Hpos2_r_sauto_d1000_F2_p10-3.txt;



mergePeaks -prefix g g_Hpos1_r_sauto_d1000_F2_p10-3.txt
g_Hpos2_r_sauto_d1000_F2_p10-3.txt;


g_Hpos1_r_sauto_d1000_F2_p10-3.txt
g_Hpos2_r_sauto_d1000_F2_p10-3.txt       Total    Name


       X        16248     g_Hpos2_r_sauto_d1000_F2_p10-3.txt


X                13327     g_Hpos1_r_sauto_d1000_F2_p10-3.txt


X        X       2188      g_Hpos1_r_sauto_d1000_F2_p10-
3.txt|g_Hpos2_r_sauto_d1000_F2_p10-3.txt



mergePeaks -prefix 1000 -d 1000 g_Hpos1_r_sauto_d1000_F2_p10-3.txt
g_Hpos2_r_sauto_d1000_F2_p10-3.txt;


g_Hpos1_r_sauto_d1000_F2_p10-3.txt
g_Hpos2_r_sauto_d1000_F2_p10-3.txt       Total    Name


       X        15920     g_Hpos2_r_sauto_d1000_F2_p10-3.txt


X                13002     g_Hpos1_r_sauto_d1000_F2_p10-3.txt
```

```
X        X        2513    g_Hpos1_r_sauto_d1000_F2_p10-
3.txt|g_Hpos2_r_sauto_d1000_F2_p10-3.txt


mv  g_g_Hpos1_r_sauto_d1000_F2_p10-
3.txt_g_Hpos2_r_sauto_d1000_F2_p10-3.txt
g_g_Hpos_r_sauto_d1000_F2_p10-3.txt;


mv  1000_g_Hpos1_r_sauto_d1000_F2_p10-
3.txt_g_Hpos2_r_sauto_d1000_F2_p10-3.txt
1000_g_Hpos_r_sauto_d1000_F2_p10-3.txt;



cp g_g_Hpos_r_sauto_d1000_F2_p10-3.txt
/Users/admin/Dropbox/JuliaW_Laengst/g_g_Hpos_r_sauto_d1000_F2_p10-
3.txt;


cp 1000_g_Hpos_r_sauto_d1000_F2_p10-3.txt
/Users/admin/Dropbox/JuliaW_Laengst/1000_g_Hpos_r_sauto_d1000_F2_p10
-3.txt;



~~~~~~~~~~~~~~~~region_s250_D1000_F2_p10-3~~~~~~~~~~~~~~~~~~~~


cd /Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/


mkdir region_s250_D1000_F2_p10-3/


cd region_s250_D1000_F2_p10-3/


 findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
-region -size 250 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s250_D1000_F2_
p10-3/HMGN5_Hpos1.1_r_s250_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
;


poisson threshold: 9 tags
```

```
 155975 peaks passed threshold

        Differential Peaks: 37779 of 155975 (24.22% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 27805 of 28040 (99.16% passed)

        Clonal filtering: 27776 of 27805 (99.90% passed)

        Total Peaks identified = 27776


findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
-region -size 250 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s250_D1000_F2_
p10-3/HMGN5_Hpos1.2_r_s250_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19

 155975 peaks passed threshold

        Differential Peaks: 46843 of 155975 (30.03% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 32894 of 33133 (99.28% passed)

        Clonal filtering: 32846 of 32894 (99.85% passed)

        Total Peaks identified = 32846
```

Appendix

```
cd
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s250_D1000_F2_
p10-3

mergePeaks -prefix g HMGN5_Hpos1.1_r_s250_d1000_F2_p10-3.txt
HMGN5_Hpos1.2_r_s250_d1000_F2_p10-3.txt;


        X        18898   HMGN5_Hpos1.2_r_s250_d1000_F2_p10-3.txt


X                13813   HMGN5_Hpos1.1_r_s250_d1000_F2_p10-3.txt


X        X       13924   HMGN5_Hpos1.1_r_s250_d1000_F2_p10-
3.txt|HMGN5_Hpos1.2_r_s250_d1000_F2_p10-3.txt



mv  g_HMGN5_Hpos1.1_r_s250_d1000_F2_p10-
3.txt_HMGN5_Hpos1.2_r_s250_d1000_F2_p10-3.txt
g_Hpos1_r_s250_d1000_F2_p10-3.txt;



 findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-region -size 250 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s250_D1000_F2_
p10-3/HMGN5_Hpos2.1_r_s250_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
;

poisson threshold: 10 tags

 276664 peaks passed threshold

        Differential Peaks: 53993 of 276664 (19.52% passed)

        Stitching together putative peaks into regions

        Checking regions against input...
```

```
        Differential Peaks: 37280 of 37592 (99.17% passed)


        Clonal filtering: 37237 of 37280 (99.88% passed)


        Total Peaks identified = 37237
```

```
findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-region -size 250 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s250_D1000_F2_
p10-3/HMGN5_Hpos2.2_r_s250_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19
;

poisson threshold: 10 tags

 276664 peaks passed threshold

 Differential Peaks: 68232 of 276664 (24.66% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 44582 of 44996 (99.08% passed)

        Clonal filtering: 44517 of 44582 (99.85% passed)

        Total Peaks identified = 44517


cd
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s250_D1000_F2_
p10-3

mergePeaks -prefix g HMGN5_Hpos2.1_r_s250_d1000_F2_p10-3.txt
HMGN5_Hpos2.2_r_s250_d1000_F2_p10-3.txt;
```

```
        X          27610    HMGN5_Hpos2.2_r_s250_d1000_F2_p10-3.txt

X                  20285    HMGN5_Hpos2.1_r_s250_d1000_F2_p10-3.txt

X       X          16874    HMGN5_Hpos2.1_r_s250_d1000_F2_p10-
3.txt|HMGN5_Hpos2.2_r_s250_d1000_F2_p10-3.txt




mv  g_HMGN5_Hpos2.1_r_s250_d1000_F2_p10-
3.txt_HMGN5_Hpos2.2_r_s250_d1000_F2_p10-3.txt
g_Hpos2_r_s250_d1000_F2_p10-3.txt;




~~~~~~~~~~~~~~~~region_s300_D1000_F2_p10-3~~~~~~~~~~~~~~~~~~~~

cd /Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/

mkdir region_s300_D1000_F2_p10-3/

cd region_s300_D1000_F2_p10-3/

 findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
-region -size 300 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s300_D1000_F2_
p10-3/HMGN5_Hpos1.1_r_s300_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
;

poisson threshold: 10 tags

 144077 peaks passed threshold

        Differential Peaks: 33356 of 144077 (23.15% passed)

        Stitching together putative peaks into regions
```

233

```
       Checking regions against input...

       Differential Peaks: 24122 of 24284 (99.33% passed)

       Clonal filtering: 24109 of 24122 (99.95% passed)

       Total Peaks identified = 24109



findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
-region -size 300 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s300_D1000_F2_
p10-3/HMGN5_Hpos1.2_r_s300_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19

 144077 peaks passed threshold

       Differential Peaks: 41966 of 144077 (29.13% passed)

       Stitching together putative peaks into regions

       Checking regions against input...

       Differential Peaks: 28935 of 29129 (99.33% passed)

       Clonal filtering: 28909 of 28935 (99.91% passed)

       Total Peaks identified = 28909



cd
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s300_D1000_F2_
p10-3

mergePeaks -prefix g HMGN5_Hpos1.1_r_s300_d1000_F2_p10-3.txt
HMGN5_Hpos1.2_r_s300_d1000_F2_p10-3.txt;
```

```
          X        16716     HMGN5_Hpos1.2_r_s300_d1000_F2_p10-3.txt

X                  11901     HMGN5_Hpos1.1_r_s300_d1000_F2_p10-3.txt

X        X         12172     HMGN5_Hpos1.1_r_s300_d1000_F2_p10-
3.txt|HMGN5_Hpos1.2_r_s300_d1000_F2_p10-3.txt
```

```
mv  g_HMGN5_Hpos1.1_r_s300_d1000_F2_p10-
3.txt_HMGN5_Hpos1.2_r_s300_d1000_F2_p10-3.txt
g_Hpos1_r_s300_d1000_F2_p10-3.txt;
```

```
 findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-region -size 300 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s300_D1000_F2_
p10-3/HMGN5_Hpos2.1_r_s300_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
;
```

```
poisson threshold: 12 tags

   171915 peaks passed threshold

        Differential Peaks: 39618 of 171915 (23.05% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 26514 of 26747 (99.13% passed)

        Clonal filtering: 26495 of 26514 (99.93% passed)

        Total Peaks identified = 26495
```

```
findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-region -size 300 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s300_D1000_F2_
p10-3/HMGN5_Hpos2.2_r_s300_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19
;

poisson threshold: 12 tags

171915 peaks passed threshold

        Differential Peaks: 49776 of 171915 (28.95% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 31630 of 31888 (99.19% passed)

        Clonal filtering: 31613 of 31630 (99.95% passed)

        Total Peaks identified = 31613




cd
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s300_D1000_F2_
p10-3

mergePeaks -prefix g HMGN5_Hpos2.1_r_s300_d1000_F2_p10-3.txt
HMGN5_Hpos2.2_r_s300_d1000_F2_p10-3.txt;

