THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)

Comprehensive analysis of human transcription factor binding sites with ChIP-seq and topological arrangements of transcription factor complexes on the DNA

by Erik Czipa

Supervisor: Dr. Endre Barta, PhD

UNIVERSITY OF DEBRECEN

DOCTORAL SCHOOL OF MOLECULAR CELL AND IMMUNE BIOLOGY

DEBRECEN, 2019

Table of content

Introduction

1.1 Chromatin structure and gene regulation in general

The DNA, as a blueprint, contains the genetic code to build an organism (Koltzoff, 1934; Levene, Phoebus; Jacobs, 1909). In a multicellular organism, every single cell contains almost the same genetic information. In contrast, the cells that cooperate in the organism show large morphological and physiological differences relative to each other. The cellular differentiation is a complex multi-stage process, which starts with organization of the active (euchromatin) and inactive (heterochromatin) DNA territories (Huisinga, Brower-Toland, & Elgin, 2006; Luger, Mader, Richmond, Sargent, & Richmond, 1997). The effects on the transcriptional activities of genes influence the proteome of the cell (Crick, 1970).

Due to advances in the field of molecular biology, regulatory mechanisms in the nucleus have been discovered. The vast majority of these can be linked to the formation and alteration of active and inactive DNA regions. Both the active and inactive stages are regulated by nuclear proteins, which modify protein-DNA interactions (e.g. nucleosomes) or the DNA (DNA methylation) itself due to their enzymatic activity. A network of nuclear proteins is required to establish these changes.

Nucleosomes form fundamental repeating units in which the DNA is coiled around histone octamers in approximately 1.7 turns (Olins & Olins, 1974). Most of the genomic DNA is wrapped into nucleosomes, which limits access to these regions. The accessibility of DNA is significantly influenced by modification of nucleosome structure (Fenley, Anandakrishnan, Kidane, & Onufriev, 2018). The stabilization of the nucleosomes with methylation of the third histone at the position 9 lysine (H3K9me histone mark) and the binding of heterochromatin protein 1 (HP1) can lead to long term gene repression and formation of heterochromatin (Zeng, Ball Jr, & Yokomori, 2010). Histones in heterochromatin are hypoacetylated and hypomethylated (Rusche, Kirchmaier, & Rine, 2003). This mechanism is extremely well conserved and maintains the gene silencing within the highly condensed chromatin structure.

An approximately 30 nm solenoid fiber forms the so-called heterochromatin, which is close to the nuclear lamina (Craig, 2005). HP1 recognizes the H3K9me with its chromodomain (CD) in a non-sequence specific manner. Most CDs of HP1 family members form a similar 3D structure (Fischle et al., 2003). ATP-dependent protein complexes, termed chromatin remodeling complexes, alter the previously mentioned HP1-histone–DNA contacts; changes in nucleosome structure occur through covalent modifications to create a less compact chromatin state (Tomschik, Zheng, van Holde, Zlatanova, & Leuba, 2005).

Less condensed regions of chromatin, named euchromatin, can be permissive to transcription when they have assumed the open-potentiated conformation requisite for gene expression. This can be viewed as an approximately 10 nm "beads-on-string" fiber (Georgel, Fletcher, Hager, & Hansen, 2003). Posttranslational modification of histones like H3K27ac or H3K4me3 can lead to loose binding of DNA or to the complete disappearance of these proteins (forming nucleosome-depleted regions (NDRs)) by chromatin remodeling complexes. These changes make them accessible and allow preinitiation complex (PIC) assembly and RNA POLII recruitment, which are required to start transcription.

1.1. Transcriptional regulation and transcription factors

1.1.1. Transcriptional initiation and preinitiation complex

Transcription is the central event of gene expression, which is preceded by many biochemical processes. Transcription starts with the assembly of the transcription initiation complex, which consists of general transcription factors (GTFs) (Kuhlman, Cho, Reinberg, & Hernandez, 1999). These factors gather around the promoter before the 5' end of the gene, which is the transcription start site (TSS) (Lagrange, Kapanidis, Tang, Reinberg, & Ebright, 1998). The promoter has a specific sequence, which can be recognized by general transcription factors (Orphanides, Lagrange, & Reinberg, 1996). Key DNA sequences, named core promoter

elements, have frequently occurring patterns. These include the TATA-box, downstream promoter elements (DPE), initiators (Inr), TCT, B recognition element (BRE), and motif ten element (MTE) (Brown, 2018). There are no universal core promoter elements represented in all promoters. The promoter element composition can associate with a special biological network. The core promoter serves as a binding platform for the transcription machinery. The general transcription factors attach to the structurally and spatially distinct promoter elements. These general transcription factors include TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, which do not function universally on all core promoters (Thomas & Chiang, 2006).

The most frequently investigated method for PIC assembly starts with the binding of the TATA-box binding protein (TBP), which is a subunit of TFIID. The TFIID is a multi-subunit complex, which consists of 14 TBP-associated factors (TAFs), in addition to TBP (Brindefalk et al., 2013). TATA is the best-known promoter element and can be found in 10 to 20 % of metazoan promoters. The TATA-box contains a consensus sequence (TATAWAAR) characterized by a repeating T and A (Reeve, 2003). The TATA-box is located approximately 30 bases upstream of the TSS (**Figure 1**).

Two types of BRE (TFIIB recognition element) can occur next to the TATA-box elements. Usually, promoters contain only an upstream (BREu) or a downstream (BREd) BRE. The contacts between BREs and TFIIB are mediated by DNA-recognition domains (**Figure 1**) (Deng & Roberts, 2006).

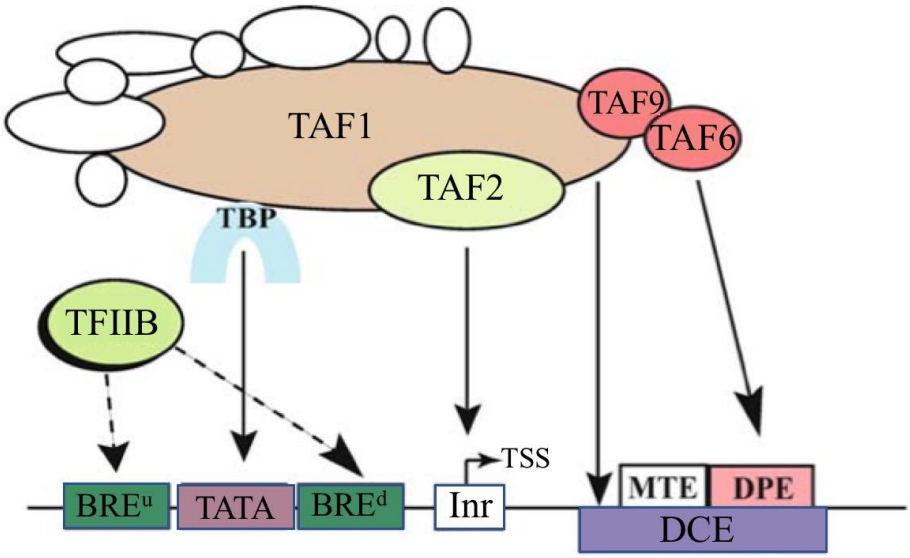
In the absence of a TATA-box, the TAF1 and TAF2 subunits of TFIID bind Inr elements. The Inr element has the consensus sequence YYANWYY in a position -2 to +5 to the TSS (**Figure 1**) (Smale & Baltimore, 1989; Xi et al., 2007).

In other TATA-less promoters, the DPE and MTE are common promoter elements, which are bound by TAF6 and TAF9 factors of TFIID. These elements are adjacent or potentially overlapping about 27-33 base pairs downstream relative to the TSS. The favored

nucleotides for each sub-regions are: 18–22 (CGANC), 27–29 (CGG), and 30–33 (WYGT) (**Figure 1**) (Lim et al., 2004).

It is worth mentioning the TCT element, which was observed in the promoters of proteins which are involved in ribosome synthesis. The TCT element has a YC+1TYTYY consensus sequence, where C+1 is the TSS. TCT is outside of the RNA polymerase II transcription system because TFIID is not able to bind promoters containing a TCT motif (Parry et al., 2010).

Core promoters and general transcription factors are essential for direct transcription initiation but generally have low basal activity, which can be further suppressed by chromatin or activated by more remote regulatory elements, such as enhancers. These regions and the promoters are occupied by other transcription factors (TFs), which can recognize specific sequences as well. Although these TFs are not part of the PIC, they play a significant role in its assembly and maintain the conditions for unperturbed transcription. TF binding occurs in nucleosome-depleted regions (NDRs), which generally encompass regions with lengths similar to those protected by nucleosomes.



Name of element	TFIIB Recognition Element	TATA Box	Initiator	Downstream promoter element
Distance from TSS	~-37 to -32 or ~-25 to -20	~-31 to -26	~-2 to +4	+28 to +32
Sequence	GGG CGCC CCA	TATAAWAR	Dm: TCA G/T TT (+1) Hs: YYAN T/A YY	A G AC G G TT A C

Figure 1. Preinitiation complex (PIC) and promoter elements. Overview of a hypothetical non-ribosomal promoter, which contains all elements and their interacting PIC subunit. Any specific core promoter may contain some, all, or none of these motifs. The table summarizes the location (relative to TSS) and the sequence preference of the elements. The Dm (Drosophila melanogaster) and Hs (Homo sapiens) show the difference between fruit fly and human initiator sequences (Butler & Kadonaga, 2002; Thomas & Chiang, 2006).

1.1.2. Transcription factors

In addition to the general transcription factors, which show DNA sequence preferences, several other proteins influence transcription to regulate gene expression. An assessment of the scientific literature resulted in the identification of 3230 “transcription factors” (in human) (Fulton et al., 2009). TFs are diverse, not only in function and cell-type specificity, but also in structure. However, all TFs have DNA-binding domains (DBDs), which recognize specific

sequences in the genome. These short genomic regions are called transcription factor binding sites (TFBSs) or responsive elements and they are frequently represented by position weight matrices (PWMs), which will be detailed later. The mechanism by which DBDs interact with their target sequences depends highly on their structural features (Wingender, Schoeps, Haubrock, & Dönitz, 2015). The current classifications of transcription factors are highly dependent on the structural specificities of the DBDs and classification of transcription factor motifs (TFBMs, simply called motifs below). Similarly (in the case of amino acid sequence), DBDs bind to related DNA sequences (Wingender, 2013).

One DNA turn is made up of 10.4 nucleotide pairs, which is about 0.34 nm (**Figure 2A**). The coiling of the two strands forms major and minor grooves (Watson & Crick, 1953). The outside of the double helix is studded with DNA sequence information: the edge of each base pair presents a distinctive pattern of hydrogen-bond donors or acceptors. The major groove displays more molecular features than the minor groove because it is wider and more accessible. Minor groove access is limited because of the sugar backbone on the outside of the double helix (**Figure 2B**). Therefore, nearly all transcription factors make contacts with major grooves (Carl-Ivar Brändén, 1999). Different amino acid side chains have different nucleic acid or phosphate backbone affinity. For instance, thymine can form a relatively high number of hydrogen bonds with arginine and lysine, cytosine with glutamate and aspartate, adenine with arginine, lysine, asparagine, glutamine, and guanine with arginine and lysine. Arginine, lysine, threonine, and asparagine shows high affinity for the DNA backbone (**Table 1**) (Luscombe, Laskowski, & Thornton, 2001).

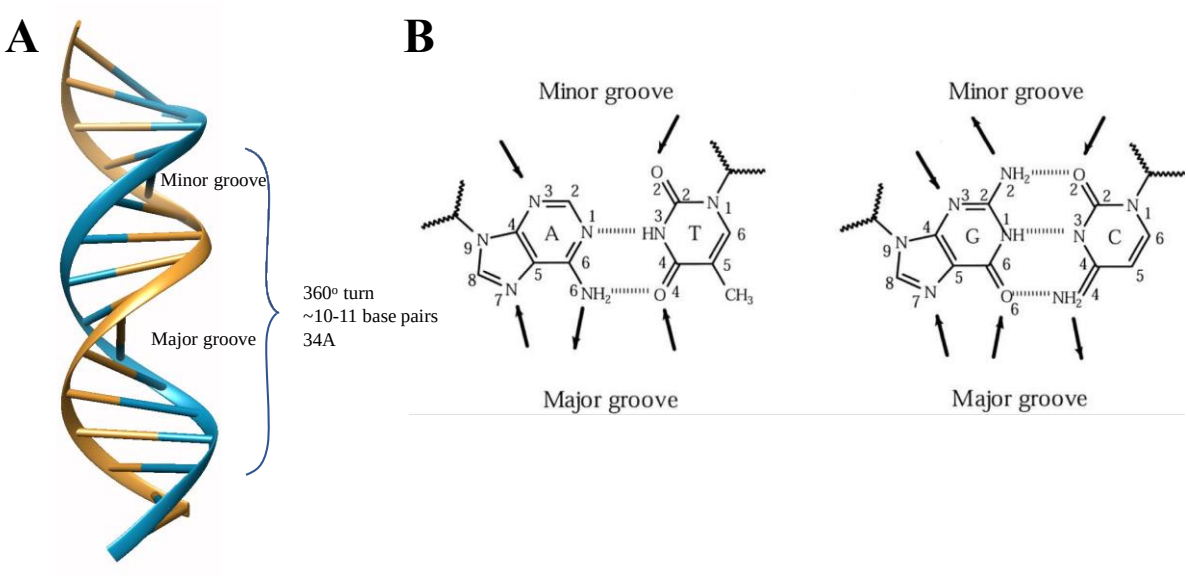


Figure 2. The DNA double helix forms a minor and major groove with differing accessibility. A) A 360º turn of DNA is 34A and involves 10.4 nucleotides. B) Structural formulas show the accessible hydrogen positions, pointing towards acceptors and away from donors. Arrows below and above the structures represent major groove access and minor groove access, respectively. (Luscombe et al., 2001)

Amino acids			DNA bases								DNA backbone				Total	
			Thymine		Cytosine		Adenine		Guanine		Sugar		Phosphate			
Arginine	ARG	R	**24**	**(2.5)**	8	(2.0)	**19**	**(4.2)**	**98**	**(4.0)**	8	(1.9)	**218**	**(49.9)**	**375**	**(64.7)**
Lysine	LYS	K	**9**	**(4.4)**	6	(3.4)	**4**	**(7.4)**	**30**	**(7.1)**	3	(3.2)	**109**	**(86.7)**	**165**	**(112.6)**
Serine	SER	S	3	(3.0)	2	(2.2)	1	(5.0)	12	(4.6)	2	(2.1)	**91**	**(57.3)**	**113**	**(74.4)**
Threonine	THR	T	5	(2.4)	3	(2.0)	4	(4.2)	-	(4.0)	1	(1.8)	**79**	**(49.2)**	**92**	**(63.9)**
Asparagine	ASN	N	7	(3.6)	10	(2.7)	**18**	**(6.0)**	7	(5.8)	3	(2.6)	**43**	**(70.7)**	88	(91.8)
Glutamine	GLN	Q	2	(2.2)	2	(1.7)	**16**	**(3.8)**	6	(3.6)	2	(1.7)	**42**	**(44.8)**	70	(58.1)
Glycine	GLY	G	1	(3.2)	4	(2.4)	-	(5.4)	6	(5.1)	1	(2.4)	29	(63.3)	**41**	**(82.2)**
Histidine	HIS	H	-	(0.8)	1	(0.6)	1	(1.5)	12	(1.4)	-	(0.7)	26	(18.3)	**40**	**(23.7)**
Tyrosine	TYR	Y	-	(1.2)	2	(1.0)	-	(2.1)	1	(2.0)	1	(1.0)	35	(25.7)	39	(33.4)
Alanine	ALA	A	1	(2.5)	1	(2.0)	-	(4.2)	1	(4.0)	-	(1.9)	**17**	**(49.8)**	**20**	**(64.6)**
Glutamate	GLU	E	-	(3.8)	**10**	**(3.0)**	1	(6.5)	1	(6.2)	-	(2.9)	**6**	**(76.2)**	**18**	**(99.0)**
Isoleucine	ILE	I	-	(0.7)	-	(0.5)	-	(1.3)	3	(1.2)	-	(0.6)	11	(15.9)	14	(20.7)
Aspartate	ASP	D	-	(4.5)	**5**	**(3.4)**	2	(7.5)	2	(7.2)	-	(3.3)	**2**	**(88.3)**	**11**	**(114.7)**
Valine	VAL	V	-	(1.2)	-	(1.0)	-	(2.0)	-	(2.0)	-	(0.9)	**8**	**(24.5)**	**8**	**(31.8)**
Cysteine	CYS	C	-	(0.2)	1	(0.2)	-	(0.5)	-	(0.5)	-	(0.3)	4	(6.7)	5	(8.7)
Phenylalanine	PHE	F	-	(0.6)	-	(0.5)	-	(1.1)	1	(1.1)	-	(0.5)	4	(14.4)	5	(18.6)
Leucine	LEU	L	-	(1.5)	-	(1.1)	-	(2.6)	-	(2.5)	-	(1.2)	**5**	**(30.8)**	**5**	**(40.0)**
Methionine	MET	M	1	(0.4)	-	(0.3)	-	(0.7)	-	(0.7)	-	(0.3)	3	(9.1)	4	(11.8)
Tryptophan	TRP	W	-	(0.3)	-	(0.2)	-	(0.7)	-	(0.6)	-	(0.3)	3	(8.7)	3	(11.3)
Proline	PRO	P	-	(3.5)	1	(2.7)	-	(6.0)	-	(5.7)	-	(2.6)	-	**(70.0)**	**1**	**(90.9)**
Total			53	(42.5)	**56**	**(33.0)**	66	(73.4)	**180**	**(69.4)**	21	(32.2)	**735**	**(860.3)**	1,111	(1,111)

Table 1. The table shows the number of observed hydrogen bonds between amino acids and DNA (backbone and bases). In parentheses, the possible numbers of hydrogen bonds between amino acids and DNA that would be expected from purely random docking are shown (Luscombe et al., 2001).

The proper orientation and position of amino acids are essential to enable the protein-DNA interaction. The following common structural motifs hold the amino acids in the right position:

- Helix-turn-helix: The helix-turn-helix motif is evolutionarily conserved. Originally, the helix-turn-helix motif was identified in bacterial transcription regulators. This motif consists of two α helices, which are connected by a short chain of amino acids (turn). The interactions between the helices hold them together at a fixed angle. The C-terminal helix is the recognition site of the motif, which binds to the major groove. The amino acids of these helices (C-terminal helices) are variable among helix-turn-helix proteins. These amino acids determine the recognized DNA sequence (Aravind, Anantharaman, Balaji, Babu, & Iyer, 2005). These proteins usually bind DNA as dimers. MYB and TRF1/2 are tri-helical transcription factors (**Figure 3A**) (Hanaoka, Nagadoi, & Nishimura, 2005; Ogata et al., 1992).
- Homeodomain: Investigations into homeotic selector genes in fruit flies led to the first discovery of the homeodomain motif. Homeotic selector genes encode DNA binding proteins, which orchestrate fly development. The homeodomain motif consists of three tightly packed α helices. Helix 2 and 3 form a helix-turn-helix. Helix 3 binds to the major groove, while helix 1 forms contacts with the minor groove (**Figure 3B**) (Yonghong Zhang, Larsen, Stadler, & Ames, 2011).
- Leucine zipper motif: Proteins with leucine zipper motifs bind the DNA as a dimer. The α helices of the leucine zipper motif form a Y-shaped coiled-coil, which attaches to the major groove like a clothespin. The name derives from the hydrophobic bound between amino acid side chains (often leucines) of the connecting proteins (**Figure 3C**) (Glover & Harrison, 1995).

- β sheet DNA recognition proteins: α helices are primarily used to recognize DNA sequences. A large group of transcription factors binds the DNA with two-stranded β sheets (**Figure 3D**) (Tian et al., 2016).
- Zinc finger proteins: Zinc finger proteins have several distinct structural groups. However, they have the following basic structure in common: a zinc atom holds an α helix and a β sheet together. Zinc fingers often form a continuous stretch with α helices. These motifs bind to the major groove as well. CTCF (CCCTC-binding factor) DBD contains 11 zinc finger domains (**Figure 3E**) (Yin et al., 2017).
- Helix-loop-helix: This protein motif is related to the leucine zipper. Two short α helices connect to each other, and both helices are connected to a longer α helix by a loop. Since most of these proteins are heterodimeric, their activity is highly regulated by the dimerization of subunits (Amoutzias, Robertson, Oliver, & Bornberg-Bauer, 2004). Examples: ARNT2, ARNTL, bHLH proteins, BMAL1, CLOCK HEY cluster, MAX, NEUROG1, and MyoD (**Figure 3F**) (Z. Wang, Wu, Li, & Su, 2013).

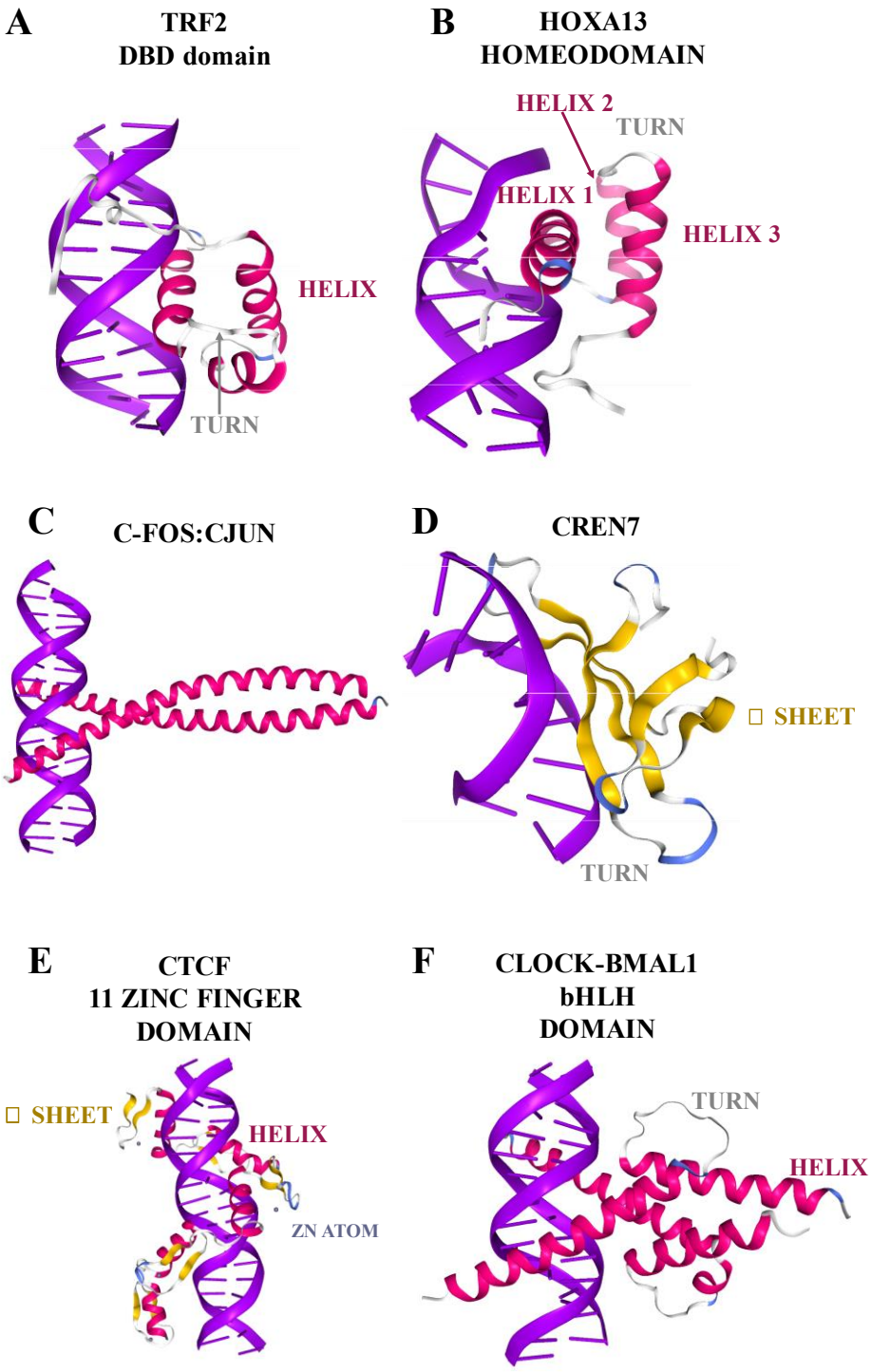


Figure 3. 3D structures of DNA binding motif-DNA complexes. The DBDs are represented by crystal structures of transcription factor examples from the RCSB PDB database: TRF2 (1VFC); HOXA13(2LD5); C-FOS:C-JUN (1FOS); CREN7 (5K07); CTCF (5YEG); CLOCK (4H10) (Berman et al., 2000; Prlić et al., 2018). A) Solution Structure of the DNA Complex Of Human Trf2 (Hanaoka et al., 2005). B) Structure of HOXA13 DNA binding domain (Yonghong Zhang et al., 2011). C) C-FOS:C-JUN:DNA complex (Glover & Harrison, 1995). D) Double Helix Bound by CREN7 (Tian et al., 2016). E) Crystal structure of CTCF ZFs4-8 (Yin et al., 2017). F) Complex structure of human CLOCK-BMAL1 basic Helix-Loop-Helix domains with E-box DNA (Z. Wang et al., 2013).

The previously described structural groups are only a schematic classification of transcription factors. A closer investigation is required for more exact clustering (**Table 2**).

TFC Class No.	**Class description**	**TFC Family No.**	**Family description**
1. Basic domains			
1.1	bZIP Factors	1.1.1	Jun-related
		1.1.2	Fos-related
		1.1.3	Maf-related
		1.1.4	B-ATF-related
		1.1.5	XBP-1-related
		1.1.6	ATF4-related
		1.1.7	CREB-related
		1.1.8	C/EBP-related
		1.1.0	ZIP only
1.2	bHLH Factors	1.2.1	E2A-related
		1.2.2	MyoD / ASC-related
		1.2.3	Tal-related
		1.2.4	Hairy-related
		1.2.5	PAS-domain factors
		1.2.6	bHLH-ZIP factors
		1.2.8	HLH only factors
1.3	bHSH Factors	1.3.1	AP-2
2. Zinc-coordinating DNA-binding domains			
2.1	C4 (nuclear receptor type)	2.1.1	Steroid hormone receptors (NR3)
		2.1.2	Thyroid hormone receptor-related factors (NR1)
		2.1.3	RXR- related receptors (NR2)
		2.1.4	NGFI-B- related receptors (NR4)
		2.1.5	FTZ-F1- related receptors (NR5)
		2.1.6	GCNF-related receptors (NR6)
		2.1.7	DAX- related receptors (NR0)
2.2	C4 (other)	2.2.1	GATA-type zinc fingers

2.3	C2H2	2.3.1	Three-zinc finger Krüppel-related factors
		2.3.2	Other factors with up to three adjacent zinc fingers
		2.3.3	More than 3 adjacent zinc finger factors
		2.3.4	Factors with multiple dispersed zinc fingers
		2.3.5	BED zinc finger factors
2.5	DM-type zinc finger factors	2.5.1	DMRT
2.6	CXXC	2.6.1	CpG-binding proteins
2.7	C2HC	2.7.1	Myelin transcription factor-related proteins
		2.7.2	Friend of GATA proteins
		2.7.3	Histone acetyltransferases with C2HC zinc finger
		2.7.4	Lethal(3)malignant brain tumor- related proteins
		2.7.5	LYAR- related proteins
2.8	C3H	2.8.1	ZC3H8-related factors
		2.8.2.	ZGPAT-related factors
		2.8.3	RC3H-related factors
		2.8.4	CNOT4-related factors
2.9	C2CH THAP	2.9.1	THAP-related factors
3. Helix-turn-helix domains			
3.1	Homeo	3.1.1	HOX-related factors
		3.1.2	NK-related factors
		3.1.3	Paired-related HD factors
		3.1.4	TALE-type homeo domain factors
		3.1.5	HD-LIM factors
		3.1.6	HD-SINE factors
		3.1.7	HD-PROS factors
		3.1.8	HD-ZF factors
		3.1.9	HD-CUT factors
		3.1.10	POU domain factors
		3.1.11	PHTF factors
3.2	Paired	3.2.1	Paired plus homeo domain

		3.2.2	Paired domain only
3.3	FKH / WH	3.3.1	Forkhead box (FOX) factors
		3.3.2	E2F-related factors
		3.3.3	RFX-related factors
3.4	HSF	3.4.1	HSF factors
3.5	W cluster	3.5.1	Myb/SANT domain factors
		3.5.2	Ets-related factors
		3.5.3	Interferon regulatory factors
3.6	TEA	3.6.1	TEF-1-related factors
3.7	ARID	3.7.1	ARID-related factors
4. All-alpha helical DNA-binding domains			
4.1	HMG	4.1.1	Sox-related factors
		4.1.2	Tox-related factors
		4.1.3	TCF-7-related factors
		4.1.4	PBRM1-related factors
		4.1.5	WHSC1-related factors
		4.1.6	UBF-related factors
		4.1.7	TFAM
4.2	CCAAT-BF	4.2.1	Heteromeric CCAAT-binding factors
5. Alpha-Helices exposed by beta-structures			
5.1	MADS	5.1.1	Regulators of differentiation
		5.1.2	Responders to external signals
5.3	SAND	5.3.1	AIRE factors
		5.3.2	DEAF factors
		5.3.3	GMEB factors
		5.3.4	Sp110 factors
		5.3.5	Sp140/Sp100-related factors
6. Immunoglobulin fold			
6.1	RHR	6.1.1	NF-kappaB-related factors
		6.1.2	Ankyrin domain-only factors
		6.1.3	NFAT-related factors
		6.1.4	CSL-related factors
		6.1.5	Early B-Cell Factor-related factors

6.2	STAT	6.2.1	STAT factors
6.3	p53	6.3.1	p53-related factors
6.4	Runt	6.4.1	Runt-related factors
6.5	T-box	6.5.1	Brachury-related factors
		6.5.2	TBrain-related factors
		6.5.3	TBX1-related factors
		6.5.4	TBX2-related factors
		6.5.5	TBX6-related factors
6.6	NDT80	6.6.1	Myelin gene regulatory factor-related factors
6.7	Grainyhead	6.7.1	Grainyhead-related factors
		6.7.2	CP2-related factors
7. Beta-Hairpin exposed by an alpha/beta-scaffold			
7.1	SMAD/NF-1	7.1.1	SMAD factors
		7.1.2	Nuclear factor 1
7.2	GCM	7.2.1	GCM factors
8. Beta-sheet binding to DNA			
8.1	TATA	8.1.1	TBP-related factors
8.2	A.T hook	8.2.1	HMGA factors
9. Beta-Barrel DNA-binding domains			
9.1	CSD	9.1.1	Dbp factors
0. Unidentified DNA-binding domains			
0.1	AXUD	0.1.1	AXUD/CSRNP domain factors
0.2	NonO	0.2.1	NonO-related factors
0.3	LRRFIP	0.3.1	LRRFIP factors
0.4	NFX1	0.4.1	NFX1
0.0	Uncharacterized	0.0.1	Nuclear localized protein 1
		0.0.2	PHF5
		0.0.3	RFXANK
		0.0.4	RFXAP
		0.0.5	PUR

Table 2. Classification of eukaryotic transcription factors based on the characteristics of their DNA-binding domains.

(Wingender, Schoeps, Haubrock, Krull, & Dönitz, 2018)

1.1.3. Transcription factor interactions

Generally, transcription factors do not act independently, but form complexes with other factors, including other transcription factors, coregulators, and chromatin remodeling complexes. Transcription factors exist predominantly as monomers in the cell, but they have low affinity for DNA. Two transcription factors can form dimers with noncovalent bonds. Depending on the cooperation partner, the dimers can be homo- or heterodimers. Dimers are generally weakly bound, but dimerization causes activation and the transcription factors bind to DNA cooperatively. Approximately 75% of metazoan transcription factors heterodimerize with other factors (Walhout, 2006). The contingency of heterodimerization is not just a matter of capability for connection but also the presence in time and space, i.e. the presence in a particular cell.

A large set of transcription factors are called “facilitators”. Facilitators are widely expressed because they are needed to facilitate transcriptional programs across many different tissues. In contrast, high specificity transcription factors called “specifiers” are expressed only in specific cell types. Generally, specifiers have well known roles in tissue differentiation, as lineage-determining factors (Ravasi et al., 2010).

The PPAR-RXR heterodimer is a well-studied example of a heterodimer. Peroxisome Proliferator-Activated Receptor Gamma (PPARγ) is restricted to adipose tissue, skin, lung, and breast (Tyagi, Gupta, Saini, Kaushal, & Sharma, 2011). Retinoid X Receptor Beta (RXRβ) is expressed ubiquitously (Watanabe & Kakuta, 2018). The interaction between RXR and PPARG is required for the regulation of adipocyte differentiation (Lefterova et al., 2010).

1.1.4. Responsive elements

Regulatory regions of the genome control the transcriptional activity of genes (Whitaker, Chen, & Wang, 2015). These regulatory sites tether protein complexes to modify

chromatin structure and initiate transcription. The regulatory regions contain specific sequences called responsive elements, which can be recognized by transcription factors to establish the protein complex recruitment (Whitaker et al., 2015).

There are different types of regulatory regions. The most ancient type of regulatory region is the promoter region, which can be found on the 5' end of the gene. The promoter region participates in the assembly of the pre-initiator complex (PIC) at the transcription start site (TSS) (Lagrange et al., 1998). The PIC assembly can be promoted or restricted in the presence of other regulator proteins, which have responsive elements in promoters (Fulton et al., 2009).

Distal DNA regions influence gene expression as well. Enhancers are located hundreds or thousands of base pairs upstream or downstream of the transcription start site. Enhancers bind transcription factors as well, which affect the corresponding gene expression (Blackwood & Kadonaga, 1998). The mechanism of closure or looping between remote DNA regions is not completely understood yet. The mediator complex may be involved in this process (Whyte et al., 2013). Recent publications about topologically associated domains and CTCF mediated chromatin looping suggest that there is a quaternary chromatin structure, which allows the enhancer-promoter regions to get close to each other (Rao et al., 2014).

Multiple enhancers can form a super-enhancer, which is collectively bound by an array of transcription factors (Blackwood & Kadonaga, 1998). Super-enhancers share the functions of enhancers.

1.1.4.1. Regulatory elements bound by dimers

Each transcription factor (which can form dimers) can be characterized by a half-site (monomer binding site). Half-sites are supplemented by another half-site. Together, these sites are called response elements (RE). The configuration of the monomer-binding sites within RE

is inherent to the heterodimerization partners. In the case of nuclear receptors, which are closely related, they often recognize similar half-sites (Pawlak, Lefebvre, & Staels, 2012). The strand specificities (orientation) of half-motifs relative to each other can be in a strand specific manner (Forman, Casanova, Raaka, Ghysdael, & Samuels, 1992):

- Direct repeat (DR): ⇒⇒
- Inverted repeat (IR): ⇒⇐
- Everted repeat (ER): ⇐⇒

In addition to the orientation, the size of the spacer is significant. The spacer is measured in nucleotides. In the naming of the responsive element of a given dimer, we use this nomenclature: a direct repeat (both motifs are located on the same strand) with a 5-bp spacer is referred to as DR5 (Harbers, Wahlström, & Vennström, 1996). Table 3 shows the classified nuclear receptors and their preferred binding sites (hormone responsive elements (HREs)) (**Table 3**).

Trivial name	**Short name**	**Half site sequence**	**configuration of HRE**
Retinoid X receptor	RXR	AGGTCA	DR1
All-trans retinoic acid receptor	RAR	AGGTCA	IR0, DR2, DR5, ER8
Androgen receptor	AR	AGAACA	IR3
Chicken ovalbumin upstream promoter transcription factor	COUP-TF	RGGTCA	DRs, IRs
Estrogen receptor	ER	RGGTCA	IR3
Farnesoid receptor	FXR	AGGTCA	IR1, DR5

Glucocorticoid receptor	GR	AGAACA	IR3
Mineralocorticoid receptor	MR	AGAACA	IR3
Peroxisome proliferator activated receptor	PPAR	AGGTCA	DR1
Pregnane X receptor	PXR	AGTTCA	DRs
Progesterone receptor	PR	AGAACA	IR3
RAR-related orphan receptor	ROR	XXCYRGGTCA	NR
Thyroid receptor	TR	RGGTCA	IR0, DR4, ER6, ER8
Vitamin D3 receptor	VDR	AGGTCA	DR1

Table 3. Nuclear receptors and their responsive elements. The table shows hormone receptors as examples to demonstrate the orientation of half-sites within hormone responsive elements (Shu, Sidell, Yang, & Kallen, 2010; Umesono, Murakami, Thompson, & Evans, 1991). The configuration of HRE is named according to the following nomenclature: direct repeat (DR); inverted repeat (IR); everted repeat (ER). The numbers, which follows the abbreviations, mark the size of spacers measured in nucleotides.

1.1.4.2. Composite elements

Functional interactions can be observed between closely related transcription factors in the phenomenon of heterodimerization. Two or more closely situated binding sites provide a way for crosstalk between distinct transcription factors. These functional units are called "composite elements" (Diamond, Miner, Yoshinaga, & Yamamoto, 1990; Kel-Margoulis, Kel, Reuter, Deineko, & Wingender, 2002). Composite elements can be synergistic or antagonistic. GATA1 and TAL1 cooperation is a good example of a synergetic relationship. Interaction of GATA1 and TAL1 was originally identified in Cas-Br-E and Akt viruses (Barat & Rassart, 1998). GATA and TAL1 can be found in the same multiprotein complexes. Their motifs are

juxtaposed along the DNA with an approximately 8 base pair spacer between them. This GATA-E-box is published and represented in the JASPAR CORE motif database. GATA and TAL1 act together as key transcription factors during erythroid development and do not interact directly. LMO2 acts as a mediator between them (**Figure 4**). The GATA-TAL collaboration is so significant in hematopoiesis, that their relationship is conserved at the DNA sequence level (Han et al., 2015; Love, Warzecha, & Li, 2014).

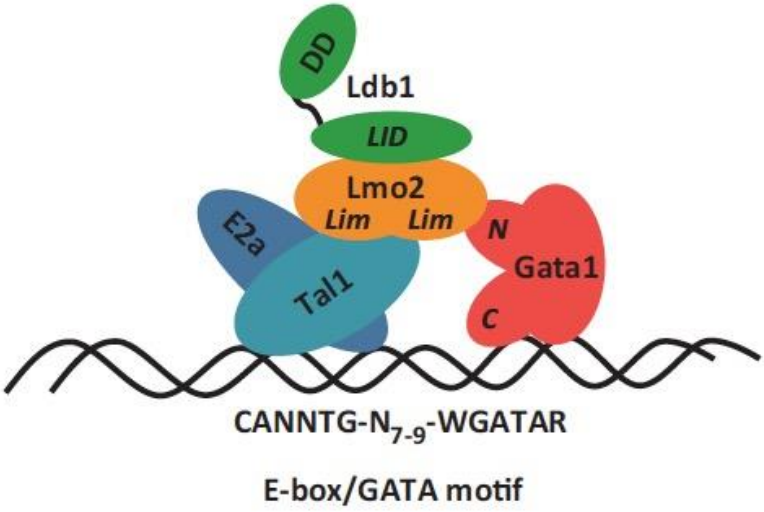


Figure 4. Structural model of erythroid Ldb1 complexes. The core erythropoietic Ldb1 complex is anchored to the E-box (WGATAR) by GATA1 and a heterodimer of TAL1 and E2A (Love et al., 2014).

In antagonistic composite elements, two transcription factors interfere with each other. Several mechanisms are involved in this interference including:

- Competition for overlapping binding sites (Casolaro et al., 1995)
- Mutually exclusive binding (structural barrier) (Takeuchi et al., 1998)
- Simultaneous binding, with repression (Diamond et al., 1990)

The competition between RORα and Rev-erbα is a good example of an antagonistic composite element. Both transcription factors are involved in circadian rhythmicity and regulate the transcription of ZT10, ZT22, and Bmal2 genes. Deletion of RORα or Rev-erbα promotes the other's recruitment to the promoter regions. Conversely, hepatic overexpression of Rev-erbα reduces RORα recruitment to *Bmal1* and *Npas2* sites in Zt22 (Yuxiang Zhang et al., 2015).

The function of interactions cannot be described only as synergistic or antagonistic. The transcription factor network is much more sophisticated and depends highly on other external factors (**Figure 5**). The interaction between two factors can be both synergistic and antagonistic depending on the circumstances.

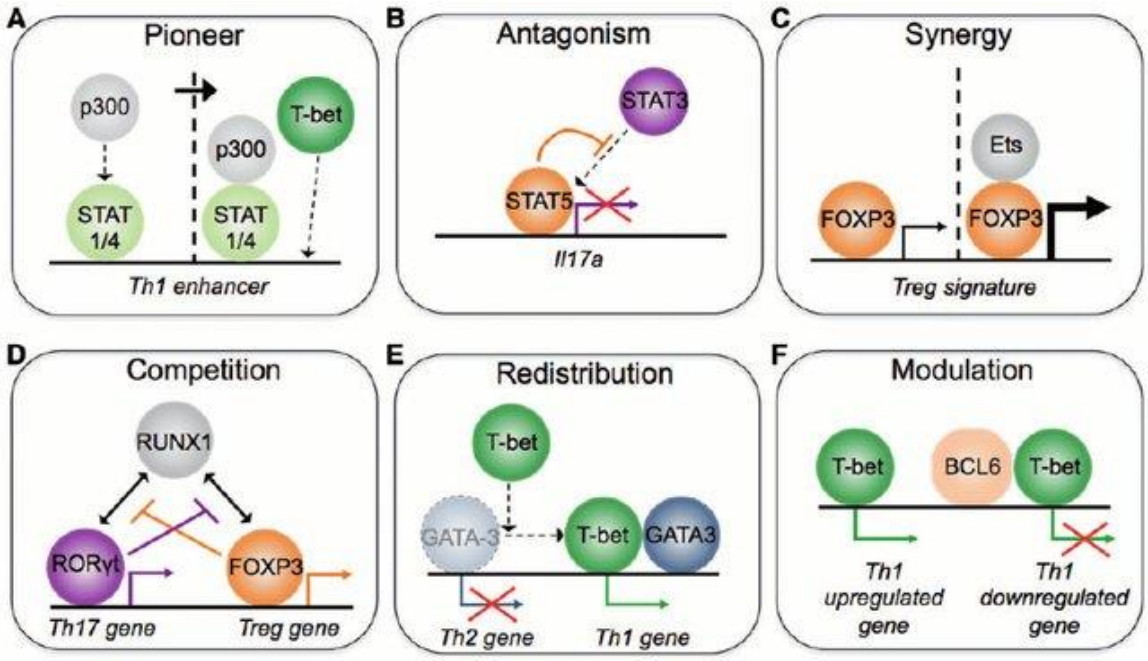


Figure 5. Transcription factor interactions. A) Pioneering transcription allows other proteins to bind to responsive elements. STAT1 and STAT4 grant the juxtaposition of T-BET at the Th1 enhancer. B) Transcription factors can prevent each other's binding in antagonism. STAT5 block STAT3 binding at the Il17a locus. C) FOXP3 proteins synergize with each other and the Ets co-activator to promote T cell differentiation. D) RORγ and FOXP3 compete to interact with RUNX1. E) Redistribution of transcription factors to new sites. T-bet segregates GATA3 from target genes (such as Th1) and readjusts it to its target genes. F) Modulation of transcription factor activity by other factors. The activity of a transcription factor can be modulated by other factors: BCL6 represses T-bet activity, but synergizes with other T-bet to upregulate gene expression again. (Evans & Jenner, 2013)

1.1.4.3. Indirect binding of transcription factors

Although transcription factors have their own DNA-binding domains, they are not bound in all circumstances to their own responsive element. The connection of a transcription factor to the DNA might be indirect through another transcription factor. The motif of liver-lineage determining TF hepatocyte nuclear factor 6 (HNF6) is commonly enriched under

predicted Rev-erbα binding sites. In contrast, these regions are not present on Rev-erbα responsive elements. Conditional deletion of the Rev-erbα DNA-binding domain has no effect on the occupancy of the protein in these DNA regions. However, a regulatory SNP (rs50735045) in the HNF6 binding site dramatically decreased the ChIP-qPCR signal of Rev-erbα compared to wild-type. This suggests that HNF6 might tether Rev-erbα to the DNA, even in the absence of Rev-erbα binding (Y Zhang et al., 2008).

