Published in final edited form as:

Semin Cell Dev Biol. 2019 June ; 90: 114–127. doi:10.1016/j.semcdb.2018.08.003.

One protein to rule them all: the role of CCCTC-binding factor in shaping Human genome in health and disease

Michał Łazniewski^{#1,3}, Wayne K. Dawson^{#1,2}, Anna Maria Rusek^{1,5}, and Dariusz Plewczynski^{1,4,6,+}

¹Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097 Warsaw, Poland

²Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 103-8657 Japan ³Department of Physical Chemistry, Faculty of Pharmacy, Medical University of Warsaw, Banacha 1, 02-097 Warsaw, Poland ⁴Centre for Innovative Research, Medical University of Białystok, Białystok, Poland ⁵Clinical Molecular Biology Department, Medical University of Białystok, Białystok, Poland ⁶Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

[#] These authors contributed equally to this work.

Abstract

The eukaryotic genome, constituting several billion base pairs, must be contracted to fit within the volume of a nucleus where the diameter is on the scale of μm . The 3D structure and packing of such a long sequence; cannot be left to pure chance, as DNA must be efficiently used for its primary roles as a matrix for transcription and replication. In recent years, methods like chromatin conformation capture (including 3C, 4C, Hi-C, ChIA-PET and Multi-ChIA) and optical microscopy have advanced substantially and have shed new light on how eukaryotic genomes are hierarchically organized; first into 10-nm fiber, next into DNA loops, topologically associated domains and finally into interphase or mitotic chromosomes. This knowledge has allowed us to revise our understanding regarding the mechanisms governing the process of DNA organization. Mounting experimental evidence suggests that the key element in the formation of loops is the binding of the CCCTC-binding factor (CTCF) to DNA; a protein that can be referred to as the chief organizer of the genome. However, CTCF does not work alone but in cooperation with other proteins, such as cohesin or Yin Yang 1 (YY1). In this short review, we briefly describe our current understanding of the structure of eukaryotic genomes, how they are established and how the formation of DNA loops can influence gene expression. We discuss the recent discoveries describing the 3D structure of the CTCF-DNA complex and the role of CTCF in establishing genome structure. Finally, we briefly explain how various genetic disorders might arise as a consequence of mutations in the CTCF target sequence or alteration of genomic imprinting.

*Correspondence: d.plewczynski@cent.uw.edu.pl.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

The human genome consists of more than 3 billion base pairs (3 Mbp) and could be stretched out to over 2m in diploid cells, yet it must pack into a nucleus with a diameter that ranges from roughly 6 to 10 μm [1]. If DNA behaved like an ideal polymer, we would expect that the volume of the nucleus (V) would increase with respect to the size of the genome (N) as $V \sim N^{3/2}$; however, based on data collected from many eukaryotic organisms, it has been shown that the volume increases linearly with respect to the genome size; i.e., $V \sim N$ [2, 3]. Another example of significant compactness comes from the calculations of the probability of a contact between two sites separated by a genomic distance s . The probability can be expressed as a power law $s^{-\alpha}$, where α equal to 1 is expected if DNA behaved like a fractal globule [4]. Based on the experimental data, however, it can be seen that α tends to vary widely between cells including 0.5 for metaphase HeLa cells, 1.2 to 1.5 for oocytes [5], 1.6 for embryonic stem cells [6], and 1.1 in lymphoblastoid cells [4]. Nevertheless, despite the wide variation of α , DNA is often significantly more compact than what would be expected for an ideal polymer ($\alpha \sim 1.5$).

This compactness combined with highly dynamic changes in the DNA structure observed during the cell cycle suggest the existence of highly organized mechanisms that bind diverse parts of the chromatin (a complex of DNA, proteins and RNA) into complex substructures. Our current knowledge indicates that a likely explanation is the tendency of chromatin fiber to form loops on a local scale up to hundreds of thousands of kilobase pairs that further compartmentalize the chromatin into topologically associated domains (TADs). For a globular polymer (or a fractal globule), loops would result from non-specific, random contacts; however, the experimental evidence strongly suggests that for DNA the locations of such loops are far from random. Thus, DNA is not a simple random walk polymer ensconced within the nucleus of a cell. The real-life structure is far more complex than might be naively assumed.