        X       18143    HMGN5_Hpos2.2_r_s300_d1000_F2_p10-3.txt


X               12994    HMGN5_Hpos2.1_r_s300_d1000_F2_p10-3.txt
```

```
X        X        13437    HMGN5_Hpos2.1_r_s300_d1000_F2_p10-
3.txt|HMGN5_Hpos2.2_r_s300_d1000_F2_p10-3.txt
```

```
mv  g_HMGN5_Hpos2.1_r_s300_d1000_F2_p10-
3.txt_HMGN5_Hpos2.2_r_s300_d1000_F2_p10-3.txt
g_Hpos2_r_s300_d1000_F2_p10-3.txt;
```

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~
```

```
~~~~~~~~~~~~~~~~region_s350_D1000_F2_p10-3~~~~~~~~~~~~~~~~~~~~
```

```
cd /Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/
```

```
mkdir region_s350_D1000_F2_p10-3/
```

```
cd region_s350_D1000_F2_p10-3/
```

```
 findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
-region -size 350 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s350_D1000_F2_
p10-3/HMGN5_Hpos1.1_r_s350_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
;
```

```
poisson threshold: 11 tags
```

```
132836 peaks passed threshold
```

```
        Differential Peaks: 30345 of 132836 (22.84% passed)
```

```
        Stitching together putative peaks into regions
```

```
        Checking regions against input...

        Differential Peaks: 21516 of 21678 (99.25% passed)

        Clonal filtering: 21512 of 21516 (99.98% passed)

        Total Peaks identified = 21512




findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
-region -size 350 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s350_D1000_F2_
p10-3/HMGN5_Hpos1.2_r_s350_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19

poisson threshold: 11 tags

 132836 peaks passed threshold

        Differential Peaks: 38420 of 132836 (28.92% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 26076 of 26250 (99.34% passed)

        Clonal filtering: 26067 of 26076 (99.97% passed)

        Total Peaks identified = 26067


cd
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s350_D1000_F2_
p10-3
```

Appendix

```
mergePeaks -prefix g HMGN5_Hpos1.1_r_s350_d1000_F2_p10-3.txt
HMGN5_Hpos1.2_r_s350_d1000_F2_p10-3.txt;


HMGN5_Hpos1.1_r_s350_d1000_F2_p10-3.txt
HMGN5_Hpos1.2_r_s350_d1000_F2_p10-3.txt Total   Name


        X       14882   HMGN5_Hpos1.2_r_s350_d1000_F2_p10-3.txt


X               10314   HMGN5_Hpos1.1_r_s350_d1000_F2_p10-3.txt


X       X       11169   HMGN5_Hpos1.1_r_s350_d1000_F2_p10-
3.txt|HMGN5_Hpos1.2_r_s350_d1000_F2_p10-3.txt




mv  g_HMGN5_Hpos1.1_r_s350_d1000_F2_p10-
3.txt_HMGN5_Hpos1.2_r_s350_d1000_F2_p10-3.txt
g_Hpos1_r_s350_d1000_F2_p10-3.txt;


 findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-region -size 350 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s350_D1000_F2_
p10-3/HMGN5_Hpos2.1_r_s350_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
;

poisson threshold: 13 tags

 177111 peaks passed threshold

        Differential Peaks: 39872 of 177111 (22.51% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 26387 of 26619 (99.13% passed)
```

```
      Clonal filtering: 26375 of 26387 (99.95% passed)


      Total Peaks identified = 26375



findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-region -size 350 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s350_D1000_F2_
p10-3/HMGN5_Hpos2.2_r_s350_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19
;

poisson threshold: 13 tags

 177111 peaks passed threshold

      Differential Peaks: 50816 of 177111 (28.69% passed)

      Stitching together putative peaks into regions

      Checking regions against input...

      Differential Peaks: 31853 of 32133 (99.13% passed)

      Clonal filtering: 31840 of 31853 (99.96% passed)

      Total Peaks identified = 31840



cd
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s350_D1000_F2_
p10-3

mergePeaks -prefix g HMGN5_Hpos2.1_r_s350_d1000_F2_p10-3.txt
HMGN5_Hpos2.2_r_s350_d1000_F2_p10-3.txt;
```

```
HMGN5_Hpos2.1_r_s350_d1000_F2_p10-3.txt
HMGN5_Hpos2.2_r_s350_d1000_F2_p10-3.txt Total    Name

        X        18268    HMGN5_Hpos2.2_r_s350_d1000_F2_p10-3.txt

X                12768    HMGN5_Hpos2.1_r_s350_d1000_F2_p10-3.txt

X        X       13539    HMGN5_Hpos2.1_r_s350_d1000_F2_p10-
3.txt|HMGN5_Hpos2.2_r_s350_d1000_F2_p10-3.txt



mv  g_HMGN5_Hpos2.1_r_s350_d1000_F2_p10-
3.txt_HMGN5_Hpos2.2_r_s350_d1000_F2_p10-3.txt
g_Hpos2_r_s350_d1000_F2_p10-3.txt;



~~~~~~~~~~~~~~~~region_s400_D1000_F2_p10-3~~~~~~~~~~~~~~~~~~~~

cd /Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/

mkdir region_s400_D1000_F2_p10-3/

cd region_s400_D1000_F2_p10-3/

 findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
-region -size 400 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s400_D1000_F2_
p10-3/HMGN5_Hpos1.1_r_s400_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
;

poisson threshold: 12 tags

 122151 peaks passed threshold

        Differential Peaks: 27892 of 122151 (22.83% passed)

        Stitching together putative peaks into regions
```

```
        Checking regions against input...

        Differential Peaks: 19396 of 19511 (99.41% passed)

        Clonal filtering: 19393 of 19396 (99.98% passed)

        Total Peaks identified = 19393


findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
-region -size 400 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s400_D1000_F2_
p10-3/HMGN5_Hpos1.2_r_s400_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19

 122151 peaks passed threshold

        Differential Peaks: 35611 of 122151 (29.15% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 23742 of 23903 (99.33% passed)

        Clonal filtering: 23738 of 23742 (99.98% passed)

        Total Peaks identified = 23738


cd
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s400_D1000_F2_
p10-3

mergePeaks -prefix g HMGN5_Hpos1.1_r_s400_d1000_F2_p10-3.txt
HMGN5_Hpos1.2_r_s400_d1000_F2_p10-3.txt;
```

```
       X       13252    HMGN5_Hpos1.2_r_s400_d1000_F2_p10-3.txt


X               8896     HMGN5_Hpos1.1_r_s400_d1000_F2_p10-3.txt


X       X       10468    HMGN5_Hpos1.1_r_s400_d1000_F2_p10-
3.txt|HMGN5_Hpos1.2_r_s400_d1000_F2_p10-3.txt




mv  g_HMGN5_Hpos1.1_r_s400_d1000_F2_p10-
3.txt_HMGN5_Hpos1.2_r_s400_d1000_F2_p10-3.txt
g_Hpos1_r_s400_d1000_F2_p10-3.txt;


 findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-region -size 400 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s400_D1000_F2_
p10-3/HMGN5_Hpos2.1_r_s400_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
;

poisson threshold: 14 tags

 177677 peaks passed threshold

        Differential Peaks: 39333 of 177677 (22.14% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 25829 of 26044 (99.17% passed)

        Clonal filtering: 25826 of 25829 (99.99% passed)

        Total Peaks identified = 25826
```

243

```
findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-region -size 400 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s400_D1000_F2_
p10-3/HMGN5_Hpos2.2_r_s400_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19
;

poisson threshold: 14 tags

 177677 peaks passed threshold

        Differential Peaks: 50297 of 177677 (28.31% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 31359 of 31627 (99.15% passed)

        Clonal filtering: 31352 of 31359 (99.98% passed)

        Total Peaks identified = 31352

cd
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s400_D1000_F2_
p10-3

mergePeaks -prefix g HMGN5_Hpos2.1_r_s400_d1000_F2_p10-3.txt
HMGN5_Hpos2.2_r_s400_d1000_F2_p10-3.txt;

        X       17986   HMGN5_Hpos2.2_r_s400_d1000_F2_p10-3.txt

X               12435   HMGN5_Hpos2.1_r_s400_d1000_F2_p10-3.txt

X       X       13324   HMGN5_Hpos2.1_r_s400_d1000_F2_p10-
3.txt|HMGN5_Hpos2.2_r_s400_d1000_F2_p10-3.txt
```

```
mv   g_HMGN5_Hpos2.1_r_s400_d1000_F2_p10-
3.txt_HMGN5_Hpos2.2_r_s400_d1000_F2_p10-3.txt
g_Hpos2_r_s400_d1000_F2_p10-3.txt;
```

~~~~~~~~~~~~~~~~region_s450_D1000_F2_p10-3~~~~~~~~~~~~~~~~~~~~

```
cd /Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/

mkdir region_s450_D1000_F2_p10-3/

cd region_s450_D1000_F2_p10-3/

 findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
-region -size 450 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s450_D1000_F2_
p10-3/HMGN5_Hpos1.1_r_s450_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
;

poisson threshold: 13 tags

112573 peaks passed threshold

        Differential Peaks: 26462 of 112573 (23.51% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 18123 of 18266 (99.22% passed)