As mentioned previously, transcription factors do not act independently, but form multiprotein complexes with other proteins. This is even more necessary if the transcription factor does not have enzymatic activity. The role of non-enzymatic transcription factor does not have enzymatic activity. The role of non-enzymatic transcription factors is regulatory region recognition, binding, and serving as a landing platform for other proteins. In a complex interaction with enzymes that modify chromatin structure, mostly by methylation and acetylation events, these transcription factors support the formation of the transcription preinitiation complex and direct the RNA polymerase to the transcription start site (TSS) (Wingender et al., 2015).

1.1.5. Co-regulators and histone modifications

Most of the cofactors are enzymes that mediate post-translational modification of target proteins. This group of enzymes is extremely diverse in structure and enzymatic activity. Most of these enzymes catalyzing post transcriptional modification to enact the transfer of specific functional groups (acetylation (lysine), methylation (lysine and arginine), phosphorylation (serine and threonine), sumoylation (lysine), ubiquitylation (lysine), ADP ribosylation, butyrylation, citrullination, crotonylation, formylation, proline isomerization, and propionylation) to histones or other regulator proteins (Müller & Muir, 2015). For instance (Marmorstein & Trievel, 2009):

- cyclin-dependent kinases (CDKs) CDK7 of the TFIIH complex
- CDK8 of the Mediator complex catalyze serine-phosphorylation
- acetyl-transferase p300 lysine-acetylation
- COMPASS proteins are methyl-transferases
- HDAC1 and HDAC2 are histone deacetylases

Biochemical changes made to specific histone tails are associated with different condensation levels of chromatin (Bannister & Kouzarides, 2011). The mono-methylation of lysine residue 9 on histone 3 (H3K9me) is the attribute of heterochromatin. This kind of modification of the histone tail is called histone code. The SUV39H-family proteins catalyze H3K9 trimethylation. H3K9me is an attribute of gene-silencing and heterochromatin. This modification is recognized by heterochromatin protein 1 (HP1) and keeps the chromatin in a condensed state. This leads to constant silencing of the affected DNA region (Zeng et al., 2010). Because of these observations, biologists hypothesized that hypermethylation of histones correlates with inactive DNA regions and hyperacetylation leads to decreased heterochromatin quantity. This hypothesis sounds logical, because acetylation removes the positive charge on histones, thereby decreasing the interaction of the N terminal of histones with the negatively charged phosphate backbone of DNA (Bowman & Poirier, 2015). On this basis, the methylation of histone proteins should reinforce the folding of DNA around nucleosomes. However, with the development of new molecular biology techniques, this hypothesis did not hold up. For instance, the H3K4me3 is strongly enriched in active promoter regions (Sims, Nishioka, & Reinberg, 2003).

The charge neutralization is an existing phenomenon with respect to nucleosome stability. The protein core of the nucleosome has a charge of +144e. The DNA charge is -294. Therefore, the nucleosome has an overall residual charge of −150e. In vitro chromatin analysis with small-angle X-ray scattering assays demonstrate that H4K16ac may lead to a massive

disruption in dense chromatin fibers (Cortini et al., 2016). In vivo, no single modification determines the net charge between DNA and histone proteins. A combination of marks works in synergy to provide sufficient proofreading so that no gene is accidentally turned on or off (Prakash & Fournier, 2017).

Not all proteins in the recruited protein complexes (to any bound transcription factors) have transferase activity. Several factors serve as structural components in complexes to help the recruitment, provide a surface for binding, or structural components to interrupt DNA organization (e.g. in CTCF mediated chromatin looping, SMC proteins embrace DNA strands to get two distal DNA regions close to each other). RE1 silencing transcription factor (REST) or CoREST proteins are described as repressors. Both proteins can connect histone deacetylase (HDAC) complexes. The N-terminal domain of the REST recruit, mSin3 HDAC, and COREST connect to the C-terminal domain. CoREST is a component of the HDAC complex, which is independent from mSin3 (You, Tong, Grozinger, & Schreiber, 2001). As we see, these proteins have indirect, but key roles in gene repression.

A wide spectrum of transcription factors associate to form different patterns of protein complexes. These complexes have a specific function in the nucleus. Previously, we discussed the machinery of transcription regulation. In the following sections, we will describe the mechanism of chromatin looping.

1.1.6. Topologically associated domains, CTCF mediated chromatin looping, and insulators

The unfolded chromatin from one single human cell would measure 2 meters end to end. This extensive length of DNA needs to be packed inside the nucleus, which is approximately 5 μm in diameter. The major features of chromosome architecture are barely known. The first level of chromosome packing is the previously described nucleosome. Nucleosomes wrap about 146 base pairs around a histone octamer. This DNA form is called

“beads on a string” according to its appearance under an electron microscope. A nucleosome is about 11 nm wide. To make the chromatin structure more compact, approximately 30 nm “zigzag” chromatin fibers are formed. These fibers need further packing; meanwhile, functionally active and inactive DNA needs to be separated (Anthony T. Annunziato, 2008).

More than a century ago, it was already hypothesized that chromosomes in interphase nuclei exist in discrete chromosome territories (Boveri, 1909; Rabl, 1885). The 3C (chromosome conformation capture) and the High Throughput Sequencing technologies provides the possibility to identify interaction (genome-wide) between two or more distal DNA regions. Due to these techniques, we have clear evidence to confirm the mentioned hypothesis. In interphase nuclei, the chromatin is divided into active and inactive territories, which highly depend on the organization of chromosomes.

There are several models of the large-scale organization of chromatin. Fractal globule (extended to tension globule) and equilibrium globule are the most avowed models. The definition of a fractal globe is “a compact polymer state that emerges during polymer condensation as a result of topological constraints which prevent one region of the chain from passing across another one” (Mirny, 2011). When applied to chromatin, this means that all chromosomes reside in a well-defined position of the nucleus without intermingling with other chromosomes. The equilibrium globule allows chromosome chains to pass across another chromosome (**Figure 6**). Both models can be confirmed with High-throughput 3C methods (HiC).

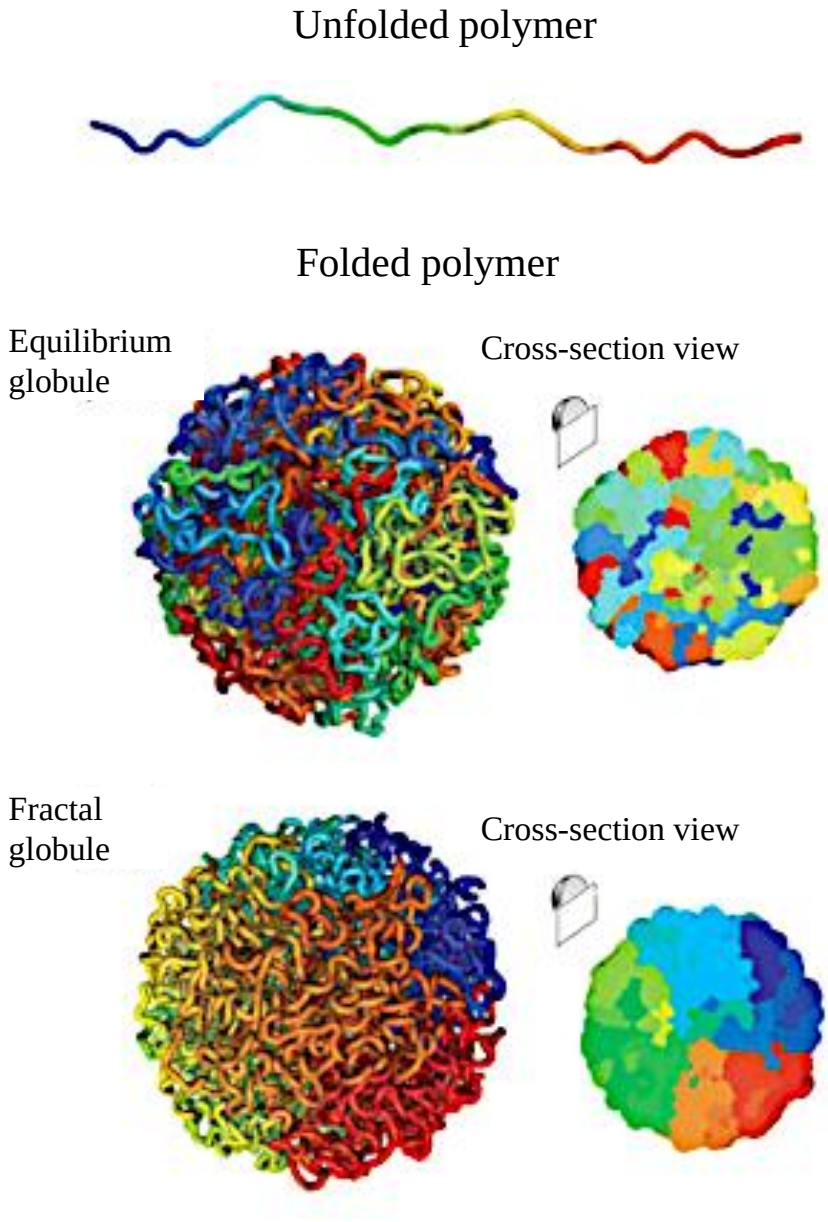


Figure 6. Visual demonstration of the difference between fractal and equilibrium globules. "Coloration corresponds to distance from one endpoint, ranging from blue to cyan, green, yellow, orange, and red. Middle: Typical example of a fractal globule drawn from our ensemble. Fractal globules lack entanglements. Loci that are nearby along the contour tend to be nearby in 3D, leading to the presence of large monochromatic blocks that are apparent on the surface and in cross-section. Bottom: An equilibrium globule. The structure is highly entangled; loci that are nearby along the contour (similar color) need not be nearby in 3D." (van Berkum et al., 2010)

Chromosome conformation capture techniques clarified the presence of chromatin domains or topologically associated domains (TAD). TADs represent hundreds of kilobases to several million bases in length. These domains are evolutionarily conserved (in related species)

and stable for many cell divisions (Dekker & Heard, 2015; Sexton & Cavalli, 2015). Their role is not completely understood, but there is convincing experimental evidence to support a simultaneous "insulator" and "co-regulator" feature. TADs may create autonomous gene regulatory domains, where genes share coordinated gene expression profiles within the same TAD (Flavahan et al., 2016; Nora et al., 2012). TADs also block the spread of activity between neighboring TADs (Grubert et al., 2015). Smaller functional units can be observed within TADs, called sub-TADs or loops (Rao et al., 2014). Sub-TADs are regions that display both self-associative and insulating properties, similar to TADs. However, what makes these domains functionally distinctive? Their length is too obvious an answer to this question. TADs are largely invariant features of genome organization that show high conservation profiles between cell types. In contrast, sub-TADs appears to vary between different cell lineages (Dixon, Gorkin, & Ren, 2016). The general mechanism of TAD formation is not fully understood. CTCF (CCCTC-binding factor) and cohesin complexes mediate chromatin looping, a physical process that brings two distal DNA regions close to each.

1.1.6.1. CTCF

For a long time, investigators speculated that chromosomes are segregated into a series of discrete, large-scale looped domains to form "insulated neighborhoods" (Udvardy, Maine, & Schedl, 1985). While this structure provides autonomous gene regulation within a domain, interdomain interactions are highly restricted/insulated. The "insulator" DNA elements block activation of a promoter by an enhancer. In 1997, the CTCF protein was shown to serve as an insulating boundary element for the chicken β-globin gene (Chung, Bell, & Felsenfeld, 1997). First, a repressor of chicken C-MYC gene was described in 1990. In these experiments, DNA footprinting analysis revealed strong DNA-protein interactions in about 200 base pairs upstream of the C-MYC transcription start site. The footprint sequence had three direct repeats

of the CCCTC sequence, which is responsible for the binding event. Therefore, the protein was designated as a CCCTC-binding factor (CTCF) (Lobanenkov et al., 1990).

High-throughput 3C methods (Hi-C) revealed the significance of the CTCF-mediated subdivision of chromatin into TADs. CTCF binding sites demarcate the individual TAD boundaries and other chromatin loop borders. The CTCF binding motif is well characterized (**Figure 7A**) and we know that there are ~80 000 sites in mammalian genomes (**Figure 7B**) (Chen, Tian, Shu, Bo, & Wang, 2012). CTCF serves as an anchor, which recognizes and binds to the DNA sequence while making connections with the proteins responsible for DNA loop formation.

The exact structure of the CTCF protein is only superficially elucidated. The complete CTCF protein consists of 727 amino acids and is approximately 82.80 kDa (Quitschke, Taheny, Fochtmann, & Vostrov, 2000). Approximately 30% of the protein structure has been determined (between 348 and 580 amino acids) (**Figure 7D**). However, from X-ray crystallographic studies, the mechanism of DNA binding is known. The DNA binding domain of CTCF contains a tandem array of eleven Cys2-His2 (C2H2) zinc fingers (ZFs), numbered from N-terminal to C-terminal (ZF1-ZF11). The fingers interact exclusively with the major groove. ZF3–7 are responsible for base-specific contacts, where ZF7 interacts with the 5' sequence (CCA), ZF6 with the second triplet (GCA), ZF5 with the third triplet (GGG), ZF4 with the fourth triplet (GGC), and ZF3 with the 3' sequence (GCT) (**Figure 7C-D**). ZF2 follows the major groove, but the interacting side chains are positioned too far away to make base-specific hydrogen bonds. ZF8-ZF11 have no effect on the binding of the core sequence, but the presence of ZF8 increase the non-specific binding (Hashimoto et al., 2017). Because of the high specificity of binding, the interaction can be disrupted by DNA methylation. Especially in the case of the aspartate residue 451 (D451) on ZF4 and the invariant cytosine at position 2 in the motif (**Figure 7B**), the presence of a methyl group at C2 sterically obstructs D451 in the

cytosine specific conformation and drastically reduces the binding affinity. In contrast, the methylation of the invariant cytosine at positions 12 does not interfere with the conformation of E362 (glutamine 362 on ZF4). This can be explained by the amino acid side chain difference between glutamine and aspartate. Glutamine shows a slightly increased affinity for methylated cytosine (Hashimoto et al., 2017). This demonstrates the dual role of DNA-methylation inside and outside of the CTCF-binding regions (Renaud et al., 2007). The CTCF also prevents spreading of methylation (De et al., 2013).

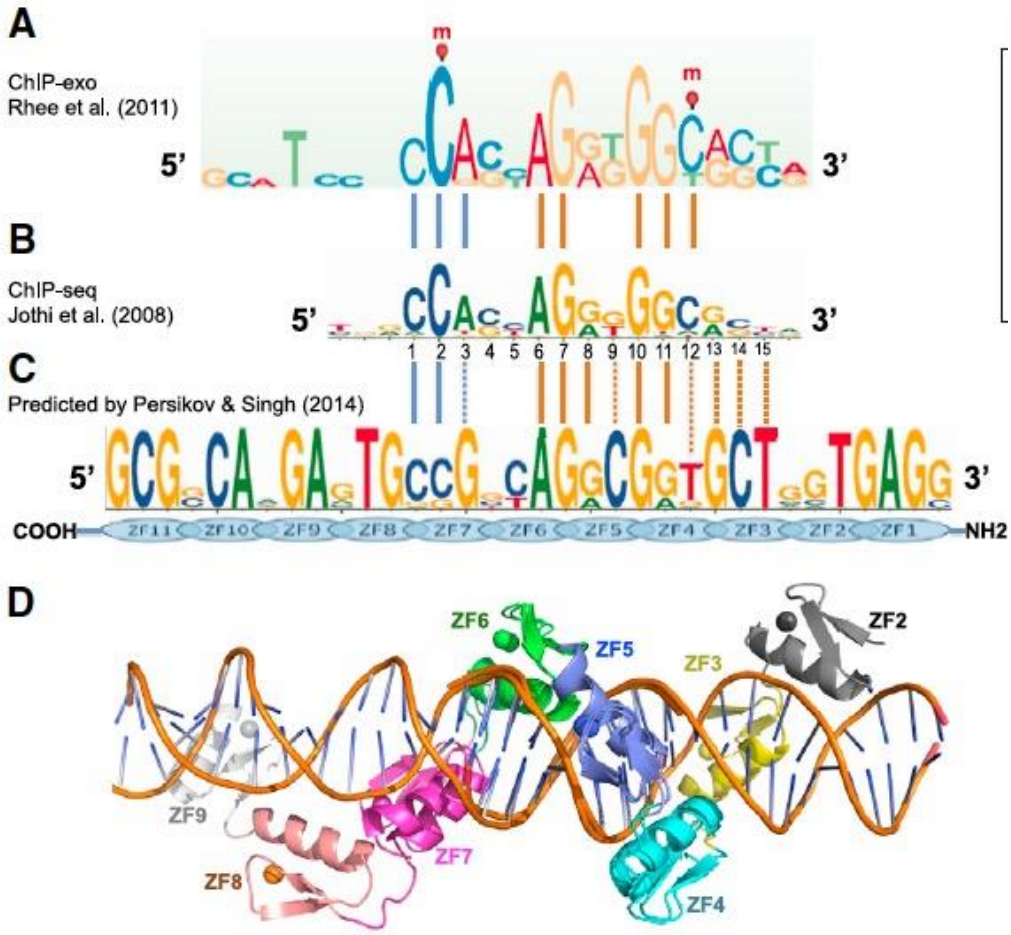


Figure 7. Mechanism of DNA binding of CTCF. A) Position weight matrix of CTCF-binding consensus sequence assigned using ChIP-exo. B) Position weight matrix of CTCF-binding consensus sequence assigned using ChIP-seq. C) Schematic interaction map between consensus sequence and 11 zinc fingers. D) Solution structure of the 11 zinc finger domain- DNA complex (Hashimoto et al., 2017).

CTCF has a wide variety of functions and the diversity of functions reflects the number of potential interacting partners. However, the mechanism of interaction and the interacting domains have not been characterized (because the vast majority of protein structure is not known). Several interacting partners have been identified (YY1, Scc1, Sin3, Kaiso, etc.) (Filippova, 2008; Ohlsson, Renkawitz, & Lobanenkov, 2001; Wallace & Felsenfeld, 2007). The discovery of a connection between CTCF and interphase cohesin broke new ground. Together with cohesin, CTCF can coordinate interactions between enhancers and their corresponding promoters by composing loops, while protecting interactions between sequences located inside and outside the loops (Dixon et al., 2012).

1.1.6.2.The Cohesin complex

The role of cohesin in cell division has been known for a long time. Cohesin mediates sister chromatid cohesion during chromosome segregation, prevents premature dissociation, and participates in post-replicative DNA repair (Mehta, Kumar, Srivastava, & Ghosh, 2013; Mehta, Rizvi, & Ghosh, 2012). The development of High-throughput Sequencing techniques revealed the frequent juxtaposition between CTCF and cohesin subunits in the interphase nucleus (Heidari et al., 2014). Numerous studies describe the correlation between cohesin and CTCF and the mechanism of CTCF-cohesin mediated chromatin loop formation.

The cohesin complex consists of Rad21 (Scc1), SA1/SA2 (Scc3, STAG1/STAG2), and Smc1/Smc3 (Feeney, Wasson, & Parish, 2010). The latter is the core of the complex, which forms a ~50 nm ring-like structure to embrace chromatin. SMC1 and SMC3 subunits fold back on each other through an antiparallel coiled-coil interaction or dimerize through a hinge–hinge interaction (Hirano, 2006). This structure is relatively rigid, with low bendability (Hirano, 2002; Shintomi & Hirano, 2007) (**Figure 8A**). The central hinge domain is crucial for the folding

reaction (Sun, Nishino, & Marko, 2013). The hinge has DNA binding and compaction capability (Gruber et al., 2006; Sun et al., 2013), which have been experimentally validated.

The heterotypic interaction between Smc1 and Smc3 dimerization domains creates a huge unclosed ring, a V-shaped structure (which is well seen on electron microscopic images) (**Figure 8B-C**) with ABC ATPase heads at the end of the long coiled-coil arms. The closure of the V-shape starts with the formation of a tripartite ring with Rad21. The C-terminal binds to the Smc1 head, while the N-terminal connects to the Smc3 ATPase domain. These three components form a closed circular structure, which maintains the attachment of sister chromatids from S phase until early anaphase (Brooker & Berkowitz, 2014). The fourth component, SA1/SA2, binds to RAD21 independently of the other subunits (Chan et al., 2003; Valdeolmillos et al., 2007). Furthermore, SA2 interacts with the C terminus of CTCF and, subsequently, the CTCF-SA2 complex recruits the cohesin ring in interphase cells (**Figure 8D**) (Xiao, Wallace, & Felsenfeld, 2011). The flexibility of chromatin fiber has been proven and the sliding of the cohesin ring on the chromatin fiber is also explained by the extrusion model (Sanborn et al., 2015). The cohesin ring moves along the DNA until it runs into a properly oriented CTCF (discussed in the results). Although much is known about this complex, the chain topology of closure is not known in sufficient detail. Current models disagree even on fundamental points, such as the number cohesin rings or number of DNA duplexes enclosed within a cohesin ring.

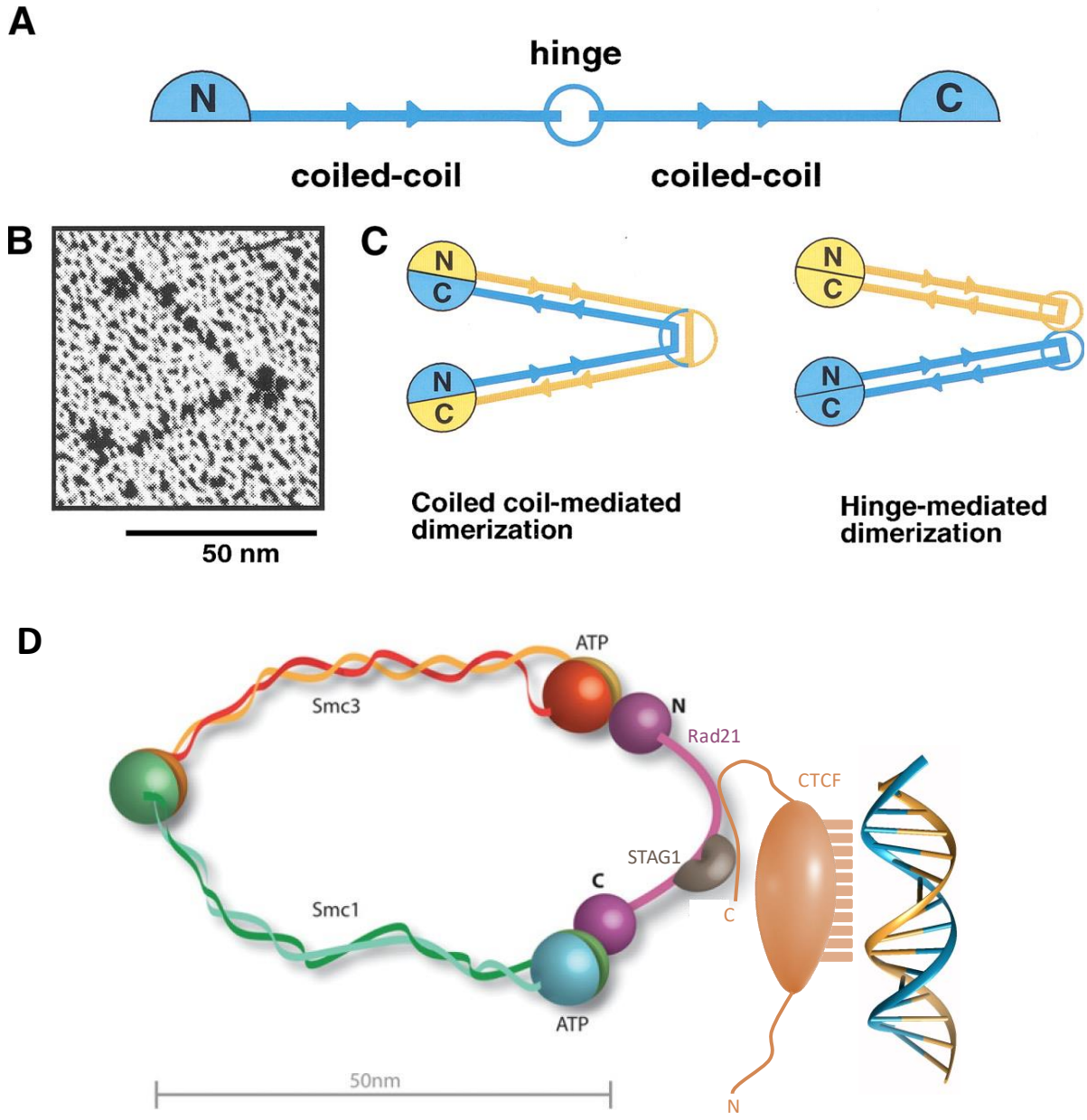


Figure 8. Structure of the cohesin complex. A) Schematic primary structure of SMC proteins. B) Electron micrographs of the Bacillus subtilis SMC homodimers with a visible 50 nm V-shape (Melby, Ciampaglio, Briscoe, & Erickson, 1998). C) Two models of SMC protein dimerization: SMC1 and SMC3 subunits fold back on each other through an antiparallel coiled-coil interaction or dimerize through a hinge–hinge interaction (Hirano, 2002). D) In the interphase nuclei, the 50 nm wide cohesin is anchored to the DNA by CTCF protein.

1.1.7. Investigation into transcriptional regulation with High-Throughput Sequencing Technologies

Research tools in functional genomics and molecular biology have developed by leaps and bounds over the last few decades. The development of high-throughput sequencing (HTS) technologies enabled relatively easy and rapid parallel DNA sequencing at a reasonable price. The combined techniques (ChIP-seq, RNA-seq, GRO-seq, ChIA-PET, HiC techniques, ATAC-seq, etc.) advanced genomic research on a global level, including gene expression profiling, chromosome counting, DNA-protein interactions, and detection of epigenetic changes. In this study, I am focusing on the processing and analysis of chromatin immunoprecipitation techniques, like ChIP-seq and ChIA-PET.

1.1.7.1. Chromatin immunoprecipitation followed by High-Throughput Sequencing (ChIP-seq)

The ChIP-seq technique involves the specific (for a protein or histone modification) antibody treatment to "fish out" the DNA-protein complex of interest, after cross-linking (with formaldehyde), followed by a random fragmentation (ultrasound sonication) step. After purification, the sequencing part of the analysis produces digitalized sequence data for millions of relatively short DNA fragments (50-100 nucleotide long) (**Figure 9A**) (Landt et al., 2012). The sequenced fragments are called sequence reads or tags. Since the fragments derive from a relatively homologous cell population, the isolated DNA fragments after ChIP are copies (with high numbers) of the same genomic regions with differing lengths and ends due to the random fragmentation. The alignment of raw reads needs to pass through many processing steps and format changes before they can be used in research. In the beginning, the reads are stored in FASTQ files, which store the identifiers (headers starts with ">" or "@" symbol), the sequences, and their quality scores. Overall, four rows belong to one read. The sequences are

aligned to a reference genome (in our case, it is the human genome hg19/GRCh37 assembly) in the mapping step, when the genomic location of reads is identified by Burrows-Wheeler algorithm (Li & Durbin, 2009). The resulted data are in SAM, BAM format (after a few transforming steps), which contains the sequence and its genomic coordinates (chromosome, start, end). This format is useful for numerous analyses, but the following "peak calling" is crucial for further investigations. Peak calling algorithms identify areas in a genome that have been enriched with aligned reads (Heinz et al., 2010; Y Zhang et al., 2008). The term "peak" is generally used to denote a loosely defined region, which predicts occupied regions by the ChIP target. The term derives from the bell-like shape of ChIP-seq signals after visualization in a genome browser (bedgraph file format) (**Figure 9B**) (Barta E., 2011). The identified "peak regions" have alternating fragment coverage within the peak, and the shape of histone mark and transcription factor peaks are largely different. Peaks for factors/cofactors are narrowly concentrated (**Figure 9B**), whereas histone peaks are broad and spread over a large region with altering peaks and valleys (**Figure 9C**).

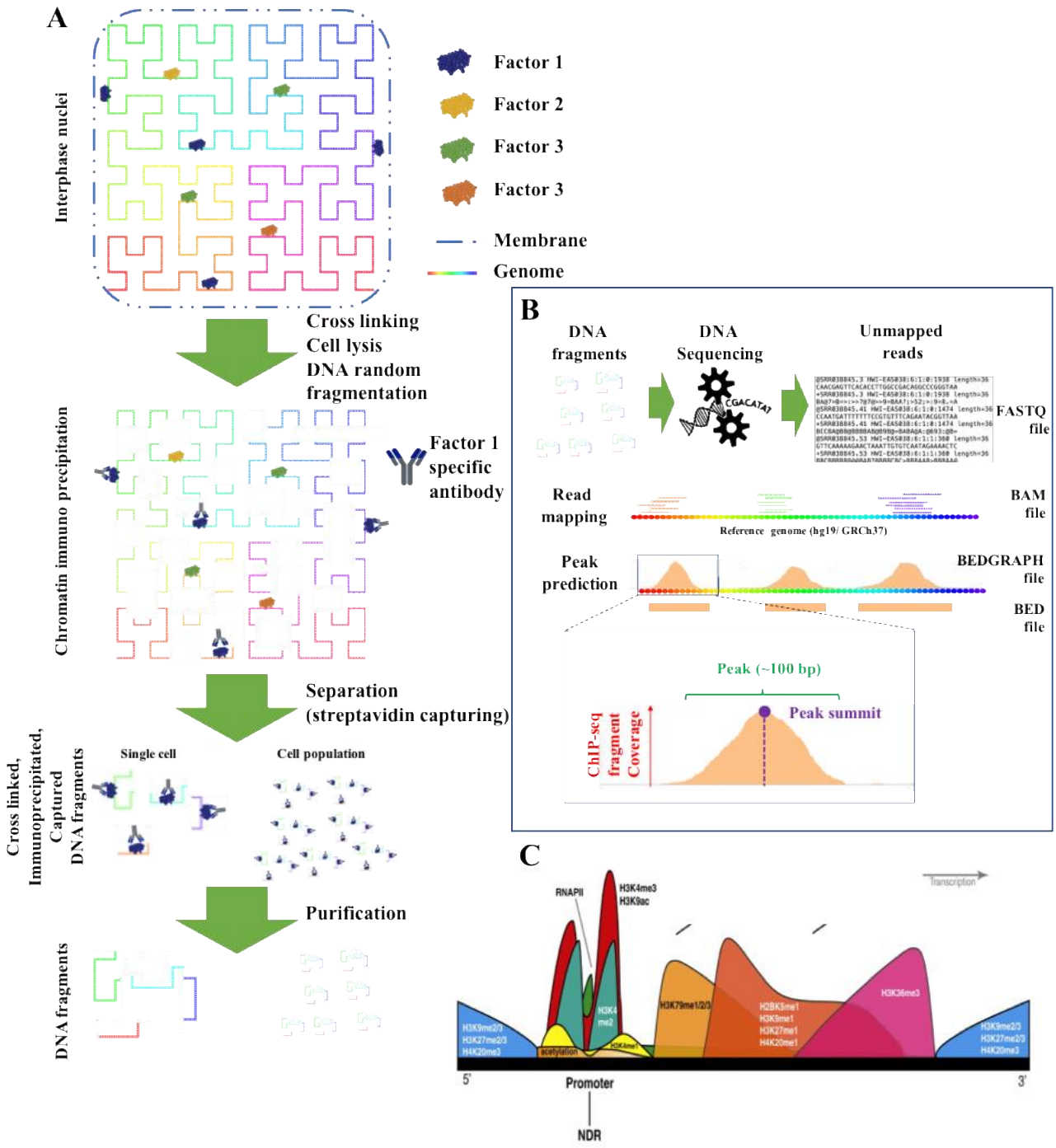


Figure 9. ChIP Sequencing overview. The essential steps of ChIP-seq from sample to data procession. A) "Wet-lab" protocol of ChIP-seq. The genome of the cell is represented as a Hilbert curve (Anders, 2009; Gu, Eils, & Schlesner, 2016; Hilbert, 1891). The ChIP-seq protocol uses millions of cells (except single cell techniques). For an easier demonstration, we present the process for a one cell example until the separation step. In the first steps, the molecules of cells are cross-linked with a reagent (formaldehyde is the most commonly used) and homogenized. Numerous methods can be used for cell lysis and fragmentation. DNA fragmentation is mostly randomized with ultrasound sonication. Specific antibodies are used to "fish out" protein-DNA complexes of interest using biotin-streptavidin separation. After reverse crosslinking and purification, the

fragmented DNA is ready for the sequencing step. B) The sequenced fragment (reads) data is stored in fastq format. In the "mapping" step, the position determination of reads (in the reference genome) is done using Burrows-Wheeler alignment (Li & Durbin, 2009). Reads with the identified position are stored in SAM/BAM files. The peak calling process is used to identify genomic regions that are enriched with aligned reads. The result is usually stored in BED format, which consists of the genomic sections (genomic coordinates: chromosome, start- and end-nucleotide) with high read enrichment (peak regions). To visualize the fragment coverages, we can create BEDGRAPH files, which display numerical values for the aligned read distribution along the peak regions. The summits are easily visible in this display mode. The visualized peak region in this example (in the genome browser) is a typical transcription factor/cofactor peak, which has a bell-like curve and is narrowly concentrated. C) The histone peaks have more fluctuating shapes, and their fragment coverage shows broad distribution. Source: (Barth & Imhof, 2010)

The Gaussian bell-like shape of the transcription factor peaks is due to the protein's bound to one genomic point (recognition site) and the previously mentioned random fragmentation. The binding site is protected by the bound protein from random fragmentation, and the DNA fragment is fished out through this protein during ChIP (Pchelintsev, Adams, & Nelson, 2016). Therefore, the end of fragments may vary (from the same genomic region), but the binding site is common because of the selection protocol. This makes these regions highly covered by tags in the vicinity of a binding event, while remote regions from the center have decreased coverage. The peak summits have the highest coverage of the region and are known to more-or-less coincide with the bound DNA elements. The use of this region is recommended in de novo motif enrichment scanning (He et al., 2015).

The Galton board is a good model for summit position appearance. This device is originally made to demonstrate the statistical concept of normal distribution (Galton, 1894). After rotating the board on its axis, the stream of beads is set into motion that rolls with equal probability to the left or right through several rows of pegs. As the beads concentrate in the bins, they come close to a bell curve, as shown by the red line (**Figure 10**). This helps to visualize the order embedded in randomness. The maxima of the distribution are in similar columns with the axis of a bottleneck at the start. The behavior of summit locations is similar

to this Galton board example. The bottleneck in our case is the binding event, which leads the majority of summits to coincide with real binding sites. Based on these, we developed a protocol for protein position detection.

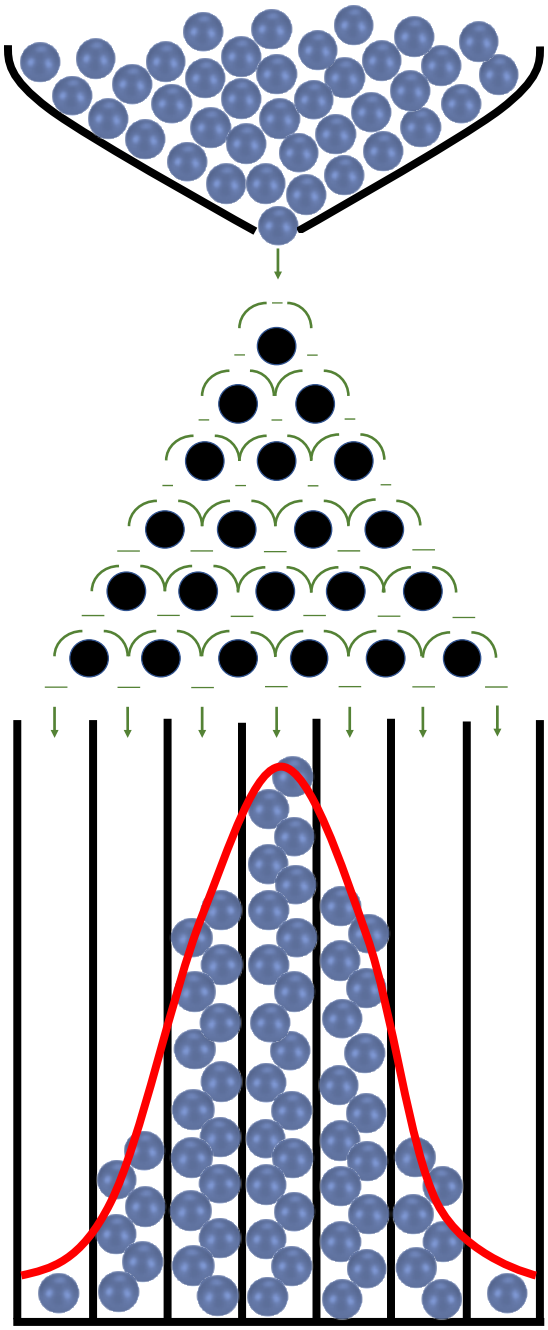


Figure 10. *The* Galton-board brings to life the normal distribution. This device can demonstrate the summit position establishment due to the random fragmentation and the relative positions of a specific factor's summits *relative* to *the* fix*ed* genomic points (e.g: the bound transcription factor binding sites) at the same time. We can consider the detected protein-DNA interaction as the bottleneck of *a* Galton-board, which keeps the maxima of fragment coverage near to *a* real interaction point.

We “fish out” the DNA fragments by the immunoprecipitated protein. Therefore, the DNA section, which *is* crosslinked with *the* protein, will be the most commonly occurring fragment (from *the* same genomic region, but from different cell*s*). However, this summit position shows loose mobility due to other background factors (e.g. cell phase, interaction proteins, and complexes). Using *a* large population of summit positions (genome-wide) can reveal the position preference of proteins (relative to the reference point) considering the most commonly preferred position (maxima of distance distribution curves).

1.1.7.2.Introduction of high-level ChIP-seq databases

The accumulation of raw ChIP-seq data made the data storage and processing challenging. Since not every lab has the computing environment (computing power and professional human resources) to process large amounts of ChIP-seq data, there is a growing needed for publicly available processed data. This prompted the development of higher level databases, which provide semi- or fully processed data to bypass the basic analysis steps (e.g. GTRD, ReMAP, ChIPBase, ChIPAtlas, and Cistrome).

The most similar database to ChIPSummitDB is the GTRD, which provides uniquely processed ChIP-seq data from human, mouse, and rat samples. The data is processed until the peak prediction level and the binding sites are clustered according to target proteins. The number of processed ChIP-seq experiments reached 6819 until the end of 2018, which includes 1367 input/control data. Compared to our dataset (3782 ChIP-seq data), the overlap is below 50 %, with 1978 instances is common.

ReMAp provides annotation tools and complex information about transcription factors. The main source of data is the ENCODE database. The processed data is directly linked to the UCSC genome browser, but the identified peaks, in BED file format, are downloadable. ReMAp processed less data then we did.

ChIPBase adapted more than 10,000 ChIP-seq data from human and mouse. The dataset includes histone, transcription factor, and co-factor ChIP-seq data. In this mixed dataset, we identified 1592 human factor ChIP-seq datasets with SRA IDs.

The largest database is ChIPAtlas, which collected more than 8368 experiments from mixed sources. Of these experiments, 7200 were no histone/input or control experiments. This huge dataset covers almost 2/3 of our collection, with 2415 experiments in common.

1.1.7.3. Introduction of Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET)

The 3C techniques were developed to analyze the spatial organization of chromatin. The 3C techniques, in combination with High-Throughput Sequencing, enable the identification of DNA loops on the genome level. In this work we used ChIA-PET technique results, which combines 3C and ChIP-based methods. The procedure has common steps with ChIP-sq. It is started with formaldehyde treatment to cross-link the DNA-protein complexes, then specific antibodies are used to fish out the protein-DNE interactions of interest. Since distal DNA regions are hold together through protein complexes, the anchor regions of DNA-loops can be ligated to each other (through linker sequences). After purification and MmeI digestion, the linked DNA fragments are sequenced with Paired-End Tag Sequencing (PET). After proper computing analysis of sequenced data, we can identify the distal DNA regions which get close to each other in association with a protein of interest (Goh et al., 2012).

2. Aims of the study

During the analysis of CTCF and cohesin ChIP-seq data, we noticed that there is a visible shift between the summit positions of these proteins. Since the CTCF is the only member which have known DNA binding domain in this complex, we wanted to identify why the summits are not located on the same genomic localization and if there is any measurable system behind this shift. We hypnotized that the juxtaposing summits referring to topological position of cohesin proteins. We assumed that there is a structural feature of complex which holds the cohesin subunits in close proximity of DNA. So thus, the non-DNA binding

proteins are crosslinked with DNA during the formaldehyde treatment of chromatin immunoprecipitation.

We downloaded and processed several CTCF and cohesin subunit ChIP-seq data from human and mouse (see in Material and methods) to answer the following questions:

- Can the observed shift be seen in all analyzed samples?
- Does the shift show any strand specificity?
- Does the shift follow the orientation of CTCF motif?
- Is there any order between the shifts of difference subunits?
- Is there any linkage between the protein positions and the known topology of CTCF-cohesin complex?
- Is there any correlation between the shift orientation and CTCF mediated chromatin looping?

After the publication of results, we extended our focus to other available human ChIP-seq data, which were analyzed using the summit-based topology analysis. We decided to collect as many human ChIP-seq data as we can and process them with our method. For unique data processing and comperability we faced with the following tasks:

- Development of a pipeline, which is able to automatically extract topological and network information from large amount of data.
- Create a database which can be used not only for data storing but for protein position analysis too.
- Discover unknown protein-protein interactions and complexes and create new topological models.
- Create a web interface which provides access to our result for other researchers.

3. Material and methods

Using the initial observations from CTCF and cohesin ChIP-seq data, we inspected CTCF mediated looping in the context of topology and transcriptional regulation. During the examination, we developed a technique which allows us to extract topological details of protein complexes from ChIP-seq data. The summit-based topology analysis was used to identify protein position preferences relative to a reference point (fixed point in the genome, e.g. CTCF motif center) and derive structural information from the results. We used the results to create a hypothetical model of CTCF-mediated chromatin looping and investigated the anchoring of CTCF binding sites. In the chromatin loop analysis, we used CTCF ChIA-PET data, which were tested from different angles. We identified the anchoring CTCF binding sites on both interaction points of the loops and found a correlation in strand specificity of binding sites. The ChIA-PET result was examined in the context of transcriptional regulation with histone ChIP-seq data and other transcription factor data to identify proteins or complexes that interact with the cohesin ring.

Then we extended our focus to other available human ChIP-seq data, which we analyzed using the summit-based topology analysis. For the large scale analysis, we developed the following pipelines and script for automatic data processing:

- Automatic naming of ChIP-seq experiment according to attributes of experiment with Perl program
- Peak filtering program
- Summit predictor and distance calculator script
- Motif optimizer and mapper script

The complete pipeline is available at https://github.com/Raziel01/SummitDB-data-prepare domain.

We collected and processed public human ChIP-seq data (transcription factor and cofactor) from numerous cell lines to create a database of transcription factor binding sites. Combining the JASPAR CORE motif set and the processed ChIP-seq data, we created a genomic map using transcription factor binding sites as reference points for summit distance calculations. We attempted to create a comprehensive map of protein networks correlating different transcription factor binding sites. Then, we tested our database with X-ray crystallography data for transcription factors. The results are publicly available at http://summit.med.unideb.hu/summitdb/index.php, where the data are downloadable and viewable.

3.1.Source of data and TF clustering

Development of protein and DNA sequencing techniques has led to an explosion of sequence data and rapidly expanding databases such as NCBI SRA (National Center for Biotechnology Information Sequence Read Archive), ENCODE, DDBJ (DNA Data Bank of Japan) etc. The accumulation of publicly available data resulted in the appearance of projects that attempted to rank all identified transcription factors. Protein sequences were retrieved from databases, such as Uniprot or RCSB Protein Data Bank (PDB), and the sequences were compared using multiple alignments (Berman et al., 2000; UniProt Consortium, 2018). PFAM and ProSite provided information about domain groups (El-Gebali et al., 2019; Sigrist et al., 2013). The DNA sequence preferences of TFs were derived from PWM databases, such as TRANSFAC, JASPAR, FootprintDB, or HOCOMOCO (Khan et al., 2018; Kulakovskiy et al., 2018; Matys et al., 2006; Sebastian & Contreras-Moreira, 2013).