Chromatin conformation capture (3C) methods have revealed that one of the primary factors responsible for the loop formation is an 11 zinc-finger protein – the CCCTC-binding factor (CTCF). This feature of CTCF is now proposed to be a causative agent for most of the previously described roles of CTCF; namely, gene repression, gene activation, chromatin insulation, alternative splicing, X-chromosome inactivation and maintenance of genomic imprinting [7, 8, 9]. The protein, besides mediating the formation of the majority of loops, was also found to maintain the borders of topologically associated-domains (TADs), as these sites were both enriched with the CTCF binding motif and the CTCF presence was confirmed with chromatin immunoprecipitation-sequencing (ChIP-seq) experiments. Because of the importance of CTCF in establishing the chromatin 3D structure the protein is sometimes referred to as the master weaver of the genome [9]. Beside CTCF, other proteins are also involved in the establishment of chromatin structure. For example, Yin Yang 1 (YY1) mediates the formation of loops that bring promoters and enhancers into close spatial proximity [10]. The CCCTC-binding factor like protein (CTCFL), which is expressed in germ cells and some cancers, mediates the formation of a subset of cell-specific loops [11,12]. Cohesin, a complex of several proteins, was originally found to mediate cohesion of

sister chromatids (for review, see [13]). More recent experiments suggest that it can traverse along DNA to actually form the DNA loops [14].

If the link between genome organization and gene expression is universal, a view supported by the current available data [15–18], it is likely that mutations targeting elements responsible for the organization of the genome might lead to pathological conditions. For example, mutations in the CTCF binding sites can lead to several types of cancer, autoimmune-based diseases, inflammatory-related conditions, or malformed limbs [101, 108, 115]. Mutations in the CTCF sequence itself can result in head and neck cancers [107]. CTCF is also involved in establishing genomic imprinting that allows only the paternal or maternal genes to be expressed. Interactions between long non-coding RNA (lncRNA) and CTCF as well as mutations and epimutations can result in imprinting-based pathologies like Russell-Silver Syndrome (RSS) [23, 122, 111].

In this work, we describe how the genome is condensed and organized at different scales starting from the nucleosome, followed by the DNA loops and, finally, the topologically associated domains (TADs). We discuss how CTCF is able to bind to a large number of DNA sequences and how the recently solved structures of F-DNA complexes have confirmed previous discoveries. We discuss the underlying experiments that explain how the chromatin domains might form and a theoretical model that is consistent with the *in vivo* observations. We also highlight the role of CTCF in human diseases like cancer and emphasize how misregulation of genomic imprinting can result in imprinting-based pathologies.

10-nm but no 30-nm fiber is observed *in vivo*

Years of studies on the DNA structure suggested a hierarchical view of its architecture. DNA wrapped around histones forms a nucleosome fiber (called a 10-nm fiber) that would next fold into the 30-nm chromatin fiber [20]. According to the hierarchical helical folding model this 30-nm fiber is next folded into larger 100–120-nm, 200-nm or even 700-nm fibers that finally form large interphase chromosomes [1, 21, 22]. The transition between the 10-nm to 30-nm fiber was believed to be essential for establishing the accessibility of DNA for the transcriptional machinery. In the 1970s, based on the results of X-ray crystallography, it was proposed that the 30-nm fiber forms a helical structure (called a solenoid or “one-start” helix) with each nucleosome wrapping around a central cavity with close to six nucleosomes per turn [23]. Nearly a decade later the “zig-zag” or “two-start helix” model was proposed based on microscopic observations of isolated nucleosomes [130]. The exact structure of the 30 nm fiber remained elusive; as changing the length of linker DNA, ion concentration, or the presence or absence of linker histones, would significantly alter the obtained results [20, 25, 26, 24]. Thus, the hierarchical helical folding model has been challenged on multiple occasions. Already in the ‘80s, using cryo-electron microscopy (cryo-EM), it was observed that the mitotic chromosomes had a homogeneous texture with a spacing of around 11-nm. No higher-order periodic structures, including the 30-nm fibers, were observed [27]. Similar conclusions were drawn from the analysis of the interphase chromosomes [28]. Results of an experiment using small-angle X-ray scattering (SAXS), which indeed showed the existence of the 30-nm fiber, might be attributed to the presence of ribosomes bound to the chromatin