        Clonal filtering: 18123 of 18123 (100.00% passed)

        Total Peaks identified = 18123
```

```
findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
-region -size 450 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s450_D1000_F2_
p10-3/HMGN5_Hpos1.2_r_s450_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19
```

```
 112573 peaks passed threshold

        Differential Peaks: 33506 of 112573 (29.76% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 21931 of 22113 (99.18% passed)

        Clonal filtering: 21930 of 21931 (100.00% passed)

        Total Peaks identified = 21930
```

```
cd
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s450_D1000_F2_
p10-3
```

```
mergePeaks -prefix g HMGN5_Hpos1.1_r_s450_d1000_F2_p10-3.txt
HMGN5_Hpos1.2_r_s450_d1000_F2_p10-3.txt;
```

```
HMGN5_Hpos1.1_r_s450_d1000_F2_p10-3.txt
HMGN5_Hpos1.2_r_s450_d1000_F2_p10-3.txtTotal     Name

        X       11983   HMGN5_Hpos1.2_r_s450_d1000_F2_p10-3.txt

X               8160    HMGN5_Hpos1.1_r_s450_d1000_F2_p10-3.txt

X       X       9918    HMGN5_Hpos1.1_r_s450_d1000_F2_p10-
3.txt|HMGN5_Hpos1.2_r_s450_d1000_F2_p10-3.txt
```

```
mv  g_HMGN5_Hpos1.1_r_s450_d1000_F2_p10-
3.txt_HMGN5_Hpos1.2_r_s450_d1000_F2_p10-3.txt
g_Hpos1_r_s450_d1000_F2_p10-3.txt;
```

```
 findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-region -size 450 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s450_D1000_F2_
p10-3/HMGN5_Hpos2.1_r_s450_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
;
```

```
poisson threshold: 15 tags
```

```
 176694 peaks passed threshold

        Differential Peaks: 38686 of 176694 (21.89% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 24927 of 25162 (99.07% passed)

        Clonal filtering: 24926 of 24927 (100.00% passed)

        Total Peaks identified = 24926
```

```
findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-region -size 450 -minDist 1000 -F 2 -poisson 0.001 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s450_D1000_F2_
p10-3/HMGN5_Hpos2.2_r_s450_d1000_F2_p10-3.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19
;
```

```
poisson threshold: 15 tags

 176694 peaks passed threshold

        Differential Peaks: 50013 of 176694 (28.30% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 30812 of 31089 (99.11% passed)

        Clonal filtering: 30810 of 30812 (99.99% passed)

        Total Peaks identified = 30810

cd
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s450_D1000_F2_
p10-3

mergePeaks -prefix g HMGN5_Hpos2.1_r_s450_d1000_F2_p10-3.txt
HMGN5_Hpos2.2_r_s450_d1000_F2_p10-3.txt;

        X       17626   HMGN5_Hpos2.2_r_s450_d1000_F2_p10-3.txt

X               11708   HMGN5_Hpos2.1_r_s450_d1000_F2_p10-3.txt

X       X       13137   HMGN5_Hpos2.1_r_s450_d1000_F2_p10-
3.txt|HMGN5_Hpos2.2_r_s450_d1000_F2_p10-3.txt



mv  g_HMGN5_Hpos2.1_r_s450_d1000_F2_p10-
3.txt_HMGN5_Hpos2.2_r_s450_d1000_F2_p10-3.txt
g_Hpos2_r_s450_d1000_F2_p10-3.txt;


~~~~~~~~~~~~~histone_F2~~~~~~~~~~~~~~~~~~~~~
```

Appendix

```
cd /Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles

mkdir histone_F2;

findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
-style histone -F 2 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/histone_F2/HMGN5_Hpos
1.1_histone.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
;

155015 peaks passed threshold

        Differential Peaks: 31879 of 155015 (20.57% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 21954 of 22143 (99.15% passed)

        Clonal filtering: 21953 of 21954 (100.00% passed)

        Total Peaks identified = 21953


findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
-style histone -F 2 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/histone_F2/HMGN5_Hpos
1.2_histone.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19
;

 155015 peaks passed threshold

        Differential Peaks: 40805 of 155015 (26.32% passed)
```

Stitching together putative peaks into regions

Checking regions against input...

Differential Peaks: 26979 of 27178 (99.27% passed)

Clonal filtering: 26978 of 26979 (100.00% passed)

Total Peaks identified = 26978


```
findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-style histone -F 2 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/histone_F2/HMGN5_Hpos
2.1_histone.txt -i
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
;
```

173466 peaks passed threshold

Differential Peaks: 37499 of 173466 (21.62% passed)

Stitching together putative peaks into regions

Checking regions against input...

Differential Peaks: 23990 of 24218 (99.06% passed)

Clonal filtering: 23988 of 23990 (99.99% passed)