TFclass is a database of classified eukaryotic TFs (Wingender et al., 2015). TFclass is a combination of the previously mentioned databases (Wingender et al., 2018). RSAT matrix-clustering uses hierarchical clustering of transcription factor binding motifs to perform the classification (Castro-Mondragon, Jaeger, Thieffry, Thomas-Chollier, & van Helden, 2017).

TFclass is comprised of four general levels (superclass, class, family, and subfamily). TFclass and RSAT classifications are differing in levels of classification trees but the identified classes are the same.

3.2.Primary analysis

The National Center for Biotechnology Information (NCBI) Sequence Read Archive served as the source of raw ChIP-seq data. The primary analysis of the downloaded raw data was carried out using an analysis pipeline developed in-house (Barta E., 2011). The read mapping was performed using the Burrows-Wheeler algorithm (bwa) program (Li & Durbin, 2009). The reads were then mapped to a reference genome hg19/GRCh37. The Hypergeometric Optimization of Motif EnRichment (HOMER) package was used in many steps (Heinz et al., 2010), including peak calling (identification), de novo motif analysis and motif optimization, motif remap, and data visualization. Next-generation sequencing experiments contains anomalous, unstructured, or high signal peaks which are independent of cell line or experiment. The removal of these false positive data is an essential quality measure. We used the ENCODE black list, which is a comprehensive set of these regions and removed the overlapping peaks with a BEDTools program (Quinlan & Hall, 2010). The filtered peak set was used for de novo motif discovery. To return the location of enriched motifs we used the annotatePeak.pl program. The summit positions were identified with PeakSplitter (Salmon-Divon, Dvinge, Tammoja, & Bertone, 2010). The parameters of the programs are indicated in **Table 4**.

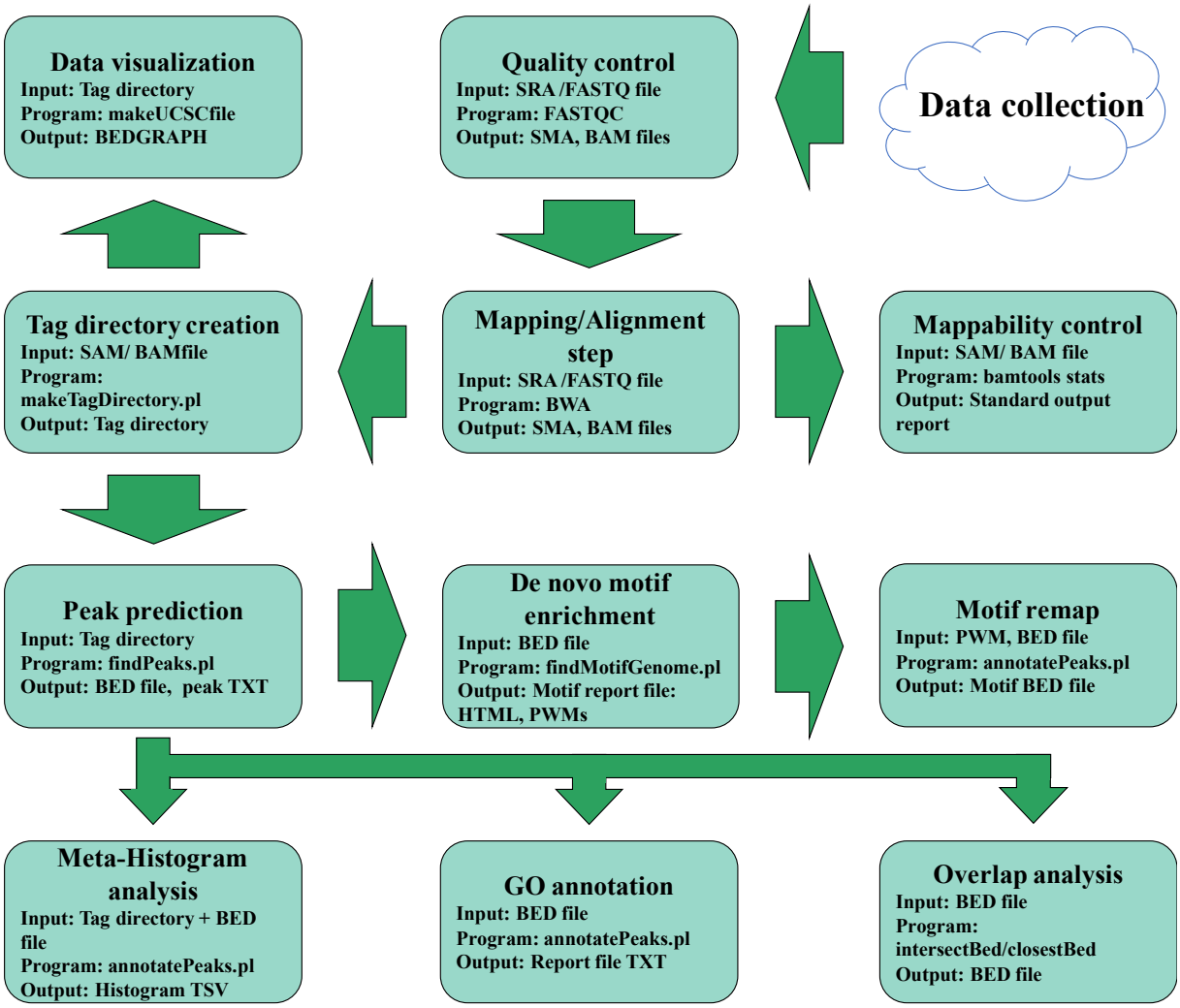


Figure 11. The bioinformatic pipeline used for primer ChIP-seq data analysis.

Step	Task	Program/Package	settings	Remarks
1	***Read alignment***	BWA version 0.7.10-r789	–B 8	
2	***Peak Identification***	HOMER	findPeaks –style factor	Peaks in the ENCODE blacklisted regions removed

3	***Peak summit identification***	PeakSplitter		
3	***Motif Enrichment/ optimization***	HOMER	findMotifsGenome.pl -mask -len 10, 12, 14, 16 -dumpFasta -bits -preparse -homer2 -size 100	Using the top 5000 Homer peaks
4	**Finding Instance of Specific Motifs**	HOMER	annotatePeaks.pl -noann -nogene -mbed -m	Using the top 5000 MACS peaks
6	***Data visualization***	HOMER	makeUCSCfile	For IGV genome viewer (6,7)

Table 4. Steps for the ChIP-seq analysis pipeline. Summary of the main steps and programs, which were used in the raw data processing.

3.3.Investigation of CTCF/cohesin co-occupied sites with ChIP-exo and DNase-seq data

To identify the genomic location and coverage of CTCF/cohesin proteins with near-single-nucleotide accuracy, we used publicly available HeLa DNase-seq and ChIP-exo data (SRX100899, SRX098243).

The single-nucleotide resolution border peak detection was executed with “model based analysis of ChIP-exo” (MACE) (L. Wang et al., 2014).

The DNase-seq bam files were downloaded directly from the ENCODE database ("An integrated encyclopedia of DNA elements in the human genome," 2012). The raw sequence reads were then aligned to the hg19 human genome. The accurate prediction of CTCF/cohesin footprints were then done with the Wellington algorithm (Piper et al., 2013). This algorithm detects characteristic depletions of DNase I cuts and compares the result with a large number of cuts in the surrounding region of open chromatin that do not harbor bound proteins.
The identified ChIP-exo and DNase-seq borders were then compared with the processed ChIP-seq data and are shown in **Figure 25**.

3.4. Data visualization and statistics

Statistical calculations were carried out in the R environment (http://www.R-project.org) using R version 3.1.2 (R Core Team, 2014). Statistics were calculated using reshape2 and PMCMR (Pohlert, 2015; Wickham, n.d.). Matched samples were analyzed with the Wilcoxon signed-rank test (a two related sample comparison) and Friedmann test using a Namenyi posthoc test (to account for multiple test attempts). The threshold for significance was $P<0.05$.

To confirm CTCF and cohesin results, we performed computer simulations to understand the interactions better. We had concrete CTCF-RAD21-SMC1/3 and SA1 position values relative to CTCF motif centers (measured in base pairs) form several samples. We pooled all of these values and randomly shuffled them with ran2 method without considering which protein the values belonged to (Press, Teukolsky, Vetterling, & Flannery, 1992). We created random clusters for all of the proteins, and we calculated their average, median and standard deviation. We were seeking for similar patterns to the experimental values. Repeating the simulation 10 million times resulted in none of the attempts showing similar patterns to those that were experimentally observed. The source code of the program is available on https://github.com/TravisCG/CTCFSim. Genome browser compatible files were made using

BEDTools and makeUCSCfile (Heinz et al., 2010; Quinlan & Hall, 2010). We used the Integrative Genomics Viewer (IGV) in the data visualization phases (Robinson et al., 2011).

The reproducibility of CTCF-cohesin peak shift values was tested on the HeLa dataset. The peak shift was measured between the peak summit of the proteins indicated (CTCF, Rad21, SMC3) and the center of the CTCF binding site (CTS). In **Figure 28**, the reproducibility was characterized using the standard deviation of the mean (Y-axis) that was determined from a number of observed peak shifts (X-axis). The inset shows the details in logarithmic scale, demonstrating that approximately 100 shift values are necessary to reach a reproducibility of +/-1 nucleotide. In our experiments, we normally used more than 5000 peaks that roughly corresponded to a reproducibility of 0.1. Naturally, the reproducibility estimates vary with the quality/coverage of the dataset. In practice, we rounded the peak shift values to one nucleotide.

3.5. Database creation

The creation of the base of ChIPSummitDB was a multilevel process consisting of the following steps:

Step 1: ChIP-seq data collection from public databases.

Step 2: Processing of collected ChIP-seq data with a custom-made in-house analysis pipeline. The pipeline includes read mapping, motif enrichment analysis, peak prediction, and generation of coverage track files (bedgraphs and bigwig). The peak region BED files and coverage file (bedgraph) are important in the following steps.

Step 3: Splitting peak regions and summit prediction. The peak region files and coverage files (bedgraph) from the previous step were used to subdivide ChIP-seq regions into discrete signals to identify summit regions.

Step 4: Peak filtering. We created a homemade script to filter peak regions, depending on their symmetry and shape. The script requires the split peak regions and coverage bedgraph files. The outcome of peak filtering is filtered peak region sets in the BED format.

Step 5: JASPAR CORE motif and ChIP-seq data pairing. We paired motifs from JASPAR CORE database to their corresponding ChIP-seq experiments. The results were stored in a table.

Step 6: Motif optimization. We performed a motif optimization of JASPAR CORE motifs. To do this we created a merged peak region set using the filtered peak regions of the corresponding ChIP-seq experiments (determined in the previous step). All JASPAR motifs were optimized using these merged genomic regions, resulting in optimized motifs. These motifs were similar to the original JASPAR core record, but their PWM values were adjusted to a more accurate (and similar) motif which is enriched in the peak regions.

Step 7: Determining motif locations. We used 3 different programs to find the instances of optimized motifs in the genome. The result is the genomic locations in BED format.

Step 8: Summit distance calculations. The centers of identified motifs (step 7.) served as reference points in the calculation of motif-protein and protein-protein distance calculations. The results are stored in MySQL data tables, which are available on the ChIPSummitDB website.

3.5.1. Large scale data collection and procession

Data from 4068 ChIP-seq experiments, covering a wide range of human proteins and cell types, were collected from the NCBI SRA and ENCODE databases (“An integrated encyclopedia of DNA elements in the human genome,” 2012; Leinonen, Sugawara, Shumway, & Collaboration, 2011; Parry et al., 2010). The naming and automatic download of experiments were performed using a homemade script. Processing of the downloaded raw data was carried

out with the previously mentioned in-house developed ChIP-seq analysis pipeline (Barta E., 2011). The primary workflow is shown in **Figure 11** and **Figure 12**. Following this analysis, the semi-processed data were further analyzed by various steps.

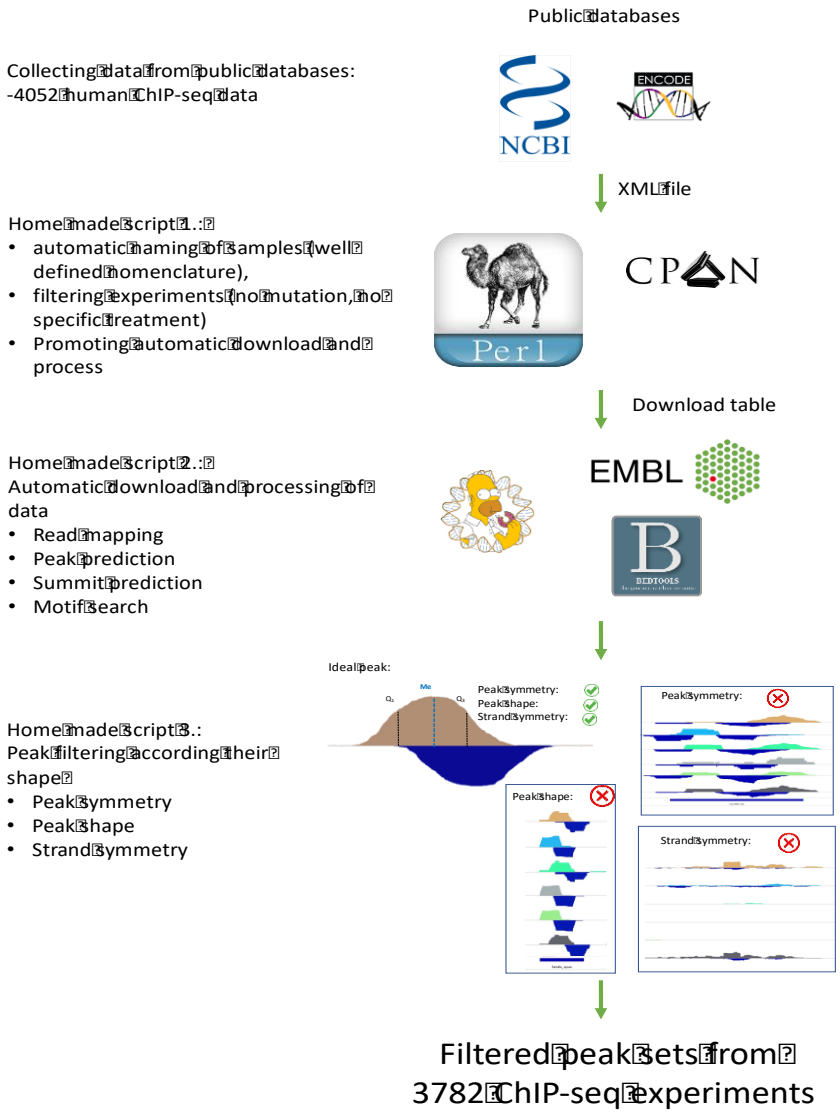


Figure 12. Schematic representation of the initial data processing. Data processing starts with data collection and proper naming. After the processing and filtering steps, we get the transcription factor binding sites in bed and bedgraph formats.

3.5.2. Peak splitting and summit prediction

We used PeakSplitter for summit predictions and, thus, more accurate identification of local maxima (**Figure 13**). PeakSplitter was developed to split subpeaks when overlapping peaks are present. The peaks for transcription factor bound sequences are usually concentrated to a narrow area showing a Gaussian distribution due to the random fragmentation and their narrow binding surface. This was especially observable after the extension of reads to the expected fragment length (Zhang et al., 2008). High signal and weak enrichment indicate insufficient discarding of read duplicates or library preparation artifacts (Star et al., 2014; Steven R. Head et al. 2014).

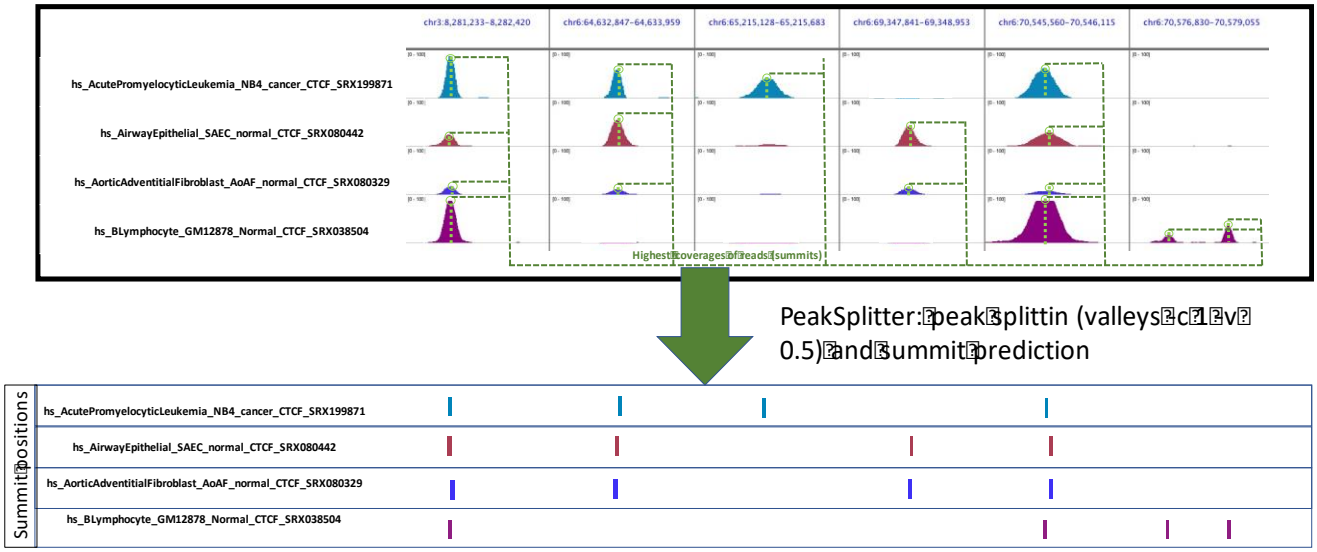


Figure13. Summit prediction and peak splitting. Identification of local maxima within peak regions.

3.5.3. Peak filtering

Identifying peaks with well-defined maxima was crucial at the early stage of data processing because false positive peaks could result in the false prediction of the protein's position. The peak summits (maxima) show the highest coverage for the peak region and coincide reasonably with the center of the corresponding DNA elements bound by transcription factors. Therefore, the identification of regions suitable for the clear determination of the

summit position(s) was required. Current software packages use different strategies, such as the evaluation of peak prediction reproducibility or false discovery rates (FDR), for peak prediction (Taslim, Huang, Huang, & Lin, 2012), which dramatically decrease the false positive rates. Unfortunately, using these methods necessitated configuring the filtering algorithms differently for each experiment, making automation of the processing of large datasets more difficult. For better filtering, we have developed a pipeline, which reduces the false positive discovery rate even further.

To avoid false positive results, we filtered out duplicated reads using a step in the ChIP-seq analysis pipeline and developed a Perl script that classified and filtered the subpeaks based on their size and shape. In the script, two parameters are responsible for the detection of the previously mentioned large signal intensity increase.

We developed a completely new method for data filtering, which is an intermediate solution between manual (checking peaks in genome browser) and automatic peak filtering. This method is not introduced yet in any ChIP-seq analysis protocol. In these analyses, the peaks are considered coverage histograms and the positions of the median, first, and third quartile values were used. The "ideal" transcription factor peak has three attributes; i) the read distribution on both strands have symmetrically curved shoulders, if the median value is the symmetry axis; ii) the peak's shape displays a bell-like curve; iii) the maxima of the ChIP-seq signal is approximately equal between the Watson and the Crick strands (**Figure 14A**). The determination of ideal peaks is derived from the following articles: Leleu, Lefebvre and Rougemont, 2010, Barth and Imhof, 2010, Pchelintsev, Adams, & Nelson, 2016, He et al., 2015.

For the first two steps, the filtering analysis is required for filtering out peak positions that have large gaps in their ChIP-seq signal intensity, even after the read extension by the peak caller software. The formula in **Figure 14B** shows the calculation of the symmetry of the two

sides of the peak. For this calculation, the maxima are used as the axis of symmetry. The second formula (**Figure 14C**) quantifies the shape of the peak based on the distances between the minimum, maximum, 2nd, and 3rd quartile values of ChIP-seq signal intensities within the peaks. This results in a value between 0 and 1. If we connect the four above-mentioned values with a straight line (where the X-axis represents the position of the signal and Y-axis represents the signal intensity), the peaks which have a "0" shape value would be shaped like a triangle. In contrast, if the value converges to 0.5, the shape of the peak would resemble a square (**Figure 14C**). Optimally, the forward and the reverse tag counts (in a peak) have, approximately, the same size due to the ChIP-seq method. The third formula calculates the symmetry between the reverse and the direct strand tag counts (**Figure 3D**).

Due to the ChIP-Seq technology, at each protein-DNA binding site, the tags from the forward strands are located on the left-hand side of the binding site and the tags observed from the reverse strand are located on the right-hand side. This is an aspect which is considered and used by several peak-calling (e.g. macs2) software to extend reads by an average value during peak identification. We used this parameter to filter data. We calculated forward-reverse maxima distances and values which could be found in the 90 percentile passed this filtering step.

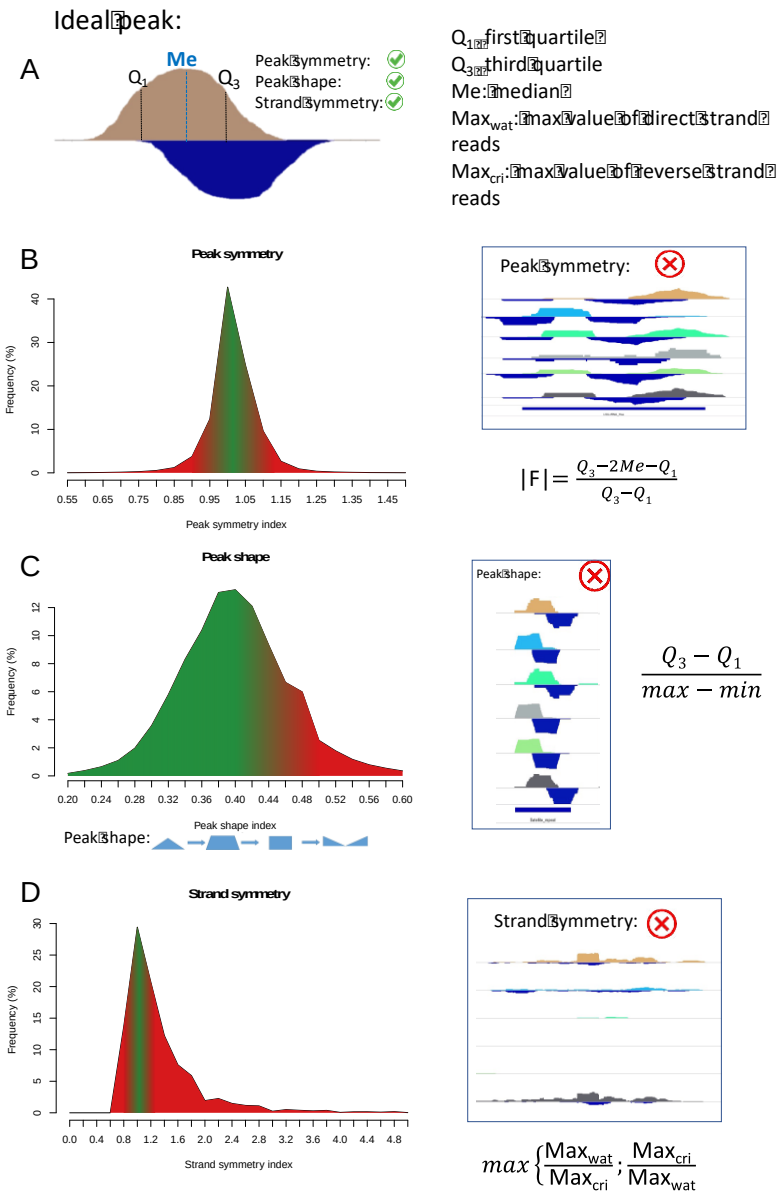


Figure 14. Peak filtering according to shape. A) Peaks of factors/cofactors are narrowly concentrated and have a bell curve shape. We filtered peaks depending on the symmetry of their two sides (summit positions serves as a symmetry axis) (B), the positions of 2nd and 3rd quartiles (C), and the symmetry between the read coverage of the two strands (D).

3.5.4. JASPAR CORE motif and ChIP-seq data pairing

Identification of the exact positions of TF binding sites is the basis of ChIPSummitDB. These motif positions are not only a collection of regulatory regions, but the motif centers are also used as reference points for summit position analysis. Our primary goal was to create consensus binding site sets for as many transcription factors as possible. To do this, we used the JASPAR CORE database, which is a "curated, non-redundant set of profiles, derived from published collections of experimentally defined transcription factor binding sites for

eukaryotes" (Khan et al., 2018) and incorporates 579 non-redundant motifs. We attempted to collect all motifs with ChIP-seq experiments from our collection. Several motifs were lacking HTS data for historical reasons, thus the JASPAR CORE was built to create families of binding profiles for as many structural transcription factor classes as possible. We were able to allocate only 338 motifs to at least one ChIP-seq experiment, because no sequence and HTS data were available for the rest of motifs in human (**Figure 15**).

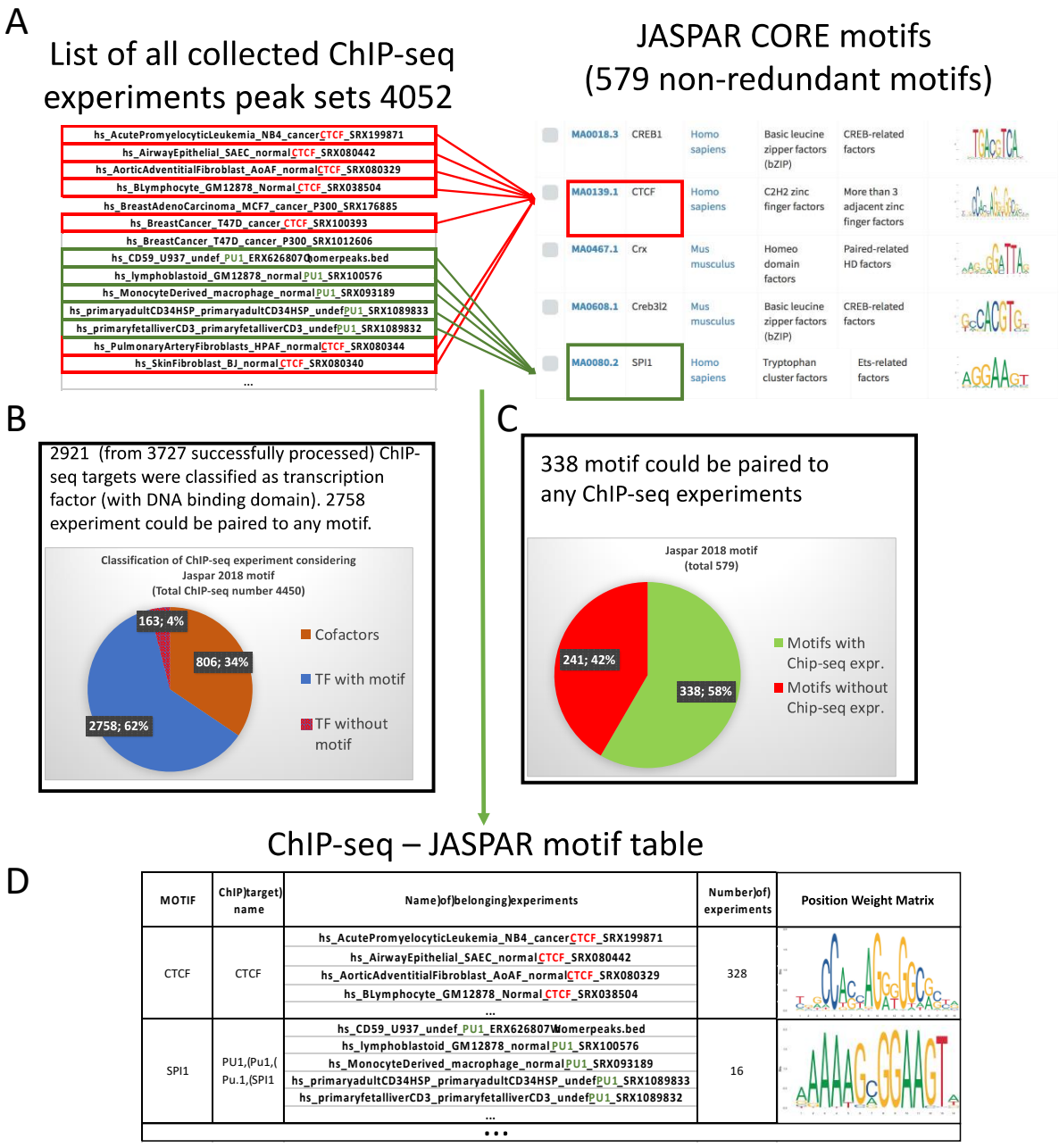


Figure 15. Pairing position weight matrices (PWMs) for processed ChIP-seq experiments. (A) 2758 experiments could be paired to a proper JASPAR motif from the downloaded and processed 3727 ChIP-seq experiments (B). This paired to 338 JASPAR CORE motifs from the 579 (C). The result was a table where the PWMs are paired to their corresponding ChIP-seq experiments (D).

3.5.5. Motif optimization and determining their locations

To optimize the allocated motifs, the peak regions of the corresponding ChIP-seq experiments were scanned for similar motif enrichments (Heinz et al., 2010). The optimized motifs were manually curated and the most similar motifs were paired with the corresponding antibodies (**Figure 16**). This step maximized the number of specific motif instances, which were identified in the next step.

Numerous tools can be used to find the occurrences of individual motifs. Instead of choosing one single tool, we combined 3 popular methods: HOMER, FIMO, and MAST (Finak et al., 2015; Grant, Bailey, & Noble, 2011; Heinz et al., 2010). The positions, which were identified for certain motifs by at least two programs, were filtered in the first step. Using the default motif scores obtained by the above-mentioned three programs and the distance of the closest summit obtained from the list of paired motif-ChIP-seq experiments, a weighted motif value was calculated. All identified ChIP-seq peaks were coupled with the closest motif possessing the highest weighted motif value. The distance cutoff was +/- 50 base pairs. Following this step, sets of non-redundant motifs were created by filtering out the motifs with identical positions and directions (**Figure 16**). Even in the case of palindromic sequences, identifying motif directions was possible due to the flanking regions and the positional preferences of the peak summits.

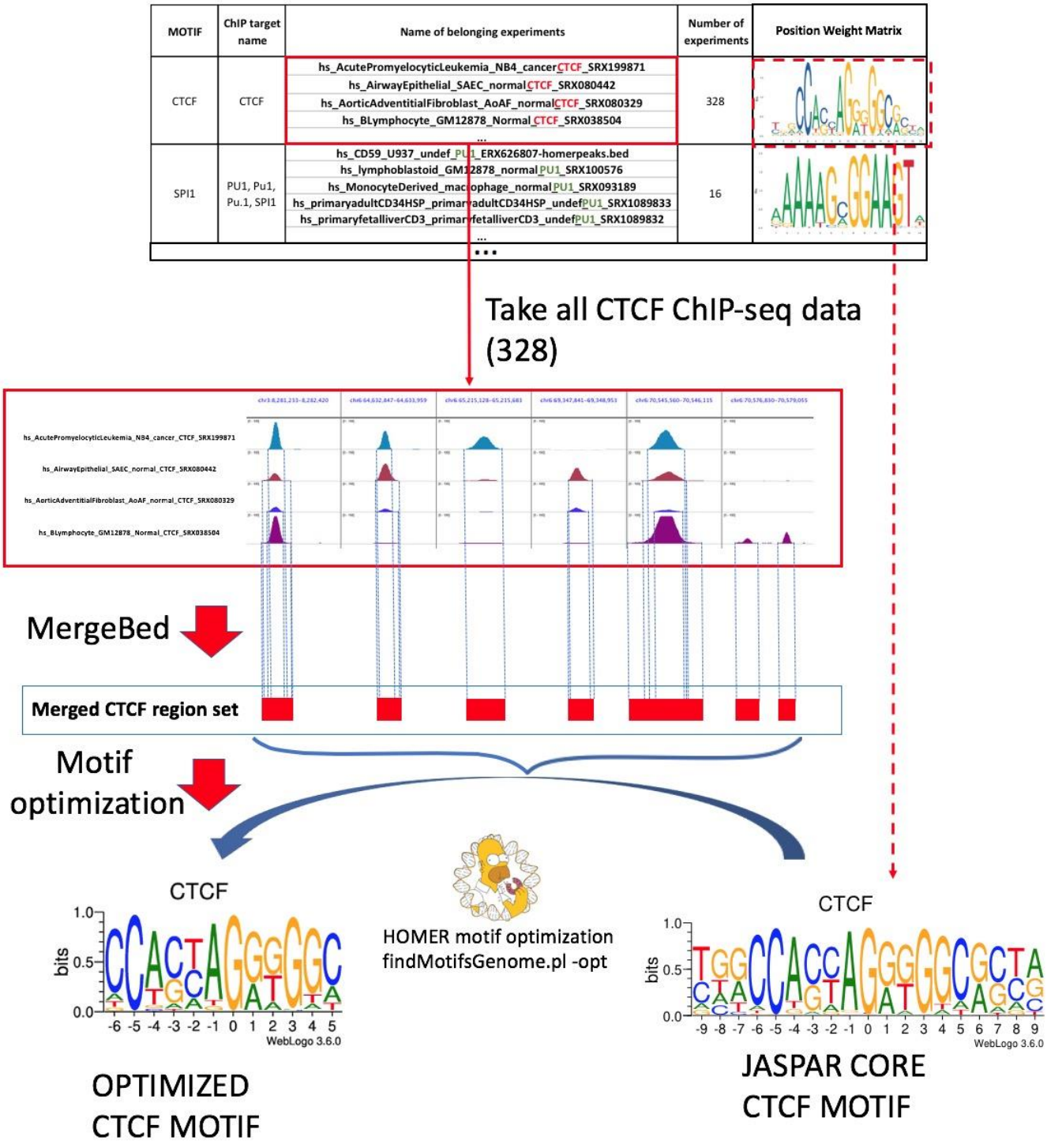


Figure 16. Motif optimization. JASPAR CORE motifs were optimized using the findMotifsGenome program, which used the original PWMs and the merged peak region set of the corresponding ChIP-seq experiments (determined in Motif- ChIP-seq experiment pairing step).

In the previously mentioned step and in the subsequent analysis, closestBed, a tool of bedtools, was used to measure the distance between the center of the motifs and the summits (Quinlan & Hall, 2010). If the length of an N bp long motif was even, then the (N/2)+1 bp from the 5' end of the sequence was considered the center of the motif. We created individual summit position pools for all motifs from their respective ChIP-seq experiments. Then, the identified motifs and summits were combined using the closestBed program. This step resulted in a table, where all of the summits positions from the proper set are shown together with one or more of the nearest (one or more) motif instances. Distances between the centers of the motifs and the summits can be calculated this way. Both this distance and the score of the motif were taken into account during the coupling of the most probable motifs with each of the summits. We combined these scores into a formula, and the motif with the calculated highest score was picked for each summit position (one summit could have more than one motif in its vicinity, but only the strongest motif was selected for the following steps). The formula for the Weighted Motif score (WMs) calculation can be found in **Figure 17.** The same motif was frequently coupled to summits from different experiments. To avoid redundancy, we removed the duplicates. Thus, we obtained non-redundant global consensus motif sets for 338 JASPAR CORE matrices.

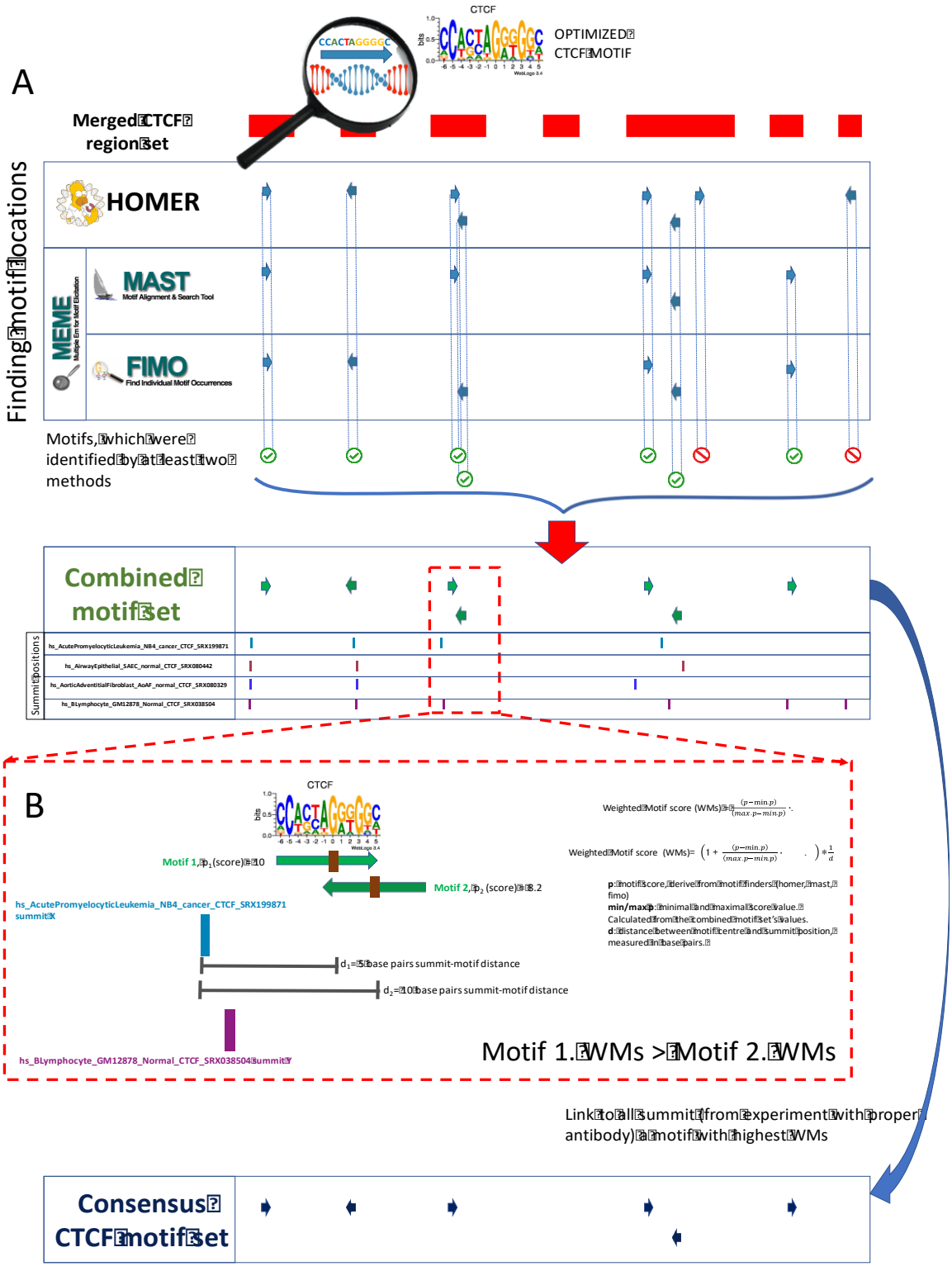


Figure 17. Determining motif locations. To identify the location of motif instances, we combined three different motif finding methods. A) The merged peak region set of the corresponding ChIP-seq experiment was used in the identification. B) To filter the identified motifs, we used the presented formula. In the case of overlapping motifs, the motif with the highest Weighted Motif score was selected.

3.5.6. Summit distance calculation

The identified consensus sequence locations are not only used to show the genome-wide distribution of transcription binding sites, but are also used as reference points for landscaping of possible co-bindings and measuring motif-protein or protein-protein distances. All motif occurrences obtained from every set were screened to identify ChIP-seq experiments containing peak summits in the +/- 50 bp vicinity of the motif center. The distances between motif centers and summit positions were calculated. The resulting distance tables can be examined for either genome-wide or local data. The genome-wide analysis highlights large-scale information about protein positioning. For example, co-location frequency, location preferences between proteins, possible members of complexes, and patterns in the protein composition of different regulatory regions can be examined. In addition to the frequency and the median/average values, both calculated from the measured distances, the standard deviation is also informative. The preferred position of a particular factor has a larger standard deviation (in relation to the positions of the motif centers) if it is physically far from the reference point (**Figure 18**).

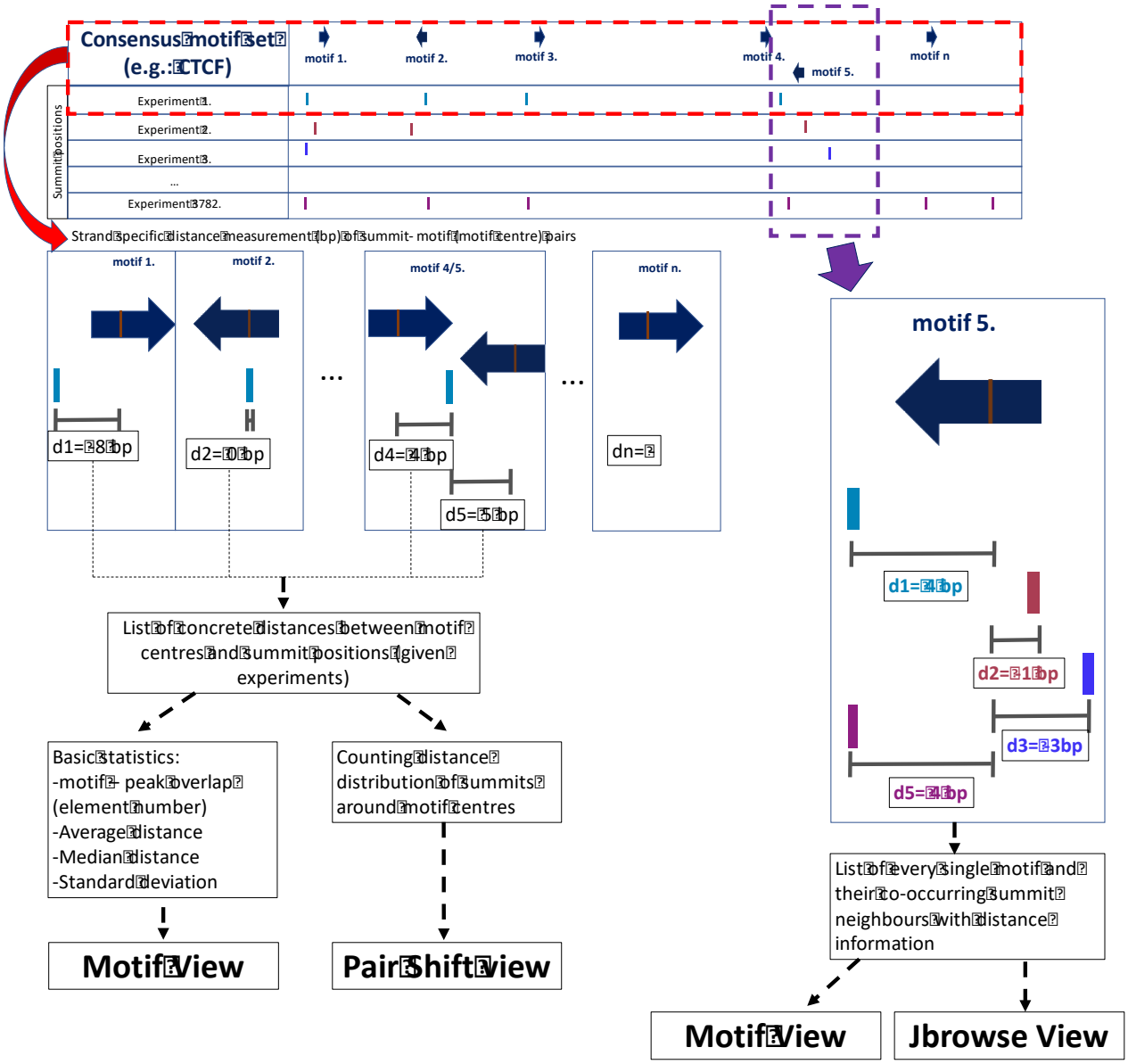


Figure 18. Measuring the distances between motif centers and the surrounding summits. We calculated the concrete distance between motifs and the neighboring summits (measured in base pairs). We took into account all of the possible summits from every experiment.

In ChIPSummitDB, examination of a specific region of the genome is also possible. Examining the summit positions at a specific motif can provide detailed information about the composition of regulatory complexes, their topology, and differences between cell lines.