fiber. After the ribosomes were washed out, the 30 nm peak was no longer observed [29, 30]. Recently Maeshima and colleagues [31] proposed an alternative model of chromosome organization. Their results suggest that in vivo the array of nucleosomes has the ability to spontaneously self-assemble through the interdigitated packaging of 10-nm fibers into globular structures with diameters of ~50–1,000 nm. The analysis of linear 12-mer 601 nucleosome arrays showed that at a low salt concentration no pelleting was observed, but it gradually appeared with the increasing concentration of MgCl₂. The 30-nm fiber was observed only under very specific salt concentration; i.e., 2.5 mM MgCl₂ and in the presence of linker histones. Also, in situ experiments using ChromEM tomography confirmed a lack of the 30-nm fiber and it was proposed that chromatin was organized into disordered 5- to 24-nm-diameter curvilinear chains [22].

How the nucleosomes interact with each other in these globular structures remains unclear; however, recently a cryo-EM study of nucleosome core particles (NCPs) assembled on a 149 bp long 601 DNA sequence addressed this issue [32]. The authors observed several possible orientations of the two NCPs that were in spatial proximity, but not directly linked by their DNA. For the so-called “class A”, which encompasses most of the observed states, the density of the NCP2 is parallel to the histone octamer core of NCP1. For class B, the position of the NCP2 is laterally shifted and tilted. In most cases, the long tail of histone H3 is in contact with the second NCP facilitating the primary interaction site. After both NCPs are connected, additional contacts are formed between the histone H4 tail and the acidic patch of DNA. These results are consistent with the proposed hypothesis that no stable 30-nm-like structures are formed; rather the globular structure, observed by Maeshima and colleagues for an array of nucleosomes, is a less defined, polymer-like structure.

The 3C methods as a source of data beyond 10-nm fiber

A growing knowledge of chromatin structures beyond the 10-nm fiber and the role of proteins in their establishment is partially due to the development of the chromatin conformation capture (3C) technique [33], and similar methods like 4C [34], 5C [35], Hi-C [4], ChIA-PET [36], etc. Although, microscopy methods like fluorescence in situ hybridization (FISH) or confocal microscopy are also widely used, they are limited to a resolution on the order of 100 nm and are restricted to analyzing only a handful of loci simultaneously [37]. All 3C methods share similar initial steps [38, 39]. First, the chromatin is fixed using a fixative agent like formaldehyde, which “freezes” the points of mutual contact between diverse regions of the chromatin polymer. Next DNA is cut into fragments using either restriction enzymes like HindIII, BamHI and other [37] or sonication [40], which is part of the ChIA-PET protocol. The later technique, according to the authors of this method, should increase the specificity of the obtained contacts by removing accidental products of self-ligation and ambiguous contacts. The following step involves ligation of the cross-linked fragment under conditions that promote intra-ligation producing circular fragments of chimeric DNA containing sequences close in 3D space. Finally, by reverse crosslinking, a linear fragment is obtained. The following downstream analyses are method-specific and were extensively covered in other reviews [37–39, 41]. In general, 3C methods can be divided into ones that focus on analyzing a pair of loci (“one-to-one” like 3C); analyzing contacts between a single locus and many other loci (“one-to-many”, like 4C), and

between many loci simultaneously (“all-to-all” like Hi-C or ChIA-PET). ChIA-PET is a unique method where, instead of analyzing all contacts observed between parts of the chromatin fiber, only the interactions mediated by specific protein factors like CTCF or RNA polymerase II (RNAPII) are captured. This can be achieved by using specific antibodies that pull down only the DNA fragments of interest while discarding the rest.