Total Peaks identified = 23988


```
findPeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-style histone -F 2 -o
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/histone_F2/HMGN5_Hpos
2.2_histone.txt -i
```

```
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19
;

 173466 peaks passed threshold

        Differential Peaks: 48611 of 173466 (28.02% passed)

        Stitching together putative peaks into regions

        Checking regions against input...

        Differential Peaks: 29697 of 30008 (98.96% passed)

        Clonal filtering: 29695 of 29697 (99.99% passed)

        Total Peaks identified = 29695


## finding common peaks (default: -d given (from v4.4))


cd /Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/histone_F2

mergePeaks -prefix 0309 HMGN5_Hpos1.1_histone.txt
HMGN5_Hpos1.2_histone.txt;

HMGN5_Hpos1.1_histone.txt       HMGN5_Hpos1.2_histone.txt
Total   Name

X       15523   HMGN5_Hpos1.2_histone.txt

X               10483   HMGN5_Hpos1.1_histone.txt

X       X       11431
HMGN5_Hpos1.1_histone.txt|HMGN5_Hpos1.2_histone.txt

mv 0309_HMGN5_Hpos1.1_histone.txt_HMGN5_Hpos1.2_histone.txt
0309_Hpos1_histone.txt;
```

```
mergePeaks -prefix 0309 HMGN5_Hpos2.1_histone.txt
HMGN5_Hpos2.2_histone.txt


HMGN5_Hpos2.1_histone.txt        HMGN5_Hpos2.2_histone.txt
Total    Name


X        16982    HMGN5_Hpos2.2_histone.txt


X                11248    HMGN5_Hpos2.1_histone.txt


X        X       12658
HMGN5_Hpos2.1_histone.txt|HMGN5_Hpos2.2_histone.txt



mv 0309_HMGN5_Hpos2.1_histone.txt_HMGN5_Hpos2.2_histone.txt
0309_Hpos2_histone.txt;


mergePeaks -prefix given 0309_Hpos1_histone.txt
0309_Hpos2_histone.txt


0309_Hpos1_histone.txt  0309_Hpos2_histone.txt  Total    Name


         X       9892    0309_Hpos2_histone.txt


X                8651    0309_Hpos1_histone.txt


X        X       2754
0309_Hpos1_histone.txt|0309_Hpos2_histone.txt



mergePeaks -prefix 1000 -d 1000 0309_Hpos1_histone.txt
0309_Hpos2_histone.txt > Hpos_histone_1000.txt;


0309_Hpos1_histone.txt  0309_Hpos2_histone.txt  Total    Name


         X       9872    0309_Hpos2_histone.txt
```

```
X                8640     0309_Hpos1_histone.txt


X        X       2781
0309_Hpos1_histone.txt|0309_Hpos2_histone.txt




-------------merging-----------


cd /Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/

mkdir merge2

cd merge2

#Hpos1

mergePeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/histone_F2/0309_Hpos1
_histone.txt
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s450_D1000_F2_
p10-3/g_Hpos1_r_s450_d1000_F2_p10-3.txt >
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450.txt


         X       2107
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s450_D1000_F2_
p10-3/g_Hpos1_r_s450_d1000_F2_p10-3.txt


X                3630
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/histone_F2/0309_Hpos1
_histone.txt


X        X       7792
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/histone_F2/0309_Hpos1
_histone.txt|/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s
450_D1000_F2_p10-3/g_Hpos1_r_s450_d1000_F2_p10-3.txt
```

```
mergePeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450.txt
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s400_D1000_F2_
p10-3/g_Hpos1_r_s400_d1000_F2_p10-3.txt >
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450_s400.txt;


        X       2062
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s400_D1000_F2_
p10-3/g_Hpos1_r_s400_d1000_F2_p10-3.txt


X               5153
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450.txt


X       X       8370
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450.txt|/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s400_
D1000_F2_p10-3/g_Hpos1_r_s400_d1000_F2_p10-3.txt


rm
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450.txt;


mergePeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450_s400.txt
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s350_D1000_F2_
p10-3/g_Hpos1_r_s350_d1000_F2_p10-3.txt
>/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone
_s450_s400_s350.txt


        X       2203
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s350_D1000_F2_
p10-3/g_Hpos1_r_s350_d1000_F2_p10-3.txt
```

```
X                6642
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge/Hpos1_histone_s
450_s400.txt


X       X        8933
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge/Hpos1_histone_s
450_s400.txt|/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s
350_D1000_F2_p10-3/g_Hpos1_r_s350_d1000_F2_p10-3.txt


rm
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450_s400.txt;


mergePeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450_s400_s350.txt
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s300_D1000_F2_
p10-3/g_Hpos1_r_s300_d1000_F2_p10-3.txt
>/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone
_s450_s400_s350_s300.txt


        X        2653
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s300_D1000_F2_
p10-3/g_Hpos1_r_s300_d1000_F2_p10-3.txt


X                8278
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge/Hpos1_histone_s
450_s400_s350.txt


X       X        9489
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge/Hpos1_histone_s
450_s400_s350.txt|/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/reg
ion_s300_D1000_F2_p10-3/g_Hpos1_r_s300_d1000_F2_p10-3.txt


rm
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450_s400_s350.txt;
```

```
mergePeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450_s400_s350_s300.txt
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s250_D1000_F2_
p10-3/g_Hpos1_r_s250_d1000_F2_p10-3.txt
>/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone
_s450_s400_s350_s300_s250.txt


        X       3517
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s250_D1000_F2_
p10-3/g_Hpos1_r_s250_d1000_F2_p10-3.txt


X               10044
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge/Hpos1_histone_s
450_s400_s350_s300.txt


X       X       10363
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge/Hpos1_histone_s
450_s400_s350_s300.txt|/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFile
s/region_s250_D1000_F2_p10-3/g_Hpos1_r_s250_d1000_F2_p10-3.txt


rm
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450_s400_s350_s300.txt;


mergePeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450_s400_s350_s300_s250.txt
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_sauto_D1000_F2
_p10-3/g_Hpos1_r_sauto_d1000_F2_p10-3.txt
>/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone
_s450_s400_s350_s300_s250_sauto.txt


        X       4266
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_sauto_D1000_F2
_p10-3/g_Hpos1_r_sauto_d1000_F2_p10-3.txt
```

```
X                    12705
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge/Hpos1_histone_s
450_s400_s350_s300_s250.txt

X        X        11206
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge/Hpos1_histone_s
450_s400_s350_s300_s250.txt|/Users/admin/NGSanalysis_JuliaW/ChIP/Pea
kFiles/region_sauto_D1000_F2_p10-3/g_Hpos1_r_sauto_d1000_F2_p10-
3.txt


rm
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450_s400_s350_s300_s250.txt;

#Hpos2

mergePeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/histone_F2/0309_Hpos2
_histone.txt
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s450_D1000_F2_
p10-3/g_Hpos2_r_s450_d1000_F2_p10-3.txt >
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450.txt

         X        3494
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s450_D1000_F2_
p10-3/g_Hpos2_r_s450_d1000_F2_p10-3.txt

X                3017
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/histone_F2/0309_Hpos2
_histone.txt

X        X        9612
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/histone_F2/0309_Hpos2
_histone.txt|/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s
450_D1000_F2_p10-3/g_Hpos2_r_s450_d1000_F2_p10-3.txt
```

```
mergePeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450.txt
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s400_D1000_F2_
p10-3/g_Hpos2_r_s400_d1000_F2_p10-3.txt >
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400.txt;


        X       2638
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s400_D1000_F2_
p10-3/g_Hpos2_r_s400_d1000_F2_p10-3.txt


X               5470
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450.txt


X       X       10646
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450.txt|/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s400_
D1000_F2_p10-3/g_Hpos2_r_s400_d1000_F2_p10-3.txt


rm
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450.txt;



mergePeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400.txt
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s350_D1000_F2_
p10-3/g_Hpos2_r_s350_d1000_F2_p10-3.txt
>/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone
_s450_s400_s350.txt


        X       2503
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s350_D1000_F2_
p10-3/g_Hpos2_r_s350_d1000_F2_p10-3.txt
```

```
X                 7745
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400.txt

X         X       10993
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400.txt|/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_
s350_D1000_F2_p10-3/g_Hpos2_r_s350_d1000_F2_p10-3.txt

rm
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400.txt;


mergePeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400_s350.txt
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s300_D1000_F2_
p10-3/g_Hpos2_r_s300_d1000_F2_p10-3.txt
>/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone
_s450_s400_s350_s300.txt


        X       2390
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s300_D1000_F2_
p10-3/g_Hpos2_r_s300_d1000_F2_p10-3.txt

X                 10234
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400_s350.txt

X         X       10990
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400_s350.txt|/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/re
gion_s300_D1000_F2_p10-3/g_Hpos2_r_s300_d1000_F2_p10-3.txt

rm
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400_s350.txt;
```

```
mergePeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400_s350_s300.txt
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s250_D1000_F2_
p10-3/g_Hpos2_r_s250_d1000_F2_p10-3.txt
>/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone
_s450_s400_s350_s300_s250.txt


        X        4626
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_s250_D1000_F2_
p10-3/g_Hpos2_r_s250_d1000_F2_p10-3.txt


X                11416
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400_s350_s300.txt


X        X       12185
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400_s350_s300.txt|/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFil
es/region_s250_D1000_F2_p10-3/g_Hpos2_r_s250_d1000_F2_p10-3.txt


rm
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400_s350_s300.txt;


mergePeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400_s350_s300_s250.txt
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_sauto_D1000_F2
_p10-3/g_Hpos2_r_sauto_d1000_F2_p10-3.txt
>/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone
_s450_s400_s350_s300_s250_sauto.txt


        X        4933
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/region_sauto_D1000_F2
_p10-3/g_Hpos2_r_sauto_d1000_F2_p10-3.txt
```

```
X                14762
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400_s350_s300_s250.txt

X        X        13443
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400_s350_s300_s250.txt|/Users/admin/NGSanalysis_JuliaW/ChIP/Pe
akFiles/region_sauto_D1000_F2_p10-3/g_Hpos2_r_sauto_d1000_F2_p10-
3.txt

rm
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400_s350_s300_s250.txt;


#Strategy1_merge2 (combination of both peak files)

mergePeaks
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450_s400_s350_s300_s250_sauto.txt
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400_s350_s300_s250_sauto.txt >
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/strategy1_merg
e2.txt

 X        26803
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos2_histone_
s450_s400_s350_s300_s250_sauto.txt

X                21806
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450_s400_s350_s300_s250_sauto.txt

X        X        6277
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/Hpos1_histone_
s450_s400_s350_s300_s250_sauto.txt|/Users/admin/NGSanalysis_JuliaW/C
hIP/PeakFiles/merge2/Hpos2_histone_s450_s400_s350_s300_s250_sauto.tx
t
```

>>54886 peaks in total

#strategy2_merge2 (only overlapping peaks)

cd /Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/

mergePeaks -prefix g
Hpos1_histone_s450_s400_s350_s300_s250_sauto.txt
Hpos2_histone_s450_s400_s350_s300_s250_sauto.txt

  X   26803   Hpos2_histone_s450_s400_s350_s300_s250_sauto.txt

X   21806
Hpos1_histone_s450_s400_s350_s300_s250_sauto.txt

X   X   6277
Hpos1_histone_s450_s400_s350_s300_s250_sauto.txt|Hpos2_histone_s450_
s400_s350_s300_s250_sauto.txt

mv
g_Hpos1_histone_s450_s400_s350_s300_s250_sauto.txt_Hpos2_histone_s45
0_s400_s350_s300_s250_sauto.txt strategy2_merge2.txt
#Annotation

cd /Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/

annotatePeaks.pl strategy1_merge2.txt hg19 -d
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Gneg1_trimmed_hg19
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Gneg2_trimmed_hg19
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Gpos1_trimmed_hg19
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-go
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/strategy1_merg

```
e2_GeneOnt -genomeOntology
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/strategy1_merg
e2_GenomeOnt -gene /Users/admin/NGSanalysis_JuliaW/ChIP/GE_info.txt
> strategy1_merge2.ann.txt
```

genomic position	Annotation	Number of peaks	Total size (bp)	Log2 Enrichment
	3UTR	589.0	21944102	0.598
	miRNA	1.0	48690	0.211
	ncRNA	239.0	5148687	1.388
	TTS	962.0	28108010	0.948
	pseudo	35.0	1865590	0.081
	Exon	1414.0	34376480	1.213
	Intron	24531.0	1194960999	0.211
	Intergenic	21419.0	1773347320	-0.554
	Promoter	5223.0	31343091	3.232
	5UTR	465.0	2505007	3.387
	snoRNA	0.0	119	-15.744
	snRNA	0.0	105	-15.744
	rRNA	0.0	10999	-15.744