4. Results

4.1.ChIP-seq data reveals the topological order of CTCF and cohesin proteins on DNA

Our first observation was that there is a visible shift between CTCF and cohesin subunit ChIP-seq peaks. Due to possible protein-protein cross-linking events, components of a protein complex that are not directly involved in specific DNA binding can produce ChIP-seq peaks that overlap with the peaks of TFs, which anchor them to DNA, and their corresponding summit positions approximately coincide (**Figure 19**).

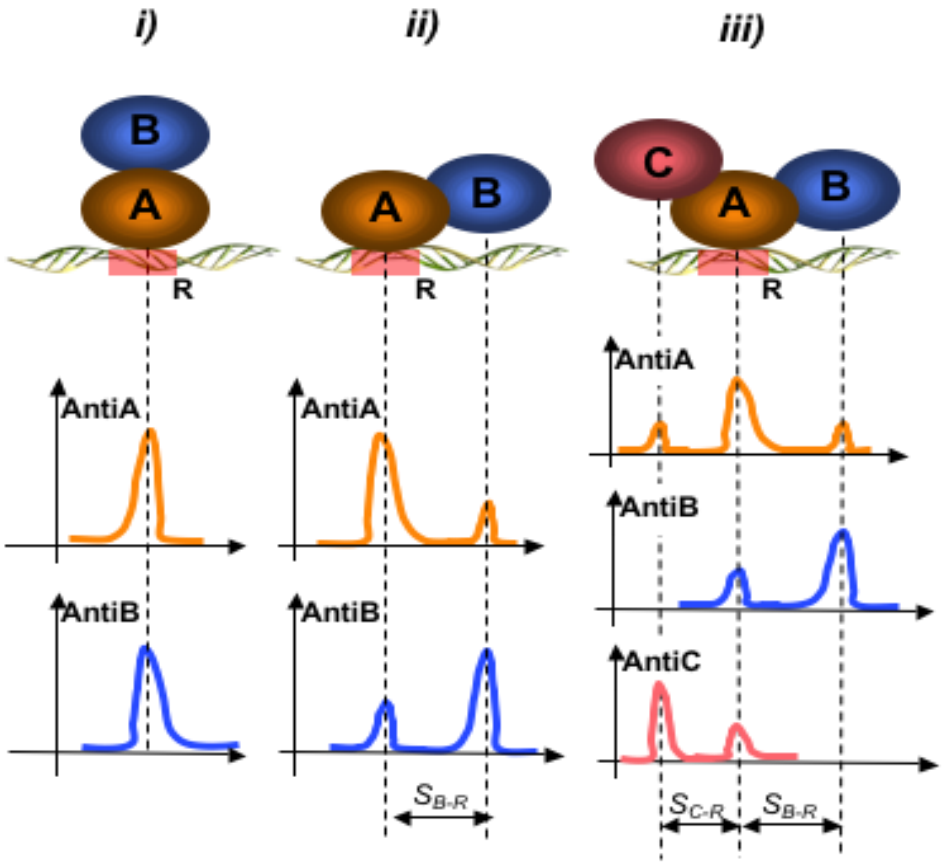


Figure 19. High-resolution Chip-seq mapping of a protein complex. Protein *A* is a DNA-binding protein that specifically recognizes site *R* in DNA. **i)** If proteins *A* and *B* form a symmetrical complex, there will be an A peak with the anti-A antibody, and a colocalizing B peak with the anti-B antibody. **ii)** If protein *B* binds to protein *A* in an asymmetrical fashion, the anti-B peak will be shifted with respect to the recognition site, by a value of $S_{B\text{-}R}$. Minor phantom peaks may also appear at the position of the other partner due to protein-protein cross-linking. **iii)** In a ternary ABC complex, we can see $S_{B\text{-}R}$ and $S_{C\text{-}R}$ shifts, as well as more phantom peaks, depending on the proximity of the constituent proteins. The shift *S* roughly indicates the positional

distance of a protein from the central residue of the binding site. If the specific DNA-binding protein binds to the recognition site in an asymmetrical way, an $S_{A\text{-}R}$ shift may also be detected, like in the case of the CTCF protein. In reality, the peaks strongly overlap and phantom peaks often appear as shoulders on the main peak.

Since CTCF is the only protein with a known DNA binding domain among the components of the CTCF/cohesin complex (**Figure 8D**), we expected that the corresponding ChIP-seq peaks will point to the same position with respect to CTCF binding site (CTS) (Chung et al., 1997);(Xiao et al., 2011). To measure the occurrence and extent of the shift, we compared summit positions relative to a reference point. The reference point needs to be fixed in the genome. The center of the CTCF motif seemed reasonable as a comparison site (**Figure 20A**). At this point, since the CTCF binding site is a non-palindromic element, we could measure the distance distribution strand specifically (**Figure 17B**). Depending on the location of the proteins and the motif orientation, the distance value can be positive or negative. The distance was measured in base pairs (bp). We extended the analysis to 237 human experiments (from 93 cell types) and 183 mouse samples (from 49 cell types) (**Figure 20B**). The overall plotting of protein positions (related to CTCF motif center) validated the observed shift. Strikingly, the average position of different proteins showed a characteristic separation around the binding sites. The plots within a cluster had the same protein target, which suggested that proteins have a position preference relative to each other and the binding site. The serial order of peak summit positions was invariably CTCF –> SMC1/3 –> RAD21, STAG1/2, irrespective of whether the average positions were calculated for a cell type or for the entire dataset (**Figure 20C-D**).

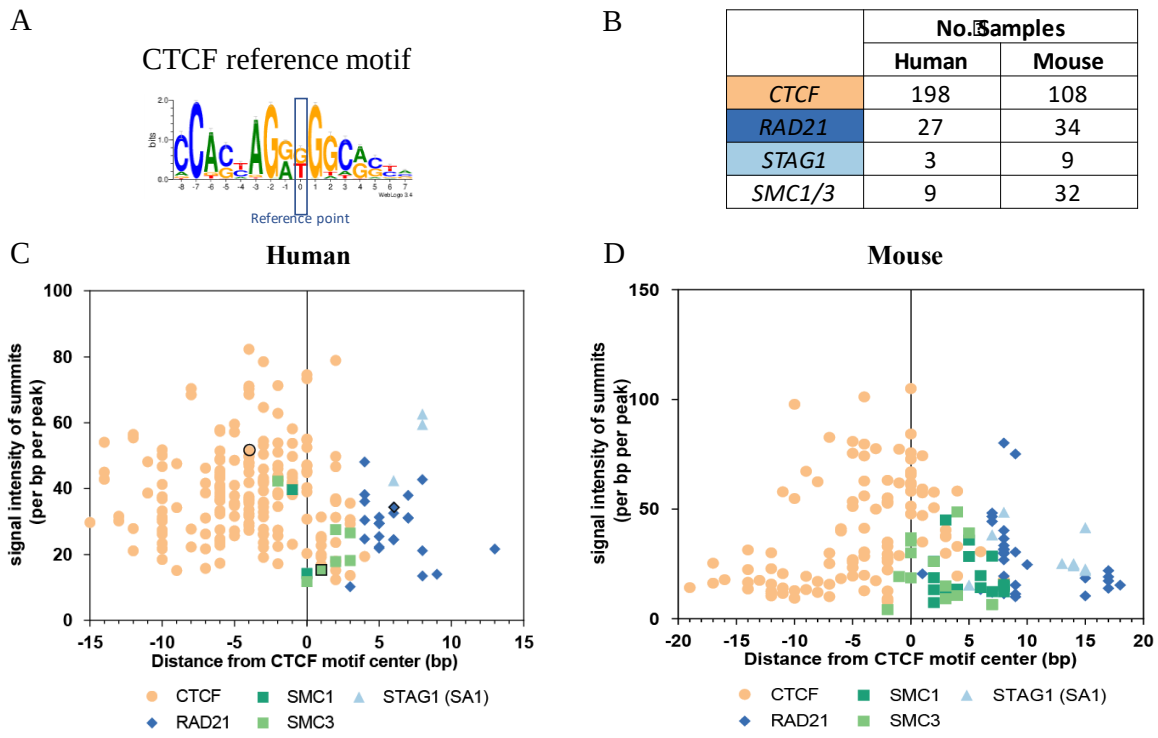


Figure 20. Shift between CTCF and cohesin bound sites in mouse cells. A) PWMs of CTCF binding sites. The presented PWM (shown as a logo) were used for the CTCF motif remap. The middle base pairs (marked as the "0" point) were used as reference points during the distance measurement between summit positions and the motif centers. B) Number of processed and plotted ChIP-seq experiments. C-D) The scatter plot shows the maxima of ChIP fragment coverage of CTCF, RAD21, SMC1/3, and STAG1 on CTSs specific for the given mouse cell or tissue type (see supplementary methods). The vertical axis shows the maxima of the average fragment depth and their positions relative to the midpoint of CTSs, which are represented on the horizontal axis. Position weight matrix of CTCF is shown at the bottom of Fig. 1B.

The shift patterns showed high conservation. We implemented a cell type specific analysis for more accuracy. We compared coherent samples, which derive from same cell line but the ChIP-target protein is differing, with paired T-test. The P value was $P<10^{-15}$ according to the Wilcoxon and Friedmann tests and $P<10^{-9}$ by simulation, even though some of the low-quality datasets gave less significant results (**Table 5-6**).

	Cell Line	Factor 1	Factor 2	Friedman Test p value	Namenyi post-hoc test p-values		
					CTCF vs Factor1	CTCF vs Factor2	Factor1 vs Factor 2
HUMAI	GM12878	SMC3	RAD21	$< 2.2x10^{-16}$	$< 2.0x10^{-16}$	$< 2.0x10^{-16}$	$< 2.0x10^{-16}$
	GP5d	SMC3	RAD21	$1.37x10^{-06}$	$3.7x10^{-6}$	0.00025	0.62614
	HEK-293	SMC3	RAD21	$< 2.2x10^{-16}$	$< 2.0x10^{-16}$	$< 2.0x10^{-16}$	$< 2.0x10^{-16}$
	HeLa	SMC3	RAD21	$< 2.2x10^{-16}$	$< 2.0x10^{-16}$	$< 2.0x10^{-16}$	$< 2.0x10^{-16}$
	Hep G2	SMC3	RAD21	$< 2.2x10^{-16}$	$1.1x10^{-1}$	$< 2.0x10^{-16}$	$< 2.0x10^{-16}$
	MCF7	RAD21	STAG1	$< 2.2x10^{-16}$	$< 2.0x10^{-16}$	$< 2.0x10^{-16}$	0.16
	SK-N-SH	SMC3	RAD21	$< 2.2x10^{-16}$	$< 2.0x10^{-16}$	$< 2.0x10^{-16}$	$< 2.0x10^{-16}$
MOUSI	Ch12	SMC3	RAD21	$< 2.2x10^{-16}$	$< 2.0x10^{-16}$	$< 2.0x10^{-16}$	$< 2.0x10^{-16}$
	Liver (C57BL/6)	RAD21	STAG1	$< 2.2x10^{-16}$	$< 2.0x10^{-16}$	$< 2.0x10^{-16}$	$2.0x10^{-11}$
	MEL	RAD21	SMC3	$2.0x10^{-14}$	$< 2.0x10^{-16}$	$< 2.0x10^{-16}$	$< 2.0x10^{-16}$
	Stem cell (C57BL/6)	SMC1	SMC3	$< 2.2x10^{-16}$	$1.2x10^{-14}$	$< 2.0x10^{-16}$	0.87

Table 5. Results of statistic analysis in case of more than two coherent samples. We used a Freidman test with a Namenyi posthoc test in the case of the CTCF samples, which had more than two available parallel cohesin ChIP-seq datasets.

SAMPLE PAIRS	HUMAN	A549	CTCF	RAD21	$< 2.2x10^{-16}$
		ECC-1	CTCF	RAD21	$< 2.2x10^{-16}$
		H1-hESC	CTCF	RAD21	$< 2.2x10^{-16}$
		THP1	CTCF	RAD21	$< 2.2x10^{-16}$
		HLS554P	CTCF	SMC1	$< 2.2x10^{-16}$
		BCBL-1	CTCF	SMC1	$< 2.2x10^{-16}$
	MOUSE	C57BL/6 Pro-B cell	CTCF	RAD21	$< 2.2x10^{-16}$
		C57bl/6 Pre-pro B cell	CTCF	RAD21	$< 2.2x10^{-16}$
		C57bl/6 Hepatocyte	CTCF	RAD21	$< 2.2x10^{-16}$
		C57bl/6 BMDM LG26	CTCF	RAD21	$< 2.2x10^{-16}$
		C57bl/6 BMDM vehicle	CTCF	RAD21	$< 2.2x10^{-16}$
		C57BL/6 and 129 Embryonic fibroblast	CTCF	SMC1	$< 2.2x10^{-16}$
		C57BL/6 Spleen B cell	CTCF	SMC1	$< 2.2x10^{-16}$
		C57BL/6 Fibroblast	CTCF	SMC3	$< 2.2x10^{-16}$

Table 6. Results of statistic analysis in case of two coherent samples. Wilcoxon signed-rank test was used in the statistical analysis comparing two matched samples.

The SMC1 and SMC3 positions showed a clear co-localization considering the distance distribution scatterplots (**Figure 20C-D**), boxplots (**Figure 21-22**), and histograms (**Figure 23-24**).

We implemented a cell type specific analysis for more accuracy. For this, we focused on CTCF/cohesin co-occupied sites and selected samples derived from the same cell line and targeting as many types of human and mouse experiments as could be obtained. The pairwise comparison of CTCF and the corresponding cohesin ChIP-seq signals was calculated on every concerned binding site per cell type. The cell specific result was then plotted on distance distribution histograms and boxplots (**Figure 21-24**). The result of this specified analysis was coinciding with the previous global analysis. The distance distribution confirmed the relative co-localization pattern between SMC1 and SMC3, just like for RAD21 and STAG1/2, while the CTCF had an easily distinguishable distribution curve.

A slight difference could be observed in the standard deviation of distances between CTCF and the cohesin subunits. The CTCF data showed a narrow distance distribution curve, while RAD21 and STAG1 had a slighter curve and SMC protein had a characteristic broad distribution (**Figure 23-24**). We investigated the network of transcription factors and cofactors on a wide spectrum of transcription factor binding sites. The distribution distance could be explained, by the physical proximity of different proteins and their DNA binding specificity. The direct and sequence specific DNA binding allowed less movability of ChIP-seq summit positions, while the indirect interaction had looser protein positioning.

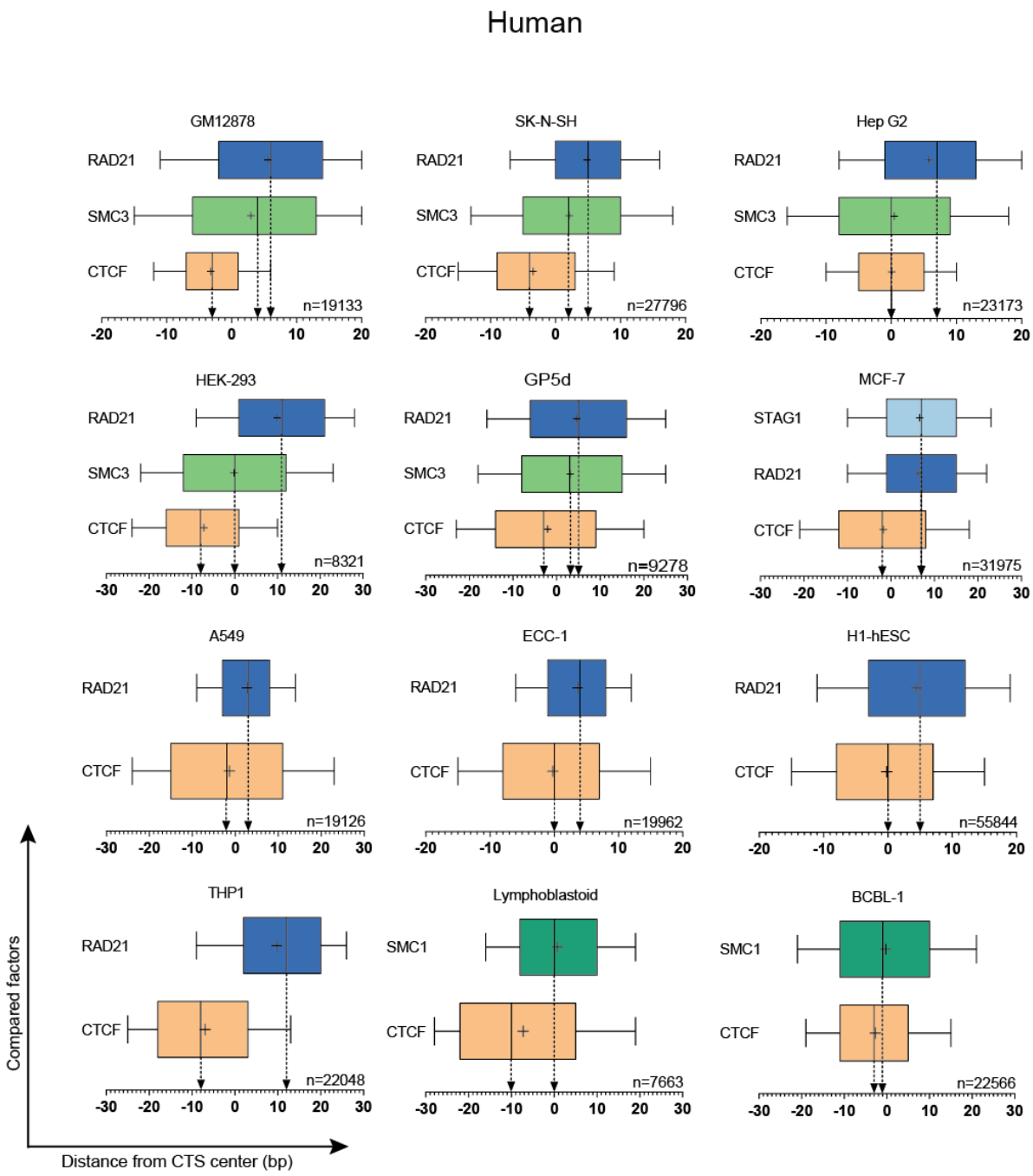


Figure 21. Box plot representation of the strand specific shift between CTCF and cohesin proteins in human cell lines. Box plots show the median positions (vertical lines), average positions ("+"), first and third quartiles (box borders), and 10-90 percentiles (whiskers) of the distributions.

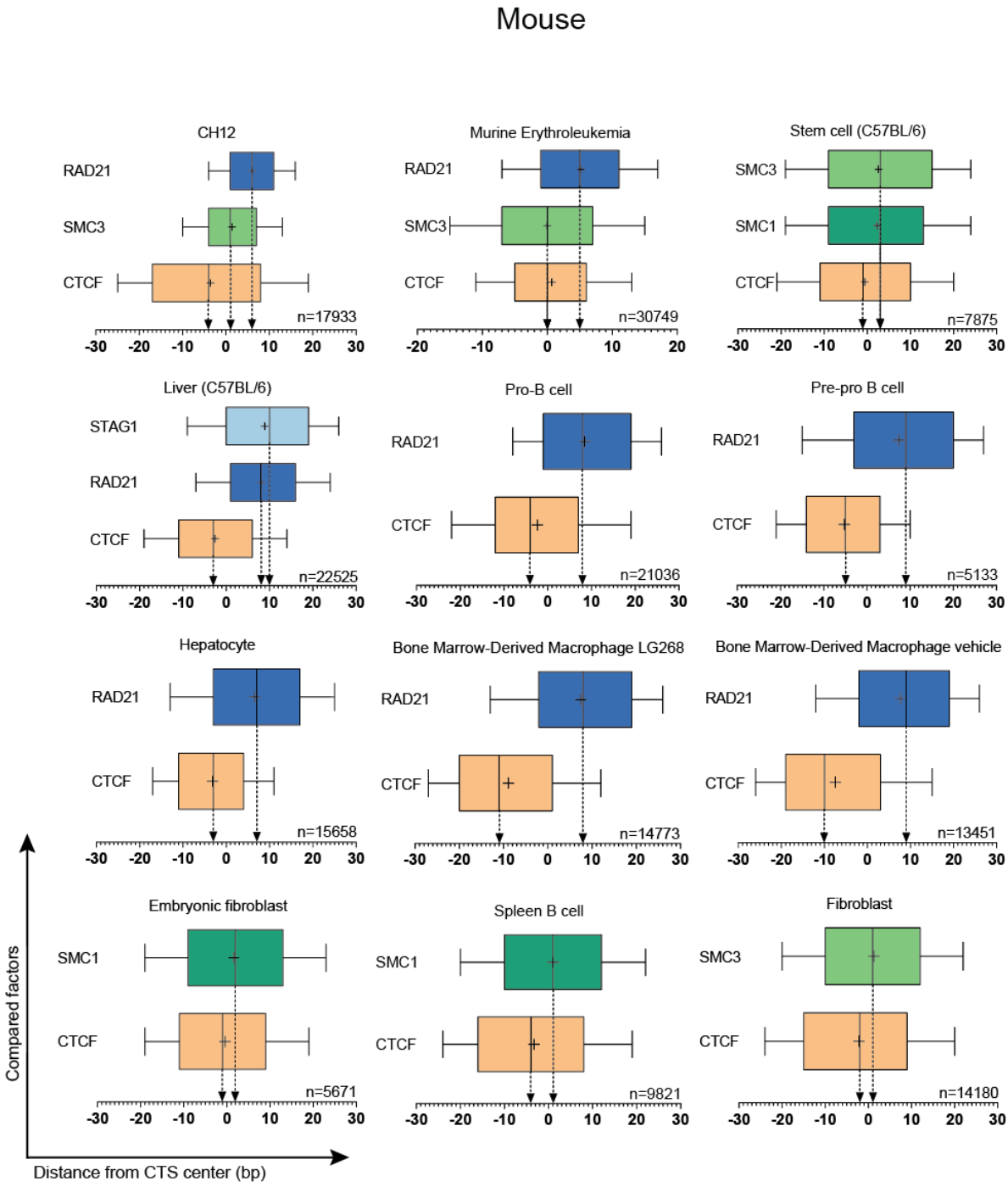


Figure 22. Box plot representation of the strand specific shift between CTCF and cohesin proteins in mouse cell and tissue types. Box plots show the median positions (vertical lines), average positions ("+"), first and third quartiles (box borders), and 10-90 percentiles (whiskers) of the distributions.

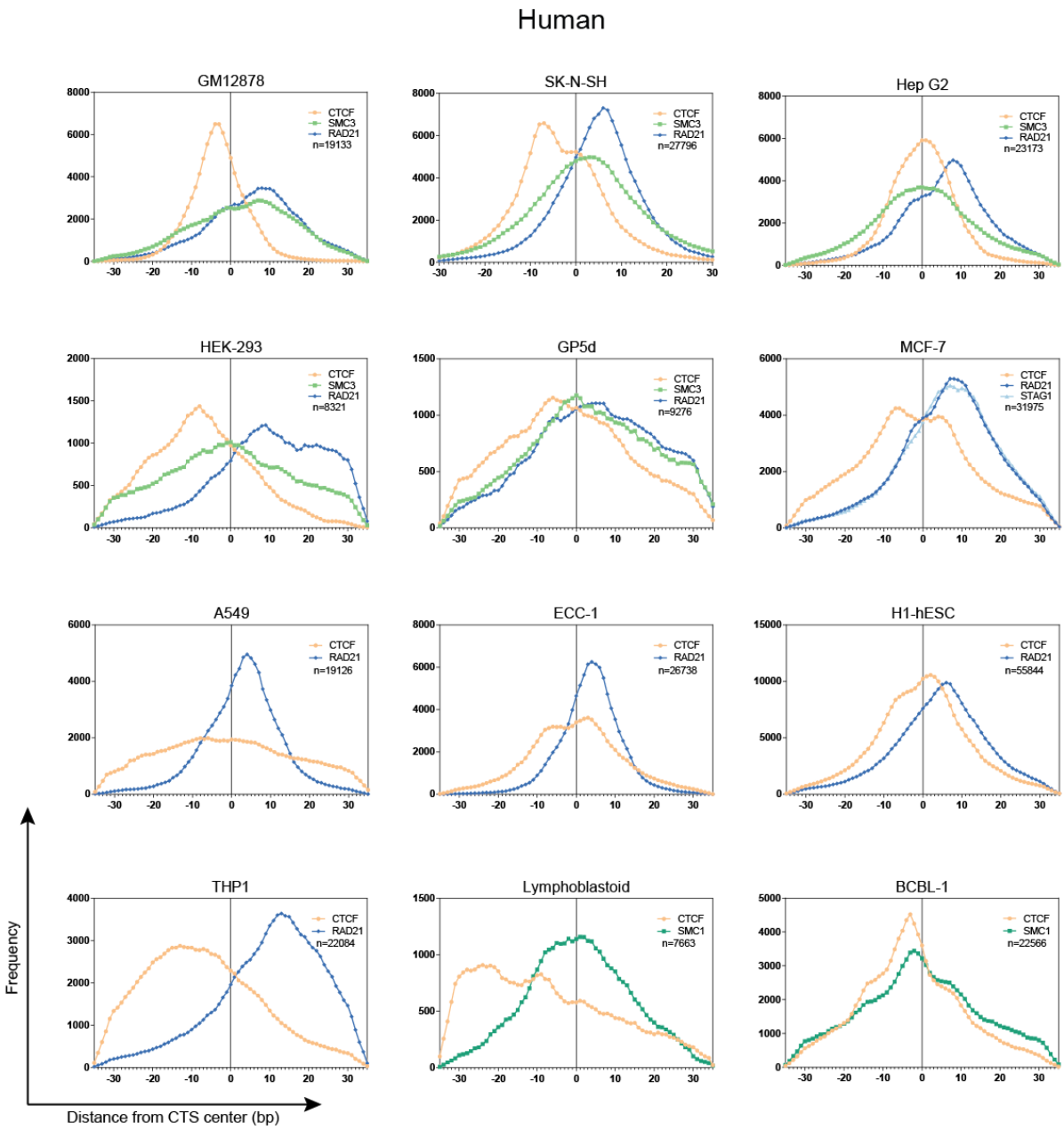


Figure 23. Shift between CTCF/cohesin proteins in human cell lines. Histograms show the distribution of the peak summits of CTCF/cohesin proteins relative to the midpoint of CTSs using a 5 bp sliding window.

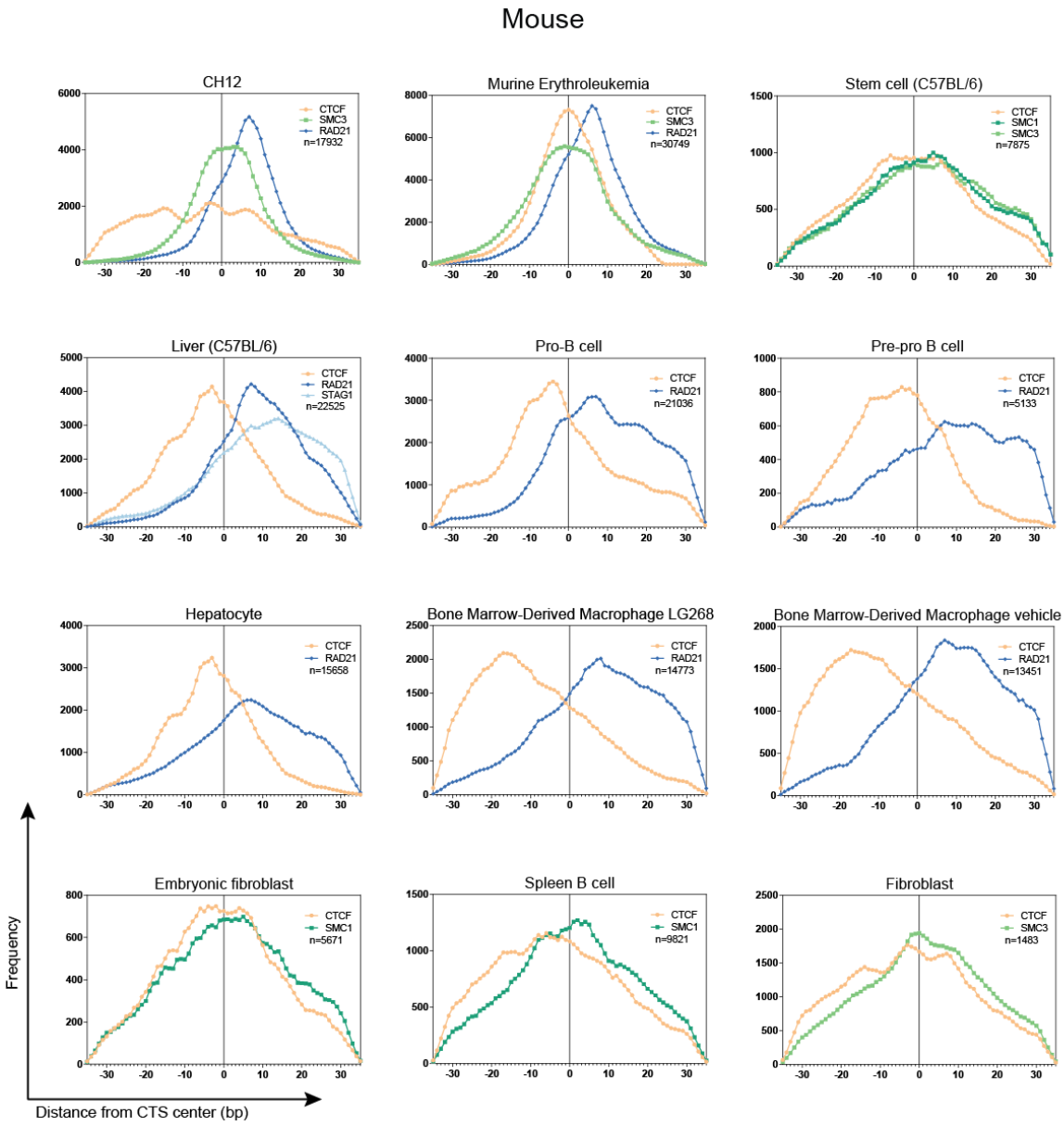


Figure 24. Shift between CTCF/cohesin proteins in mouse cell and tissue types. Histograms show the distribution of the peak summits of CTCF/cohesin proteins relative to the midpoint of CTSs by using a 5 bp sliding window.

We have also re-analyzed the available HeLa DNase-seq and CTCF ChIP-exo datasets and found that they exactly mark the borders of the region we had found to be occupied on the DNA by CTCF/cohesin proteins (**Figure 25**).

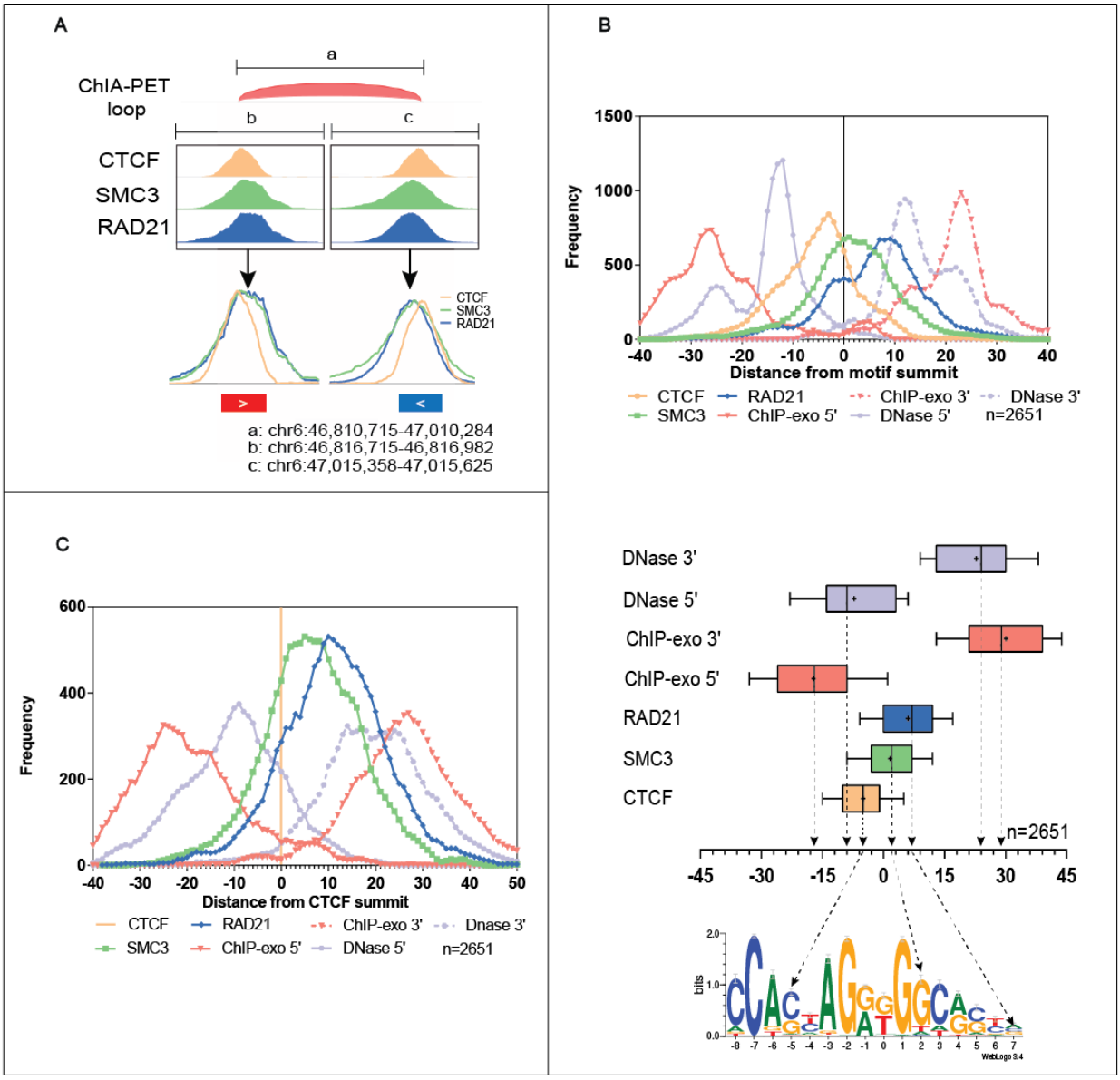


Figure 25. The boundaries of genomic regions covered by CTCF/cohesin. (A) Representative example of the strand specific CTCF-cohesin shift derived from CTCF, SMC3, and RAD21 ChIP-seq data in HeLa cells. Available ChIA-PET (target: CTCF) was used to identify CTCF binding sites that were involved in the chromatin loop (SRX160885). The red and blue boxes indicate the CTCF elements on the forward and reverse strand, respectively. (B) ChIP-seq peak summit positions of CTCF/cohesin complex components show a conserved, strand-specific distance pattern relative to both the CTS center and the DNase-seq footprint and ChIP-exo borders in HeLa cells (SRA datasets SRX080392, SRX150650, SRX150464, SRX098243, SRX100899). Top: histogram of summit distribution of the CTCF and cohesin bound sites using a 5 bp sliding window. Middle: box plots show the median positions (vertical lines), average positions ("+"), first and third quartiles (box borders), and 10-90 percentiles (whiskers) of the distributions. The bottom panel shows the mapping on the CTCF motif logo. (C) Distance distribution of cohesin proteins and "edge markers" relative to CTCF. The horizontal axis represents the

distance of peak summits and ChIP-exo and DNase footprint borders relative to the CTCF summits (orange line). The vertical axis represents the distance frequency. A rolling mean with a 5 bp window was applied to smooth the frequency.

Since we know their structural characteristics and their heterodimerization domains, this observation was not surprising. In contrast, their signal position between CTCF and RAD21-STAG1 raised questions. As previously mentioned, the CTCF-cohesin subunit interaction order has the following sequence: CTCF-> STAG1/2-> RAD21-> SMC1/3. To understand the identified serial order between ChIP-seq signals, we converted the positional distances (shifts) into approximate 3D spatial constraints. For this, we took all average positions from all human and mouse cells (**Table 7**) and chose the human median positions for CTCF (-4), SMC1/3 (+1) and RAD21 (+6), and for STAG1 (+7) (**Table 8**).

	Cell line	Cohesin 1.	Cohesin 2.	Median position			Mean position			motif number
				CTCF	Cohesin 1.	Cohesin 2.	CTCF	Cohesin 1.	Cohesin 2.	
HUMAN	A549	RAD21	-	-2	3	-	-1.32	2.78	-	19126
	BCBL-1	SMC1	-	-3	-1	-	-2.70	-0.26	-	22566
	ECC-1	RAD21	-	0	4	-	-0.26	3.65	-	26738
	GM12878	SMC3	RAD21	-3	4	6	-3.22	2.98	5.54	19133
	GP5d	SMC3	RAD21	-3	3	5	-2.04	3.13	4.55	9276
	H1-hESC	RAD21	-	0	5	-	-0.19	4.37	-	55844
	HEK-293	SMC3	RAD21	-8	0	11	-7.20	-0.11	9.97	8321
	HeLa	SMC3	RAD21	-4	1	6	-4.61	1.50	5.22	21994
	HepG2	SMC3	RAD21	0	0	7	0.09	0.47	5.80	23173
	Lymphoblastoid	SMC1	-	-10	0	-	-7.24	0.80	-	7663
	MCF-7	RAD21	STAG1 (SA1)	-2	7	7	-1.78	6.26	6.57	31975
	SK-N-SH	SMC3	RAD21	-4	2	5	-3.44	2.10	4.83	27796
	THP1	RAD21	-	-8	12	-	-6.95	9.75	-	22048
MOUSE				-4	1	6	-3.59	1.41	6.02	17932
	Embryonic fibroblast	SMC1	-	-1	2	-	-0.42	1.65	-	5671
	Fibroblast	SMC3	-	-2	1	-	-2.16	1.29	-	14830
	Hepatocyte	RAD21	-	-3	7	-	-3.16	6.63	-	15658
	Liver (C57BL/6)	RAD21	STAG1 (SA1)	-3	9	11	-2.98	8.81	9.87	22525
	MEL	SMC3	RAD21	0	0	5	0.70	-0.01	5.17	30749
	Pro-B cell	RAD21	-	-4	8	-	-2.33	8.44	-	21036
	Pre-pro B cell	RAD21	-	-5	9	-	-5.21	7.38	-	5133
	Spleen B cell	SMC1	-	-4	1	-	-3.28	1.01	-	9821
	Stem cell (C57BL/6)	SMC1	SMC3	-1	3	3	-0.60	2.28	2.61	7875
	BMDM LG268	RAD21	-	-11	8	-	-8.90	7.47	-	14773
	BMDM vehicle	RAD21	-	-10	9	-	-7.44	7.69	-	13451

Table 7. Summary table of CTCF-cohesin samples. The table summarizes the average and median positions of factor summits relative to the center of the CTS.

		Mean					Median		
		CTCF	RAD21	SMC1/3	STAG1	CTCF	RAD21	SMC1/3	STAG1
SCATTER PLOT VALUES	MOUSE	-4.29	9.47	3.63	12.00	-4	8	3	14
	HUMAN	-3.83	6.00	0.63	7.33	-4	5	0.5	8
	HUMAN + MOUSE	-3.99	7.89	3.02	11.06	-4	8	3	13
BOX PLOT VALUES	MOUSE	-3.00	7.07	1.07	9.87	-3	7	1	11
	HUMAN	-2.38	4.97	1.34	6.57	-3	6	1	7
	HUMAN + MOUSE	-2.62	5.70	1.23	7.93	-3	6	1	9

Table 8. Average shift values of CTCF/cohesin summits related to CTS. The table shows the calculated mean and median position of the samples.

This pattern was then mapped on the surface of a B-DNA model that we built using a sequence dependent modeling procedure (**Figure 26**). Normally, 10/11 base pairs are involved in one turn of B-DNA (Watson & Crick, 1953). The summit distance gap between CTCF and RAD21-STAG1/2 is approximately equal to one turn of the DNA (between position -4 and position 6-7). This makes the peak summits of CTCF, STAG1/2, and RAD21 map to one face of the double helix, while the intermediate contact sites of SMC1/3 (position +1) are located on the opposite face (Phillips-Cremins & Corces, 2013; West, Gaszner, & Felsenfeld, 2002).

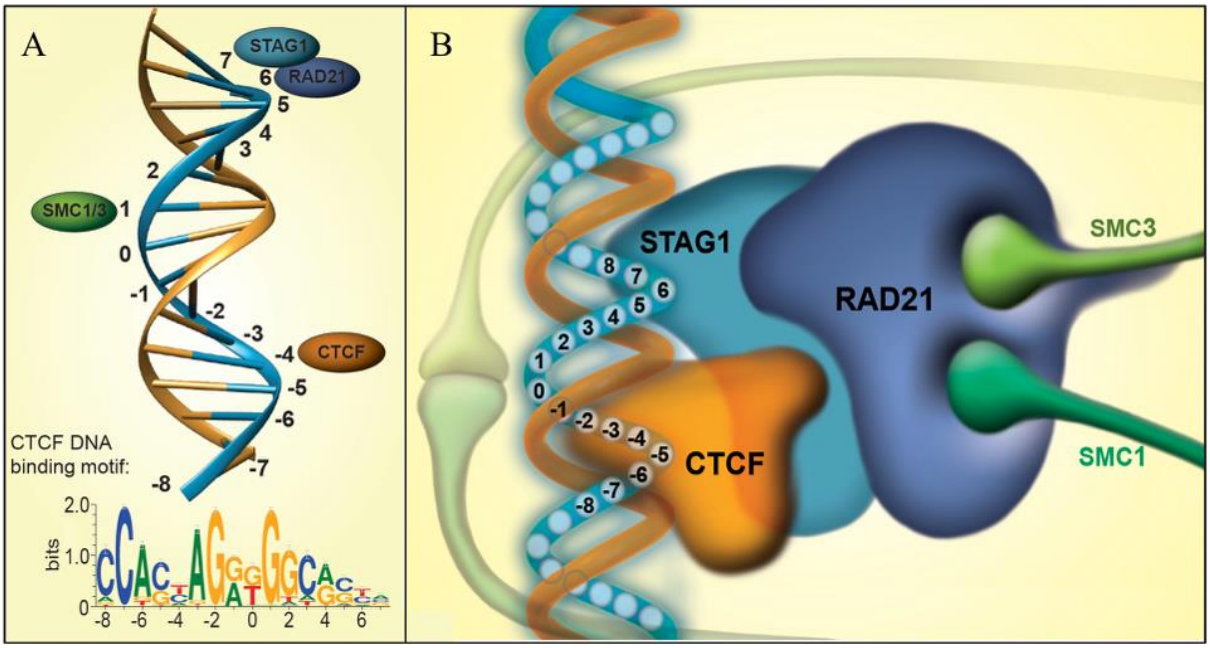


Figure 26. Visualizing the shift values on the B-DNA model. A) Mapping the summit shift values of the cohesin subunits onto a schematic model of B-DNA shows that CTCF, RAD21, and STAG1/2 are located on one side of the helix, while SMC1 and SMC3 are on the other face. B) 3D topology of CTCF/cohesin complex on the DNA. The positions correspond to the median values indicated of human average positions.

Atomic details for parts of the cohesin complex are known (Hashimoto et al., 2017; Melby et al., 1998). Combining the published structural models and our position data, we could give a possible explanation of the protein positions and create a hypothetical model for CTCF mediated chromatin looping (**Figure 27**). The double embrace model was integrated into our topological model that involves two cohesin rings. This model can help explain the unusual positioning of ChIP-seq summits of SMC proteins (Nasmyth & Haering, 2009). The SMC hinge domain has nonspecific DNA binding capability, which stabilizes the chromatin loop formation. The hinge and head domains are separated by a relatively long rod like structure, which embraces the chromatin (**Figure 8**). The head connects directly to other members of the cohesin ring and indirectly to the CTCF protein. On the opposite side of the rod, the hinge domain can form a non-specific bond with the distal DNA region, which comes close during loop formation (Gruber et al., 2006; Sun et al., 2013). This means that the detected SMC signals

belong to the other anchor region's cohesin ring (**Figure 27**). This explains the opposite localization of SMC on B-DNA relative to CTCF-RAD21 and STAG1. Our results clearly suggest that RAD21 and STAG are in contact with each other (Haering, Lowe, Hochwagen, & Nasmyth, 2002) and also either one or both proteins are in close contact with DNA, around the 3' end of CTCF binding site (**Figure 26B**). Their position preference is unspecific and may be the consequence of physical proximity to DNA and the cross-linking procedure during chromatin immunoprecipitation. The double embrace arrangement provides a testable hypothesis that may help to clarify several, seemingly contradictory features, of loop closure.