Several issues arise when analyzing 3C data. Because a sufficiently high coverage is rarely achieved from the sequencing procedure, the reads are usually binned together in some cases using a simple summation. For Hi-C data the often-used term ‘map resolution’, like 1 or 10 kbp, reflects the minimum scale at which the local contacts are discernable and assigned to specific regions of the genome [42]. Another issue is the likelihood of observing the so-called self-ligation products. The chromatin fiber might be considered a long, continuous, flexible cord that is stiff in bp scale but at the scale of several kbp capable of significant twisting and bending. It is easy to imagine that some parts of this cord would be in contact with each other simply by chance and the accidental origin of that observation would have no biological meaning. For example, for the CTCF ChIA-PET experiments, instead of observing a contact mediated by a pair of CTCFs, a contact between a single CTCF and a DNA fragment located in close spatial proximity could be captured. To solve this issue, contacts that join fragments that are less than 8 kbp apart that have a different distribution of frequencies, were discarded as mostly products of this self-ligation [43]. Another major issue is that 3C methods typically require the analysis of reads from millions of cells all fixed at the same time. However, chromatin is quite dynamic and with bulk Hi-C it is difficult to say if any particular observed set of contacts are present in all the cells or represent the state in some sub-populations. The bulk Hi-C might thus represent an “average state” that does not correspond to any specific *in vitro* (or *in vivo*) situation. Only recently, studies designed to capture the chromatin structure of a single cell have been conducted using modified Hi-C protocol and they might provide better understanding of the differences between the chromatin structure observed for a population or particular cells in specific states or conditions [5, 131].

The genome is hierarchically partitioned into chromosomes, topologically associated domains and DNA loops.

The FISH and other microscopy experiments allowed us to understand that chromosomes are not randomly arranged within the nucleus, but rather occupy well-defined chromosomal territories [44–47]. The larger chromosomes tend to be localized closer to the nuclear envelope while smaller ones occupy the nucleus center. The DNA fragments in direct contact with the nuclear lamina form the so-called lamina-associated domains (LADs). The LADs are heterochromatin-rich, their median size is about 0.5 Mbp and more than 1000 such LADs are found in most mammals. The position within the nucleus can influence gene expression as transcription would typically happen in regions rich in euchromatin that are located toward the center of the nucleus [48, 49]

Initial Hi-C data obtained for the human genome [4] at a megabase resolution show that chromatin within a chromosome is hierarchically partitioned into large “megadomains” of size ranging from 5 to 20 Mbp. Megadomains represent the regions of DNA with an

enhanced number of contacts within and a reduced number of intradomain contacts. Individual 1Mb loci can be also classified, based on the contact patterns at this length scale, into two compartments named A and B. The A compartment bears the marks of euchromatin as it is usually found in regions that are genes-rich and carries characteristic histone modifications like H3K36me3. The B compartment might represent heterochromatin (densely packed regions of chromatin) because at a given genomic distance it showed a higher interaction frequency than pairs of loci in compartment A. The data also showed that the number of contacts observed within chromosomes is much larger than that observed between chromosomes, which is in agreement with the notion of chromosomal territories.