genomic features	Annotation	Number of peaks	Total size (bp)	Log2 Enrichment
	3UTR	589.0	21944102	0.599
	Other	3.0	3962446	-4.549
	Unknown?	0.0	18108	-15.744
	RNA	5.0	115354	1.290
	miRNA	1.0	48690	0.213
	ncRNA	239.0	5148687	1.389
	TTS	962.0	28108010	0.949
	LINE	4850.0	625086598	-1.192
	LINE?	1.0	10448	2.433
	srpRNA	2.0	255222	-1.177
	SINE	6742.0	381200056	-0.003
	RC	4.0	443678	-0.975

```
tRNA        10.0     93275    2.597
DNA?        4.0      265124   -0.232
pseudo      35.0     1865590  0.082
DNA         1064.0   96207793        -0.680


Exon        1414.0   34376480        1.215
Intron      14818.0  634436369       0.398
Intergenic           11650.0 904233348       -0.460
Promoter             5221.0  31343091        3.232
5UTR        463.0    2505007  3.383
snoRNA      0.0      119      -15.744
LTR?        1.0      21177    1.414
scRNA       2.0      116995   -0.052
CpG-Island           1349.0   9368574 3.022
Low_complexity       152.0    15514766        -0.855
LTR         4863.0   260567553       0.075
Simple_repeat        377.0    24952436        -0.230
snRNA       4.0      323087   -0.518
Unknown     21.0     1252703  -0.080
SINE?       1.0      43187    0.386
Satellite            38.0     12902670        -2.589
rRNA        1.0      182734   -1.695
```

```
annotatePeaks.pl strategy2_merge2.txt hg19 -d
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Gneg1_trimmed_hg19
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Gneg2_trimmed_hg19
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Gpos1_trimmed_hg19
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg1_trimmed_hg19
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hneg2_trimmed_hg19
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos1_trimmed_hg19
/Users/admin/NGSanalysis_JuliaW/ChIP/TagDir/HMGN5_Hpos2_trimmed_hg19
-go
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/strategy2_merg
e2_GeneOnt -genomeOntology
/Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/strategy2_merg
e2_GenomeOnt -gene /Users/admin/NGSanalysis_JuliaW/ChIP/GE_info.txt
> strategy2_merge2.ann.txt
```

genomic position Annotation	Number of peaks	Total size (bp)	Log2 Enrichment
3UTR	60.0	21896602	0.406
miRNA	0.0	48690	−12.616
ncRNA	38.0	5106375	1.847
TTS	138.0	27943868	1.255
pseudo	8.0	1831063	1.079
Exon	248.0	34274169	1.806
Intron	2174.0	1190135564	−0.179
Intergenic	1620.0	1718905554	−1.134
Promoter	1791.0	31179196	4.795
5UTR	199.0	2495214	5.269
snoRNA	0.0	119	−12.616
snRNA	0.0	105	−12.616
rRNA	0.0	10999	−12.616

genomic features Annotation	Number of peaks	Total size (bp)	Log2 Enrichment
3UTR	60.0	21896602	0.406
Other	1.0	3945746	−3.028
Unknown?	0.0	17982	−12.616
RNA	1.0	115110	2.071
miRNA	0.0	48690	−12.616
ncRNA	38.0	5106375	1.848
TTS	138.0	27943868	1.256
LINE	219.0	618432922	−2.545
LINE?	0.0	10448	−12.616
srpRNA	1.0	254103	0.929
SINE	314.0	378546582	−1.317
RC	0.0	439839	−12.616
tRNA	4.0	92339	4.389
DNA?	1.0	264912	0.869
pseudo	8.0	1831063	1.080
DNA	61.0	95742637	−1.698
Exon	248.0	34274169	1.807

```
              Intron   1412.0   632193295           0.112
              Intergenic        1055.0   863261291         -0.758
              Promoter          1791.0   31179196          4.796
              5UTR     197.0    2495214 5.255
              snoRNA   0.0      119      -12.616
              LTR?     0.0      21177    -12.616
              scRNA    0.0      116686   -12.616
              CpG-Island        468.0    9275326 4.609
              Low_complexity    28.0     15301963          -0.176
              LTR      194.0    256011178         -1.448
              Simple_repeat     35.0     24458261          -0.531
              snRNA    1.0      320159   0.595
              Unknown  1.0      1249709 -1.369
              SINE?    0.0      42649    -12.616
              Satellite         1.0      11207736          -4.534
              rRNA     0.0      179404   -12.616
```

```
cd /Users/admin/NGSanalysis_JuliaW/ChIP/PeakFiles/merge2/

cp strategy1_merge2.txt /Users/admin/Dropbox/JuliaW_Laengst/

cp strategy2_merge2.txt /Users/admin/Dropbox/JuliaW_Laengst/

> create bed files for UCSC

cp strategy1_merge2.ann.txt /Users/admin/Dropbox/JuliaW_Laengst/

cp strategy2_merge2.ann.txt /Users/admin/Dropbox/JuliaW_Laengst/

cp strategy1_merge2_GeneOnt/geneOntology.html
/Users/admin/Dropbox/JuliaW_Laengst/strategy1_merge2_GeneOntology.ht
ml
cp strategy2_merge2_GeneOnt/geneOntology.html
/Users/admin/Dropbox/JuliaW_Laengst/strategy2_merge2_GeneOntology.ht
ml
cp strategy1_merge2_GenomeOnt/GenomeOntology.html
/Users/admin/Dropbox/JuliaW_Laengst/strategy1_merge2_GenomeOntology.
html
```

```
cp strategy2_merge2_GenomeOnt/GenomeOntology.html
/Users/admin/Dropbox/JuliaW_Laengst/strategy2_merge2_GenomeOntology.
html
```

```
####motif analysis with strat2_merge2 peak file (annotated)
```

```
cd /Volumes/Storage2/HMGN5_0817/
```

```
mkdir ChIP_oldAnalysis
```

```
cd C:\Users\JuliaWimmer\Dropbox\Share
```

```
pscp.exe
C:\Users\JuliaWimmer\Dropbox\Collaborations\L‰ngst\HMGN5\ChIP-
seq\strategy2_merge2.ann.txt admin@pc1011202646b.uni-
regensburg.de:/Volumes/Storage2/HMGN5_0817/ChIP_oldAnalysis/strategy
2_merge2.ann.txt
```

```
#Run 'findMotifsGenome.pl' on hg19 genome, assume 1000bp size (peaks
of varying size (200/250/300/350/400/450/500bp) were stitched into
regions if closer than 1000bp - alignemnt viz showed broad regions
of enrichment with several peaks within this region)
```

```
findMotifsGenome.pl
/Volumes/Storage2/HMGN5_0817/ChIP_oldAnalysis/strategy2_merge2.ann.t
xt hg19 /Volumes/Storage2/HMGN5_0817/ChIP_oldAnalysis/motifs_s1000 -
size 1000 -p 20
```

```
#also test '-size 200' (smallest peak finding size) and compare
```

```
findMotifsGenome.pl
/Volumes/Storage2/HMGN5_0817/ChIP_oldAnalysis/strategy2_merge2.ann.t
xt hg19 /Volumes/Storage2/HMGN5_0817/ChIP_oldAnalysis/motifs_s200 -
size 200 -p 20
```