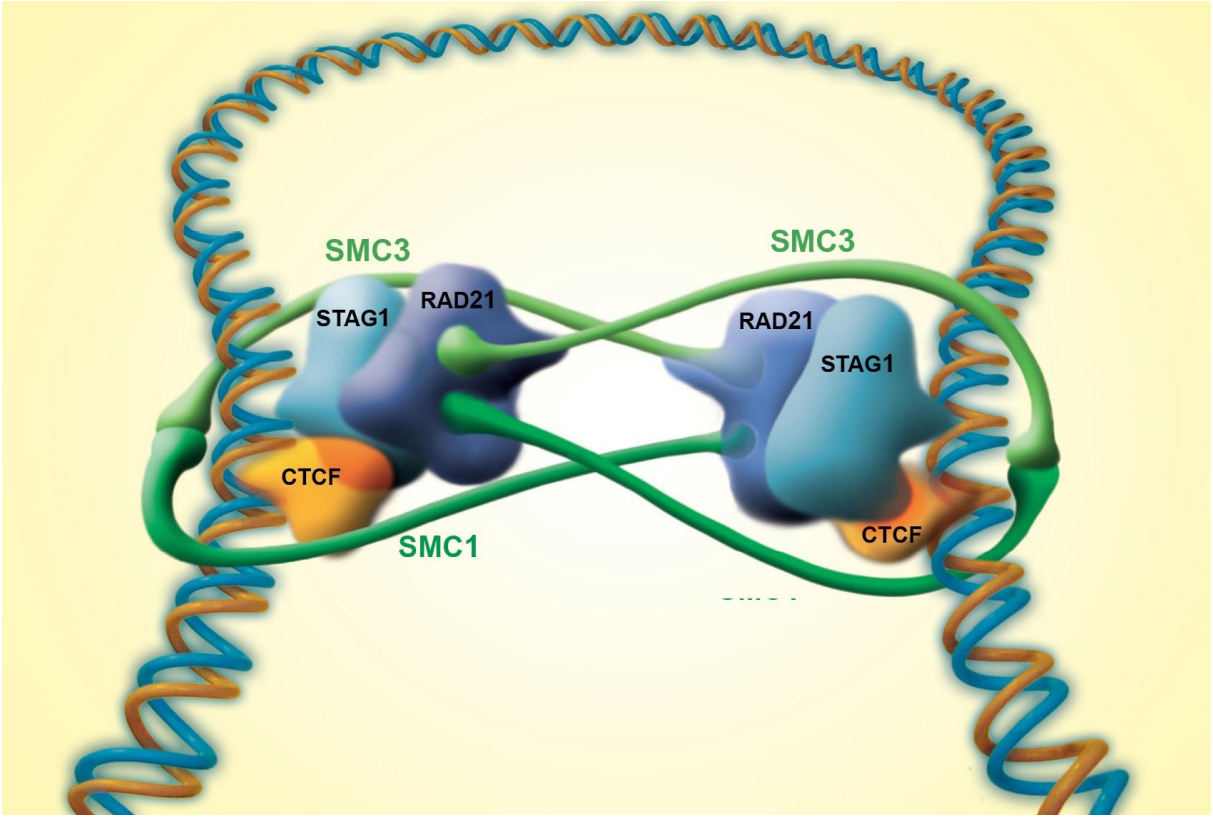


Figure 27. Double embrace model of CTCF mediated chromatin looping. We combined our measurement results and already existing molecular structure data to create a hypothetical model for CTCF mediated chromatin looping. The model explains how a DNA-loop is fixed by flanking the CTCF/cohesin complexes.

4.2.Computer simulation and experimental validation

For validation purposes, we performed computer simulations to test the reproducibility of our data by chance. The reproducibility of peak shift values was tested on the HeLa dataset. The peak shift was measured between the peak summit of the proteins indicated (CTCF, Rad21, SMC3) and the center of the CTCF binding site. In **Figure 28**, the reproducibility was characterized using the standard deviation of the mean (Y-axis) that was determined from a number of observed peak shifts (X-axis). The inset shows that approximately 100 shift values are necessary to reach a reproducibility of +/-1 nucleotide. In our experiments, we normally used more than 5000 peaks that roughly corresponded to a reproducibility of 0.1. Naturally, the reproducibility estimates vary with the quality/coverage of the dataset. In practice, we rounded the peak shift values to one nucleotide.

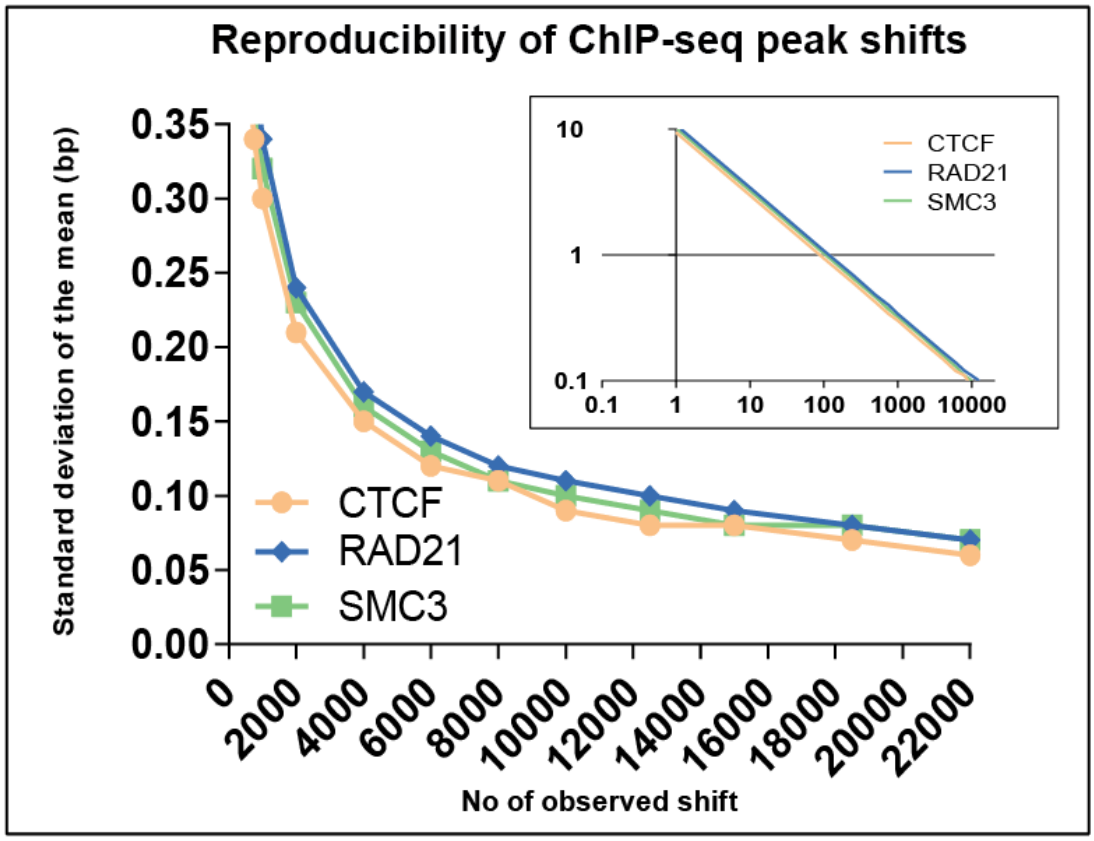


Figure 28. The reproducibility of ChIP-seq peak shifts in a HeLa cell experiment. The reproducibility was characterized as the standard deviation of the mean (Y-axis) determined from a number of observed peak shifts (X-axis). The peak shift was measured between the peak summit of the proteins indicated in the figure (CTCF-SRX102984, Rad21-SRX150650, SMC3-SRX150464) and the center of the CTCF binding site (CTS). The inset shows that approximately 100 shift values are necessary to reach a reproducibility of +/-1 nucleotide. In our experiments, we used typically more than 5000 peaks, which roughly correspond to a reproducibility of 0.1. These estimates vary with the quality/coverage of the dataset.

For experimental validation, we applied our method to known interactions between protein complexes and DNA. We examined DNA binding events with symmetric and asymmetric properties. In this context, asymmetric binding properties resulted in non-zero ChIP-seq peak shifts, which could serve as positive controls. On the other hand, no peak shift was expected for symmetrical binding complexes; thus, symmetrical binding complexes could serve as negative controls.

Positive control. Forkhead (FOX) proteins have been described as pioneer factors, which can open up compact DNA by superseding linker histones (Sahu et al., 2011). Thus,

FOX proteins are key transcription factors (TFs) in the development of tumorigenesis, e.g. in steroid hormone dependent cancers. The FOXA1::AR (androgen receptor) composite element was discovered in human prostate cancer derived cells (Sahu et al., 2011). In this element, there are spacer of four nucleotides between the recognition sites of FOXA1 and AR (**Figure 29**). Thus, we expected a shift between the co-binding proteins on the DNA. We detected a significant number of sites bound by both FOXA1 and AR in VCaP (3477) and a feasible number of such sites in LNCaP prostate cancer cells (713), where a ~4 bp shift was observed (at the level of $P<2.2x10^{-16}$ and $P=5.06x10^{-08}$, respectively, according to the Wilcoxon signed-rank test) (Toropainen et al., 2015) (**Table 3**). These data matched well with our preliminary expectations and showed that this number of ChIP-seq summits is sufficient to determine the strand specific relative location of the co-binding TFs.

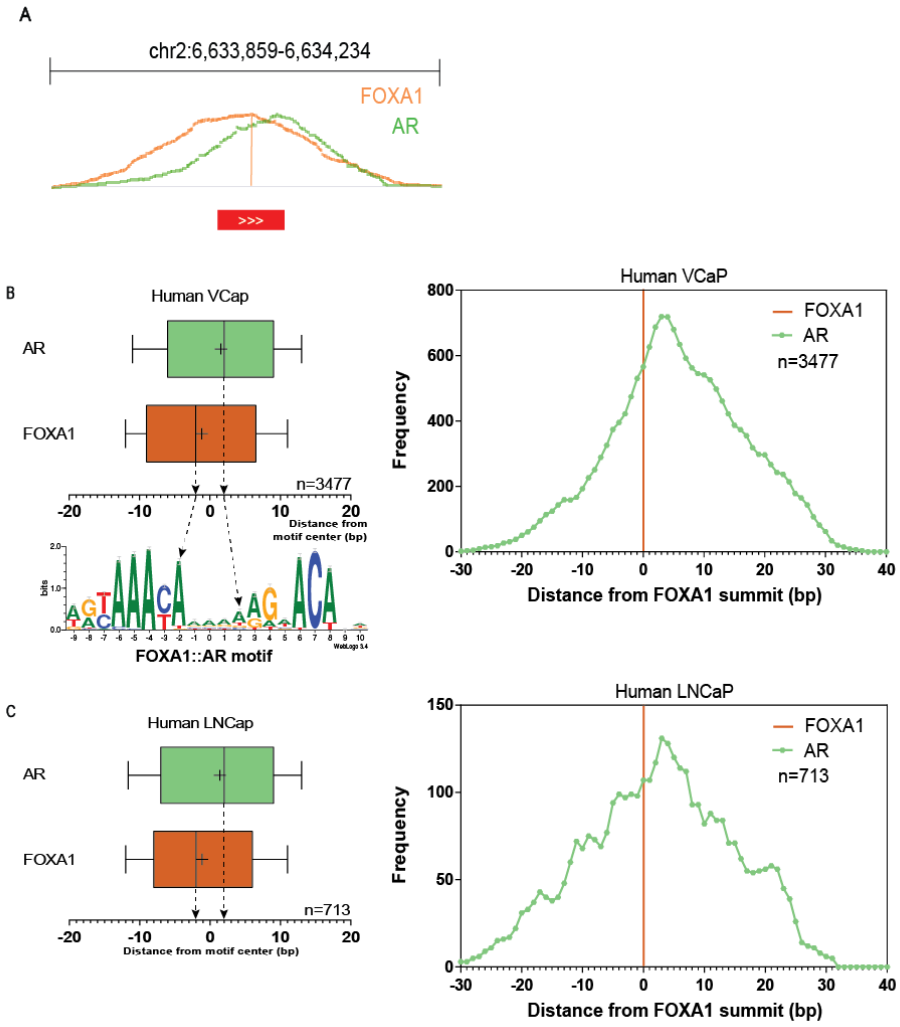


Figure 29. Shift between interacting transcription factors (positive control). (A) Representative example of the strand specific FOXA1/AR shift derived from ChIP-seq data in a human VCaP cell line (Toropainen et al., 2015). The red box indicates the FOXA1::AR composite element. (B) Box plots indicate the mean (shown as "+") and median (vertical line) peak summit positions of AR and FOXA1 in 3477 bound regions in the VCaP cell line. The bottom panel shows the mapping on the FOXA1:AR motif logo. The histogram (at right) shows the distance distribution of AR relative to the FOXA1 on their common binding sites. The horizontal axis represents the distance of AR summits (green curve) relative to the FOXA1 summits (orange line), while the vertical axis represents the distance frequency. A rolling mean with a 5 bp window was applied to smooth the frequency curves. (C) The relationship of FOXA1 and AR on their 713 composite elements (shown in Figure S12B) in the LNCaP cell line.

Negative controls: Coregulators that do not bind to DNA directly but through TF(s). Thus, coregulators are expected to show no shift compared to their respective DNA-binding factor(s). Retinoid X receptor (RXR) is the heterodimerizing partner of class II nuclear

receptors (NRs) (**Table 3**), which are typically activated by lipid molecules (Mangelsdorf et al., 1995). For this, they need coregulators, such as P300 (Daniel et al., 2014). Although the NR binding site has direction, one can expect the RXR and P300 at the same location. Indeed, P300 did not show any strand specific shifts relative to the location of RXR at 779 commonly bound NR (half) sites in mouse 3T3-L1 preadipocytes (P=0.5287) (**Figure 30A-B**) (Siersbaek et al., 2014).

We tested a third complex containing stem cell specific TFs (OCT3/4, SOX2, KLF4, and cMYC), which are involved in the dedifferentiation of many cell types (Takahashi & Yamanaka, 2006). These complexes also include the NANOG homeodomain protein, which is a pluripotent factor responsible for self-renewal (Chambers et al., 2003). For the comparison, we used the composite element of OCT3/4 (POU5F1) and SOX2, which showed the co-occurrence of POU5F1 and NANOG proteins at 2206 sites. As NANOG binding – in a similar manner to the coregulators – is secondary, a minimal, statistically insignificant shift (P=0.1776) was detected relative to the location of POU5F1 in mouse embryonic stem cells (**Figure 30C**) (**Table 9**)(Galonska, Ziller, Karnik, & Meissner, 2015).

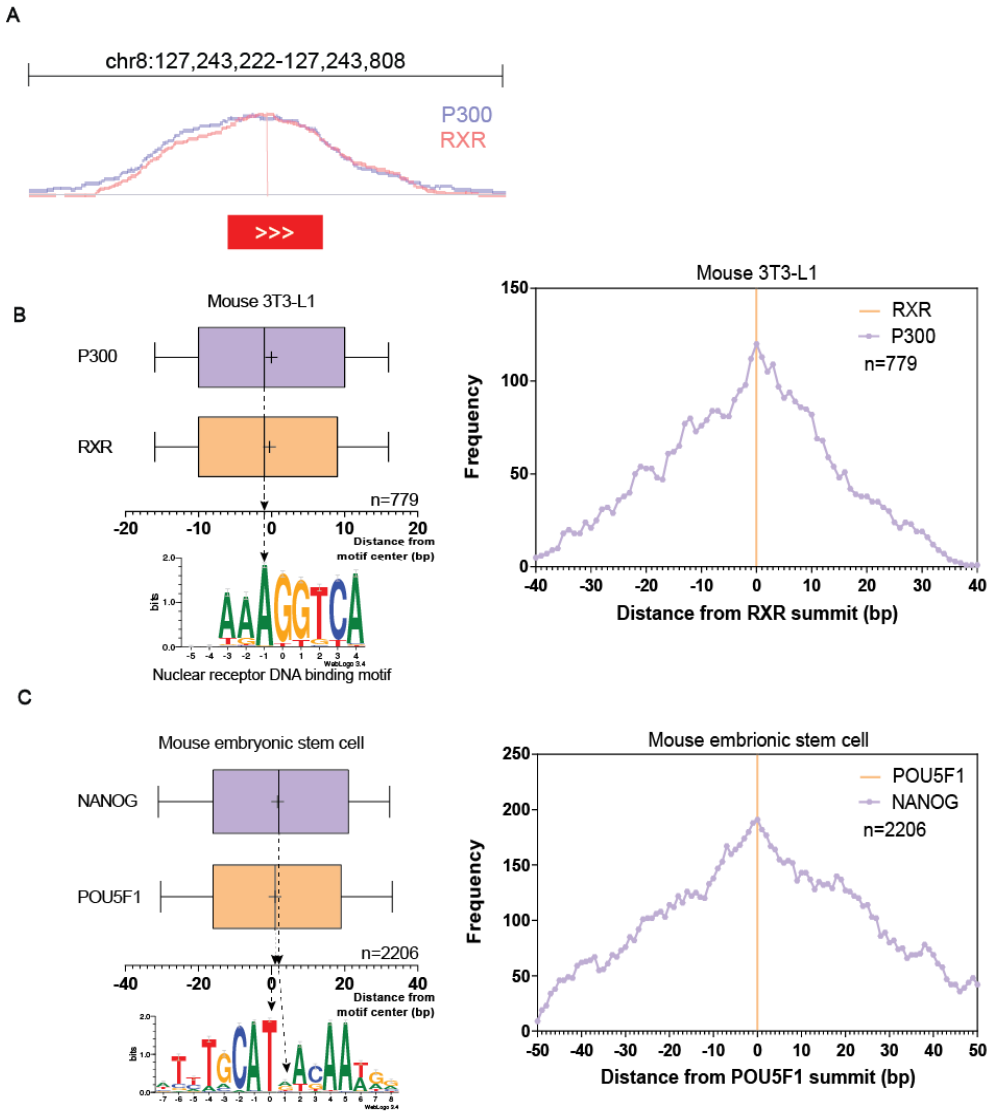


Figure 30. Lack of shift between interacting transcriptional regulator proteins (negative control). (A) Representative example of the lack of a strand specific RXR/P300 shift in the mouse 3T3-L1 cell line (Siersbaek et al., 2014). The red box indicates the nuclear receptor half site. (B) Box plots indicate the mean (shown as "+") and median (vertical line) peak summit positions of P300 and RXR in 779 bound regions in the 3T3-L1 cell line. The bottom panel shows the mapping on the nuclear receptor motif logo. The histogram (at right) shows the distance distribution of P300 relative to the RXR on commonly occupied regions. The horizontal axis represents the distance of P300 summits (purple curve) relative to the RXR summits (orange line) and the vertical axis represents the distance frequency. A rolling mean with a 5 bp window was applied to smooth the frequency curves. (C) The relationship of NANOG and POU5F1 on 2206 POU5F1 binding sites in mouse embryonic stem cells (Galonska et al., 2015).

	Cell line	Factor 1	Factor2	Wilcoxon signed rank p-value
Negative control	3T3L1	RXR	P300	0.5287
	KH2 mESCs	POU5F1	NANOG	0.1776
Positive control	LNCap	FOXA1	AR	$5.06x10^{-8}$
	VCap	FOXA1	AR	$< 2.2x10^{-16}$

Table 9. Results of statistic analysis in case of two coherent samples. Wilcoxon signed-rank test was used in the statistical analysis comparing two matched samples.

4.3.CTCF-binding site orientation shapes the chromatin loops

In light of the previous result, we wanted to look into the correlation between topological data and chromatin looping. The 3C techniques were developed to analyze the spatial organization of chromatin. The 3C techniques, in combination with High-Throughput Sequencing, enable the identification of DNA loops on the genome level. We can rely on chromatin immunoprecipitation for TAD mapping (ChIA-PET for TAD associated proteins) and HiC data (Davies, Oudelaar, Higgs, & Hughes, 2017). We used publicly available ChIA-PET data (prepared with the CTCF antibody) to identify the orientation of CTCF motifs, which are involved in chromatin looping. First, we downloaded prepared MCF7 CTCF ChIA-PET data from the ENCODE database, because it has biological replicates (GEO database GSM970215). After the complex processing of these data, interaction tables were generated to store information about interacting distal DNA regions. To simplify the prepared data, the identified interaction can be divided into 3 parts: i) first anchor, ii) interior loop region, and iii) second anchor. The anchor regions frame the chromatin interaction and serve as the binding platform for the loop mediating proteins. In terms of 3C data processing, the anchor regions represent regions with variable length (depending on the resolution of technique) that contain the possible interaction points. Most available Hi-C datasets have relatively low resolution, between 25 to 40 kb, and the most advanced procedures produce 5kb data. The resolution could

be improved to 1 kb in the case of ChIA-PET. The MCF7 CTCF ChIP-seq have an average anchor length of ~850 nucleotides (**Table 10**). This provides more accurate identification of the anchoring of the CTCF binding site.

sample_name	**All_interactions number**	**SD of anchor length**	**Filtered interaction number 200kbp <**	**Filtered 300kbp <**
MCF7_ChIA-PET_rep1	53762	285	31681	**37829**
MCF7_ChIA-PET_rep2	22099	251	12536	14909
MCF7_ChIA-PET_consensus	8522		6910	**7234**

Table 10. The number of identified interactions in the MCF7 cell line. In the ENCODE database, two replicas of MCF7 CTCF ChIA-PET (GSM970215) data are presented. There is a large difference between the replicas with respect to the identified CTCF interaction numbers. We created a consensus dataset, which contains only the interactions present in both replicas, to extract permanent loops.

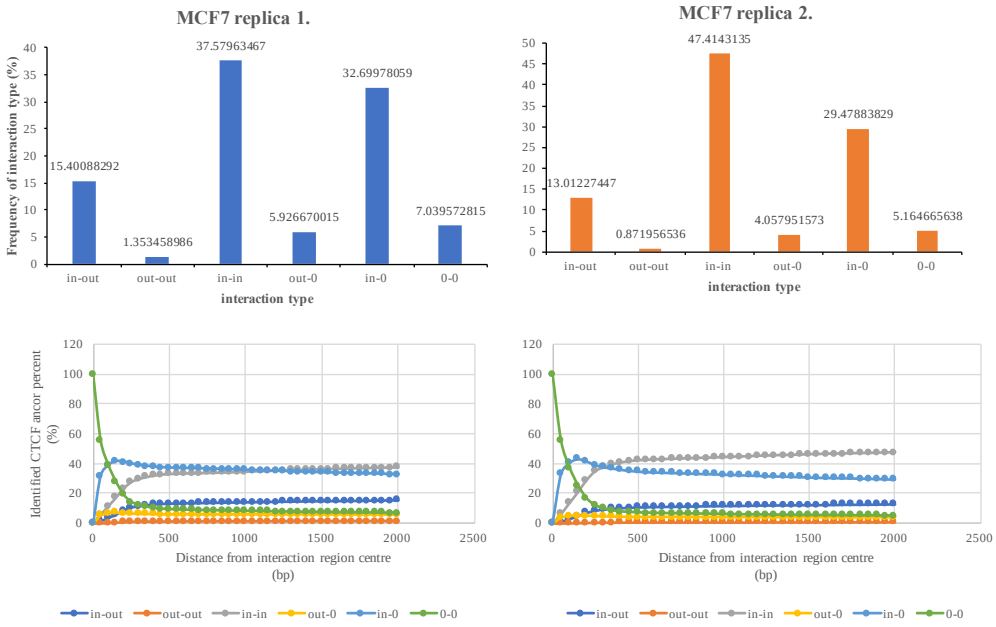
To find the exact interaction point, we created a pipeline that scans anchor regions. The pipeline requires CTCF ChIP-seq data for experimental validation and motif enrichment analyses. The enriched de novo CTCF motif was remapped to the MCF7 CTCF peak regions to find the motif instances. Then, a home-made scanner program found the most proximal CTCF motifs relative to the midpoint of the anchor region (**Figure 31A**). This allowed us to distinguish different types of loops. Since one loop has two anchor regions and we paired every anchor with a CTCF, we can cluster the loops depending on the CTCF binding site orientation as follows (**Figure 31A, Figure 31C**):

- Convergent: the motifs face each other and are directed inside the loop
- Divergent: the motifs are oriented in opposite directions towards the outside of the loop

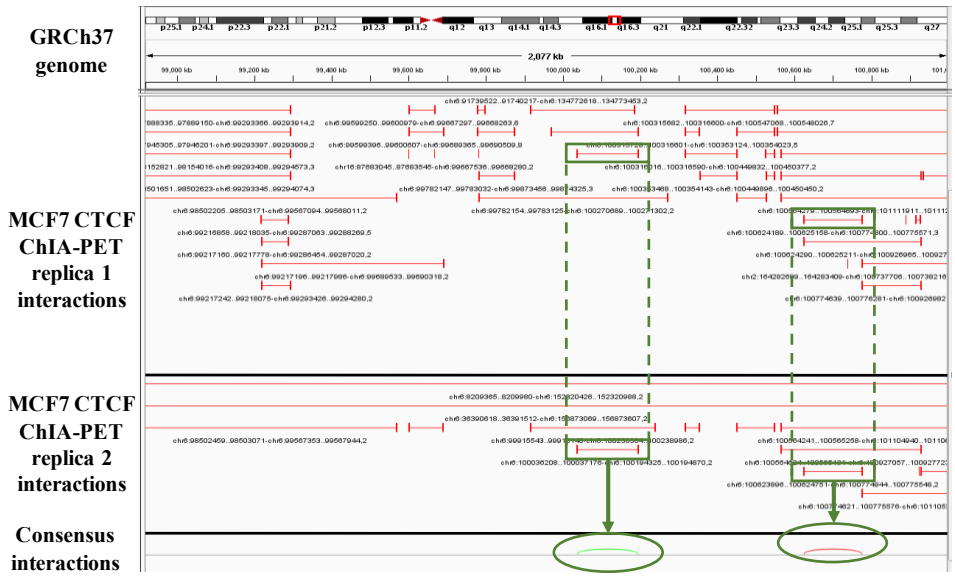
- Same direction on the strand: both motifs are on the same strand
- Convergent with unidentified pair: One anchor's motif is not identified, but the other motif anchor faces the inside of the loop
- Divergent with unidentified pair: One anchor's motif is not identified, but the other motif anchor faces the outside of the loop

The frequency analysis showed a clear enrichment of convergent motif orientations. We compared this data with ChIA-PET results from other cell types. First, we created a consensus MCF7 loop set, considering the interactions with overlapping CTCF binding sites on both anchor regions (**Figure 31B**). Then, we processed the consensus loop set and the other downloaded ChIA-PET data (K562, mouse limbbud) with the previously described procedure. The results were congruent (data not shown).

A


MCF7 replica 1.
Frequency of interaction type (%)
15.40088292
1.353458986
37.57963467
5.926670015
32.69978059
7.039572815
in-out
out-out
in-in
out-0
in-0
0-0
interaction type
MCF7 replica 2.
13.01227447
0.871956536
47.4143135
4.057951573
29.47883829
5.164665638
in-out
out-out
in-in
out-0
in-0
0-0
interaction type
Identified CTCF ancor percent (%)
Distance from interaction region centre (bp)
in-out
out-out
in-in
out-0
in-0
0-0

B


GRCh37 genome
MCF7 CTCF ChIA-PET replica 1 interactions
MCF7 CTCF ChIA-PET replica 2 interactions
Consensus interactions

C

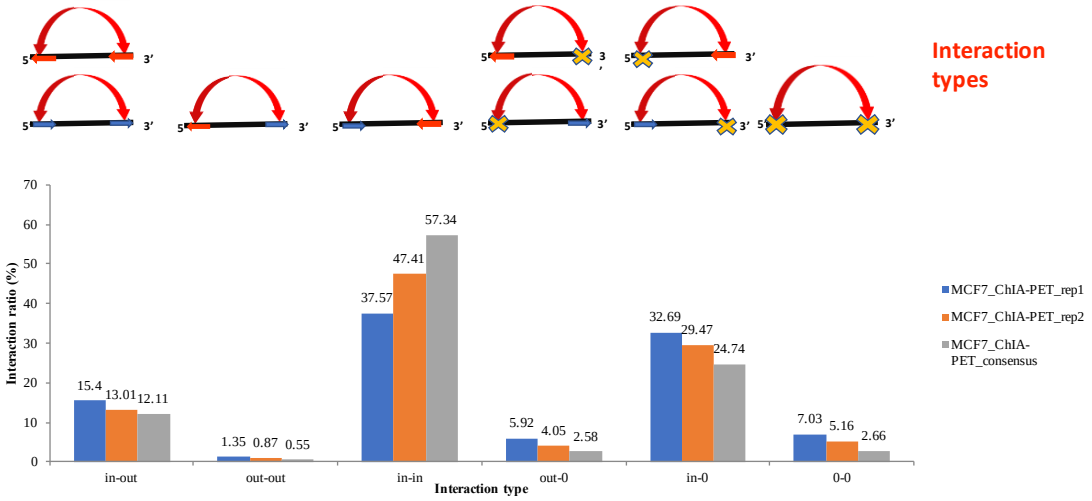

Interaction types
Interaction ratio (%)
15.4
13.01
12.11
1.35
0.87
0.55
37.57
47.41
57.34
5.92
4.05
2.58
32.69
29.47
24.74
7.03
5.16
2.66
in-out
out-out
in-in
out-0
in-0
0-0
Interaction type
MCF7_ChIA-PET_rep1
MCF7_ChIA-PET_rep2
MCF7_ChIA-PET_consensus

Figure 31. Investigation of MCF7 CTCF ChIA-PET data in the context of motif orientation. A) We created an analysis pipeline, which identifies the most probable anchoring CTCF motif within interacting DNA regions (which were identified with ChIA-PET). The interacting regions have an approximate 2000 bp width. We took the middle of these regions and, converging from the middle to the edges, we attempted to identify the anchoring motifs. The line charts represent the change of identified anchor types along with the analysis (converging to edges). We clustered the interactions according to motif orientation at the two-anchor region of the loop. The bar charts show the ratio of the identified loop types. B) We created a consensus MCF7 CTCF loop set according to their presence in both replicates. C) Loop type frequency in MCF7 interaction sets. The ratio of in-in type loops significantly increased in the case of permanent loop sets (consensus set) (however it should be emphasized that this set represents a smaller interaction set (**Table 10**)).

The role of CTCF motif orientation in chromatin looping did not evade the observation of other working groups. Rao identified (and published) motif orientation with in situ HiC analysis (Rao et al., 2014) in 2014, before we could finish our study. Our results were coinciding with the Rao working group's publication. Since then, several models have been created, which explain the formation of cohesin mediated loops and the loop extrusion formation. The convergent CTCF motif orientation is already validated with CRISPR-Cas systems. In one study, the orientation of the CTCF motif in the Sox2 super-enhancer and Malt1 locus was inverted, which led to the loss of loops in these regions and to the formation of other non-specific sub-TAD interactions (de Wit et al., 2015). Combining the topological data and the dominant motif orientation of the CTCF motif suggests that the cohesin ring is in the proximal position of DNA loops.

4.4.Histone modifications in CTCF mediated chromatin looping

As mentioned previously, substructures within TAD are associated with cohesin and form functional units like enhancer-promoter loops. The substructures within TAD (sub-TAD domains) are also associated with CTCF, whose function is difficult to define. CTCF sites facilitate gene activation, while other CTCF sites function as insulators (Dixon et al., 2012).

To identify correlations between transcriptional regulation and chromatin looping, we investigated histone modifications in the vicinity of the loop anchor regions. We used the previously defined consensus CTCF interaction set from the MCF7 cell line and seven publicly available MCF7 histone ChIP-seq datasets (H3K4me1, H3K4me2, H3K4me3, H3K9me3, H3K27ac, H3K27me, and H3K36me).

We examined the histone ChIP-Fragment Coverage (with HOMER annotatePeaks.pl program), which indicates the density of aligned tags relative to CTCF anchor regions. The average tag coverage of different histone experiments was calculated within 1000 bp frame (+/- 500 bp) relative to CTCF centers. The overall meta profile of histone occupancy followed the previously described peak-valley-peak pattern (Calo & Wysocka, 2013), where the center valley contains the cis-regulatory element (Hoffman et al., 2010), the CTCF motif in our case. Remarkably, signal intensities for H3K4 methylations were higher compared to other histone modifications (**Figure 32A**).

Thus, the analysis was focused on the H3K4 methylation data. MCF7 consensus interactions were used to profile the histone ChIP-seq fragment coverage of every CTCF anchor position. Anchor regions were occupied by at least one H3K4 methylation type in the 4188 interactions (from 7234). Heatmaps were centered on the anchoring CTCF motif centers (with 1000 bp frame). Sides of the loops (identified by CTCF ChIA-PET) were paired next to each other. This approach facilitated the simultaneous investigation of histone coverage on both anchor regions. We clustered the profile with k-Means clustering. The most prominent phenomenon was the characteristic asymmetry of histone signals on opposite sides of the loops. The signals were high for only one anchor region in each cluster. In the first round, we distinguished 6 clusters, which were reduced to 3 with side ordering. The “downstream and upstream side of loops” are artificial constructs and there are large similarities between cluster

2-4, 1-3, and 5-6; thus, we ordered the anchors with "strong signal" into one side of the heatmap, resulting in 3 clusters (**Figure 32B**) as follows:

- Cluster 1: strong H3K4me2 and H3K4me3 signals with relatively low H3K4me1
- Cluster 2: High H3K4me1, medium H3K4me2, and low H3K4me3 signals
- Cluster 3: Strong H3K4me2 signal and medium H3K4me1 and H3K4me3 signals

Cluster 1 is mostly involved in promoter specific interactions, while clusters 2 and 3 interact with bridge enhancer regions for introns or other enhancer regions (**Figure 32C**).

An article was published with results consistent with our studies (Dowen et al., 2014). In this study, ESC ChIA-PET data were processed. The investigators introduced the definition of polycomb domains, which are characterized by a particularly high presence of the polycomb proteins, like EZH2 and SUZ12, in association with H3K27me3 histone modification. This complex structure represses lineage-specifying developmental regulators. Their meta-analysis had a similar result as ours. Taken together these studies indicate that the CTCF loops have structural roles in both gene activation and repression, which enable the physical proximity between enhancer and promoter regions.

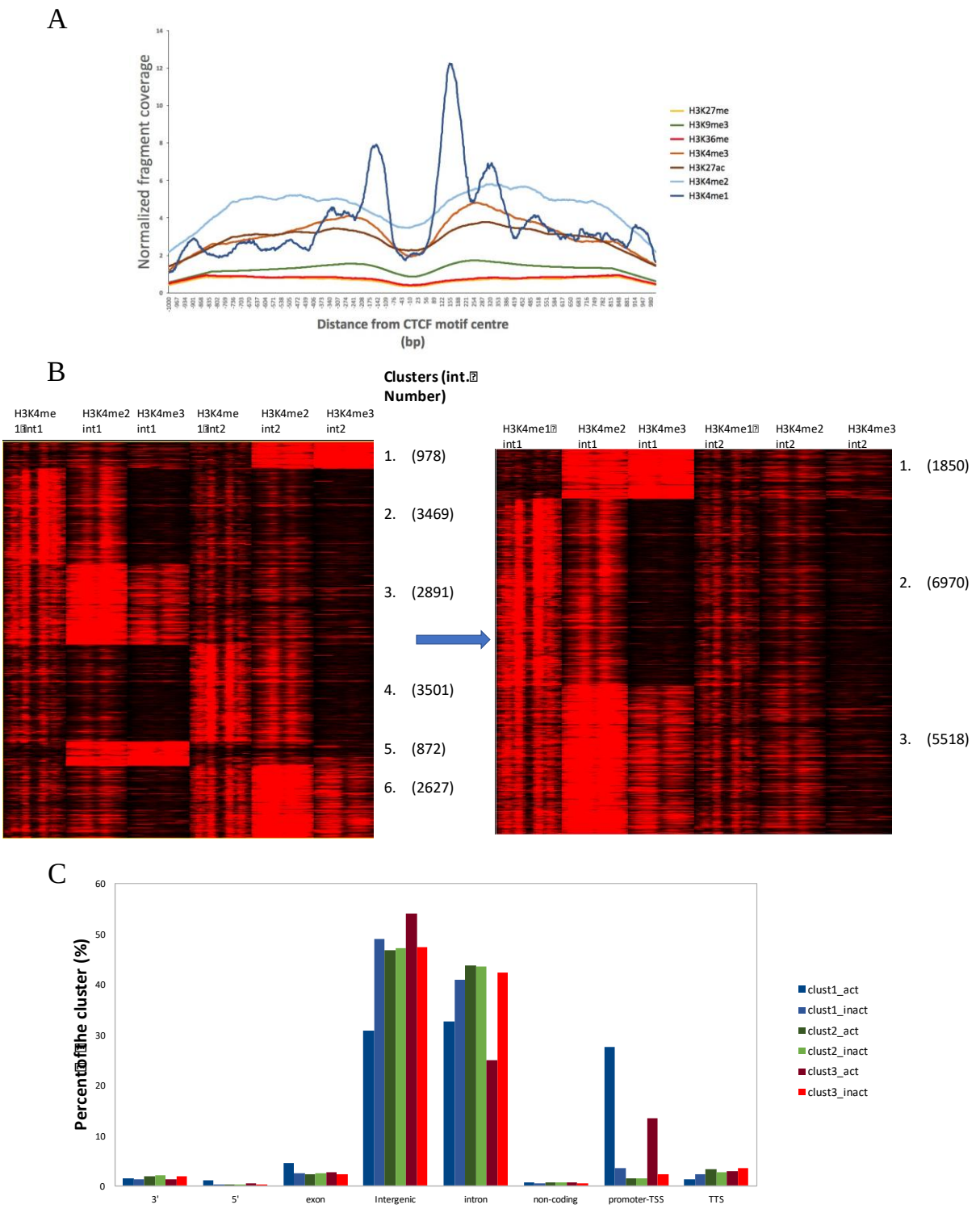


Figure 32. Investigation of CTCF mediated looping in the context of transcriptional regulation. A) We took CTCF interactions from the MCF7 cell line's consensus set (**Table 10**) and examined the ChIP-seq fragment coverage of different histone marks. The Y-axis represents the average (normalized) ChIP-seq fragment coverage of the indicated histone mark ChIP-seq. The X-axis shows the distance from the CTCF motif center (the analysis was centralized to the CTCF motif centers). The H3K4me modification showed a remarkably high signal relative to H3K27me and H3K9me. B) We performed a k-Means clustering on the H3K4me modification ChIP-seq fragment coverage intensities at the anchoring CTCF binding sites. The

results were visualized using a heatmap diagram. We considered the ChIP-seq signal differences between the two loop anchors. We analyzed the two anchors separately but we kept the corresponding anchors next to each other. The "downstream and upstream side of loops" are artificial constructs and there are large similarities between cluster 2-4, 1-3, and 5-6; thus, we ordered the anchors with "strong signal" into one side of the heatmap. C) The annotation of loops within different clusters showed that cluster 1 is mostly involved in promoter specific interactions, while cluster 2 and 3 interactions bridge enhancer regions to introns or other enhancer regions.

4.5.The establishment of ChIPSummitDB

4.5.1. ChIPSummitDB:

The main goal of analyzing ChIP-seq experiments is to identify regions in the genome where we find more sequencing reads (tags) than we would expect to see by chance. These regions are called peak regions due to the appearance of the visualized distribution of mapped tags (Albert, Wachi, Jiang, & Pugh, 2008). Our goal was to create a global database based on combining the location of identified transcriptional regulatory elements (TREs) with the positional information of the co-bound regulatory proteins (using publicly available ChIP-seq data, targeting as many proteins as we could). By investigating a global picture of different transcription factors and cofactors, we can identify previously unknown transcriptional regulatory networks. Using the database, we can browse co-bound proteins on TREs and acquire information about their positioning relative to each other and the bound transcription factor motif.

The comparison between experiments requires uniform processing of data. Several databases contain pre-processed ChIP-seq data. They differ in the stage and the approach to data processing. Their downloadable content makes the gene regulatory research work easier by providing information about identified transcription factor binding sites or motif enrichment (GTRD, ReMap, FACTORbook, HOCOMOCO) (Cheneby, Gheorghe, Artufel, Mathelier, & Ballester, 2018; Kulakovskiy et al., 2018; J. Wang et al., 2013; Yevshin, Sharipov, Valeev, Kel, & Kolpakov, 2017).

The previously mentioned SRA and ENCODE databases are the most common source of raw ChIP-seq data, which is processed with mostly uniform workflows. Our database consists of consistently processed ChIP-seq data ("An integrated encyclopedia of DNA elements in the human genome," 2012; Bethesda (MD): National Center for Biotechnology Information (US), 2011). The represented downloadable content includes motif enrichment, peak prediction, motif remap, and overlap information. The JASPAR CORE contains a large collection of curated, non-redundant transcription factor profiles (PWMs) (Khan et al., 2018). The TFBS profiles are the basis of our motif instance identification in combination with the collected ChIP-seq data for experimental validation. The motif occurrences were used in further protein position analyses.

The basic analysis workflow contains mapping of sequence reads and prediction of putative TFBSs as peaks and summit predictions. As was previously mentioned, we developed an analysis pipeline, which is applicable for transforming ChIP-seq summit data to topological information about the positioning of DNA-binding protein complexes (Gergely Nagy et al., 2016). The summit based analysis reveals the relative position of subunits with respect to a reference point (fix genomic point, a transcription factor binding site) and to each other.

In the first phase of the analysis, we prepared for large scale data analysis and prepared scripts and pipelines for a probe set. This set contained more than human ChIP-seq data. As a result, we get an automatized pipeline, which can be used for further data collection and deep analysis (to extend the database) with minimal manual interaction. The goals were unified data processing, the identification of as many transcription factor binding sites as possible (used as reference points), and topological data extraction.

For easier data collection, we created a program that processes SRA xml records to uniformly named experiments and download them (https://github.com/Raziel01/SummitDB-data-prepare/blob/master/SRA_XML_process.pl). For the naming of sample extracts, the

following information from the SRA record was collected: organism name, original tissue, name of cell type or cell line, library selection, target protein name, SRA ID, and sample preparation method. The program also considers special treatment protocols (e.g. reagent treatment, mutation, knock-out, and knock down), which can influence transcription factor binding events. In this stage, we avoided the collection of special reagent treated and mutated samples, but we plan on extending our database with this type of data. The collected data were uniquely processed using a custom made pipeline (Barta E., 2011). The analysis includes a peak filtering step for more exact summit analysis, a binding site prediction, a motif optimization step to maximize the number of identified motif instances, and summit distance calculations (**Figure 14**).

After the developmental phase, we extended the data collection and processed it with the custom made pipeline. We collected 4052 human ChIP-seq experiments and successfully analyzed 3782 of them. The remaining 270 experiments lacked identified peak regions due to the low quality of the data. Overall, more than 93.4 million peaks were used in the database creation, which covered more than 1 billion base pairs (Gbp) in the human genome (**Figure 33A**). A total of 2659 ChIP-seq targets (from 3782) were classified as transcription factors (66 %); the others were cofactors (1397, 34%) to whom binding site could not be paired . A total of 2496 experiments belonged to transcription factors that have described motifs in the JASPAR CORE database (**Figure 15B**). These experiments were used for motif optimization and binding site prediction (consensus binding site creation). The JASPAR CORE database stores 579 non-redundant motifs (Khan et al., 2018). From the 579 non-redundant motifs, 338 PWMs could be paired to at least one ChIP-seq experiment from the 2496 experiments with described motifs. We could find motif instances for 280 JASPAR CORE motifs (**Figure 15C**). This kind of reduction is not striking, because many motifs are theoretical and cannot be linked to a ChIP-seq experiment (they derive from DNAse footprinting assay or Electrophoretic Mobility Shift

Assay (EMSA)). A few older CORE models were also lacking sequence data, for historical reasons: “CORE was originally built in order to create familial binding profiles for as many structural classes of transcription factor classes as possible. In some experimental literature, only matrices and not sequences are available. For this project, we were forced to include some matrices to gain coverage of certain binding site classes. For recent additions, it is a requirement to have the sequences available” (Khan et al., 2018).