Using 5C and FISH, which yielded data of higher resolution, for the mouse X-inactivation center Nora and colleagues showed that this fragment of the genome can be partitioned into smaller structures, called topologically associated domains (TADs) [50]. The TADs had lengths ranging between 0.2 to 1 Mb and their formation was found to be independent of the underlying histone modifications. *In situ* Hi-C analysis of human lymphoblastoid cells at 1kb resolution permitted a deeper insight into the nature of TADs [42]. In the human genome, TAD sizes vary between 40kb to 3Mb (median size 185kb). Based on the long-range interaction patterns, these domains can be classified into at least six sub-compartments – A1, A2, B1, B2, B3, B4 – each representing a different chromatin “state” (Figure 1A). Rao and colleagues also noted that certain TAD loci at the edge of domains have a significantly higher number of contacts than loci within domains (Figure 1B). These regions of increased number of contacts were called peak loci and at 5kb resolution 9,448 peaks associated with 12,903 distinct peak loci were identified. Peak loci were assumed to form the base of the loops between distant part of the DNA and, therefore, domains with peak loci were called loop domains. Loop domains have some very interesting characteristics. First, on average they join fragments that are less than 2 Mb apart, some of them are conserved between tissues [51] and to a smaller extent even between species [52]. These phenomena may be associated with the putative role of the loops, as more than 30% of loop domains bring into close spatial proximity known promoters and enhancers. The presence or absence of peak loci in data from several cell lines was also associated with changes in gene expressions. For example, the promoter region of the ADAMTS1 gene that is expressed in the IMR90 cell line, is involved in 6 loops. On the other hand, in the GM12878 cell line, this gene is not expressed and its promoter is involved in the formation of only 2 loops. Another interesting fact is that the peak loci correspond to known protein factors; e.g., 86% of these peak loci, overlapped CTCF ChIP-seq peaks and, to the same extent, subunits of the cohesin complex, specifically SMC3 and RAD21. Furthermore, 54% of peak loci were associated with the CTCF binding motif.

ChIA-PET experiments on several human cell lines provide a complementary picture of chromatin structure (Figure 2), despite the fact that, for this method, only contacts mediated by specific protein factors like CTCF or RNA polymerase II were analyzed [43]. With CTCF ChIA-PET for the GM1278 cell line, more than 50k unique interactions could be observed that were used to establish that the genome is partitioned into 2,267 CTCF-mediated chromatin contact domains (CCDs). Anchors of the CTCF-mediate loops usually colocalize with subunits of the cohesin complex, which is reminiscent of the situation observed for Hi-C loop domains. Among the 42,297 interactions for which cohesin subunits were found to be

present in both anchors, 83% had the CTCF-binding motif. How the remaining 17% of the loops were formed is unknown; however, it should be noted that CTCF can bind to a diverse number of sequences, some of which significantly differ from the proverbial CTCF-binding motif. Along with the binding of CTCF, other proteins like RNAPII complex also appears to shape the human genome. For example, the DNA loops mediated by RNAPII are shorter than the CTCF-mediated loops and most of them lie within CCDs defined via CTCF ChIA-PET [43]. Genes can be divided into those that are proximal to CTCF anchors and those that are located within the loop region, based on the given gene location relative to the CTCF. The expression of “anchor” genes was significantly less tissue-specific than “loop” genes. Anchor genes were found to be almost exclusively housekeeping genes further highlighting the importance of loop formation.

Multiple studies highlight the fact that contact domains tend to be relatively stable when analyzing cells in bulk. The recent single-cell Hi-C study shed new light on the genome organization of individual cells [5]. It was found that the 3D structure of the genome of individual oocytes cells can differ in both the number of observed TADs and their respective boundaries when compared with observations reported from *in situ* Hi-C data. Thus, TADs observed in bulk Hi-C and ChIA-PET appear to reflect the general tendency for contact enrichment arising from a diverse conformational ensemble, rather than being isolated blocks of DNA present in individual cells. Nevertheless, Figure 2 shows persuasively that the two methods generate comparable, correlated, and complementary results.

The number of TADs is a matter of debate. Initially, Dixon et al. reported 2200 in mice, while Rao and colleagues identified between 4000 and 9000 domains depending on the analyzed cell line. More recent studies of mouse embryonic stem cells showed around 3400 domains [42, 51, 53, 54]. In addition to discrepancies between studies of various kinds, another source of variation is the algorithms used to analyze the data that can influence the number of reported TADs [55].