9.3.2 CLIP-seq analysis command lines.

A detailed description of the command lines used for CLIP-seq analysis and output is indicated in the following script. Brief descriptions of the command lines are indicated after "##". Output and explanation of usage are described after"#".

```
##Detailed analysis bowtie, homer

wget --no-check-certificate
https://sourceforge.net/projects/bowtie-
bio/files/bowtie/1.2.1.1/bowtie-1.2.1.1-macos-x86_64.zip
wget --no-check-certificate
ftp://ftp.ccb.jhu.edu/pub/data/bowtie_indexes/GRCh38_no_alt.zip

/Volumes/Storage2/HTS_software/bowtie-1.2.1.1/bowtie

-best -chunkmbs 512 -n 1 -S -M 100.
mkdir /Volumes/Storage2/HMGN5_0817/CLIP/bowtie
mkdir /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-1
mkdir /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-2
mkdir /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-3
mkdir /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-1
mkdir /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-2
mkdir /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-3

mv /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
1/GCA_000001405.15_GRCh38_no_alt_analysis_set.1.ebwt
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/
mv /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
1/GCA_000001405.15_GRCh38_no_alt_analysis_set.2.ebwt
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/
mv /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
1/GCA_000001405.15_GRCh38_no_alt_analysis_set.3.ebwt
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/
mv /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
1/GCA_000001405.15_GRCh38_no_alt_analysis_set.4.ebwt
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/
```

```
mv /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
1/GCA_000001405.15_GRCh38_no_alt_analysis_set.rev.1.ebwt
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/
mv /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
1/GCA_000001405.15_GRCh38_no_alt_analysis_set.rev.2.ebwt
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/
cd /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/
/Volumes/Storage2/HTS_software/bowtie-1.2.1.1/bowtie --best --
chunkmbs 512 -n 1 -S -M 100 -p 8
GCA_000001405.15_GRCh38_no_alt_analysis_set
/Volumes/Storage2/HMGN5_0817/CLIP/rawdata_Project_S145/Sample_GFP1/G
FP1_GCACTA_L002_R1_001.fastq.gz
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-1.bowtie.sam

# reads processed: 26530237
# reads with at least one reported alignment: 16153689 (60.89%)
# reads that failed to align: 5411828 (20.40%)
# reads with alignments sampled due to -M: 4964720 (18.71%)
#Reported 16153689 alignments to 1 output stream(s)

/Volumes/Storage2/HTS_software/bowtie-1.2.1.1/bowtie --best --
chunkmbs 512 -n 1 -S -M 100 -p 8
GCA_000001405.15_GRCh38_no_alt_analysis_set
/Volumes/Storage2/HMGN5_0817/CLIP/rawdata_Project_S145/Sample_GFP2/G
FP2_ACCTCA_L002_R1_001.fastq.gz
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-2.bowtie.sam
/Volumes/Storage2/HTS_software/bowtie-1.2.1.1/bowtie --best --
chunkmbs 512 -n 1 -S -M 100 -p 8
GCA_000001405.15_GRCh38_no_alt_analysis_set
/Volumes/Storage2/HMGN5_0817/CLIP/rawdata_Project_S145/Sample_GFP3/G
FP3_GTGCTT_L002_R1_001.fastq.gz
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-3.bowtie.sam
/Volumes/Storage2/HTS_software/bowtie-1.2.1.1/bowtie --best --
chunkmbs 512 -n 1 -S -M 100 -p 8
GCA_000001405.15_GRCh38_no_alt_analysis_set
/Volumes/Storage2/HMGN5_0817/CLIP/rawdata_Project_S145/Sample_HMGN51
/HMGN51_AACCAG_L002_R1_001.fastq.gz
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-1.bowtie.sam
```

```
/Volumes/Storage2/HTS_software/bowtie-1.2.1.1/bowtie --best --
chunkmbs 512 -n 1 -S -M 100 -p 8
GCA_000001405.15_GRCh38_no_alt_analysis_set
/Volumes/Storage2/HMGN5_0817/CLIP/rawdata_Project_S145/Sample_HMGN52
/HMGN52_TGGTGA_L002_R1_001.fastq.gz
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-2.bowtie.sam

# reads processed: 33310255
# reads with at least one reported alignment: 16298962 (48.93%)
# reads that failed to align: 4750503 (14.26%)
# reads with alignments sampled due to -M: 12260790 (36.81%)
#Reported 16298962 alignments to 1 output stream(s)


/Volumes/Storage2/HTS_software/bowtie-1.2.1.1/bowtie --best --
chunkmbs 512 -n 1 -S -M 100 -p 8
GCA_000001405.15_GRCh38_no_alt_analysis_set
/Volumes/Storage2/HMGN5_0817/CLIP/rawdata_Project_S145/Sample_HMGN53
/HMGN53_AGTGAG_L002_R1_001.fastq.gz
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-3.bowtie.sam

# reads processed: 24727981
# reads with at least one reported alignment: 16265546 (65.78%)
# reads that failed to align: 3232836 (13.07%)
# reads with alignments sampled due to -M: 5229599 (21.15%)
#Reported 16265546 alignments to 1 output stream(s)


##Uniquely mapped reads

samtools flagstat /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
1.bowtie.sam
#21118409 + 0 mapped (79.60%)
samtools flagstat /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
2.bowtie.sam
#16050306 + 0 mapped (68.95%)
samtools flagstat /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
3.bowtie.sam
#18191290 + 0 mapped (77.13%)
```

```
samtools flagstat /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
1.bowtie.sam
#17519792 + 0 mapped (84.70%)
samtools flagstat /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
2.bowtie.sam
#28559752 + 0 mapped (85.74%)
samtools flagstat /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
3.bowtie.sam
#21495145 + 0 mapped (86.93%)



##Converting to bam format & indexing for IGV visualization

wget --no-check-certificate
ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_26/gencode.
v26.primary_assembly.annotation.gtf.gz
cd /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/
samtools view -bh -o GFP-1.bowtie.bam
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-1.bowtie.sam
samtools view -bh -o GFP-2.bowtie.bam
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-2.bowtie.sam
samtools view -bh -o GFP-3.bowtie.bam
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-3.bowtie.sam
samtools view -bh -o HMGN5-1.bowtie.bam
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-1.bowtie.sam
samtools view -bh -o HMGN5-2.bowtie.bam
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-2.bowtie.sam
samtools view -bh -o HMGN5-3.bowtie.bam
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-3.bowtie.sam
samtools sort -o GFP-1.bowtie.sorted.bam
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-1.bowtie.bam
samtools sort -o GFP-2.bowtie.sorted.bam
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-2.bowtie.bam
samtools sort -o GFP-3.bowtie.sorted.bam
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-3.bowtie.bam
samtools sort -o HMGN5-1.bowtie.sorted.bam
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-1.bowtie.bam
samtools sort -o HMGN5-2.bowtie.sorted.bam
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-2.bowtie.bam
```

```
samtools sort -o HMGN5-3.bowtie.sorted.bam
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-3.bowtie.bam
samtools index GFP-1.bowtie.sorted.bam

samtools index GFP-2.bowtie.sorted.bam
samtools index GFP-3.bowtie.sorted.bam
samtools index HMGN5-1.bowtie.sorted.bam
samtools index HMGN5-2.bowtie.sorted.bam
samtools index HMGN5-3.bowtie.sorted.bam


cd C:\Users\JuliaWimmer\Dropbox\Share
pscp.exe admin@pc1011202646b.uni-
regensburg.de:/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
1.bowtie.sorted.bam C:\Users\JuliaWimmer\
pscp.exe admin@pc1011202646b.uni-
regensburg.de:/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
2.bowtie.sorted.bam C:\Users\JuliaWimmer\
pscp.exe admin@pc1011202646b.uni-
regensburg.de:/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
3.bowtie.sorted.bam C:\Users\JuliaWimmer
pscp.exe admin@pc1011202646b.uni-
regensburg.de:/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
1.bowtie.sorted.bam C:\Users\JuliaWimmer
pscp.exe admin@pc1011202646b.uni-
regensburg.de:/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
2.bowtie.sorted.bam C:\Users\JuliaWimmer
pscp.exe admin@pc1011202646b.uni-
regensburg.de:/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
3.bowtie.sorted.bam C:\Users\JuliaWimmer
pscp.exe admin@pc1011202646b.uni-
regensburg.de:/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
1.bowtie.sorted.bam.bai C:\Users\JuliaWimmer
pscp.exe admin@pc1011202646b.uni-
regensburg.de:/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
2.bowtie.sorted.bam.bai C:\Users\JuliaWimmer\
pscp.exe admin@pc1011202646b.uni-
regensburg.de:/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
3.bowtie.sorted.bam.bai C:\Users\JuliaWimmer\
```

```
pscp.exe admin@pc1011202646b.uni-
regensburg.de:/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
1.bowtie.sorted.bam.bai C:\Users\JuliaWimmer\
pscp.exe admin@pc1011202646b.uni-
regensburg.de:/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
2.bowtie.sorted.bam.bai C:\Users\JuliaWimmer\
pscp.exe admin@pc1011202646b.uni-
regensburg.de:/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
3.bowtie.sorted.bam.bai C:\Users\JuliaWimmer\


bedtools bamtobed -i /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
1.bowtie.bam > /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
1.bowtie.bed
bedtools bamtobed -i /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
2.bowtie.bam > /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
2.bowtie.bed
bedtools bamtobed -i /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
3.bowtie.bam > /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
3.bowtie.bed
bedtools bamtobed -i /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
1.bowtie.bam > /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
1.bowtie.bed
bedtools bamtobed -i /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
2.bowtie.bam > /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
2.bowtie.bed
bedtools bamtobed -i /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
3.bowtie.bam > /Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
3.bowtie.bed



##Motif analysis with homer


cd /Users/admin/Homer/
sudo mv configureHomer.pl configureHomer_v4.1.pl
cp /Volumes/Storage2/HTS_software/homer_v4.9/configureHomer.pl
/Users/admin/Homer/configureHomer_v4.9.pl
```

```
perl /Users/admin/Homer/configureHomer_v4.9.pl -install homer
perl /Users/admin/Homer/configureHomer_v4.9.pl -list
perl /Users/admin/Homer/configureHomer_v4.9.pl -install hg38


##Creating tag directory

mkdir /Volumes/Storage2/HMGN5_0817/CLIP/homer
cd /Volumes/Storage2/HMGN5_0817/CLIP/homer

makeTagDirectory /Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP1_tagdir
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-1.bowtie.bam
makeTagDirectory /Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP2_tagdir
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-2.bowtie.bam
makeTagDirectory /Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP3_tagdir
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-3.bowtie.bam
makeTagDirectory
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51_tagdir
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-1.bowtie.bam
makeTagDirectory
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN52_tagdir
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-2.bowtie.bam
makeTagDirectory
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN53_tagdir
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-3.bowtie.bam
makeTagDirectory
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP2+3_tagdir
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-2.bowtie.bam
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-3.bowtie.bam
makeTagDirectory
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51+2+3_tagdir
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-1.bowtie.bam
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-2.bowtie.bam
/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-3.bowtie.bam

makeUCSCfile /Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP1_tagdir -o
auto -strand separate
makeUCSCfile /Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP2_tagdir -o
auto -strand separate
```

274

```
makeUCSCfile /Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP3_tagdir -o
auto -strand separate
makeUCSCfile /Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51_tagdir -
o auto -strand separate
makeUCSCfile /Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN52_tagdir -
o auto -strand separate
makeUCSCfile /Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN53_tagdir -
o auto -strand separate
```

```
##Identify HMGN5-specific peaks (GFP as input)
```

```
findPeaks /Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51_tagdir -
style factor -o auto -i
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP2+3_tagdir
```

```
 #Expected tags per peak = 0.068200 (tbp = 0.000304)
```