Palindromic binding sites provide an opportunity for tentative strand specific distance measurements. We attempted to create a non-redundant transcription factor binding site dataset. Since many factors bind to palindromic sequences, avoiding the bilateral motif identification was challenging. Most of these factors have a position preference, which shifts their relative summit positions away from the motif center. The zero point (middle of the motif center) differs between the direct strand and reverse strand motifs. They are not located at the same position in the genome. During the creation of the consensus dataset creation, we used a summit selection to choose the motifs which were closest to the summit positions. If summits of a given factor gathered around a given point of the genome, which was relatively far from the motif center, distinguishing the closest binding site from the bilateral motifs was easy (**Figure 33**). In some cases, the summits did not show position preference or did not gather in the proximity of zero points. In this event, both motifs were added to the consensus motif set. Using the summit selection, we could reduce the ratio of bilateral motifs below 30%, except in the cases of the FOS and JUN motifs, which had an approximately 50 % dual motif ratio. This problem concerned 1 in 7 of the JASPAR motifs.

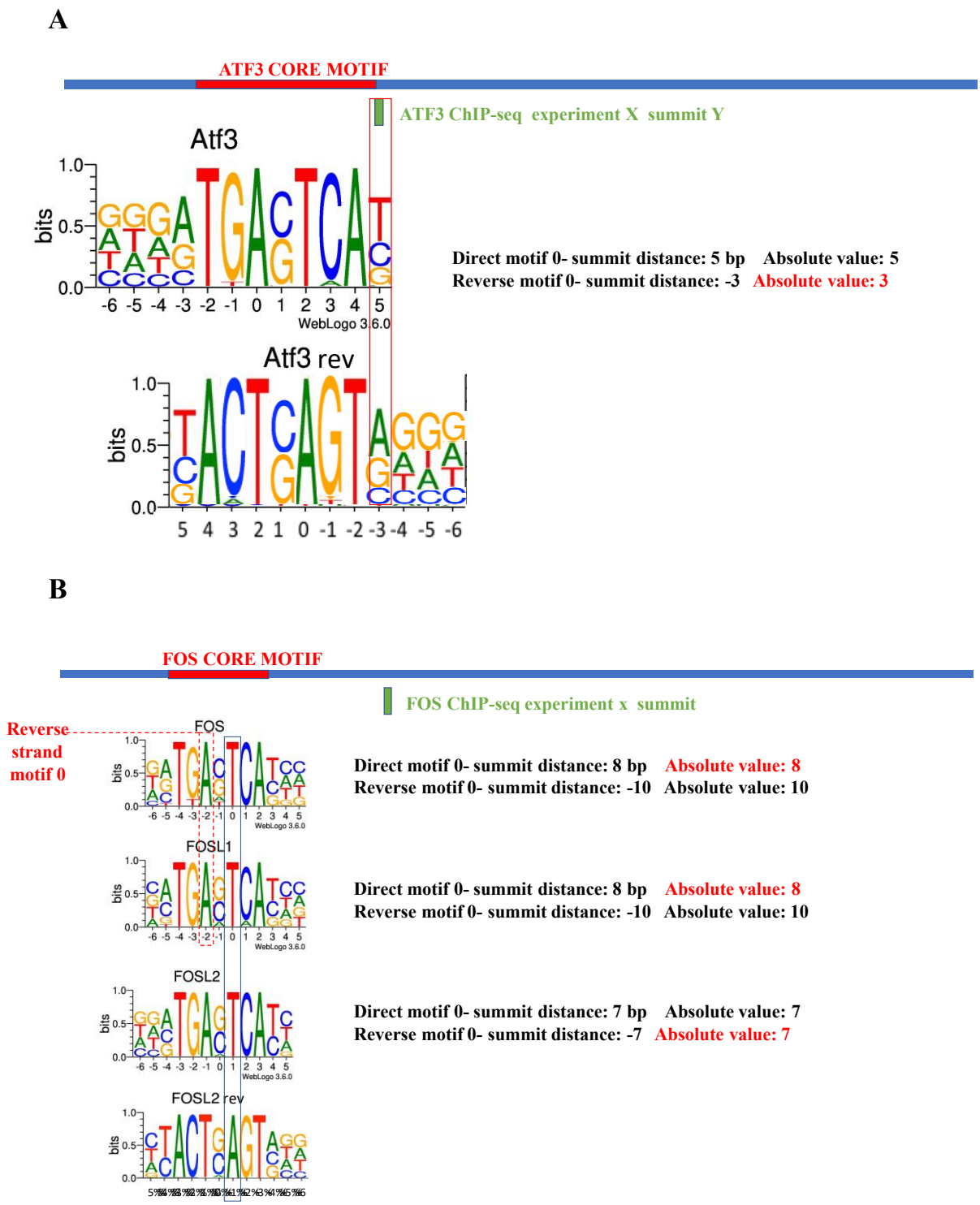


Figure 33. Summit based selection to avoid bilateral motif identification in the case of palindromic sequences. A) The ATF3 protein recognizes the palindromic sequence in the genome (Gargiulo et al., 2013). Due to the motif center shift, the motif with the lower motif center-summit distance absolute value was selected for the consensus ATF3 motif set. The selected ("winner") motif's absolute value was marked with red color. B) The different FOS binding sites were similar to each other (Rodriguez-Martinez, Reinke, Bhimsaria, Keating, & Ansari, 2017). The FOS and FOSL1's 0 positions coincided. Both motifs was identified as direct strand motifs, because they are proximal to the identified FOS ChIP-seq summit. The FOSL2 0 point

is shifted, which puts the reverse strand's 0 point closer to the summit. Thus, the direct strand motifs are selected in the case of FOS and FOSL1, and the reverse strand motif is selected in the case of FOSL2.

Finally, more than 5 million transcription factor binding sites were identified, which cover 40.8 Megabase pairs (Mbp) in the genome. The identified binding sites represent valuable information in themselves. However, a comparison of all motif data with complete ChIP-seq dataset provides information about the protein complexes and transcription factor network in correlation with specific binding sites. We calculated the protein position preferences with respect to the corresponding motif centers and compared the results to identify topological relationships between proteins. The complete dataset is stored in the database (MySQL), which is publicly available on the ChIPSummitDB's web interface (http://summit.med.unideb.hu/summitdb/index.php).

4.5.2. Database and web interface

To reach the viewable results from raw sequence data requires a large investment in time and computing resources. Currently, transcriptional regulation related studies are efficient due to the ChIP-seq technique. This requires unified data processing for comprehensive analysis. The comparison of two or more samples can provide large scale biological correlation, but this data is still insufficient for regulatory network mapping. We aimed to collect data about transcription factor/cofactor occupancy on different types of transcription factor binding sites. We established ChIPSummitDB, a web interface to browse processed ChIP-seq data, and identified transcription factor binding sites in a global manner.

The website provides information about:

- Transcription factor binding site profiles: JASPAR CORE motifs are optimized with HOMER analysis (-opt) of ChIP-seq experiments to extract accurate PWMs. The motifs are carefully paired with available ChIP-seq experiments (e.g. CTCF motif- CTCF

ChIP-seq; RAR:RXR motif- RXR and RAR CHIP-seq). The matrices represent frequently presented sequence motifs in the peak regions, which resemble the original JASPAR motif (Heinz et al., 2010; Khan et al., 2018).

- Occupancy patterns of regulatory elements: Obtained ChIP-seq peaks within 100 bp frame around different types of motifs are included. In different display modes, overall and detailed motif occupancy (by transcription factors) can be visualized. After choosing a motif type, the adjacent proteins can be displayed according to the ChIP-seq experiments. The number of overlapping peaks, the origin (tissue and cell type), and the preferred summit positions are viewable. In a different display mode, the detailed position distribution of summits of a given experiment around the motif can be identified.
- Topological data: Spatial organization of DNA bound protein complexes is based on summit analysis.
- The overlap between ChIP-seq peaks: The juxtaposing ChIP-seq signal in correlation with a given motif can be examined.
- Regulatory SNP: Using dbSNP and ClinVar databases, an SNP finder has been integrated. The identified transcription factor binding sites can be scanned for regulatory SNPs (Landrum et al., 2014; Sherry et al., 2001).
- Genomic map: All data are viewable in a genome browser format (Skinner, Uzilov, Stein, Mungall, & Holmes, 2009).
- Source of data: Detailed information about the origin of downloaded ChIP-seq data with external links is included (Bethesda (MD): National Center for Biotechnology Information (US), 2011; Khan et al., 2018).

The six display modes are provided to visualize the data during different approaches.

4.5.3. Overlap with other processed ChIP-seq databases

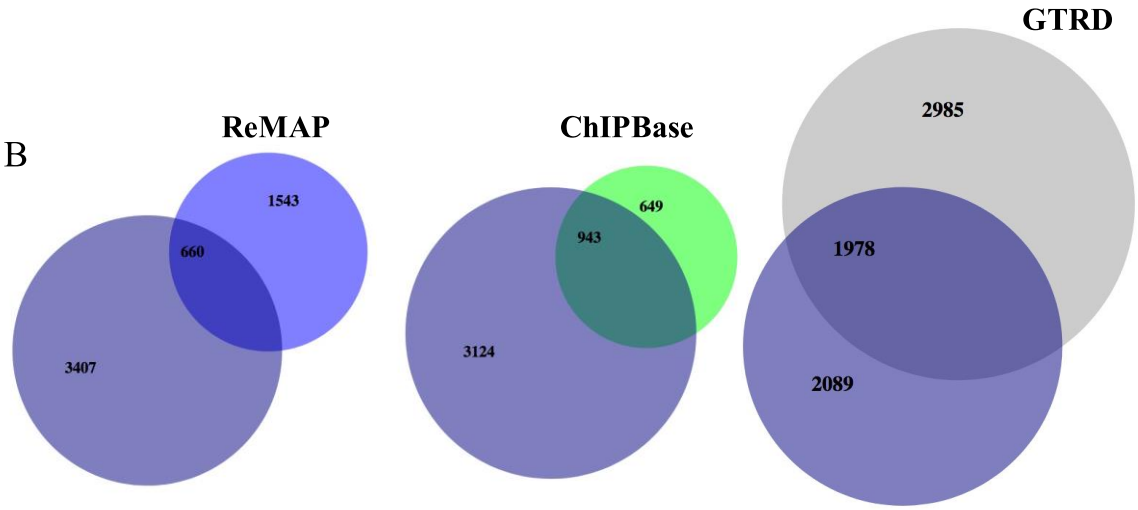
We have attempted to compare our database with existing. Most raw data are uniformly uploaded to GEO, NCBI SRA, or ENCODE. As the number of data sources is limited, we were curious about the overlap of processed experiments between databases. The comparison was limited to human ChIP-seq data, which have SRA ID, because our data were collected solely from NCBI SRA (**Figure 34A**). Using SRA ID, 1204 ChIP-seq experiments were identified from their full dataset (2829 experiments) and only 660 overlapped with our data (**Figure 34B**). The number of the identified peaks is ~90,000,000 in all databases. As the redundant coverage of peaks is variable, and the different peak callers produce peak regions with differing widths, we created a non-redundant regulatory region set for all databases. To accomplish this, we took the center of peaks and their 100 bp frame (+/- 50 base pairs) and merged them with the bedtools mergeBed program. GTRD and our database showed the highest genome coverage, which was above 10 Gbp.

We investigated the intersection between non-redundant peak regions in databases. We integrated the Fantom5 enhancer set into the analysis to approach the identified enhancer set from a different perspective. Using a bidirectional CAGE pattern, 162,819 human enhancers were identified. We compared coverages of non-redundant peaks and the number of overlapping base pairs between databases. The result was plotted on a heatmap, where the overlap is measured in percentage. On the horizontal axis, the percent of the database is shown on every square. Fantom5 shows a high overlap rate with other databases. Among ChIP-seq databases, ReMap and ChIPAtlas overlap the most with our database, but the similarity is still below 65 % (**Figure 34C**).

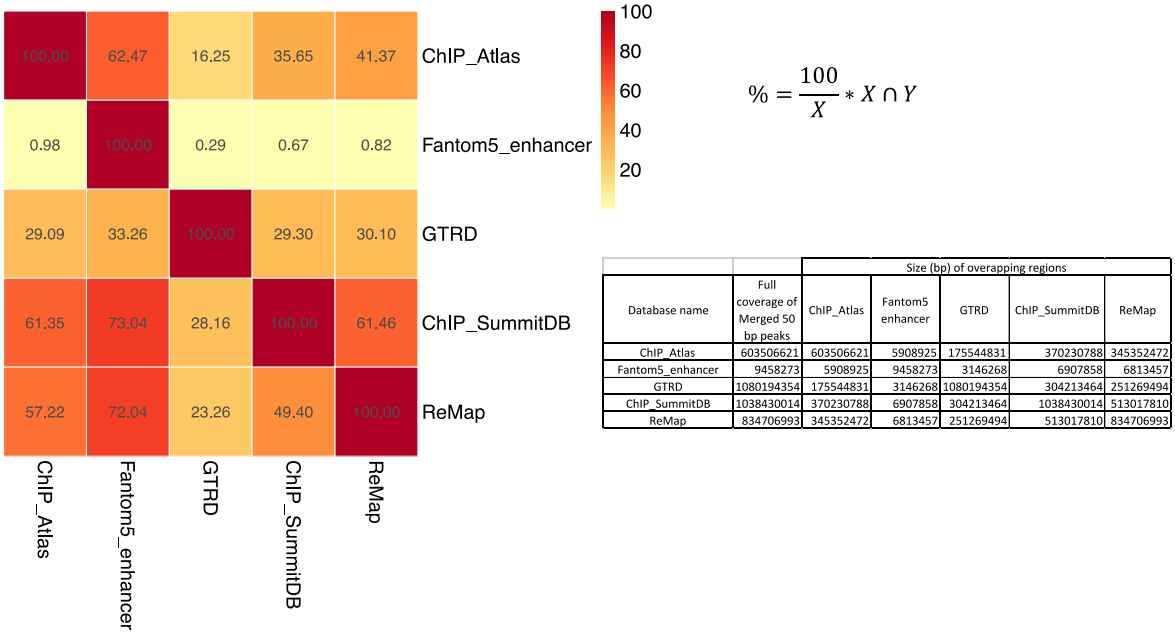
A

	Downloaded TF ChIP-seq + Contol (IgG, input)	Downloaded TF ChIP-seq no Contol (IgG, input)	Found SRA IDs	Source	Number of all peaks	Number of 50 bp merged peaks	Genome coverage of peaks
ChIPSummitDB	4067	4067	4067	SRA	93 445 309	11 584 327	10 384 30 014
ChIP-Atlas	8368	7200	7200	SRA	91 224 465	6 208 846	603 506 621
GTRD	6819	5098	4963	ENCODE, GEO, SRA	99 644 796	10 711 153	1 080 194 354
ChIPBase	2498	2498	1592	GEO, ENCODE	-		
ReMap	2829	2829	2203	GEO	79 994 115	8 525 656	834 706 993

B



C



Database name	Full coverage of Merged 50 bp peaks	Size (bp) of overapping regions				
		ChIP_Atlas	Fantom5 enhancer	GTRD	ChIP_SummitDB	ReMap
ChIP_Atlas	603506621	603506621	5908925	175544831	370230788	345352472
Fantom5_enhancer	9458273	5908925	9458273	3146268	6907858	6813457
GTRD	1080194354	175544831	3146268	1080194354	304213464	251269494
ChIP_SummitDB	1038430014	370230788	6907858	304213464	1038430014	513017810
ReMap	834706993	345352472	6813457	251269494	513017810	834706993

Figure 34. Comparison of ChIP-seq databases. There are many databases that store semi-processed or completely processed ChIP-seq data. On this figure, the overlap between our ChIPSummitDB and the most popular databases is shown. A) The table summarizes the processed human ChIP-seq experiments and their genome coverage. In the case of ChIP-seq, we accessed the list of processed data, but the experimental results were temporarily unavailable. The number of processed data is highly

variable, but the identified peaks and their non-redundant genome coverage are closely related. B) Proportional Venn diagrams represent the overlap between processed ChIP-seq experiments (which have SRA record) from different databases. The difference is significant, as it is visible on the low number of common experiments. C) We created non-redundant genome coverage sets by merging the center of the peaks and their +/- 25 bp flanking regions. We measured the overlapping regions between databases (in base pairs) and the ratio of similarity was plotted on a heatmap diagram. The complete genome coverage of databases is shown on the horizontal axis and the percentage of commonly covered genomic regions is shown on the vertical axis. The table shows the concrete number of overlapping base pairs.

4.5.4. ChIPSummitDB Views

4.5.4.1.MotifView

In this viewing mode, the average distances between the read peaks from the ChIP-seq experiments is obtained and the given consensus motifs are visualized on a scatterplot. Each scatter represents an experiment. Circles represent transcription factors with defined binding sites, while triangles represent co-factors and other indirectly bound proteins. Different colors indicate the antibodies used in the immunoprecipitation. The X-axis shows the average distance between peak summits and the center of the binding sites overlapped by the peaks. The Y-axis shows either the number of the peaks (elements) overlapping the center of the binding sites or, in the default mode, the standard deviation of the shift values (distances) between the peak summits and motif centers. Such a scatterplot representation is available for every consensus binding motif set. The displayed data can be filtered for the number of the peaks (element number) or for the standard deviation. Data can also be displayed based on the ChIP antibody or cell type. The average data obtained by the same antibody in different experiments can also be calculated and shown.

If browsing the standard deviation scatterplots, the dots show a visible clustering among factors. Since the dots are colored according to the experimental antibody, different color groups are distinguishable around the preferred position (relative to the motif center) and standard deviation. The standard deviation showed unexpected correlation with factor-DNA

proximity. Apparently, the factor that is responsible for the motif binding has a significantly lower standard deviation than other associated proteins. The fixation of DNA binding proteins to DNA limits the variability of summit positions relative to the bond sequence. In contrast, the spatial distance between DNA and indirectly attached factors (cofactors) causes higher mobility, which is restricted only by the structural characteristics of protein-protein interactions. Approximate position preferences of cofactors can be tracked and their connectivity order (levels/layers) can be distinguished.

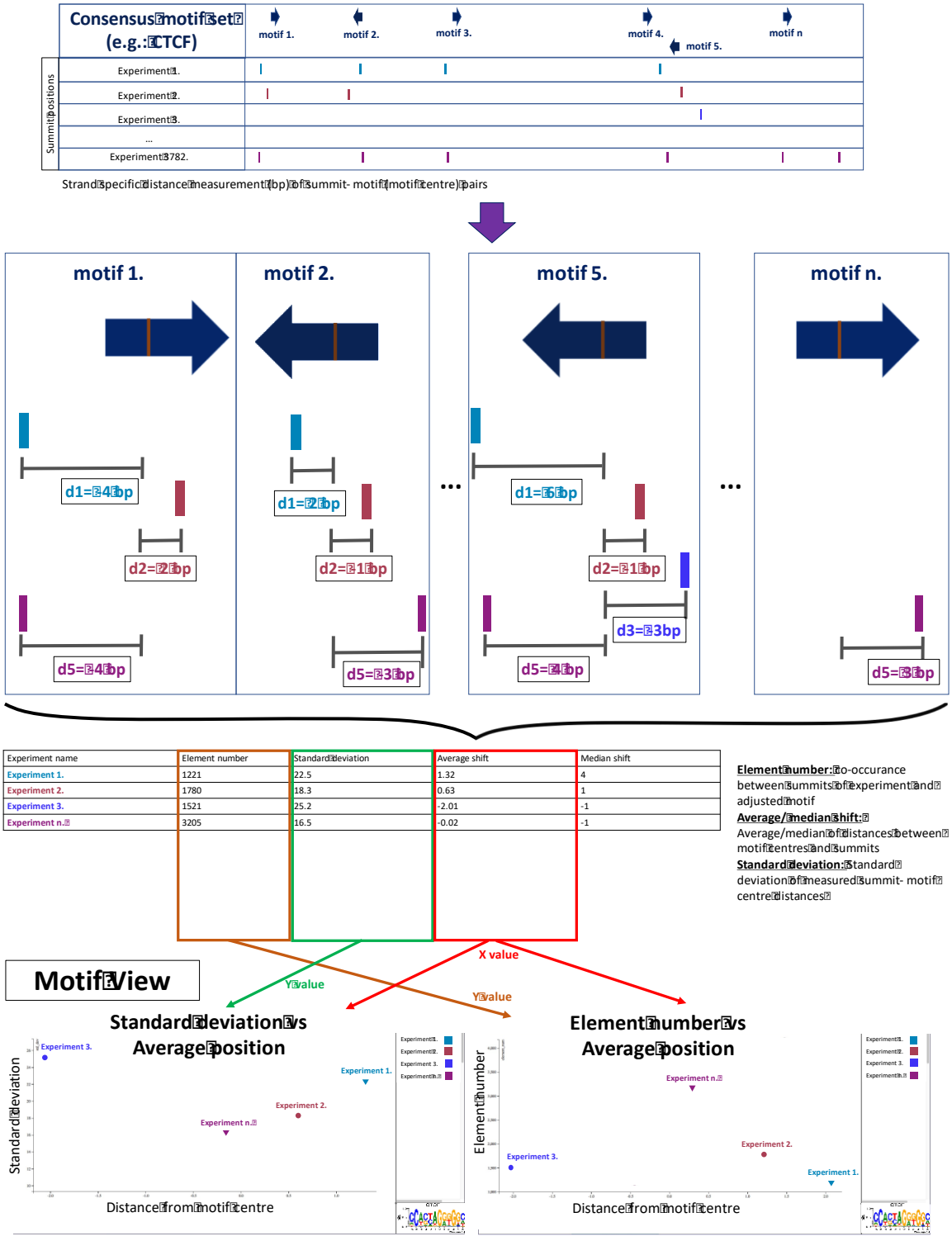


Figure 35. Motif view. The motif view appears as an interactive scatterplot, which gives information about the given JASPAR core motif instances in the genome, the overlap frequency between motifs and ChIP-seq peaks, and the average/median positioning of summits from different experiments. Every dot on the chart belongs to a single ChIP-seq experiment. The dots are placed depending on the relation between the positioning information and the adjusted motif center. The position weight matrix of the adjusted motif is shown in the bottom-right corner. The center of the motif is marked by "0" on the scale below.

The motif of interest can be set in the "Set a motif" dropdown box. In the boxes on the right, you can modify the data that is displayed and the minimum and maximum values of the Y-axis. Important: All of the changes will only be displayed after clicking on the "Resend Data" button below. If the cursor hovers over a given dot, a tool-tip will appear that gives us information about the ChIP-seq experiment, such as the name of the experiment, cell type, target protein, and quantified information about summit positions (average/median distance, standard deviation of distances, overlap number). The dots are colored according to the type of target protein. The legend with color codes is visible on the right-hand side of the chart, which is also interactive. Clicking on a specific target protein name in the legend section can hide the respective dots from the chart. A large amount of displayed dots can be overwhelming in the data review, so we created a "Clear all dots out" button to hide all of the points of the chart. The specific spots can be called back one-by-one by clicking on the factor name in the legend. Using this process, we can compare the positioning of proteins of interest. All of these steps are revocable by clicking on the "Show all dots" button. The X-axis is constant and represents the distance from the center of the adjusted JASPAR CORE motif (the distance is measured in base pairs). The Y-axis is adjustable. The number of summit-motif overlaps or the standard deviation of summit positions can be displayed.

4.5.4.2.Pair Shift View

The pair shift view shows the summit distance distributions of the selected ChIP-seq data (a maximum of 3) related to the motif as a histogram. The X-axis represents the distance (measured in base pairs) from the middle of the given motif, which is marked as the "0" point. The numbering of the axis is consistent with the position weight matrix below the diagram. The Y-axis shows the frequency of summit occurrences at the positions (at the given base pair) relative to the motif center. In the case of a well-defined protein topology, high overlap frequency, and close DNA localization, the curve has a bell-like pattern (normal distribution-like) (**Figure 36**). According to our observations, the narrowness of the curve is inversely proportional to the protein's physical distance from the DNA (direct or indirect binding). This relationship can be detected when looking at the standard deviations as well (motif view). Factors with low overlap frequency and no position preference show plateau distribution. Setting the parameters on the drop down boxes, we can check the summit distribution of 1 to 3 experiments around an adjusted motif. The minimum and maximum values of the axes are configurable as well, in the text boxes below the diagram. A rolling mean with a 5 bp frame

was applied to smooth the frequency curves. There is a possibility to select an experiment in this view and see it in the ExperimentView.

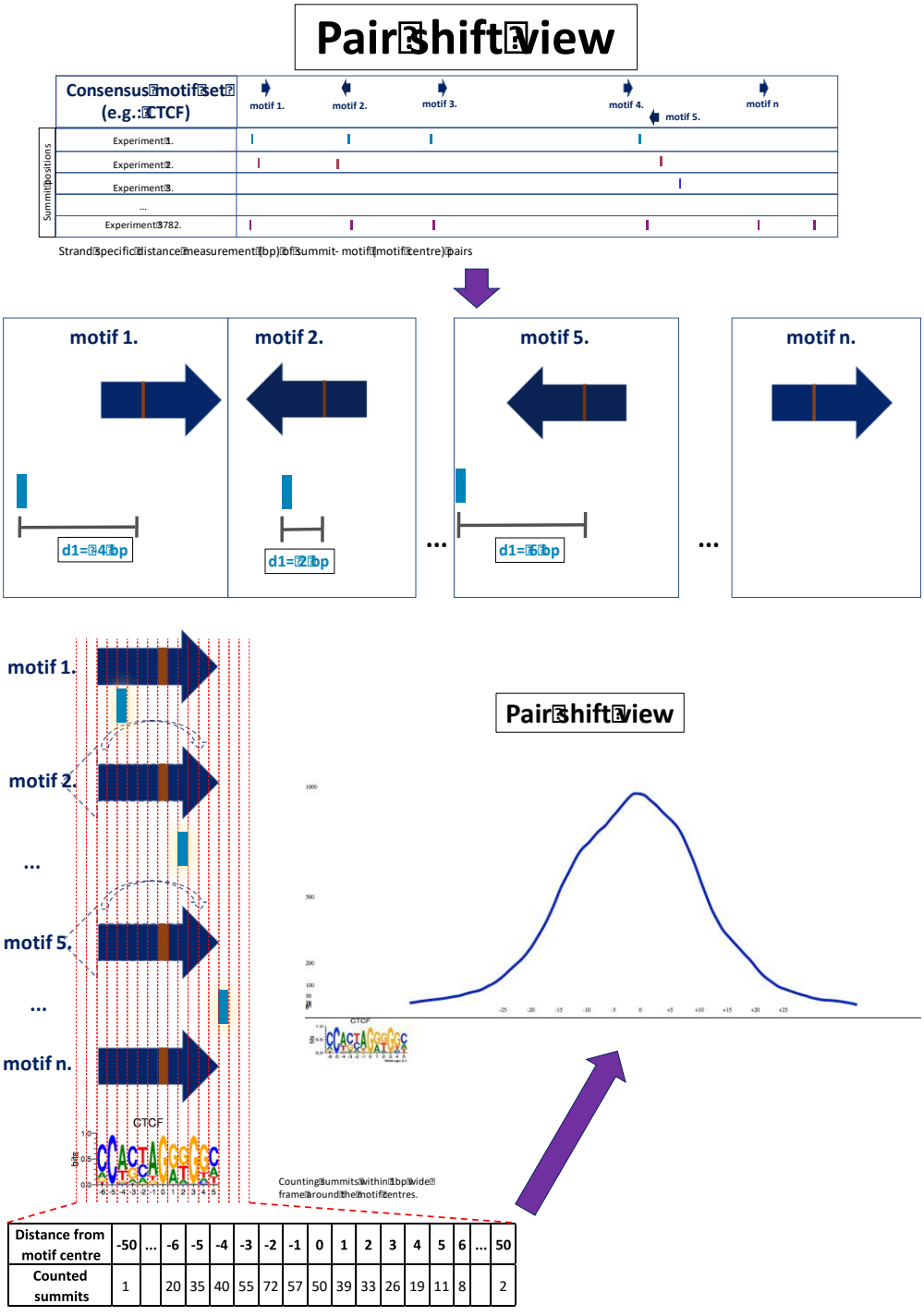


Figure 36. Pair Shift view. The summit-motif center distance distributions are calculated. In this mode, the frequencies of the different distance values between the motif and peak summit pairs for a given consensus binding site set are displayed in a histogram. To smooth the graph, a 5 bp rolling average is applied. No more than three different experiments can be compared. The maximum value of the curves shows the most frequent distance.

4.5.4.3.Experiment view

At the early stage of our work, we collected 4068 human ChIP-seq data from public databases (NCBI SRA, ENCODE) (Leinonen et al., 2011; Davis et al., 2018). From this data, 3782 experiments were successfully processed and used in the following steps of the analysis. The basic information from this data is at least as crucial as the final results. As previously mentioned, we tried to use a wide variety of ChIP-seq data considering both the origins (cell type, tissue) and the target proteins. To track the source of the data, we created an "Experiment view", which is a more manageable and readable way to browse essential information about the distinct experiments by putting all of the data into a simple table.

4.5.4.4.VennView

The diagrams of the motif view cumulatively represent the statistical data of all occupying ChIP-seq experiments (occurrence frequency, average/median distance related to the motif, and distance standard deviation) on all instances (consensus motif set) of an adjusted motif type. The co-occurrence frequency of distinct ChIP-seq summits from different experiments is not taken into account here. To fill this gap, we created a Venn diagram view. The Venn diagram displays all possible logical relationships between different sets. In our case, the sets are the motifs, which overlap with the peaks of a chosen ChIP-seq experiment, and the relationship is the number of common motifs that are simultaneously occupied in these experiments.

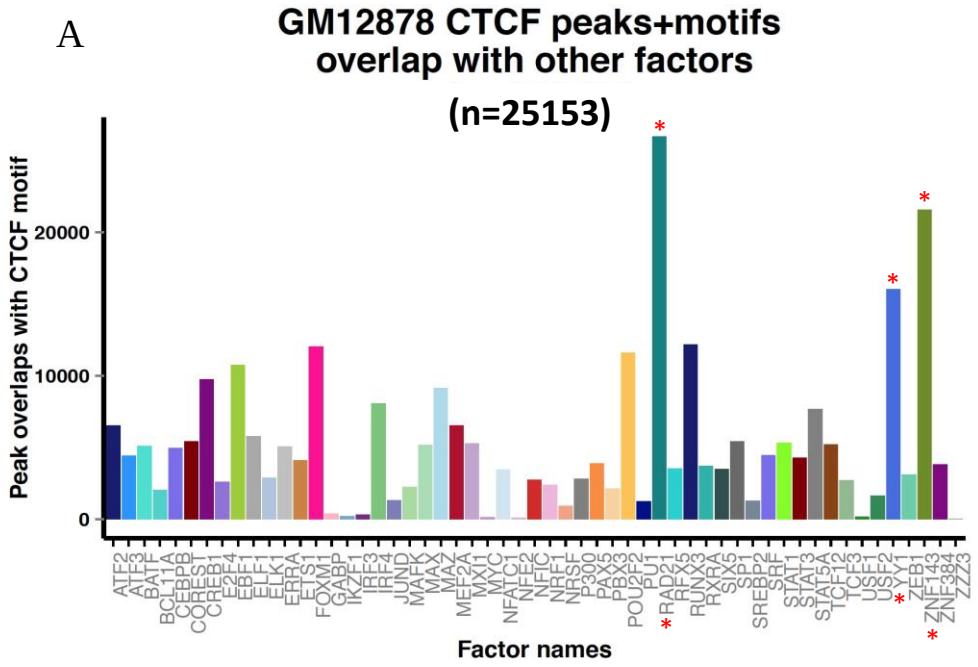
4.5.5. CTCF binding sites in genome regulation and gene expression

Using the motif view of ChIPSummitDB provides not only protein positioning information around the adjusted motif type, but also an extensive picture of the occupied protein network. The number of summit-motif co-occurrences can be tracked and visualized on a

scatterplot (Y-axis value). Using the co-occurrence frequency (element number) as the Y value, we can browse the most frequently occurring ChIP-seq experiments around a motif type. This highlights the members of commonly assembled protein complexes.

As we were familiar with the CTCF and cohesin complex, we started to analyze their related network. The identified CTCF motifs showed frequent ChIP-seq occupancy with various transcription factors. Strikingly, in addition to cohesin signals, other factors were also located downstream of the CTCF element. In the order of co-occurrence frequency, the YY1 and ZNF143 signals were the most enriched factors at the CTCF motifs next to the cohesin ChIP-seq signals (**Figure 37A**). The list and frequency of juxtaposing factors were variable, while the abundant presence of YY1 and ZNF143 was relatively constant between different cell types. Interestingly, the global (in the case of binding sites and transcription factors) hierarchical clustering (Manhattan distance) analysis of CTCF elements revealed that other peaks could be observed in close proximity to CTCF binding sites only in the presence of YY1 and/or ZNF143 (**Figure 37B**). We performed hierarchical clustering with Manhattan distance on CTCF binding sites in a cell type specific manner. ChIP-seq experiments were collected from the same cell line but with different antibodies. The signal from these experiments was compared on experimentally validated CTCF binding sites (with CTCF ChIP-seq from the same cell type). Four cell lines were used in the analysis (GM12878, H1hESC, HeLA, and K562). The results were consistent in all cell lines; details are shown for GM12878 only (**Figure 37B**).

A


GM12878 CTCF peaks+motifs
overlap with other factors
(n=25153)
Peak overlaps with CTCF motif
20000
10000
0
Factor names
ATF2
ATF3
BATF
BCL11A
CEBPB
CORESTE
CREB1
E2F4
EBF1
ELF1
ELK1
ERRA
ETS1
FOXM1
GABP
IKZF1
IRF3
IRF4
JUND
MAFK
MAX
MAZ
MEF2A
MXI1
MYC
NFATC1
NFE2
NFIC
NRF1
NRSF
P300
PAX5
PBX3
POU2F2
PU1
RAD21
RFX5
RUNX3
RXRA
SIX5
SP1
SREBP2
SRF
STAT1
STAT3
STAT5A
TCF12
TCF3
USF1
USF2
YY1
ZEB1
ZNF143
ZNF384
ZZZ3

B

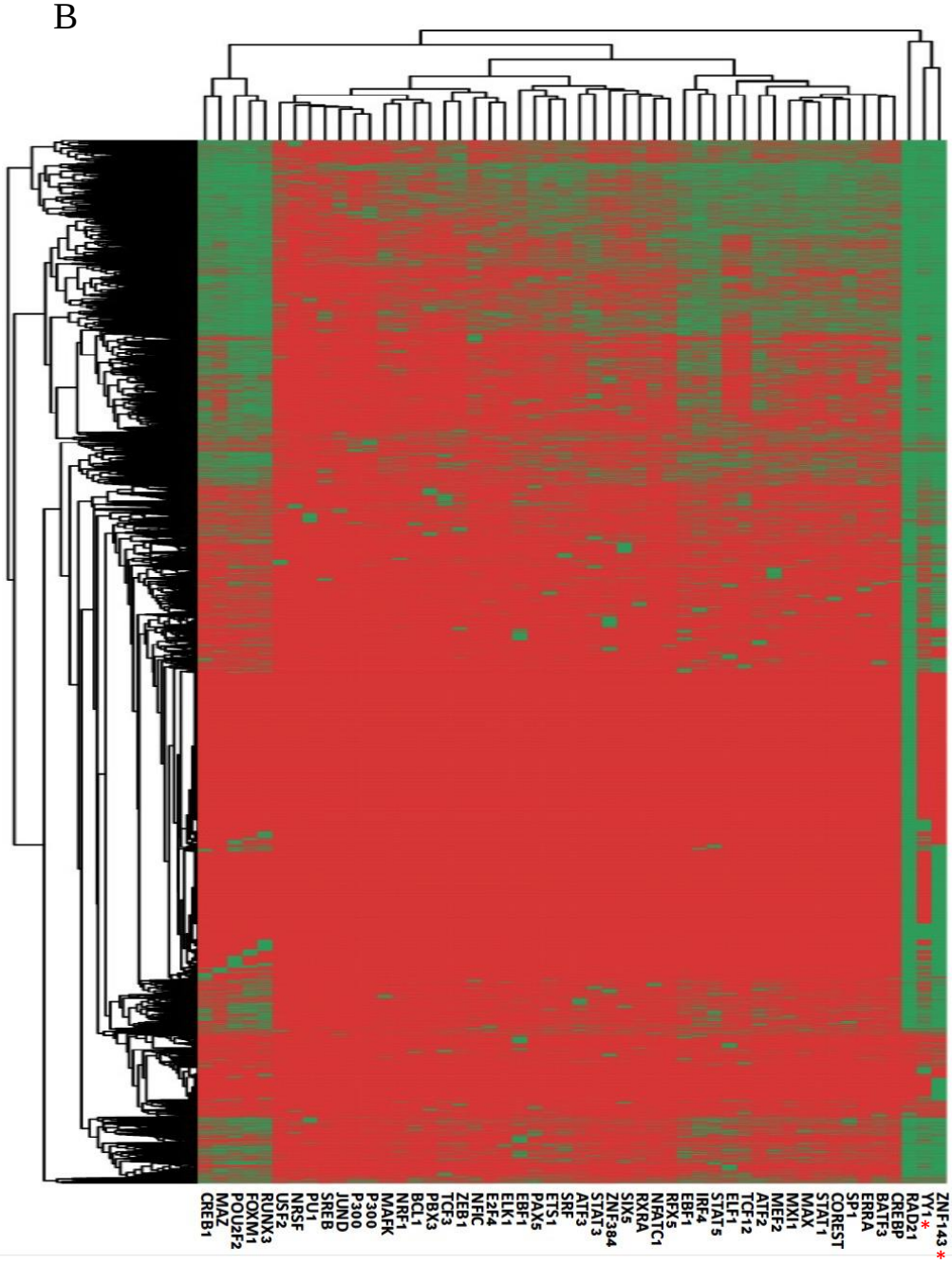
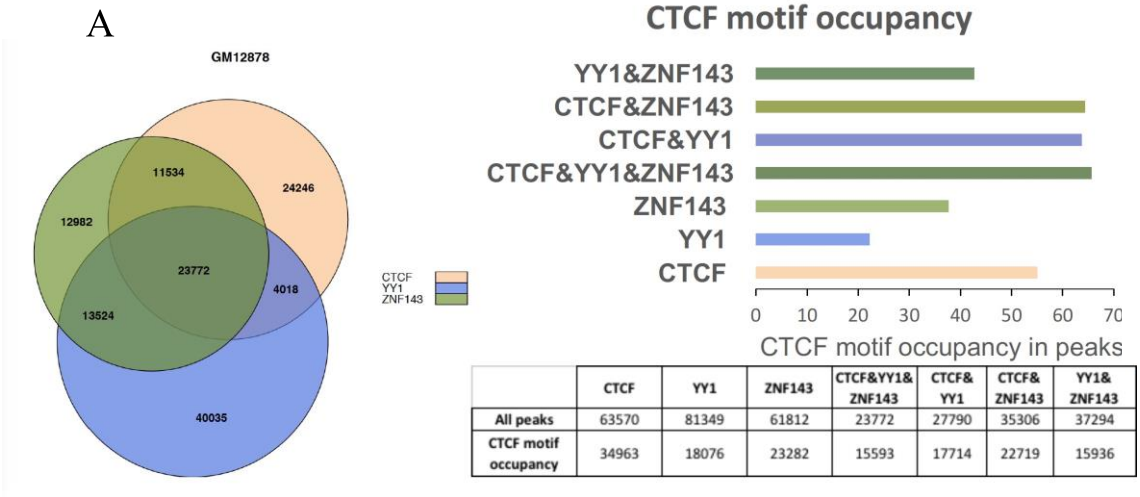

CREB1
MAZ
POU2F2
FOXM1
RUNX3
USF2
NRSF
PU1
SREB
JUND
P300
MAFK
NRF1
BCL1
PBX3
TCF3
ZEB1
NFIC
E2F4
ELK1
EBF1
PAX5
ETS1
SRF
ATF3
STAT3
ZNF384
SIX5
NFATC1
RXRA
RFX5
EBF1
IRF4
STAT5
ELF1
TCF12
ATF2
MEF2
MXI1
MAX
STAT1
COREST
SP1
ERRA
BATF3
CREBP
RAD21
YY1
ZNF143

Figure 37. The CTCF motif is highly enriched for ChIP-seq signals of different factors in the GM12878 cell line. A) The bar chart represents the frequency (with concrete overlap numbers) of transcription factor occupancy on CTSs. B) The CTCF binding sites and their transcription factor occupancy were hierarchically clustered (Manhattan distance) and plotted on a heatmap. The green color represents the presence of a given factor on a given CTS, the red shows the absence of the factor. The heatmap consists of 25153 lines. Every line represents one CTS in the genome. The heatmap contains 50 columns, which belong to one-one ChIP-seq experiments. The heatmap highlights the appearance of other transcription factor CTSs, only in the presence of YY1 and ZNF143 (green line).

We examined the complete peak sets of CTCF, YY1, and ZNF143. All factors have a remarkable peak number in the GM12878 cell line (CTCF= 63,570; YY1= 81,349; and ZNF143= 61,812). We investigated the overlap between factors. The analysis distributed the peaks into subsets, which represent large populations (**Figure 38A**). We checked the presence of CTSs in the subsets. The results suggest that all of the peak sets (all CTCF-YY1 and ZNF143 peaks) overlap with CTSs up to 55 %. In the case of CTCF, this means that the CTCF ChIP-seq signal appears not only on CTSs binding sites, but several phantom peaks can also be observed in the genome. The phantom peaks may represent genomic regions, which connect to CTCF indirectly, through other factors. The large overlap indicates a common co-occurrence between ZNF143, YY1, and CTCF; however, the low CTS presence under ZNF143 and YY1 peaks highlights that these two factors have several other interaction sites in the genome. Thus, this relationship between factors is not mutually exclusive between ZNF143-YY1 and CTCF. However, the ratio of CTS was higher in common binding sites (**Figure 38A**).

We separated the CTSs which had CTCF ChIP-seq signals in GM12878 into two populations. In the first population, YY1 ChIP-seq signals were also present, while the second population lacked YY1 ChIP-seq signals. The results showed that CTSs are highly enriched in other transcription factor ChIP-seq signals in the presence of YY1 (**Figure 38B**). This enrichment almost vanished in the absence of YY1 (**Figure 38C**).