Nevertheless, from the above mentioned works, a hierarchical chromatin structure emerges with the DNA loops forming TADs and finally compartments and chromosomal territories. DNA loops, beside a purely structural role, have an additional purpose as they would bring in close spatial proximity-specific DNA elements and at the same time “shield” them from regulatory elements localized outside the loops [56]. Equilibrium simulations of a confined polymer chain confirmed that elements on the loops are less frequently in contact with elements located outside the loops [57]. Based on the available data, several types of DNA loops can be named; each having a particular role and a different level of conservation among cells and organism. The first group, cohesin CTCF-mediate loops found at the borders of TADs, are the most stable and conserved between species. A second group constitutes cohesin-associated enhancer-promoter loops that might originate from small loops formed when the mediator complex joins the transcription factors bound to enhancers and promoters [58,59]. These small loops may be transiently stabilized by cohesin as the Nipped-B-like protein (NipbL), a mediator-associated protein, loads the cohesin complex at active enhancers and promoters [17]. Other proteins involved in stabilization of these loops include, as already mentioned, CTCF and YY1, a ubiquitously expressed protein. It was recently shown that YY1-YY1 interactions occur predominantly

between enhancers and promoters and to a lesser extent between insulators. It was also shown that depletion of YY1 influences the expression level of more than 8200 genes [10]. Since YY1 binds to a poorly conserved sequence motif, this may partially explain why enhancer-promoter loops are usually cell-specific. Thus, enhancer-promoter loops, contrary to loops located at the TAD boundaries, are shorter and most likely cell-specific [17, 60]. A different subset of loops, identified in yeast, can be involved in bridging a gene promoter and its transcription termination sites, thus enforcing transcriptional directionality [61]. Another group are polycomb-mediate loops identified for example in *Drosophila* [62]. Strikingly, for this organism, only 120 chromatin loops were detected – a number significantly lower than that in the human genome or even the number of TADs observed for *Drosophila*. Unexpectedly, CTCF does not seem to play a role in the formation of these loops as only 28% of loops overlap the CTCF-binding sites; however, cohesin is still involved in these loop-formations because 72% of loop anchors overlap RAD21 ChIP-seq peaks. Instead, the loop anchors were found to be associated with the polycomb repressive complex 1 (PRC1) and 32.7% of these polycomb loops were found at promoters of important developmentally regulated genes. Interestingly, promoters associated with these loops were found to be less likely expressed than promoters associated with other types of loops. The polycomb loops are probably not restricted to *Drosophila* as, for example, PRC1 was found to be involved in organizing genome architecture for mouse embryonic stem cells [63].

CTCF a chief organizer of the genome

All 3C experiments highlight the role of CTCF in the establishment of 3D structure of the human genome. CTCF is a multidomain protein present in most bilaterian phyla [65] and displays a high conservation among all vertebrate with over 95% amino acid sequence identity within its DNA binding domains [66]. Several experiments proved the protein's importance because CTCF-knockouts lead to embryonic lethality in mice [9] and, within the thymus, CTCF depletion deregulated cell-cycle progression during T lymphocyte lineage commitment [67]. The protein consists of 11 zinc finger domains (named ZF1 to ZF11) and long, disordered N- and C-termini. The zinc fingers domains are responsible for interactions with DNA while the N-termini might have an additional function like promoting dimerization of the protein [68]. ChIP-seq data identify that CTCF binds to around 50 000 DNA sequences; the majority of which have a central 15–20 bp core motif 5'-NCA-NNA-G(G/A)N-GGC-(G/A)(C/G)(T/C)-3' (also called the M1 motif) [69, 70]. For other analyzed species, some of which are evolutionary distant from humans, a similar number of peaks ranging in tens of thousands has been established. Interestingly, around 30% of these sites are shared between different organisms. The high number of DNA sites recognized by CTCF is associated with retrotransposon repeat element expansions [71].