Threshold	Peak Count	Expected Peak Count	FDR
4	13729.000	15.243	0.001110
3	25933.000	897.104	0.034593
2	62357.000	39688.484	0.636472
1	281180.000	1177261.018	4.186859
0	281180.000	17857142.857	63.507870

```
        #0.10% FDR Threshold set at 5.0 (poisson pvalue ~ 1.16e-08)
        #8532 peaks passed threshold
        #Differential Peaks: 3534 of 8532 (41.42% passed)
        #Local Background Filtering: 3246 of 3534 (91.85% passed)
        #Clonal filtering: 3216 of 3246 (99.08% passed)
        #Total Peaks identified = 3216
        #Centering peaks of size 224 using a fragment length of 61
```

```
findPeaks /Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN52_tagdir -
style factor -o auto -i
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP2+3_tagdir
```

```
#Expected tags per peak = 0.065731 (tbp = 0.000307)
```

Threshold	Peak Count	Expected Peak Count	FDR
4	14822.000	13.794	0.000931
3	28336.000	842.213	0.029722
2	67531.000	38651.984	0.572359
1	294228.000	1189099.809	4.041423
0	294228.000	18691588.785	63.527566

```
#0.10% FDR Threshold set at 4.0 (poisson pvalue ~ 7.38e-07)
#14822 peaks passed threshold
#Differential Peaks: 5754 of 14822 (38.82% passed)
#Local Background Filtering: 5444 of 5754 (94.61% passed)
#Clonal filtering: 5340 of 5444 (98.09% passed)
#Total Peaks identified = 5340
#Centering peaks of size 214 using a fragment length of 61


findPeaks /Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN53_tagdir -
style factor -o auto -i
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP2+3_tagdir


 #Expected tags per peak = 0.105069 (tbp = 0.000370)
```

Threshold	Peak Count	Expected Peak Count	FDR
5	11588.000	1.377	0.000119
4	19579.000	65.764	0.003359
3	38149.000	2516.984	0.065978
2	89141.000	72505.915	0.813385
1	89141.000	1404753.908	15.758786
0	89141.000	14084507.042	158.002569

```
#0.10% FDR Threshold set at 5.0 (poisson pvalue ~ 9.78e-08)
#11588 peaks passed threshold
#Differential Peaks: 3054 of 11588 (26.35% passed)
```

```
#Local Background Filtering: 2689 of 3054 (88.05% passed)
#Clonal filtering: 2686 of 2689 (99.89% passed)
#Total Peaks identified = 2686
#Centering peaks of size 284 using a fragment length of 60
```

findPeaks /Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51+2+3_tagdir
-style factor -o auto -i
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP2+3_tagdir

#Expected tags per peak = 0.253292 (tbp = 0.001481)

Threshold	Peak Count	Expected Peak Count	FDR
5	90790.000	164.668	0.001814
4	172508.000	3278.805	0.019007
3	203752.000	52457.332	0.257457
2	203752.000	634928.800	3.116184
1	203752.000	5234131.229	25.688735
0	203752.000	23391812.865	114.805317

```
#0.10% FDR Threshold set at 6.0 (poisson pvalue ~ 2.95e-07)
#76939 peaks passed threshold
#Differential Peaks: 34969 of 76939 (45.45% passed)
#Local Background Filtering: 30533 of 34969 (87.31% passed)
#Clonal filtering: 30458 of 30533 (99.75% passed)
#Total Peaks identified = 30458
#Centering peaks of size 171 using a fragment length of 60
```

##Finding motifs

##Against whole hg38 in individual HMGN5vsGFP2+3 peaks

findMotifsGenome.pl
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51_tagdir/peaks.txt hg38

```
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51_tagdir/motifvshg38 -
rna -p 10
findMotifsGenome.pl
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN52_tagdir/peaks.txt hg38
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN52_tagdir/motifvshg38 -
rna -p 10
findMotifsGenome.pl
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN53_tagdir/peaks.txt hg38
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN53_tagdir/motifvshg38 -
rna -p 10
```

```
##against whole hg38 in overlapping HMGN5vsGFP2+3 peaks (from all
three HMGN5 samples)

mkdir
/Volumes/Storage2/HMGN5_0817/CLIP/homer/mergedPeaks_HMGN5overGFP2+3
cd
/Volumes/Storage2/HMGN5_0817/CLIP/homer/mergedPeaks_HMGN5overGFP2+3
sudo cp
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51_tagdir/peaks.txt
HMGN51.txt
sudo cp
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN52_tagdir/peaks.txt
HMGN52.txt
sudo cp
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN53_tagdir/peaks.txt
HMGN53.txt
mergePeaks -d given HMGN51.txt HMGN52.txt -prefix overlap
sudo mv overlap_HMGN51.txt_HMGN52.txt HMGN512.txt
mergePeaks -d given HMGN512.txt HMGN53.txt -prefix overlap
sudo mv overlap_HMGN512.txt_HMGN53.txt HMGN5123.txt
findMotifsGenome.pl
/Volumes/Storage2/HMGN5_0817/CLIP/homer/mergedPeaks_HMGN5overGFP2+3/
HMGN5123.txt hg38
/Volumes/Storage2/HMGN5_0817/CLIP/homer/mergedPeaks_HMGN5overGFP2+3/
motifvshg38 -rna -p 10
```

```
##against whole hg38 in HMGN5vsGFP2+3 peaks from combined tag
directory

findMotifsGenome.pl
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51+2+3_tagdir/peaks.txt
hg38
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51+2+3_tagdir/motifvshg3
8 -rna -p 10

###Annotate the resulting peak file (annotatePeaks.pl)

##Configure homer to annotate peaks on the basis of human genome
information

perl /Users/admin/Homer/configureHomer_v4.9.pl -list

perl /Users/admin/Homer/configureHomer_v4.9.pl -install hg38

sudo perl /Users/admin/Homer/configureHomer_v4.9.pl -install human-o

sudo perl /Users/admin/Homer/configureHomer_v4.9.pl -install human-p

sudo perl /Users/admin/Homer/configureHomer_v4.9.pl -install human-
mRNA

sudo perl /Users/admin/Homer/configureHomer_v4.9.pl -install human-
mRNA-5UTR

sudo perl /Users/admin/Homer/configureHomer_v4.9.pl -install human-
mRNA-3UTR


##Usage

% annotatePeaks.pl <peak/BED file> <genome> <other options> >
<output file>
```

```
##Options

-d <tag dir1> #Adds tag information from listed tag directories
-go  <output dir>    #Adds gene ontology analysis
-genomeOntology <output dir>    #Adds genome ontology analysis

cd /Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51+2+3_tagdir/
annotatePeaks.pl peaks.txt hg38 -d
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN52_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN53_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP1_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP2_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP3_tagdir/ -go
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51+2+3_tagdir/GeneOnt -
genomeOntology
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51+2+3_tagdir/GenomeOnt
> HMGN51+2+3peaks.ann.txt


cd
/Volumes/Storage2/HMGN5_0817/CLIP/homer/mergedPeaks_HMGN5overGFP2+3_
d100/
annotatePeaks.pl HMGN5123.txt hg38 -d
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN52_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN53_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP1_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP2_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP3_tagdir/ -go
/Volumes/Storage2/HMGN5_0817/CLIP/homer/mergedPeaks_HMGN5overGFP2+3_
d100/GeneOnt -genomeOntology
/Volumes/Storage2/HMGN5_0817/CLIP/homer/mergedPeaks_HMGN5overGFP2+3_
d100/GenomeOnt > HMGN5123_merged100_peaks.ann.txt
```

```
cd
/Volumes/Storage2/HMGN5_0817/CLIP/homer/mergedPeaks_HMGN5overGFP2+3_
d50/
annotatePeaks.pl HMGN5123.txt hg38 -d
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN52_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN53_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP1_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP2_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP3_tagdir/ -go
/Volumes/Storage2/HMGN5_0817/CLIP/homer/mergedPeaks_HMGN5overGFP2+3_
d50/GeneOnt -genomeOntology
/Volumes/Storage2/HMGN5_0817/CLIP/homer/mergedPeaks_HMGN5overGFP2+3_
d50/GenomeOnt > HMGN5123_merged50_peaks.ann.txt


cd
/Volumes/Storage2/HMGN5_0817/CLIP/homer/mergedPeaks_HMGN5overGFP2+3/

annotatePeaks.pl HMGN5123.txt hg38 -d
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN52_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN53_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP1_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP2_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP3_tagdir/ -go
/Volumes/Storage2/HMGN5_0817/CLIP/homer/mergedPeaks_HMGN5overGFP2+3/
GeneOnt -genomeOntology
/Volumes/Storage2/HMGN5_0817/CLIP/homer/mergedPeaks_HMGN5overGFP2+3/
GenomeOnt > HMGN5123_mergedgiven_peaks.ann.txt


cd /Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51_tagdir/

annotatePeaks.pl peaks.txt hg38 -d
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN52_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN53_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP1_tagdir/
```

```
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP2_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP3_tagdir/ -go
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51_tagdir/GeneOnt -
genomeOntology
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51_tagdir/GenomeOnt >
HMGN51peaks.ann.txt


cd /Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN52_tagdir/


annotatePeaks.pl peaks.txt hg38 -d
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN52_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN53_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP1_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP2_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP3_tagdir/ -go
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN52_tagdir/GeneOnt -
genomeOntology
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN52_tagdir/GenomeOnt >
HMGN52peaks.ann.txt


cd /Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN53_tagdir/


annotatePeaks.pl peaks.txt hg38 -d
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN51_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN52_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN53_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP1_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP2_tagdir/
/Volumes/Storage2/HMGN5_0817/CLIP/homer/GFP3_tagdir/ -go
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN53_tagdir/GeneOnt -
genomeOntology
/Volumes/Storage2/HMGN5_0817/CLIP/homer/HMGN53_tagdir/GenomeOnt >
HMGN53peaks.ann.txt




##Differential gene expression analysis
```

```
source("http://bioconductor.org/biocLite.R")
biocLite("TxDb.Hsapiens.UCSC.hg38.knownGene")
library(GenomicFeatures)
library(GenomicAlignments)
library(TxDb.Hsapiens.UCSC.hg38.knownGene)



setwd("/Volumes/Storage2/HMGN5_0817/CLIP/R_Analysis")


outputDir <-
file.path("/Volumes/Storage2/HMGN5_0817/CLIP/R_Analysis")



#Genes for DGE analysis were extracted from
TxDb.Hsapiens.UCSC.hg38.knownGene by
txdb <- TxDb.Hsapiens.UCSC.hg38.knownGene  # create transcription
database based on hg38 known genes from ucsc
eByg <- exonsBy(txdb, by="gene") # defines features to look for
overlap later



bfl <-
BamFileList(c("/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
1.bowtie.sorted.bam",
"/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
2.bowtie.sorted.bam",
"/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/HMGN5-
3.bowtie.sorted.bam", "/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
2.bowtie.sorted.bam", "/Volumes/Storage2/HMGN5_0817/CLIP/bowtie/GFP-
3.bowtie.sorted.bam")) # list of bam files



counteByg <- summarizeOverlaps(eByg, bfl, mode="Union",
ignore.strand=FALSE, inter.feature=FALSE, singleEnd=TRUE) #creates
an object including a matrix with all known UCSC genes and the
corresponding read counts. Counting option is 'union according to
HTseq-count: only reads mapping to unique features (i.e. not-
overlapping gene regions) are counted.
```

```
countDFeByg <- assays(counteByg)$counts #extracts read count table
as matrix

write.table(countDFeByg, file=file.path(outputDir,
"countDFeByg_2+3.txt"), sep="\t")

#list of Gene Symbols (downloaded from http://www.genenames.org/cgi-
bin/download [Curated by the HGNC > Approved Symbol // Downloaded
from external sources > Entrez ID],

# create data structures; DGEList object contains $counts (count
matrix), $sample (sample information such as names, group, lib.size,
norm.factors)

library(edgeR)

grp <- factor(c(1,1,1,2,2))

dge <- DGEList(counts=countDFeByg, genes=countDFeByg[,0], group=grp)
#create data structure for normalisation, diff gene expr analysis
and so on... (type 'dge' to look at structure)


#Normalization

dge_norm <- dge

dge_norm <- calcNormFactors(dge_norm) # calculates a normalization
factor based on the total number of Read Counts and inserts this
into the sample information.


##Data exploration

pdf("plotMDS_dge_norm_2+3.pdf",width=7,height=5)

plotMDS_dge_norm <- plotMDS.DGEList(dge_norm)

dev.off()
```

```
## filter un-informative genes (with low CPM by threshold) (EdgeR
UserGuide -2.6: usually a gene is required to have a count of 5-10
in a library to be considered expressed in that library.  Users
should also filter with count-per-million (CPM) rather than
filtering on the counts directly, as the latter does not account for
differences in library sizes between samples.


mcpm_norm <- 1e6 * t(t(dge_norm$counts) / dge_norm$samples$lib.size)
#Defines counts per million to be used as threshold for read counts
in count matrix


ridx5 <- rowSums(mcpm_norm > 5) >= 2 #keep only those genes for
further processing that have at least 5 cpms in at least 2 samples
(setting >2 samples is a simplified assumption that all 3 replicates
have cpm >5)

ridx20 <- rowSums(mcpm_norm > 20) >= 2

dge_norm_f5<- dge_norm[ridx5,] #only genes with at least 3 samples
with cpm >5 are further processed

dge_norm_f20<- dge_norm[ridx20,]

dim(dge_norm$counts)  #returns dimension of the count matrix in the
dge_s_norm data.set

#[1] 24943     5#all 24943 genes of initial count table

dim(dge_norm_f5$counts)

#[1] 10235     5 #only 10235 genes 'survive'

dim(dge_norm_f20$counts)

#[1] 4686     5
```