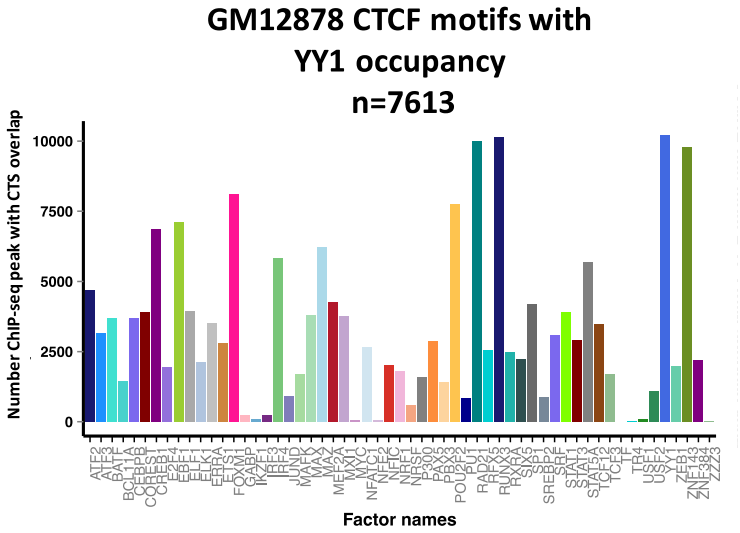
A



CTCF motif occupancy in peaks

	CTCF	YY1	ZNF143	CTCF&YY1& ZNF143	CTCF& YY1	CTCF& ZNF143	YY1& ZNF143
All peaks	63570	81349	61812	23772	27790	35306	37294
CTCF motif occupancy	34963	18076	23282	15593	17714	22719	15936

B



C

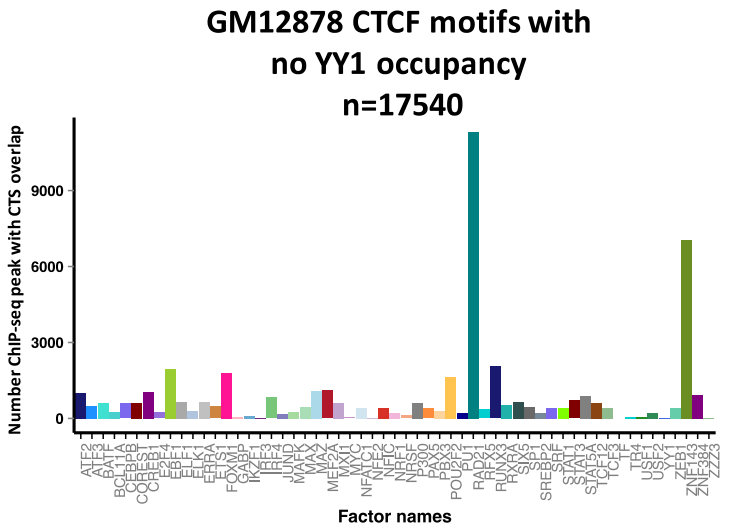
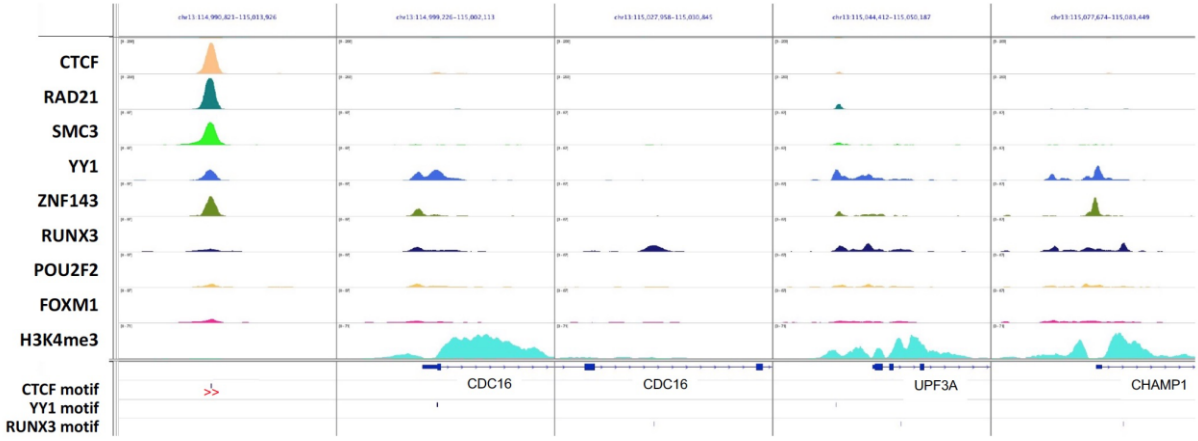


Figure 38. ChIP-seq signals on CTSs in the GM12878 cell line. A) Venn diagram showing the overlap between CTCF, ZNF143, and YY1 ChIP-seq peaks regardless of the presence of CTS. All subsets contain a large number of peaks. The bar

chart on the right displays the ratio (in percentage) of the peaks in different subsets that contain CTS. The table introduces the concrete number of CTS in the different subsets. B) CTSs in the presence of YY1 show high occupancy of different transcription factors. C) In the absence of YY1, the ChIP-seq signals of other transcription factors vanish from CTSs.

Other factor peaks (than CTCF, cohesin, YY1, and ZNF143) usually show relatively lower intensities than their instances in promoter or enhancer regions (Figure 6 poster). We can link these regions to each other: CTCF anchor regions with the presence YY1 and low factor signal (e.g. FOXM1, RUNX3) and transcription factor binding sites with strong ChIP-seq signals and motif presences (**Figure 39A**). The RUNX3 binding sites are located within 100 kbp of the CTCF binding site (**Figure 39B**).

A



B

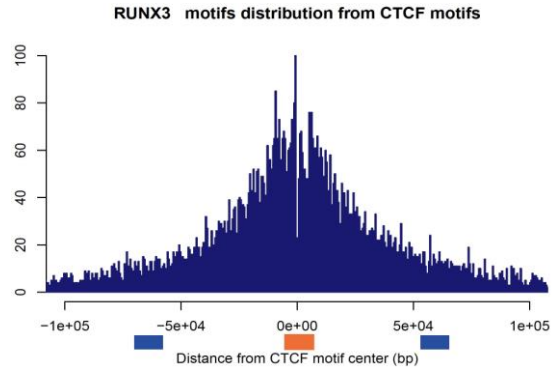


Figure 39. Relationship between CTCF and other transcription factor binding sites. A) The genome browser screenshot shows an example of a CTCF anchor region with RUNX3 and FOXM1 occupancy and nearby genes with active histone marks. The promoters of genes show signs of transcription factor binding (high signal intensity, identified motif instances). The significantly lower signal intensity at the anchor region might be the result of indirect binding between CTCF and

RUNX3/FOXM1 (mediated by YY1) (Robinson et al., 2011). B) RUNX1 ChIP-seq signal with the RUNX motif can be found within 100 kbp of CTSs.

These observations lead us to conclude that ZNF143 and YY1 not only co-occupy binding sites with CTCF and cohesin, but they also establish a connection between cohesin and transcription factors on regulatory regions of nearby genes. It is worth mentioning that there are differences between the YY1 and ZNF143 overlap with CTCF. ZNF143 generally shows more frequent juxtaposition with CTCF than YY1. The occupied ZNF143 summit displays a narrow distance distribution curve with a lower standard deviation relative to the CTCF motif center and CTCF summit positions (**Figure 40A**). In contrast, YY1 displays a broad summit distance distribution curve with higher SD (**Figure 40B**). The maxima of both factors are co-located. This position coincides with the predicted SMC1/3 positions.

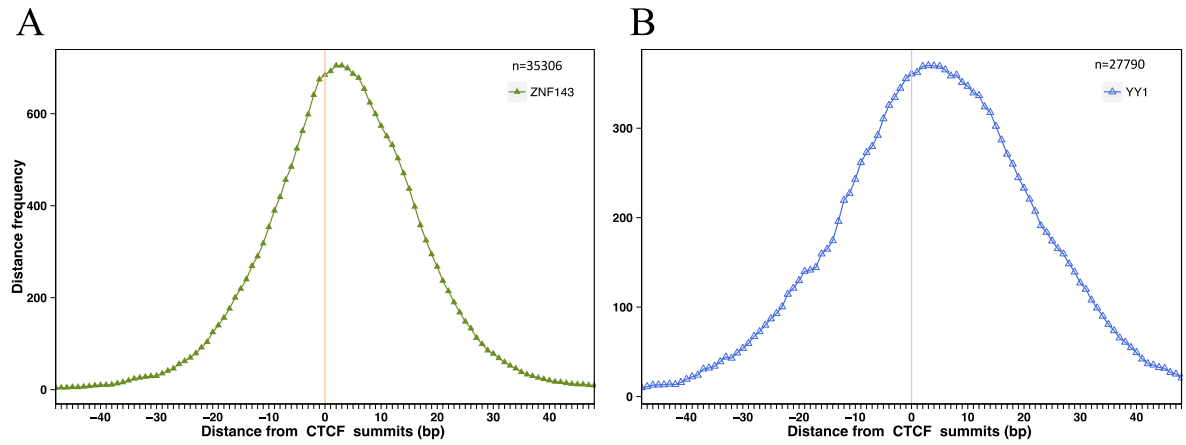


Figure 40. YY1 and ZNF143 positions relative to CTCF summits. Distance distribution of YY1 and ZNF143 proteins relative to the CTCF in the GM12878 cell line. A-B) The horizontal axes represent the distance of ZNF143 (green curve) and YY1 summits (blue curve) relative to the CTCF summits (orange line) and the vertical axes represent the distance frequencies. A rolling mean with a 5 bp window was applied to smooth the frequency curves.

Considering the overlap frequency and the differences in standard deviations among factors, we hypothesize that ZNF143 is in closer proximity to the cohesin ring then YY1. The maxima position and low SD indicate that ZNF143 may interact with the hinge domain of SMC

subunits directly. (We predicted that the hinge domain is at this exact position in our topological CTCF-cohesin mediated loop model.) The YY1 may interact with the SMC hinge domain via ZNF143 to provide a connection and close physical proximity between distal DNA regions (enhancers-promoters-TAD anchors). Numerous studies discuss the significance of YY1 and ZNF143 in the stabilization of the cohesin ring and in the transcriptional regulation of different genes (Bailey et al., 2015; Weintraub et al., 2017). Recently, a publication was released with a similar conclusion to ours (Beagan et al., 2017). This publication reinforces our hypothesis that YY1 links the enhancers to the anchor regions of the CTCF-mediated loops. In their model, other transcription factors are not mentioned. As we discussed, the separated CTSs (according to the YY1 presence) showed that in the absence of YY1, no other factor was enriched (**Figure 38B-C**).

4.5.6. Investigation of GATA1 and TAL1 binding events with summit analysis

As mentioned previously, two or more closely situated binding sites form a composite element (CE). CEs are well studied and their PWMs are represented in motif databases, e.g. TAL1:TCF3, MAX:MYC, POU5F1:SOX2, RXR:VDR FOS:JUN, etc. (Khan et al., 2018). The collaboration between GATA1 and TAL1 in erythroid development and differentiation from multiprogenitor cells into red blood cells is well studied, and their CE is represented in the JASPAR database (Han et al., 2015; Khan et al., 2018). If we separately have a look at GATA proteins in the vicinity of GATA1 binding sites, we can observe a large population of summits, which are situated 7 base pairs upstream from the GATAA sequence's guanine nucleotide (**Figure 41A-B**). The summit position enrichment is located approximately 9-10 base pairs downstream, relative to the GATA1 consensus motif center (**Figure 41A**). This observation is valid for all investigated GATA proteins (GATA1, GATA2, GATA3, GATA4, and GATA6).

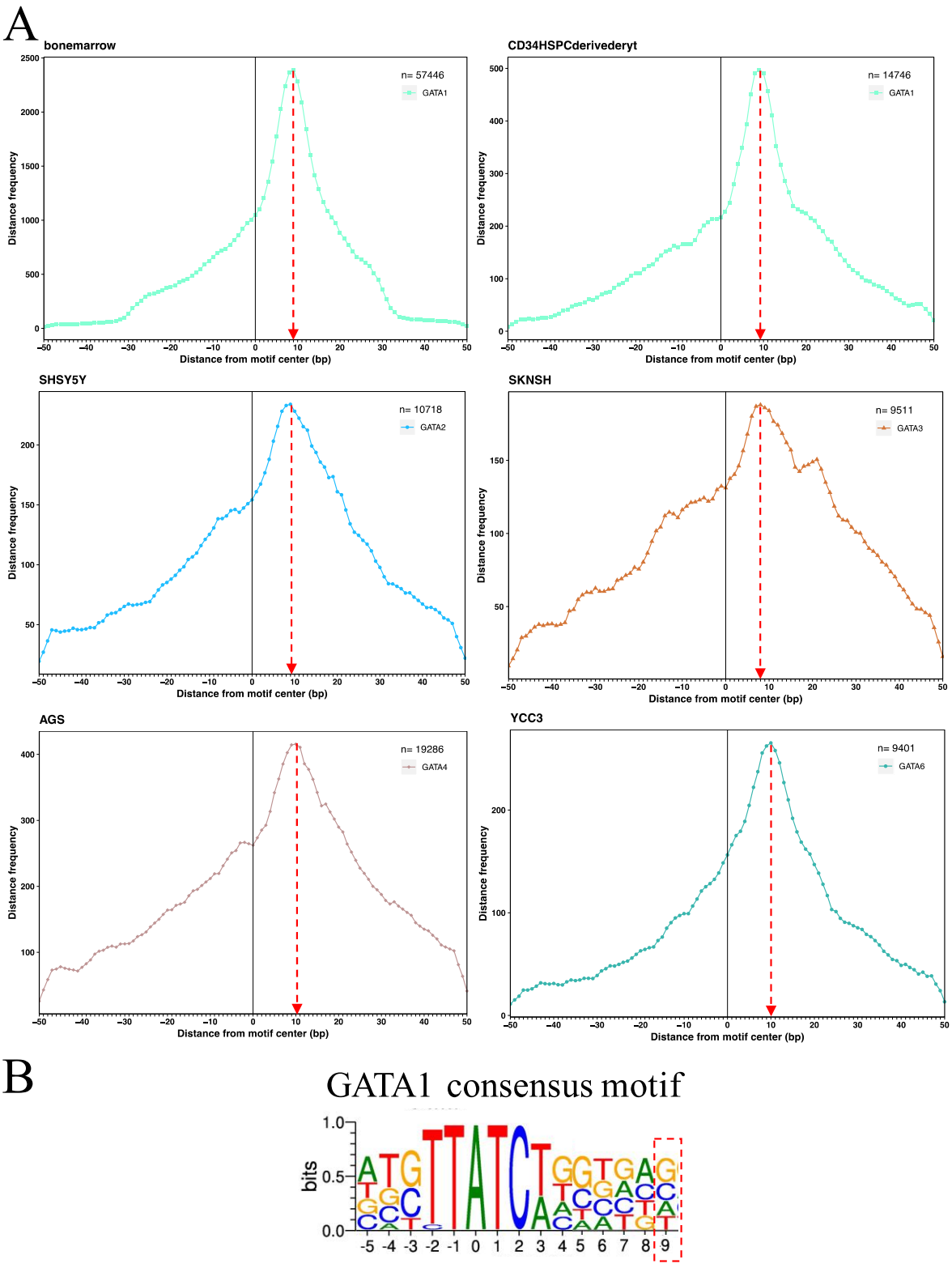


Figure 41. GATA summit position preferences relative to GATA1 binding sites. A) Histograms show the distribution of the peak summits of different GATA proteins in various cell lines relative to the midpoint of the GATA1 motif using a 5 bp sliding window. The maxima of proteins are located 9 basepairs downstream relative to the motif center. B) Motif logo of the

consensus GATA1 PWMs. The logo represents the remapped recognition sequence of GATA1. The shown motif is manually extended to mark the maxima position of summit enrichment.

In the case of TAL1, the summit positions cannot be sharply delineated. Due to the well-known heterodimerization between TAL1 and TCF3 (E2A) proteins, we can study their CE (**Figure 42**). Both proteins bind to the CAG DNA sequence in convergent orientations, which makes the CE palindromic. Unfortunately, the palindromic nature of the CAGCTG sequence affects the definition of protein positions despite the elimination of redundancy. The maxima of the distribution curve are located around position 1-2 (in most cell lines), which represents strong TG base pairs (the reverse complement of the CAG sequence). Interestingly, the shoulders of the distribution curves are at -20 and +20 base pair away from the motif center (**Figure 42**). These represent a summit population with a position away from the core motif. Because of the disturbance of the palindrome sequence, the opposite shoulders can be considered one population. Their remote location remains unclear, but we hypothesize that this is caused by the co-binding events of other factors. We investigated the TCF3 motif also, which is quite similar to the TAL1::TCF3 CE. The summit distances of TAL1 and TCF3 proteins show a congruent distribution on both motifs.

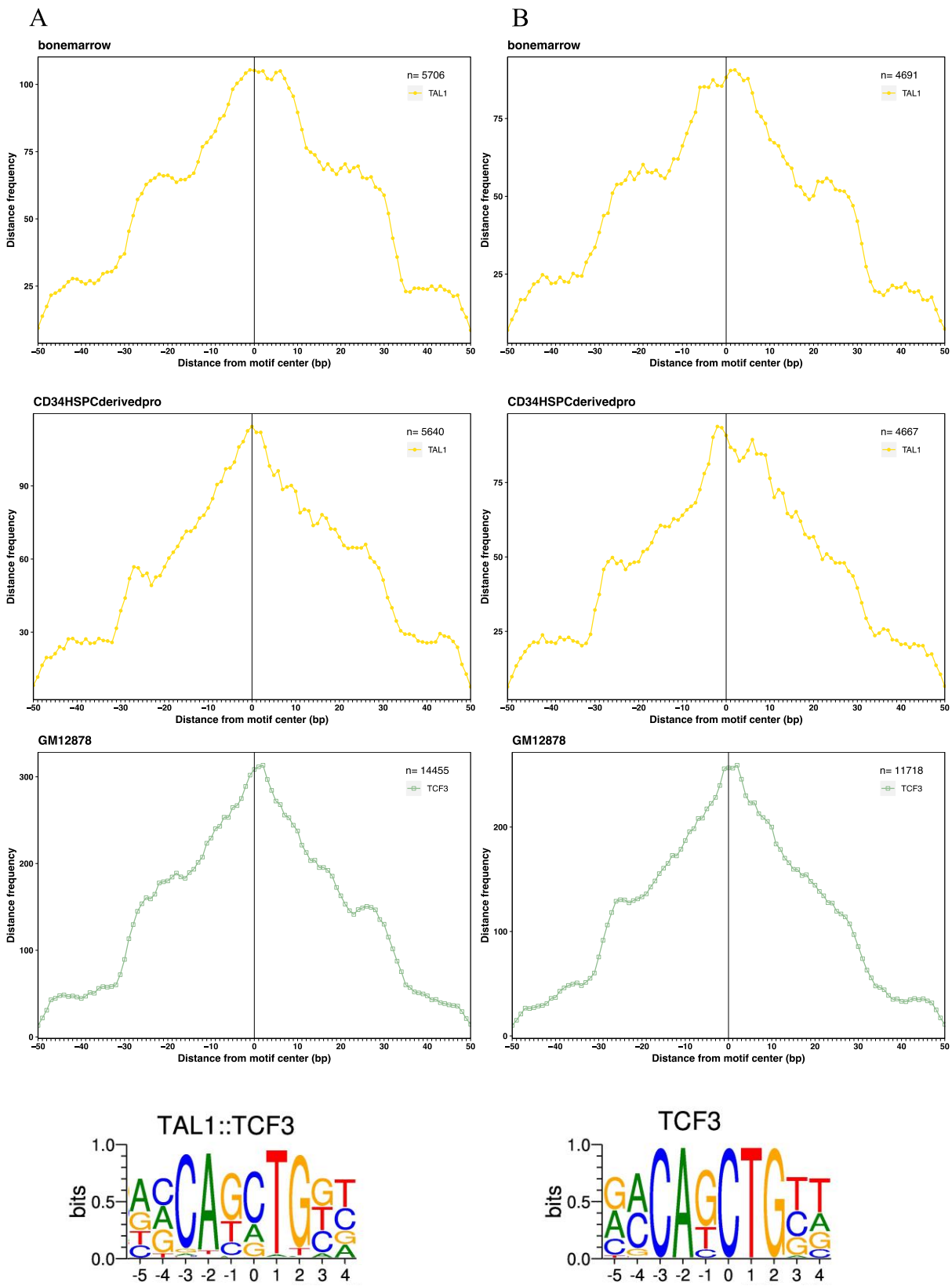


Figure 42. TAL1 and TCF3 summit position preferences relative to TAL1::TCF3 and TCF3 motifs. A) Histograms show the distribution of the peak summits of different TAL1 and TCF3 (E2A) proteins in various cell lines relative to the midpoint

of the motif (which is represented as 0 point) using a 5 bp sliding window. The motif logos represents the consensus binding sites and show the reference points.

Since GATA1 and TAL1 bind non-palindromic sequences, they are useful for further study and validation of our technique and database (Khan et al., 2018). The scatterplot for the GATA1:TAL1 composite element indicates discernible segregation between TAL1 and GATA1 proteins. Both individual signals are located in the proximity of their binding site with a relatively low standard deviation. However, we can observe the same position preference for GATA1 summits as in the case of GATA1 motif, which is not situated directly on their binding site, with an average signal around 5-6 base pairs downstream of the GATA1:TAL1 motif center (**Figure 43A**). In contrast, the TAL1 signal is overlapping with the TAL1 motif at position 8-10, which does not correlate with the observed GATA-like shift. However, the TAL1 signal can be compared to a non-palindromic motif; the summit distance distribution shows a broad enrichment around the mentioned position. The maxima of distance distribution have similar locations as observed in the case of the TAL1::TCF3 motifs. The TAL1 summits, which are juxtaposing with GATA1:TAL1 motifs or TAL1:TCF3 motifs, represent two different clusters. The two CEs are barely overlapping the same summits from the same experiment. The overlap ratio is less than 20 % of the TAL1 summits from bone marrow samples (**Figure 43B**).

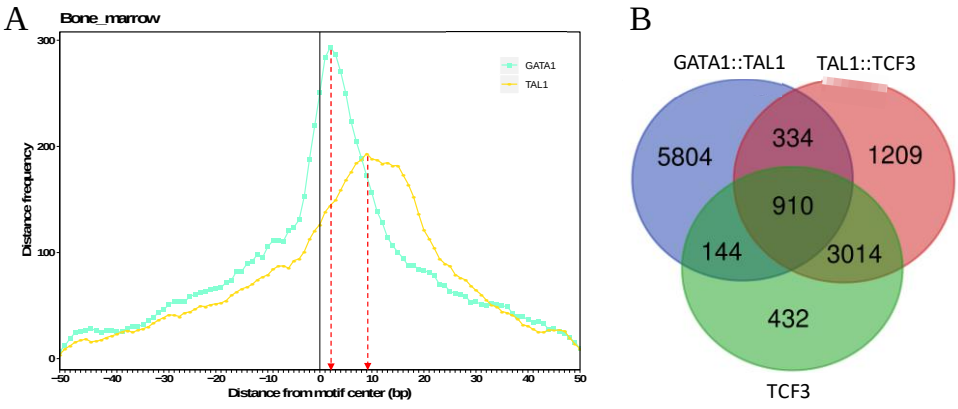


Figure 43. GATA1 and TAL1 position preferences relative to the GATA1::TCF3 motif center. A) The histogram shows the distribution of the peak summits of different GATA1 and TAL1 proteins relative to the center of the GATA1::TAL1

composite element, using a 5 bp sliding window. The representative GATA1 and TAL1 summits were derived from bone marrow cells. (SRA IDs: SRX386202, SRX386203). The maxima positions follow the binding site order within the composite element (GATA1 > TAL1). The TAL1 maximum is located on the TAL1 core motif, while the GATA1 is shifted about 7 bps upstream relative to the GATA1 core motif. B) We investigated the common TAL1 bone marrow ChIP-seq (SRX386203) summits in the vicinity of the three motifs: GATA1::TAL1, TAL1::TCF3, and TCF3. The Venn diagram shows that the surrounding summits of TCF3 and TAL1::TCF3 are common to both motifs. In contrast, the summits in the vicinity of the GATA1::TAL1 motif represent a completely different population and the common summit ratio is below 20 %.

4.5.7. Regulatory SNP analysis in ChIPSummitDB

Sequence variation in the regulatory region can interrupt the expression of the corresponding gene. These disruptions can lead to both increased or decreased expression profiles (Burkhardt et al., 2015). Like malfunction or over-function of a protein due to an exonic mutation, sequence variations in the regulatory region can cause disease. The identification of regulatory SNPs is challenging. The exact regulatory regions are still not well landscaped, which makes the identification difficult. In contrast to exonic mutations, which display well characterized codon changes, the surrounding rSNPs do not have obvious functional consequences. As mentioned previously, there are “spacer” nucleotides in transcription factor binding sites; changes in the spacer region are silent. However, a single nucleotide modification in a recognition site can lead to the loss of transcription factor binding, which can prevent gene expression (Ponomarenko et al., 2003).

We integrated data from the human archive of Single Nucleotide Polymorphism Database (dbSNP), a broad collection of simple genetic polymorphisms containing more than 893 million submissions covering as many rSNPs as possible, and looked into their relationship with responsive elements (Sherry et al., 2001). The overlapping motifs with a specific SNP are viewable in base pair resolution, which allows the examination of the modified nucleotide and its significance (concluded from the PWM score) in DBD recognition. The most efficient way to simultaneously examine rSNPs and TFSBs is via the genome browser (rSNP view). We can

use a dbSNP ID (rs code) to investigate specific SNPs or we can examine a genomic region (maximum 1 Mbp frame) and the list of involved TFBSs and rSNPs (Sherry et al., 2001). The binding sites are displayed as PWMs, which facilitates the assessment of the effects of a specific nucleotide change. The dbSNP view provides a graphical interface, which displays the list the SNPs and their overlapping motifs with PWM scores. The dbSNP view also provides a list of ChIP-seq experiments, which have overlapping signals with the SNP. Therefore, we can collect data about the interacting proteins from different cell lines and highlight which cell lines have binding proteins at the SNP position. Thus, we can assess how the direct binding event is affected by the SNP. In loss of binding, other factors in the complex (which have ChIP-seq signals at this position) can vanish from this region. In the list, we can find information about the disturbed factors in different cell lines.

SNP rs2742624 is an A to G transition located in an intergenic region, approximately 4100 base pairs upstream of the UPK3A gene (on chromosome 22 at position 45676678 bp in GRCh37 genome). UPK3A is expressed in the inner membrane of the urinary bladder and contributes to the strength of that membrane. The absence or loss of a functioning UPK3A protein leads to renal dyplasia. In recent studies, GATA1 was identified as a regulator of the UPK3A gene. A mutation in the GATA responsive element (in the mentioned enhancer region) leads to decreased expression of the gene (Jin, Jung, DebRoy, & Davuluri, 2016). According to their results, the presence of rs2742624 decreased GATA2 ChIP-qPCR and UPK3A mRNA expression in LNCaP cell lines. The mutation occurs at the last A nucleotide of the GATAA core motif, which leads to lower binding affinity. We also identified the GATA1 binding site in our database. The rSNP influences not only a GATA2 motif, but also affects predicted GATA1, GATA3, and GATA5 binding sites. The overlapping ChIP-seq peaks indicate that the SNP can disorientate GATA1 binding in pro-erythroblasts, GATA2 binding in HUVECs, and SHSY5Y and K562 and GATA3 binding in MCF7 cells.

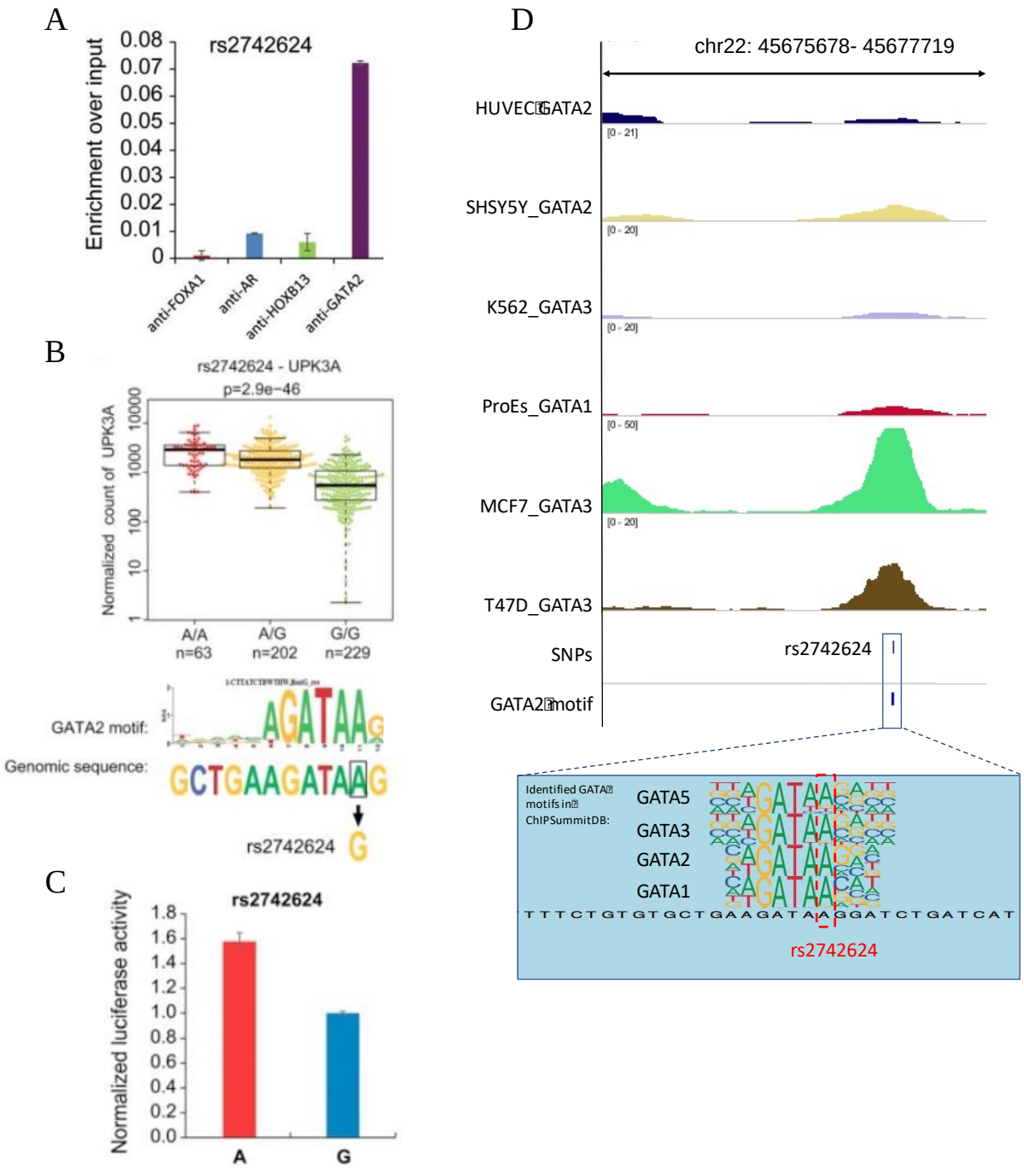


Figure 44. Investigation of rs2742624 SNP. A) GATA2 chromatin occupancy of rs2742624 (position: chr22-45676678; change: A>G) region in normal (A wild type) in LNCaP cells, measured with ChIP–qPCR analysis. B) The expression of UPK3A RNA for patients having homozygous or heterozygous alleles for regulatory SNP rs2742624. The G allele disrupts a GATA2 motif and decreases enhancer activity compared to the A allele of SNP rs2742624. C) GATA2 ChIP was followed by PCR amplification and Sanger sequencing of the rs2742624 region. An input sample from a ChIP assay was used as a control (Jin et al., 2016). D) In ChIPSummitDB, we identified GATA1, GATA2, GATA3, and GATA5 motifs, which were experimentally validated (with ChIP-seq).

5. Discussion

The development of HTS platforms and the related molecular biological techniques resulted in the accumulation of DNA sequencing data. As a measure of its popularity, the amount of sequencing data is growing exponentially, hence there is a growing need for storing and sharing this information. Databases like ENCODE, SRA, and DDBJ solve this problem ("An integrated encyclopedia of DNA elements in the human genome," 2012; Bethesda (MD): National Center for Biotechnology Information (US), 2011; Mashima et al., 2016). These are not just public repositories, but can be used as a path of reference in publications. The stored unprocessed data can serve as a baseline to repeat the experiment, or used for further investigations.

The data producing laboratories and authors are focusing on specific examinations, which support their projects. However, there is more information in HTS data, which has not been revealed yet. As the data requires large computing resources, especially in large scale comparisons, only a few laboratories undertake this challenge. Several processing steps are needed to extract the necessary information from raw data (Barta, 2011). Downstream analysis of raw HTS data is required for extracting meaningful information. The rapid maturation of HTS techniques is a result of the development of processing software and protocols (Heinz et al., 2010). Numerous groups are working simultaneously on the challenges of data processing, resulting in the appearance of distinct software, which combines already existing and newly developed algorithms. The programs often solve the same problem with slightly different methods. The selection of processing programs is crucial in the construction of the pipeline (Koohy, Down, Spivakov, & Hubbard, 2014). Since our investigations often require the comparison of several ChIP-seq experiments, we created a uniform data processing protocol (Barta, 2011) (https://github.com/Raziel01/SummitDB-data-prepare). During our investigation, we observed that the ChIP-seq summit positions can reveal protein position

information in complexes, which is demonstrated through the example of CTCF-cohesin (G. Nagy et al., 2016). The visible shift and order between CTCF and cohesin subunit summit locations are related to the connection order and position preferences of the different proteins relative to each other and to a fixed genomic point. We used the center of the CTCF motifs (as reference points) and published protein structure data to understand the topology of CTCF-cohesin complexes. The extended analysis revealed the strand specific orientation of proteins, which follows a CTCF-SMC1/3-RAD21-STAG1 sequence (**Figure 20-25**). This locates the cohesin ring to the downstream of the CTCF motif. Further investigation with CTCF ChIA-PET data revealed that the CTCF motifs of chromatin loops within TADS face each other. Observation of the convergent orientation of two anchoring CTCF motifs was concurring with the results from other labs (**Figure 31**) (Rao et al., 2014). Relying on this observation, we assumed that the cohesin ring has a proximal position in the DNA loops. We combined the appearance order of cohesin subunit ChIP-seq signals and the published protein structure data, which was plotted on a B-DNA model (**Table 7-8**) (**Figure 26**). Using the result, we created a hypothetical topology model of CTCF mediated chromatin looping that integrates the double embrace model (**Figure 26-27**). This explains the opposite position of SMC1/3 on the double helix relative to other subunits and supports the hypothesis of the hinge domain's nonspecific binding to DNA (G. Nagy et al., 2016).

The cohesin ring, situated inside the loop, enables the physical proximity between inter loop enhancer-promoter regions. We created a "permanent" loop set from MCF7 ChIA-PET parallel replicas (**Figure 31**). About 60 % of the consensus loops could be linked to an active promoter/enhancer mark (H3K4me, H3K4me2, and H3K4me3). The signal intensities showed high asymmetry between opposite sides of the loop (**Figure 32**), indicating that CTCF-cohesin mediated loops have a structural role in the formation of enhancer-promoter looping (in addition to the CTCF insulator function). Furthermore, a global analysis revealed transcription factor/

co-factor ChIP-seq signal enrichment in the vicinity of CTCF binding sites when ZNF143 and YY1 proteins are present. This observation indicates the complex role of CTCF loops in transcriptional regulation (**Figure 37**). The high transcription factor population completely vanishes when YY1 and ZNF143 proteins are not present (**Figure 38**). The summit positions (ZNF143 and YY1 summit position relative to CTCF motif center) and their standard deviations highlight some interesting correlation between ChIP-seq signal and protein topology. The low standard deviation of ZNF143 can be explained by direct binding between ZNF143 and cohesin (**Figure 40**). In contrast, binding seems to be looser as the distance from DNA increases (the crosslink during ChIP), as for indirectly bound factors. The higher mobility of ChIP-seq summit positions increases the standard deviation. However, the maxima of distance distribution curves indicate the approximate position of indirectly bound factors. In the case of YY1 and ZNF143, the maxima are overlapping with SMC proteins, suggesting that ZNF143 and YY1 interact with the hinge domain of SMC proteins and close enhancers through YY1, ZNF143, and other transcription factors at the cohesin ring (**Figure 39-40**). As a large number of correlations could be revealed with extended analyses of these factors, we decided to expand our focus to other regulator proteins and apply this technique. The large amount of data was provided by public databases ("An integrated encyclopedia of DNA elements in the human genome," 2012; Bethesda (MD): National Center for Biotechnology Information (US), 2011).

The accumulation of freely available high throughput sequencing data and the wide variety of processing software resulted in the appearance of secondary databases, which contain processed and semi-processed HTS data. The previously mentioned databases are the source of secondary databases, which have large, continuously expanding content. The database sources are similar, but the processed data is differing due to different processing methods.

Numerous projects accomplished the identification of transcription factor binding sites with developing ChIP-seq processing pipelines (ENCODE, ChIP-Atlas, UCSC, Cistrome etc.)

("An integrated encyclopedia of DNA elements in the human genome," 2012; Karolchik, Hinrichs, & Kent, 2009; Mei et al., 2017; Oki, Ohta, Shioi, Hatanaka, & Ogasawara, 2018). The pipelines mostly include read mapping, peak prediction, and motif enrichment scanning. ENCODE and GEO provide semi-processed data, in addition to the original files (Clough & Barrett, 2016). These are produced using well-defined analysis pipelines, and usually involve mapping and peak prediction results. The available data are not completely consistent and the list of provided files is variable. UCSC extracts information from whole genome sequencing, RNA-seq, and ChIP-seq data, among others (Karolchik et al., 2009). The versatile genome browser visualizes the genomic data and the table browser makes these data downloadable. These data are the result of a well-defined processing protocol of source data from variable origins. The genomic location of transcription factor binding sites is a downloadable result of peak predictions.

TFBS identification has never been so intense. Databases like CR Cistrome, ChIP-Atlas, ReMap, and BloodChIP are attempting to maximize the number of collected data to create a global map of transcription factor binding sites (regions in the genome which are enriched in ChIP-seq aligned reads) (Chacon, Beck, Perera, Wong, & Pimanda, 2014; Cheneby et al., 2018; Mei et al., 2017; Oki et al., 2018). The choice of appropriate processing can be crucial when comparing two or more experiments. Some databases use uniformly processed data, bearing in mind the best comparability (e.g. GTRD). The semi-processed data is usually freely downloadable from the official websites (Yevshin et al., 2017). This data includes regulatory region prediction (BED files), motif enrichment report, and annotation data (gene regulation).

Peak prediction gives us information about the approximate position of TFBS, but this provides only a blurry picture (Heinz et al., 2010; Y Zhang et al., 2008). Several factors can influence the width of predicted peaks, such as biological factors, like the cobinding of other proteins, or technical issues, like the selection of the peak prediction program (differing

algorithms) (Koohy et al., 2014). If investigating transcription factors with known DNA binding domains, the motif enrichment analysis can reveal the preferred sequence. A combination of known preferred sequences and summit positions can give us a more precise prediction of the concrete protein location (Salmon-Divon et al., 2010; Y Zhang et al., 2008). Large collections of transcription factor DNA-binding preferences are published, like JASPAR or HOCOMOCO (Khan et al., 2018; Kulakovskiy et al., 2018). These databases contain curated sets of profiles, collected from literature data. Motif finding software frequently uses these databases to find the most similar profiles to the identified motifs. The profiles are stored as position weight matrices, which can be used to find the occurrences of the TFBSs in the genome. These motif centered databases usually store transcription factor profiles (as PWM or logo) or motif enrichment reports of HTS data, but often lack positioning information for the TFBSs.

The ChIPSummitDb and the previously mentioned databases share common features:

- Large scale data collection
- Uniformly processed ChIP-seq data
- Transcription factor binding site prediction
- Comparable binding sites
- Combinable file formats
- Downloadable content
- Motif report to ChIP-seq experiments

But, only a few databases work with occurrences of a given motif. The difference between our database and other databases based on motif localization centered approach is that we use the JASPAR database as a source of motif matrices, and we attempted to identify their genome wide localization of motif instances. The collected ChIP-seq data provide experimental validation (**Figure 12; Figure 15**). The motif information is represented, not just as a motif

enrichment report, but concrete genomic locations (**Figure 16-17**). We examine the occupancy of different ChIP-seq experiment signals on the identified motifs. Thus, we can map the protein network, which is connected to distinct regulatory sequences. The identified motifs serve as fixed reference points in the genome. This makes the positions of connecting proteins measurable relative to each other proteins or to other motifs (motif protein distance) (**Figure 18**).

The developed summit position based topology prediction was applied in our database (**Figure 18-19**). In addition to TFBS identification, we created a network map of the factors that show overlap with different types of binding sites, based on the downloaded ChIP-seq experiments and the presence of their signals in the vicinity of the binding sites. We can get a comprehensive view of the protein complexes and involved proteins for a motif type with the display modes of ChIPSummitDB. The provided information and features include:

- Occurrence of specific proteins (in different cell types) on the adjusted motif (**Figure 35**)
- The preferred position of proteins relative to the center of a motif (**Figure 35-36**)
- Detailed histogram about the summit- motif center distance distribution for an adjusted experiment (**Figure 36**)
- Overlap between ChIP-seq experiments (peak overlap) in correlation with a given motif
- All produced data is viewable in a genome browser
- Downloadable content
- SNP scan on transcription factor binding sites (**Figure 44**)
- Detailed information about processed ChIP-seq experiments, including origin, link to other databases, and motif enrichment

The database was tested with characterized transcription factor interactions. The FOXA1 and AR were used as positive controls, where we could observe the expected shift

between two proteins with respect to the FOXA1 motif (**Figure 29**). RXR binding sites were investigated with P300 (as co-regulator protein, without a DNA binding domain) as a negative control (**Figure 30**). Due to the indirect binding of P300 to DNA through the RXR protein, there was no observable juxtaposition between the factors. P300 did not show any position preference in addition to the RXR position.

The GATA1:TAL1 motif was treated as a composite element and the juxtaposition could be clearly tracked between GATA1 and TAL1. Thus, our results were congruent with the published structural data **(Figure 43)**. The distance distribution maxima order (of GATA1 and TAL1) followed the pattern in the CE (GATA1 >TAL1). The TAL1 maxima coincided with the TAL1 core motif in the CE, while the GATA1 maxima signal showed a 7 bp shift relative to the start of the GATA1 motif (guanine of GATAA sequence; reverse complement CTATT). The shift can be explained with the structural characteristics of the GATA1 protein. However, the protein structure is barely charted with X-ray crystallography and the vast majority of missing structure (N terminus) is facing in the direction of the shift (data not shown). The shift maxima are located almost one DNA turn away, relative to the core motif. This suggests that the detected ChIP-seq summit position preference may be a result of cross-linking between GATA1 protein non-DNA binding regions and the proximal DNA region during preparation of the ChIP.

To sum up, we proved our initial hypothesis: The unusual CTCF and cohesin ChIP-seq positions refer to the structural features of the protein complex. Using large scale data analysis we were able to create a hypothetical model of CTCF mediated chromatin looping, which support the so-called "double-embrace model". It also proves, that ChIP-seq technique is suitable to deduce the local topology of the protein complexes. During this work, we developed a pipeline, which can automatically perform summit-based topology analysis. This pipeline was used to create a large human ChIP-seq database, which contains genome-wide binding sites for

292 transcription factors and the topological arrangement of their associated protein complexes from more than 3700 experiments. The database is open access and completed with several features to make available the higher-level analysis of binding sites.

The database is still expanding, and we are planning to complete it with mouse ChIP-seq data and evolutionary conservation information. In the future we would like to supplement our database with transcription related information from histone ChIP-seq, RNA-seq and GRO-seq data.

6. Keywords

ChIP-seq, transcription factor, binding site, peak, summit, database, transcription factor binding site (TFBS), High-Throughput Sequencing (HTS), CTCF, chromatin looping, cohesin

7. Summary

The ChIP-seq technique can be used to extract topological information about protein complexes. We developed a summit position based technique, which was used to identify protein positioning relative to a fixed genomic point. To do this, we identified transcription factor binding sites genome-wide with ChIP-seq experimental validation. The identified TFBSs were used as reference points, to measure motif-protein and protein-protein distances. The technique was tested with a CTCF-cohesin complex analysis. The results revealed the cohesin subunit internal orientation in chromatin loops and its structural support in transcriptional regulation and insulation within TADs. The mediator function of YY1 and ZNF143 was also identified between cohesin and transcription factors.

The analysis was extended and we tracked several proteins with published ChIP-seq experiments. We created a database from the results, which contains more than 3702 processed ChIP-seq data. The results have been made publicly available through the http://summit.med.unideb.hu/summitdb/index.php domain. The web interface provides a surface to download and visualize data. Different display modes are provided to investigate transcription factor binding sites and their protein networks in detail. The database was tested with published structural data, including GATA1:TAL1.

8. Glossary

AR: Androgen receptor

ATAC-seq: Assay for Transposase Accessible Chromatin with high-throughput sequencing

BRE: B (TFIIB) recognition element

C-terminal: Carboxy-terminal

CD: chromodomain

COUP-TF: Chicken ovalbumin upstream promoter transcription factor

CTCF: CCCTC-binding factor

ChIA-PET: Chromatin Interaction Analysis by Paired-End Tag Sequencing. This technique is used to identify distal DNA regions which get close proximity to each other. The method is completed with chromatin immunoprecipitation to identify interactions which are associated with a specific protein.

ChIP-exo: ChIP-exonuclease. This technique is used to identify genomic localization (genome wide) of a specific protein. The resolution is improved with exonuclease treatment which degrades strands of the protein-bound DNA in the 5'-3' direction.

ChIP-seq: Chromatin immunoprecipitation followed by sequencing. This technique is used to identify genomic localization (genome wide) of a specific protein.

DBD: DNA-binding domain

DNA: Deoxyribonucleic acid is a molecule

DPE: TATA-box downstream element

DR: Direct repeat

ER: Estrogen receptor

ER: Everted repeat

FR: Farnesoid receptor

GATA1: GATA-binding factor 1

GR: Glucocorticoid receptor

GRO-seq: Global Run-On Sequencing

GTF: General transcription factors

HLH: Helix-loop-helix

HNF6: Hepatocyte nuclear factor 6

HP1: heterochromatin protein 1

HRE: Hormone responsive element

HTH: Helix-turn-helix

HTS: High-Throughput Sequencing

HeLa: cervical cancer cells

HiC: High-throughput 3C methods. This global technique is used to identify distal DNA regions which get close proximity to each other (genome-wide).