Appendix

###Estimating dispersion ("The more the members of a collection differ to larger extents from each other, the more dispersed or spread-out the variation pattern is judged to be. This roughly describes the dispersion aspect of variation." > taken from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4373960/ >>> the smaller the better)
Strictly speaking, it captures all sources of the inter-library variation between replicates, including perhaps contributions from technical causes such as library preparation as well as true biological variation between samples. (Extracted from EdgeR UserGuide)

##Pairwise comparisons between two or more groups (classic): edgeR uses the quantile-adjusted conditional maximum likelihood (qCML) method for experiments with single factor (see chapter 2.9.1 of EdgeR UserGuide)

#Overall level of biological variability

dge_norm <- estimateCommonDisp(dge_norm, verbose=TRUE)

Disp = 1.43326, BCV = 1.1972   #The square root of the common dispersion gives the coefficient of variation of biological variation

dge_norm_f5 <- estimateCommonDisp(dge_norm_f5, verbose=TRUE)

Disp = 0.87898, BCV = 0.9375

dge_norm_f20 <- estimateCommonDisp(dge_norm_f20, verbose=TRUE)

Disp = 0.3421 , BCV = 0.5849

#>> Removing lowly expressed genes from the count table reduces the variability between individual samples.

Appendix

```
#Repeating Data Exploration with filtered data

pdf("plotMDS_dge_norm_f5_2+3.pdf",width=7,height=5)

plotMDS_dge_norm_f5 <- plotMDS.DGEList(dge_norm_f5)

dev.off()

pdf("plotMDS_dge_norm_f20_2+3.pdf",width=7,height=5)

plotMDS_dge_norm_f20 <- plotMDS.DGEList(dge_norm_f20)

dev.off()
```

#Gene-wise dispersion (applicable on datasets with a single factor design (such as here: GFP vs HMGN5) since it fails to take into account the effects from multiple factors in a more complicated experiment.)

```
dge_norm <- estimateTagwiseDisp(dge_norm)
dge_norm_f5 <- estimateTagwiseDisp(dge_norm_f5)
dge_norm_f20 <- estimateTagwiseDisp(dge_norm_f20)
```

##Differential Expression (EdgeR UsersGuide-2.9.2: the exact test is only applicable to experiments with a single factor - such as here: GFP vs HMGN5)

```
et_norm <- exactTest(dge_norm)

et_norm_f5 <- exactTest(dge_norm_f5)  #the (gene-wise) exact test is
comparable to t-test

et_norm_f20 <- exactTest(dge_norm_f20)

tt_et_norm <- topTags(et_norm)
```

```
tt_et_norm_f5 <- topTags(et_norm_f5)  #sorts entries of
"et_s_norm_f5" > #The top DE tags have tiny p-values and FDR values,
as well as large fold changes.
tt_et_norm_f20_100 <- topTags(et_norm_f20, n=100)
tt_et_norm_24943 <- topTags(et_norm, n=24943)
tt_et_norm_f5_10235 <- topTags(et_norm_f5, n=10235)
tt_et_norm_f20_4686 <- topTags(et_norm_f20, n=4686)


#summary(de_norm <- decideTestsDGE(et_norm))
-1      17
0   24925
1       1


#summary(de_norm_f5 <- decideTestsDGE(et_norm_f5))
-1      29
0   10205
1       1


#summary(de_norm_f20 <- decideTestsDGE(et_norm_f20))
-1   28
0   4657
1       1



write.table(tt_et_norm, file=file.path(outputDir,
"tt_et_norm_2+3.txt"), sep="\t")
write.table(tt_et_norm_f5, file=file.path(outputDir,
"tt_et_norm_f5_2+3.txt"), sep="\t")



write.table(tt_et_norm_f20_100, file=file.path(outputDir,
"tt_et_norm_f20_2+3_100.txt"), sep="\t")


tt_et_norm_100 <- topTags(et_norm, n=100)
```

```
tt_et_norm_f5_100 <- topTags(et_norm_f5, n=100)


write.table(tt_et_norm_100, file=file.path(outputDir,
"tt_et_norm_2+3_100.txt"), sep="\t")
write.table(tt_et_norm_f5_100, file=file.path(outputDir,
"tt_et_norm_f5_2+3_100.txt"), sep="\t")
write.table(tt_et_norm_24943, file=file.path(outputDir,
"tt_et_norm_2+3_24943.txt"), sep="\t")
write.table(tt_et_norm_f5_10235, file=file.path(outputDir,
"tt_et_norm_f5_2+3_10235.txt"), sep="\t")
write.table(tt_et_norm_f20_4686, file=file.path(outputDir,
"tt_et_norm_f20_2+3_4686.txt"), sep="\t")
```

10 Acknowledgements