IR: Inverted repeat

Inr: Initiator

LNCap: Lymph Node Carcinoma of the prostate

MCF7: human breast adenocarcinoma cell line

MR: Mineralocorticoid receptor

MTE: Motif ten element

N-terminal: Amino-terminal

NCBI: National Center for Biotechnology Information

NDR: Nucleosome-depleted region

NR: Nuclear receptor

PDB: Protein Data Bank

PIC: Preinitiation complex

PPAR: Peroxisome proliferator activated receptor

PPARgamma: Peroxisome Proliferator-Activated Receptor Gamma

PR: Progesterone receptor

PWM: Position weight matrix

PXR: Pregnane X receptor

RAD21: Double-strand-break repair protein (Scc1, Mcd1)

RAR: All-trans retinoic acid receptor

RNA POLII: RNA polymerase 2

RNA-Seq: RNA sequencing

ROR: RAR-related orphan receptor

RXR: 9-cis Retinoic acid receptor

RXRbeta: Retinoid X Receptor Beta

SMC: Structural maintenance of chromosomes proteins

SNP: Single-nucleotide polymorphism

SRA: Sequence Read Archive

TAD: Topological associated domain

TAL1: T-cell acute lymphocytic leukemia protein 1

TBP: TATA-box binding protein

TCT: Polypyrimidine initiator

TF: Transcription factor

TFBM: Transcription factor binding matrix

TFBS: Transcription factor binding site

TR: Thyroid receptor

TRE: Transcriptional regulatory elements

TSS: Transcription start site

UPK3A: Uroplakin-3a is a protein

VDR: Vitamin D3 receptor

ZF: Zinc finger

YY1: Yin Yang 1

9. Acknowledgement

The work presented in this thesis would not have been possible without my close association with many people. I take this opportunity to extend my sincere gratitude and appreciation to all those who made this Ph.D thesis possible

First and foremost, I would like to extend my sincere gratitude to my research guide Dr. (Mrs.) Mala Rao for introducing me to this exciting field of science and for her dedicated help, advice, inspiration and encouragement , throughout my Ph.D.

continuous support,

I gratefully acknowledge Prof. Dr. Fésüs László and Prof. Dr. József Tőzsér the former and recent head of the Department of Biochemistry and Molecular Biology for the opportunity to work in a professional, internationally recognized work environment.

I would like to express my deepest gratitude to my beloved family; my mother, my grandparents and Orsolya Filep for their love and infinite patience.

Funding

This work was supported by the GINOP-2.3.2-15-2016-00044, the 2017-1.3.1-vke-2017-00026 and the FIKP_20428-3_2018_FELITSTRAT grants.

10. References

Amoutzias, G. D., Robertson, D. L., Oliver, S. G., & Bornberg-Bauer, E. (2004). Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. *EMBO Reports*, *5*(3), 274 LP – 279. https://doi.org/10.1038/sj.embor.7400096

An integrated encyclopedia of DNA elements in the human genome. (2012). *Nature*, *489*. https://doi.org/10.1038/nature11247

Anders, S. (2009). Visualization of genomic data with the Hilbert curve. *Bioinformatics (Oxford, England)*, *25*(10), 1231–1235. https://doi.org/10.1093/bioinformatics/btp152

Anthony T. Annunziato. (2008). DNA Packaging: Nucleosomes and Chromatin. *Nature Education*, *1*(26).

Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M., & Iyer, L. M. (2005). The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiology Reviews*, *29*(2), 231–262. https://doi.org/10.1016/j.femsre.2004.12.008

Bailey, S. D., Zhang, X., Desai, K., Aid, M., Corradin, O., Cowper-Sal Lari, R., … Lupien, M. (2015). ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nature Communications*, *2*, 6186. https://doi.org/10.1038/ncomms7186

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, *21*(3), 381–395. https://doi.org/10.1038/cr.2011.22

Barat, C., & Rassart, E. (1998). Members of the GATA family of transcription factors bind to the U3 region of Cas-Br-E and graffi retroviruses and transactivate their expression. *Journal of Virology*, *72*(7), 5579–5588. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9621016

Barta, E. (2011). Command line analysis of ChIP-seq results. *EMBnet.Journal; Vol 17, No 1:*

Next Generation Sequencing Data AnalysisDO - 10.14806/Ej.17.1.209 . Retrieved from http://journal.embnet.org/index.php/embnetjournal/article/view/209/480

Barth, T. K., & Imhof, A. (2010). Fast signals and slow marks: the dynamics of histone modifications. *Trends in Biochemical Sciences*, *35*(11), 618–626. https://doi.org/10.1016/j.tibs.2010.05.006

Beagan, J. A., Duong, M. T., Titus, K. R., Zhou, L., Cao, Z., Ma, J., … Phillips-cremins, J. E. (2017). YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment, 1–14. https://doi.org/10.1101/gr.215160.116.Freely

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242.

Bethesda (MD): National Center for Biotechnology Information (US). (2011). *Sequence Read Archive Submissions Staff. Using the SRA Toolkit to convert .sra files into other formats.*

Blackwood, E. M., & Kadonaga, J. T. (1998). Going the distance: a current view of enhancer action. *Science (New York, N.Y.)*, *281*(5373), 60–63.

Boveri, T. (1909). Die Blastomerenkerne von Ascaris megalocephala und die Theorie der Chromosomenindividualität. *Archiv Für Zellforschung*, (3), 181–268.

Bowman, G. D., & Poirier, M. G. (2015). Post-translational modifications of histones that influence nucleosome dynamics. *Chemical Reviews*, *115*(6), 2274–2295. https://doi.org/10.1021/cr500350x

Brindefalk, B., Dessailly, B. H., Yeats, C., Orengo, C., Werner, F., & Poole, A. M. (2013). Evolutionary history of the TBP-domain superfamily. *Nucleic Acids Research*, *41*(5), 2832–2845. https://doi.org/10.1093/nar/gkt045

Brooker, A. S., & Berkowitz, K. M. (2014). The roles of cohesins in mitosis, meiosis, and human health and disease. *Methods in Molecular Biology (Clifton, N.J.)*, *1170*, 229–266. https://doi.org/10.1007/978-1-4939-0888-2_11

Brown, J. C. (2018). Control of human gene expression: High abundance of divergent transcription in genes containing both INR and BRE elements in the core promoter. *PloS One*, *13*(8), e0202927–e0202927. https://doi.org/10.1371/journal.pone.0202927

Burkhardt, R., Kirsten, H., Beutner, F., Holdt, L. M., Gross, A., Teren, A., … Scholz, M. (2015). Integration of Genome-Wide SNP Data and Gene-Expression Profiles Reveals Six Novel Loci and Regulatory Mechanisms for Amino Acids and Acylcarnitines in Whole Blood. *PLoS Genetics*, *11*(9), e1005510–e1005510. https://doi.org/10.1371/journal.pgen.1005510

Butler, J. E. F., & Kadonaga, J. T. (2002). The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes & Development*, *16*(20), 2583–2592. https://doi.org/10.1101/gad.1026202

Calo, E., & Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? *Molecular Cell*, *49*(5), 825–837. https://doi.org/10.1016/j.molcel.2013.01.038

Carl-Ivar Brändén, J. T. (1999). *Introduction to protein structure* (2nd ed.). New York (N.Y.) : Garland. Retrieved from http://lib.ugent.be/catalog/rug01:000455626

Casolaro, V., Georas, S. N., Song, Z., Zubkoff, I. D., Abdulkadir, S. A., Thanos, D., & Ono, S. J. (1995). Inhibition of NF-AT-dependent transcription by NF-kappa B: implications for differential gene expression in T helper cell subsets. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(25), 11623–11627.

Castro-Mondragon, J. A., Jaeger, S., Thieffry, D., Thomas-Chollier, M., & van Helden, J. (2017). RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Research*, *45*(13), e119. https://doi.org/10.1093/nar/gkx314

Chacon, D., Beck, D., Perera, D., Wong, J. W. H., & Pimanda, J. E. (2014). BloodChIP: a database of comparative genome-wide transcription factor binding profiles in human

blood cells. *Nucleic Acids Research*, *42*(Database issue), D172-7. https://doi.org/10.1093/nar/gkt1036

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., & Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell*, *113*(5), 643–655.

Chan, R. C., Chan, A., Jeon, M., Wu, T. F., Pasqualone, D., Rougvie, A. E., & Meyer, B. J. (2003). Chromosome cohesion is regulated by a clock gene paralogue TIM-1. *Nature*, *423*(6943), 1002–1009. https://doi.org/10.1038/nature01697

Chen, H., Tian, Y., Shu, W., Bo, X., & Wang, S. (2012). Comprehensive Identification and Annotation of Cell Type-Specific and Ubiquitous CTCF-Binding Sites in the Human Genome. *PLOS ONE*, *7*(7), e41374. Retrieved from https://doi.org/10.1371/journal.pone.0041374

Cheneby, J., Gheorghe, M., Artufel, M., Mathelier, A., & Ballester, B. (2018). ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Research*, *46*(D1), D267–D275. https://doi.org/10.1093/nar/gkx1092

Chung, J. H., Bell, A. C., & Felsenfeld, G. (1997). Characterization of the chicken beta-globin insulator. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(2), 575–580.

Clough, E., & Barrett, T. (2016). The Gene Expression Omnibus Database. *Methods in Molecular Biology (Clifton, N.J.)*, *1418*, 93–110. https://doi.org/10.1007/978-1-4939-3578-9_5

Cortini, R., Barbi, M., Caré, B. R., Lavelle, C., Lesne, A., Mozziconacci, J., & Victor, J.-M. (2016). The physics of epigenetics. *Reviews of Modern Physics*, *88*(2), 25002. https://doi.org/10.1103/RevModPhys.88.025002

Craig, J. M. (2005). Heterochromatin--many flavours, common themes. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *27*(1), 17–28. https://doi.org/10.1002/bies.20145

Crick, F. (1970). Central dogma of molecular biology. *Nature*, *227*(5258), 561–563.

Daniel, B., Nagy, G., Hah, N., Horvath, A., Czimmerer, Z., Poliska, S., … Nagy, L. (2014). The active enhancer network operated by liganded RXR supports angiogenic activity in macrophages. *Genes & Development*, *28*(14), 1562–1577. https://doi.org/10.1101/gad.242685.114

Davies, J. O. J., Oudelaar, A. M., Higgs, D. R., & Hughes, J. R. (2017). How best to identify chromosomal interactions: a comparison of approaches. *Nature Methods*, *14*(2), 125–134. https://doi.org/10.1038/nmeth.4146

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., … Cherry, J. M. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*, *46*(D1), D794–D801. https://doi.org/10.1093/nar/gkx1081

De, S., Shaknovich, R., Riester, M., Elemento, O., Geng, H., Kormaksson, M., … Michor, F. (2013). Aberration in DNA Methylation in B-Cell Lymphomas Has a Complex Origin and Increases with Disease Severity. *PLOS Genetics*, *9*(1), e1003137. Retrieved from https://doi.org/10.1371/journal.pgen.1003137

de Wit, E., Vos, E. S. M., Holwerda, S. J. B., Valdes-Quezada, C., Verstegen, M. J. A. M., Teunissen, H., … de Laat, W. (2015). CTCF Binding Polarity Determines Chromatin Looping. *Molecular Cell*, *60*(4), 676–684. https://doi.org/10.1016/j.molcel.2015.09.023

Dekker, J., & Heard, E. (2015). Structural and functional diversity of Topologically Associating Domains. *FEBS Letters*, *589*(20 Pt A), 2877–2884. https://doi.org/10.1016/j.febslet.2015.08.044

Deng, W., & Roberts, S. G. E. (2006). Core promoter elements recognized by transcription

factor IIB. *Biochemical Society Transactions*, *34*(Pt 6), 1051–1053. https://doi.org/10.1042/BST0341051

Diamond, M. I., Miner, J. N., Yoshinaga, S. K., & Yamamoto, K. R. (1990). Transcription factor interactions: selectors of positive or negative regulation from a single DNA element. *Science (New York, N.Y.)*, *249*(4974), 1266–1272.

Dixon, J. R., Gorkin, D. U., & Ren, B. (2016). Chromatin Domains: The Unit of Chromosome Organization. *Molecular Cell*, *62*(5), 668–680. https://doi.org/10.1016/j.molcel.2016.05.018

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., … Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380. https://doi.org/10.1038/nature11082

Dowen, J. M., Fan, Z. P., Hnisz, D., Ren, G., Abraham, B. J., Zhang, L. N., … Young, R. A. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell*, *159*(2), 374–387. https://doi.org/10.1016/j.cell.2014.09.030

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., … Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Evans, C. M., & Jenner, R. G. (2013). Transcription factor interplay in T helper cell differentiation. *Briefings in Functional Genomics*, *12*(6), 499–511. https://doi.org/10.1093/bfgp/elt025

Feeney, K. M., Wasson, C. W., & Parish, J. L. (2010). Cohesin: a regulator of genome integrity and gene expression. *The Biochemical Journal*, *428*(2), 147–161. https://doi.org/10.1042/BJ20100151

Fenley, A. T., Anandakrishnan, R., Kidane, Y. H., & Onufriev, A. V. (2018). Modulation of nucleosomal DNA accessibility via charge-altering post-translational modifications in

histone core. *Epigenetics & Chromatin*, *11*(1), 11. https://doi.org/10.1186/s13072-018-0181-5

Filippova, G. N. (2008). Genetics and epigenetics of the multifunctional protein CTCF. *Current Topics in Developmental Biology*, *80*, 337–360. https://doi.org/10.1016/S0070-2153(07)80009-3

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., … Gottardo, R. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, *16*, 278. https://doi.org/10.1186/s13059-015-0844-5

Fischle, W., Wang, Y., Jacobs, S. A., Kim, Y., Allis, C. D., & Khorasanizadeh, S. (2003). Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. *Genes & Development*, *17*(15), 1870–1881. https://doi.org/10.1101/gad.1110503

Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., … Bernstein, B. E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, *529*(7584), 110–114. https://doi.org/10.1038/nature16490

Forman, B. M., Casanova, J., Raaka, B. M., Ghysdael, J., & Samuels, H. H. (1992). Half-site spacing and orientation determines whether thyroid hormone and retinoic acid receptors and related factors bind to DNA response elements as monomers, homodimers, or heterodimers. *Molecular Endocrinology (Baltimore, Md.)*, *6*(3), 429–442. https://doi.org/10.1210/mend.6.3.1316541

Fulton, D. L., Sundararajan, S., Badis, G., Hughes, T. R., Wasserman, W. W., Roach, J. C., & Sladek, R. (2009). TFCat: the curated catalog of mouse and human transcription factors. *Genome Biology*, *10*(3), R29–R29. https://doi.org/10.1186/gb-2009-10-3-r29

Galonska, C., Ziller, M. J., Karnik, R., & Meissner, A. (2015). Ground State Conditions Induce Rapid Reorganization of Core Pluripotency Factor Binding before Global Epigenetic Reprogramming. *Cell Stem Cell*, *17*(4), 462–470. https://doi.org/10.1016/j.stem.2015.07.005

Galton, F. (1894). *Natural Inheritance*. Macmillan and Company. Retrieved from https://books.google.hu/books?id=a51UeN5hsEQC

Gargiulo, G., Cesaroni, M., Serresi, M., de Vries, N., Hulsman, D., Bruggeman, S. W., … van Lohuizen, M. (2013). In vivo RNAi screen for BMI1 targets identifies TGF-beta/BMP-ER stress pathways as key regulators of neural- and malignant glioma-stem cell homeostasis. *Cancer Cell*, *23*(5), 660–676. https://doi.org/10.1016/j.ccr.2013.03.030

Georgel, P. T., Fletcher, T. M., Hager, G. L., & Hansen, J. C. (2003). Formation of higher-order secondary and tertiary chromatin structures by genomic mouse mammary tumor virus promoters. *Genes & Development*, *17*(13), 1617–1629. https://doi.org/10.1101/gad.1097603

Glover, J. N., & Harrison, S. C. (1995). Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature*, *373*(6511), 257–261. https://doi.org/10.1038/373257a0

Goh, Y., Fullwood, M. J., Poh, H. M., Peh, S. Q., Ong, C. T., Zhang, J., … Ruan, Y. (2012). Chromatin Interaction Analysis with Paired-End Tag Sequencing (ChIA-PET) for mapping chromatin interactions and understanding transcription regulation. *Journal of Visualized Experiments : JoVE*, (62), 3770. https://doi.org/10.3791/3770

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, *27*(7), 1017–1018. https://doi.org/10.1093/bioinformatics/btr064

Gruber, S., Arumugam, P., Katou, Y., Kuglitsch, D., Helmhart, W., Shirahige, K., &

Nasmyth, K. (2006). Evidence that loading of cohesin onto chromosomes involves opening of its SMC hinge. *Cell*, *127*(3), 523–537. https://doi.org/10.1016/j.cell.2006.08.048

Grubert, F., Zaugg, J. B., Kasowski, M., Ursu, O., Spacek, D. V, Martin, A. R., … Snyder, M. (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, *162*(5), 1051–1065. https://doi.org/10.1016/j.cell.2015.07.048

Gu, Z., Eils, R., & Schlesner, M. (2016). HilbertCurve: an R/Bioconductor package for high-resolution visualization of genomic data. *Bioinformatics (Oxford, England)*, *32*(15), 2372–2374. https://doi.org/10.1093/bioinformatics/btw161

Haering, C. H., Lowe, J., Hochwagen, A., & Nasmyth, K. (2002). Molecular architecture of SMC proteins and the yeast cohesin complex. *Molecular Cell*, *9*(4), 773–788.

Han, G. C., Vinayachandran, V., Bataille, A. R., Park, B., Chan-Salis, K. Y., Keller, C. A., … Pugh, B. F. (2015). Genome-Wide Organization of GATA1 and TAL1 Determined at High Resolution. *Molecular and Cellular Biology*, *36*(1), 157–172. https://doi.org/10.1128/MCB.00806-15

Hanaoka, S., Nagadoi, A., & Nishimura, Y. (2005). Comparison between TRF2 and TRF1 of their telomeric DNA-bound structures and DNA-binding activities. *Protein Science : A Publication of the Protein Society*, *14*(1), 119–130. https://doi.org/10.1110/ps.04983705

Harbers, M., Wahlström, G. M., & Vennström, B. (1996). Transactivation by the thyroid hormone receptor is dependent on the spacer sequence in hormone response elements containing directly repeated half-sites. *Nucleic Acids Research*, *24*(12), 2252–2259. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8710493

Hashimoto, H., Wang, D., Horton, J. R., Zhang, X., Corces, V. G., & Cheng, X. (2017). Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA.

Molecular Cell, *66*(5), 711-720.e3. https://doi.org/https://doi.org/10.1016/j.molcel.2017.05.004

He, X., Cicek, A. E., Wang, Y., Schulz, M. H., Le, H.-S., & Bar-Joseph, Z. (2015). De novo ChIP-seq analysis. *Genome Biology*, *16*(1), 205. https://doi.org/10.1186/s13059-015-0756-4

Heidari, N., Phanstiel, D. H., He, C., Grubert, F., Jahanbani, F., Kasowski, M., … Snyder, M. P. (2014). Genome-wide map of regulatory interactions in the human genome. *Genome Research*, *24*(12), 1905–1917. https://doi.org/10.1101/gr.176586.114

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., … Glass, C. K. (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. https://doi.org/10.1016/j.molcel.2010.05.004

Hilbert, D. (1891). Über die stetige Abbildung einer Linie auf ein Flächenstück. *Mathematische Annalen*. Retrieved from http://eudml.org/doc/157555

Hirano, T. (2002). The ABCs of SMC proteins: two-armed ATPases for chromosome condensation, cohesion, and repair. *Genes & Development*, *16*(4), 399–414. https://doi.org/10.1101/gad.955102

Hirano, T. (2006). At the heart of the chromosome: SMC proteins in action. *Nature Reviews. Molecular Cell Biology*, *7*(5), 311–322. https://doi.org/10.1038/nrm1909

Hoffman, B. G., Robertson, G., Zavaglia, B., Beach, M., Cullum, R., Lee, S., … Hoodless, P. A. (2010). Locus co-occupancy, nucleosome positioning, and H3K4me1 regulate the functionality of FOXA2-, HNF4A-, and PDX1-bound loci in islets and liver. *Genome Research*, *20*(8), 1037–1051. https://doi.org/10.1101/gr.104356.109

Huisinga, K. L., Brower-Toland, B., & Elgin, S. C. R. (2006). The contradictory definitions of heterochromatin: transcription and silencing. *Chromosoma*, *115*(2), 110–122.

https://doi.org/10.1007/s00412-006-0052-x

Jin, H.-J., Jung, S., DebRoy, A. R., & Davuluri, R. V. (2016). Identification and validation of regulatory SNPs that modulate transcription factor chromatin binding and gene expression in prostate cancer. *Oncotarget*, *7*(34), 54616–54626. https://doi.org/10.18632/oncotarget.10520

Karolchik, D., Hinrichs, A. S., & Kent, W. J. (2009). The UCSC Genome Browser. *Current Protocols in Bioinformatics*, *Chapter 1*, Unit1.4-Unit1.4. https://doi.org/10.1002/0471250953.bi0104s28

Kel-Margoulis, O. V, Kel, A. E., Reuter, I., Deineko, I. V, & Wingender, E. (2002). TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Research*, *30*(1), 332–334. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11752329

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

Koltzoff, N. (1934). THE STRUCTURE OF THE CHROMOSOMES IN THE SALIVARY GLANDS OF DROSOPHILA. *Science (New York, N.Y.)*, *80*(2075), 312–313. https://doi.org/10.1126/science.80.2075.312

Koohy, H., Down, T. A., Spivakov, M., & Hubbard, T. (2014). A Comparison of Peak Callers Used for DNase-Seq Data, (May). https://doi.org/10.1371/journal.pone.0096303

Kuhlman, T. C., Cho, H., Reinberg, D., & Hernandez, N. (1999). The general transcription factors IIA, IIB, IIF, and IIE are required for RNA polymerase II transcription from the human U1 small nuclear RNA promoter. *Molecular and Cellular Biology*, *19*(3), 2130–2141. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10022900

Kulakovskiy, I. V, Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., … Makeev, V. J. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research*, *46*(D1), D252–D259. https://doi.org/10.1093/nar/gkx1106

Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D., & Ebright, R. H. (1998). New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes & Development*, *12*(1), 34–44. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9420329

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, *42*(Database issue), D980–D985. https://doi.org/10.1093/nar/gkt1113

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., … Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, *22*(9), 1813–1831. https://doi.org/10.1101/gr.136184.111

Lefterova, M. I., Steger, D. J., Zhuo, D., Qatanani, M., Mullican, S. E., Tuteja, G., … Lazar, M. A. (2010). Cell-specific determinants of peroxisome proliferator-activated receptor gamma function in adipocytes and macrophages. *Molecular and Cellular Biology*, *30*(9), 2078–2089. https://doi.org/10.1128/MCB.01651-09

Leinonen, R., Sugawara, H., Shumway, M., & Collaboration, on behalf of the I. N. S. D. (2011). The Sequence Read Archive. *Nucleic Acids Research*, *39*(Database issue), D19–D21. https://doi.org/10.1093/nar/gkq1019

Leleu, M., Lefebvre, G., & Rougemont, J. (2010). Processing and analyzing ChIP-seq data:

from short reads to regulatory interactions. *Briefings in Functional Genomics*, *9*(5–6), 466–476. https://doi.org/10.1093/bfgp/elq022

Levene, Phoebus; Jacobs, W. (1909). *Über Inosinsäure*. Berichte der deutschen chemischen Gesellschaft.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lim, C. Y., Santoso, B., Boulay, T., Dong, E., Ohler, U., & Kadonaga, J. T. (2004). The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes & Development*, *18*(13), 1606–1617. https://doi.org/10.1101/gad.1193404

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.

Love, P. E., Warzecha, C., & Li, L. (2014). Ldb1 complexes: the new master regulators of erythroid gene transcription. *Trends in Genetics : TIG*, *30*(1), 1–9. https://doi.org/10.1016/j.tig.2013.10.001

Luger, K., Mader, 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. https://doi.org/10.1038/38444

Luscombe, N. M., Laskowski, R. A., & Thornton, J. M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Research*, *29*(13), 2860–2874.

Mangelsdorf, D. J., Thummel, C., Beato, M., Herrlich, P., Schutz, G., Umesono, K., … Evans, R. M. (1995). The nuclear receptor superfamily: the second decade. *Cell*, *83*(6),

835–839.

Marmorstein, R., & Trievel, R. C. (2009). Histone modifying enzymes: structures, mechanisms, and specificities. *Biochimica et Biophysica Acta*, *1789*(1), 58–68. https://doi.org/10.1016/j.bbagrm.2008.07.009

Mashima, J., Kodama, Y., Kosuge, T., Fujisawa, T., Katayama, T., Nagasaki, H., … Takagi, T. (2016). DNA data bank of Japan (DDBJ) progress report. *Nucleic Acids Research*, *44*(D1), D51–D57. https://doi.org/10.1093/nar/gkv1105

Matys, V., Kel-Margoulis, O. V, Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., … Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, *34*(Database issue), D108-10. https://doi.org/10.1093/nar/gkj143

Mehta, G. D., Kumar, R., Srivastava, S., & Ghosh, S. K. (2013). Cohesin: functions beyond sister chromatid cohesion. *FEBS Letters*, *587*(15), 2299–2312. https://doi.org/10.1016/j.febslet.2013.06.035

Mehta, G. D., Rizvi, S. M. A., & Ghosh, S. K. (2012). Cohesin: a guardian of genome integrity. *Biochimica et Biophysica Acta*, *1823*(8), 1324–1342. https://doi.org/10.1016/j.bbamcr.2012.05.027

Mei, S., Qin, Q., Wu, Q., Sun, H., Zheng, R., Zang, C., … Liu, X. S. (2017). Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Research*, *45*(D1), D658–D662. https://doi.org/10.1093/nar/gkw983

Melby, T. E., Ciampaglio, C. N., Briscoe, G., & Erickson, H. P. (1998). The symmetrical structure of structural maintenance of chromosomes (SMC) and MukB proteins: long, antiparallel coiled coils, folded at a flexible hinge. *The Journal of Cell Biology*, *142*(6), 1595–1604.

Mirny, L. A. (2011). The fractal globule as a model of chromatin architecture in the cell. *Chromosome Research*, *19*(1), 37–51. https://doi.org/10.1007/s10577-010-9177-0

Müller, M. M., & Muir, T. W. (2015). Histones: at the crossroads of peptide and protein chemistry. *Chemical Reviews*, *115*(6), 2296–2349. https://doi.org/10.1021/cr5003529

Nagy, G., Czipa, E., Steiner, L., Nagy, T., Pongor, S., Nagy, L., & Barta, E. (2016). Motif oriented high-resolution analysis of ChIP-seq data reveals the topological order of CTCF and cohesin proteins on DNA. *BMC Genomics*, *17*(1). https://doi.org/10.1186/s12864-016-2940-7

Nagy, Gergely, Czipa, E., Steiner, L., Nagy, T., Pongor, S., Nagy, L., & Barta, E. (2016). Motif oriented high-resolution analysis of ChIP-seq data reveals the topological order of CTCF and cohesin proteins on DNA. *BMC Genomics*, *17*(1), 637. https://doi.org/10.1186/s12864-016-2940-7

Nasmyth, K., & Haering, C. H. (2009). Cohesin: its roles and mechanisms. *Annual Review of Genetics*, *43*, 525–558. https://doi.org/10.1146/annurev-genet-102108-134233

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., … Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), 381–385. https://doi.org/10.1038/nature11049

Ogata, K., Hojo, H., Aimoto, S., Nakai, T., Nakamura, H., Sarai, A., … Nishimura, Y. (1992). Solution structure of a DNA-binding unit of Myb: a helix-turn-helix-related motif with conserved tryptophans forming a hydrophobic core. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(14), 6428–6432. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/1631139

Ohlsson, R., Renkawitz, R., & Lobanenkov, V. (2001). CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends in Genetics : TIG*, *17*(9), 520–527.

Oki, S., Ohta, T., Shioi, G., Hatanaka, H., & Ogasawara, O. (2018). ChIP-Atlas : a data-mining suite powered by full integration of public ChIP-seq data, 1–10. https://doi.org/10.15252/embr.201846255

Olins, A. L., & Olins, D. E. (1974). Spheroid chromatin units (v bodies). *Science (New York, N.Y.)*, *183*(4122), 330–332.

Orphanides, G., Lagrange, T., & Reinberg, D. (1996). The general transcription factors of RNA polymerase II. *Genes & Development*, *10*(21), 2657–2683.

Parry, T. J., Theisen, J. W. M., Hsu, J.-Y., Wang, Y.-L., Corcoran, D. L., Eustice, M., … Kadonaga, J. T. (2010). The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes & Development*, *24*(18), 2013–2018. https://doi.org/10.1101/gad.1951110

Pawlak, M., Lefebvre, P., & Staels, B. (2012). General molecular biology and architecture of nuclear receptors. *Current Topics in Medicinal Chemistry*, *12*(6), 486–504. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/22242852

Pchelintsev, N. A., Adams, P. D., & Nelson, D. M. (2016). Critical Parameters for Efficient Sonication and Improved Chromatin Immunoprecipitation of High Molecular Weight Proteins. *PloS One*, *11*(1), e0148023. https://doi.org/10.1371/journal.pone.0148023

Phillips-Cremins, J. E., & Corces, V. G. (2013). Chromatin insulators: linking genome organization to cellular function. *Molecular Cell*, *50*(4), 461–474. https://doi.org/10.1016/j.molcel.2013.04.018

Piper, J., Elze, M. C., Cauchy, P., Cockerill, P. N., Bonifer, C., & Ott, S. (2013). Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Research*, *41*(21), e201. https://doi.org/10.1093/nar/gkt850

Pohlert, T. (2015). *PMCMR: Calculate Pairwise Multiple Comparisons of Mean Rank Sums (Version 4.0)*.

Ponomarenko, J. V, Merkulova, T. I., Orlova, G. V, Fokin, O. N., Gorshkova, E. V, Frolov, A. S., … Ponomarenko, M. P. (2003). rSNP_Guide, a database system for analysis of transcription factor binding to DNA with variations: application to genome annotation. *Nucleic Acids Research*, *31*(1), 118–121. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/12519962

Prakash, K., & Fournier, D. (2017). Deciphering the histone code to build the genome structure. *BioRxiv*, 217190. https://doi.org/10.1101/217190

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical Recipes in C (2Nd Ed.): The Art of Scientific Computing*. New York, NY, USA: Cambridge University Press.

Prlić, A., Bradley, A. R., Duarte, J. M., Rose, P. W., Rose, A. S., & Valasatava, Y. (2018). NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics*, *34*(21), 3755–3758. https://doi.org/10.1093/bioinformatics/bty419

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*. https://doi.org/10.1093/bioinformatics/btq033

Quitschke, W. W., Taheny, M. J., Fochtmann, L. J., & Vostrov, A. A. (2000). Differential effect of zinc finger deletions on the binding of CTCF to the promoter of the amyloid precursor protein gene. *Nucleic Acids Research*, *28*(17), 3370–3378. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10954607

Rabl, C. (1885). Über Zelltheilung. *Morphologisches Jahrbuch Band*, (10), 214–330.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., … Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

Ravasi, T., Suzuki, H., Vittorio Cannistraci, C., Katayama, S., Bajic, V., Tan, K., … Hayashizaki, Y. (2010). *An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man*. *Cell* (Vol. 140). https://doi.org/10.1016/j.cell.2010.01.044

Reeve, J. N. (2003). Archaeal chromatin and transcription. *Molecular Microbiology*, *48*(3), 587–598.

Renaud, S., Loukinov, D., Abdullaev, Z., Guilleret, I., Bosman, F. T., Lobanenkov, V., & Benhattar, J. (2007). Dual role of DNA methylation inside and outside of CTCF-binding regions in the transcriptional regulation of the telomerase hTERT gene. *Nucleic Acids Research*, *35*(4), 1245–1256. https://doi.org/10.1093/nar/gkl1125

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011, January). Integrative genomics viewer. *Nature Biotechnology*. United States. https://doi.org/10.1038/nbt.1754

Rodriguez-Martinez, J. A., Reinke, A. W., Bhimsaria, D., Keating, A. E., & Ansari, A. Z. (2017). Combinatorial bZIP dimers display complex DNA-binding specificity landscapes. *ELife*, *6*. https://doi.org/10.7554/eLife.19272

Rusche, L. N., Kirchmaier, A. L., & Rine, J. (2003). The establishment, inheritance, and function of silenced chromatin in Saccharomyces cerevisiae. *Annual Review of Biochemistry*, *72*, 481–516. https://doi.org/10.1146/annurev.biochem.72.121801.161547

Sahu, B., Laakso, M., Ovaska, K., Mirtti, T., Lundin, J., Rannikko, A., … Janne, O. A. (2011). Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. *The EMBO Journal*, *30*(19), 3962–3976. https://doi.org/10.1038/emboj.2011.328

Salmon-Divon, M., Dvinge, H., Tammoja, K., & Bertone, P. (2010). PeakAnalyzer: Genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics*, *11*(1), 415. https://doi.org/10.1186/1471-2105-11-415

Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., … Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 201518552. https://doi.org/10.1073/pnas.1518552112

Sebastian, A., & Contreras-Moreira, B. (2013). footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics*, *30*(2), 258–265. https://doi.org/10.1093/bioinformatics/btt663

Sexton, T., & Cavalli, G. (2015). The role of chromosome domains in shaping the functional genome. *Cell*, *160*(6), 1049–1059. https://doi.org/10.1016/j.cell.2015.02.040

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11125122

Shintomi, K., & Hirano, T. (2007). How are cohesin rings opened and closed? *Trends in Biochemical Sciences*, *32*(4), 154–157. https://doi.org/https://doi.org/10.1016/j.tibs.2007.02.002

Shu, F.-J., Sidell, N., Yang, D., & Kallen, C. B. (2010). The tri-nucleotide spacer sequence between estrogen response element half-sites is conserved and modulates ERalpha-mediated transcriptional responses. *The Journal of Steroid Biochemistry and Molecular Biology*, *120*(4–5), 172–179. https://doi.org/10.1016/j.jsbmb.2010.04.009

Siersbaek, R., Rabiee, A., Nielsen, R., Sidoli, S., Traynor, S., Loft, A., … Mandrup, S. (2014). Transcription factor cooperativity in early adipogenic hotspots and super-enhancers. *Cell Reports*, *7*(5), 1443–1455. https://doi.org/10.1016/j.celrep.2014.04.042

Sigrist, C. J. A., de Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., … Xenarios, I. (2013). New and continuing developments at PROSITE. *Nucleic Acids Research*, *41*(Database issue), D344-7. https://doi.org/10.1093/nar/gks1067

Sims, R. J. 3rd, Nishioka, K., & Reinberg, D. (2003). Histone lysine methylation: a signature for chromatin function. *Trends in Genetics : TIG*, *19*(11), 629–639. https://doi.org/10.1016/j.tig.2003.09.007

Skinner, M. E., Uzilov, A. V, Stein, L. D., Mungall, C. J., & Holmes, I. H. (2009). JBrowse: a next-generation genome browser. *Genome Research*, *19*(9), 1630–1638. https://doi.org/10.1101/gr.094607.109

Smale, S. T., & Baltimore, D. (1989). The "initiator" as a transcription control element. *Cell*, *57*(1), 103–113.

Sun, M., Nishino, T., & Marko, J. F. (2013). The SMC1-SMC3 cohesin heterodimer structures DNA through supercoiling-dependent loop formation. *Nucleic Acids Research*, *41*(12), 6149–6160. https://doi.org/10.1093/nar/gkt303

Takahashi, K., & Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, *126*(4), 663–676. https://doi.org/10.1016/j.cell.2006.07.024

Takeuchi, A., Reddy, G. S., Kobayashi, T., Okano, T., Park, J., & Sharma, S. (1998). Nuclear factor of activated T cells (NFAT) as a molecular target for 1alpha,25-dihydroxyvitamin D3-mediated effects. *Journal of Immunology (Baltimore, Md. : 1950)*, *160*(1), 209–218.

Taslim, C., Huang, K., Huang, T., & Lin, S. (2012). Analyzing ChIP-seq data: preprocessing, normalization, differential identification, and binding pattern characterization. *Methods in Molecular Biology (Clifton, N.J.)*, *802*, 275–291. https://doi.org/10.1007/978-1-61779-400-1_18

Team, R. C. (2014). R: A Language and Environment for Statistical Computing. Retrieved from http://www.r-project.org/

Thomas, M. C., & Chiang, C.-M. (2006). The general transcription machinery and general cofactors. *Critical Reviews in Biochemistry and Molecular Biology*, *41*(3), 105–178.

https://doi.org/10.1080/10409230600648736

Tian, L., Zhang, Z., Wang, H., Zhao, M., Dong, Y., & Gong, Y. (2016). Sequence-Dependent T:G Base Pair Opening in DNA Double Helix Bound by Cren7, a Chromatin Protein Conserved among Crenarchaea. *PloS One*, *11*(9), e0163361. https://doi.org/10.1371/journal.pone.0163361

Tomschik, M., Zheng, H., van Holde, K., Zlatanova, J., & Leuba, S. H. (2005). Fast, long-range, reversible conformational fluctuations in nucleosomes revealed by single-pair fluorescence resonance energy transfer. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(9), 3278 LP – 3283. Retrieved from http://www.pnas.org/content/102/9/3278.abstract

Toropainen, S., Malinen, M., Kaikkonen, S., Rytinki, M., Jääskeläinen, T., Sahu, B., … Palvimo, J. J. (2015). SUMO ligase PIAS1 functions as a target gene selective androgen receptor coregulator on prostate cancer cell chromatin. *Nucleic Acids Research*, *43*(2), 848–861. https://doi.org/10.1093/nar/gku1375

Tyagi, S., Gupta, P., Saini, A. S., Kaushal, C., & Sharma, S. (2011). The peroxisome proliferator-activated receptor: A family of nuclear receptors role in various diseases. *Journal of Advanced Pharmaceutical Technology & Research*, *2*(4), 236–240. https://doi.org/10.4103/2231-4040.90879

Udvardy, A., Maine, E., & Schedl, P. (1985). The 87A7 chromomere. Identification of novel chromatin structures flanking the heat shock locus that may define the boundaries of higher order domains. *Journal of Molecular Biology*, *185*(2), 341–358.

Umesono, K., Murakami, K. K., Thompson, C. C., & Evans, R. M. (1991). Direct repeats as selective response elements for the thyroid hormone, retinoic acid, and vitamin D3 receptors. *Cell*, *65*(7), 1255–1266. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/1648450

UniProt Consortium, T. (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, *46*(5), 2699. https://doi.org/10.1093/nar/gky092

Valdeolmillos, A. M., Viera, A., Page, J., Prieto, I., Santos, J. L., Parra, M. T., … Rufas, J. S. (2007). Sequential loading of cohesin subunits during the first meiotic prophase of grasshoppers. *PLoS Genetics*, *3*(2), e28–e28. https://doi.org/10.1371/journal.pgen.0030028

van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., … Lander, E. S. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. *Journal of Visualized Experiments : JoVE*, (39). https://doi.org/10.3791/1869

Walhout, A. J. M. (2006). Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. *Genome Research*, *16*(12), 1445–1454. https://doi.org/10.1101/gr.5321506

Wallace, J. A., & Felsenfeld, G. (2007). We gather together: insulators and genome organization. *Current Opinion in Genetics & Development*, *17*(5), 400–407. https://doi.org/10.1016/j.gde.2007.08.005

Wang, J., Zhuang, J., Iyer, S., Lin, X.-Y., Greven, M. C., Kim, B.-H., … Weng, Z. (2013). Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Research*, *41*(Database issue), D171–D176. https://doi.org/10.1093/nar/gks1221

Wang, L., Chen, J., Wang, C., Uuskula-Reimand, L., Chen, K., Medina-Rivera, A., … Li, W. (2014). MACE: model based analysis of ChIP-exo. *Nucleic Acids Research*, *42*(20), e156. https://doi.org/10.1093/nar/gku846

Wang, Z., Wu, Y., Li, L., & Su, X.-D. (2013). Intermolecular recognition revealed by the complex structure of human CLOCK-BMAL1 basic helix-loop-helix domains with E-

box DNA. *Cell Research*, *23*(2), 213–224. https://doi.org/10.1038/cr.2012.170

Watanabe, M., & Kakuta, H. (2018). Retinoid X Receptor Antagonists. *International Journal of Molecular Sciences*, *19*(8), 2354. https://doi.org/10.3390/ijms19082354

Watson, J. D., & Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, *171*, 737. Retrieved from https://doi.org/10.1038/171737a0

Weintraub, A. S., Li, C. H., Zamudio, A. V, Sigova, A. A., Hannett, N. M., Day, D. S., … Young, R. A. (2017). YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell*, *171*(7), 1573-1588.e28. https://doi.org/10.1016/j.cell.2017.11.008

West, A. G., Gaszner, M., & Felsenfeld, G. (2002). Insulators: many functions, many mechanisms. *Genes & Development*, *16*(3), 271–288. https://doi.org/10.1101/gad.954702

Whitaker, J. W., Chen, Z., & Wang, W. (2015). Predicting the human epigenome from DNA motifs. *Nature Methods*, *12*(3), 265–272, 7 p following 272. https://doi.org/10.1038/nmeth.3065

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., … Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, *153*(2), 307–319. https://doi.org/10.1016/j.cell.2013.03.035

Wickham, H. (n.d.). Reshaping Data with the reshape Package, (November 2007).

Wingender, E. (2013). Criteria for an updated classification of human transcription factor DNA-binding domains. *Journal of Bioinformatics and Computational Biology*, *11*(1), 1340007. https://doi.org/10.1142/S0219720013400076

Wingender, E., Schoeps, T., Haubrock, M., & Dönitz, J. (2015). TFClass: a classification of human transcription factors and their rodent orthologs. *Nucleic Acids Research*,

43(Database issue), D97-102. https://doi.org/10.1093/nar/gku1064

Wingender, E., Schoeps, T., Haubrock, M., Krull, M., & Dönitz, J. (2018). TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Research, 46*(D1), D343–D347. Retrieved from http://dx.doi.org/10.1093/nar/gkx987

Xi, H., Yu, Y., Fu, Y., Foley, J., Halees, A., & Weng, Z. (2007). Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome Research, 17*(6), 798–806. https://doi.org/10.1101/gr.5754707

Xiao, T., Wallace, J., & Felsenfeld, G. (2011). Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity. *Molecular and Cellular Biology*, *31*(11), 2174–2183. https://doi.org/10.1128/MCB.05093-11

Yevshin, I., Sharipov, R., Valeev, T., Kel, A., & Kolpakov, F. (2017). GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Research, 45*(D1), D61–D67. https://doi.org/10.1093/nar/gkw951

Yin, M., Wang, J., Wang, M., Li, X., Zhang, M., Wu, Q., & Wang, Y. (2017). Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites. *Cell Research*, *27*(11), 1365–1377. https://doi.org/10.1038/cr.2017.131

You, A., Tong, J. K., Grozinger, C. M., & Schreiber, S. L. (2001). CoREST is an integral component of the CoREST- human histone deacetylase complex. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(4), 1454–1458. https://doi.org/10.1073/pnas.98.4.1454

Zeng, W., Ball Jr, A. R., & Yokomori, K. (2010). HP1: heterochromatin binding proteins working the genome. *Epigenetics*, *5*(4), 287–292. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20421743

Zhang, Y, Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., … Li, W. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biol, 9*. https://doi.org/10.1186/gb-2008-9-9-r137

Zhang, Yonghong, Larsen, C. A., Stadler, H. S., & Ames, J. B. (2011). Structural basis for sequence specific DNA binding and protein dimerization of HOXA13. *PloS One*, *6*(8), e23069. https://doi.org/10.1371/journal.pone.0023069

Zhang, Yuxiang, Fang, B., Emmett, M. J., Damle, M., Sun, Z., Feng, D., … Lazar, M. A. (2015). GENE REGULATION. Discrete functions of nuclear receptor Rev-erbα couple metabolism to the clock. *Science (New York, N.Y.)*, *348*(6242), 1488–1492. https://doi.org/10.1126/science.aab3021