CTCF chip-seq suggest that the protein has more than one sequence motif

An interesting feature of CTCF is the ability to recognize long and remarkably divergent DNA sequences. To analyze how CTCF interacts with DNA, Nakahashi et al [69] obtained the binding profile of mutated CTCF proteins where, for each zinc finger domain, one of the histidines responsible for binding the zinc ion was mutated to glycine. This led to disruption in the formation of the corresponding zinc finger domain, diminishing the interaction between this part of CTCF and DNA. Interestingly regardless of which zinc finger domain

was mutated, CTCF could still bind to DNA; albeit the number of recognized sites was substantially smaller than for the wild type protein. Additionally, there was a substantial variation in number of peaks between the mutants. The protein with mutations targeting ZF4–7 domains had the lowest number of peaks whilst mutations of the peripheral zinc finger domains had a lesser effect. Changes in the CTCF sequence could be directly attributed with binding to DNA based on the residence time of the mutated proteins obtained with Fluorescence Recovery After Photobleaching (FRAP). The complete recovery time dropped from 11 min (80% recovery close to 20s) observed for the native CTCF to only 15s for CTCF that carried a mutation targeting ZF6. The more recent estimates suggest even longer residence time of the wild type CTCF with 80s recovery after 1–2 min. The difference between the two studies could be attributed to overexpression of CTCF in the work from 2013 [72]. Nevertheless, both estimates agree that CTCF is bound to exposed DNA longer than the average transcription factor, but substantially shorter than it is observed for the histones or cohesin complex.

Analysis of the binding profiles of the mutated CTCF showed that the ZF domains do not act individually, but rather cooperatively. Nakahashi and colleagues identified 3 groups, ZF4/5/6/7, ZF1/2 and ZF8/9/10/11; each recognizing specific DNA sequences. The first group, due to the profound impact on the CTCF binding, would most likely recognize a core motif. ZF 8–11 would interact with an additional 10 bp-long upstream motif (also called M2) that has been associated with the increased residence time of CTCF. This motif is present only on a subset of binding sites (around 13%) but was most likely present even in the last common ancestor of all vertebrates and was recently identified in jawless fish [73]. The M1 and M2 motifs are separated by a linker of varying length of 5 to 6 bp. Additionally, a downstream motif, also 10 bp-long, has been localized 6–8 bps from the core motif. It was identified in 8% of the binding sites and was either the only motif present or was associated with the core motif or together with the core and upstream motifs. The role of these additional motifs seem to be regulation of the occupancy of the CTCF. It was found that the downstream motif decreases while the upstream motif increases the CTCF residence time [69].

Conserved nucleotides in CTCF binding motif form stable interactions with CTCF

All zinc finger domains of CTCF have a similar structure constituting a single helix and an antiparallel beta hairpin. The structure of this domain in complex with DNA has been known for more than 25 years; however, how multiple zinc finger proteins interact with DNA remains only partially understood [68, 74]. Knowledge regarding the interactions between CTCF and DNA was mostly a result of sequence analysis and mutation studies and the question how a single protein could recognize such a diverse set of DNA sequences remains largely unanswered. Another unanswered question was why methylation of the core motif involving cytosine at position 2 but not 12 resulted in decreased binding of CTCF to DNA. Only recently, the structures of multiple zinc finger domains from CTCF were solved, bound to either DNA containing the CORE sequence (5–CCA GCA GGG GGC GCT-3) [75] or the sequence of the core motif from the CTCF-binding site (CBS) in the HS5-1a enhancer of the *Pcdh_l* cluster [68].

When combining the separately obtained structures of ZF2–7 and ZF4–9, it can be clearly seen that CTCF binds to the major groove of DNA with most interactions involving DNA from one of the two strands (Figure 3B). As expected from ChIP-seq analysis, the ZF3–7 segments interact with the CORE sequence and each ZF domain interacts with a triplet of nucleotides. The N-terminal ZFs interact with the downstream 3'-end of the DNA sequence, while the C-terminal domains interact with the upstream 5'-part of DNA. Interestingly, there is strong agreement between nucleotide conservation in the motif and the interactions visible for the CTCF-DNA complex. The 4th triplet (GGC), which constitutes nucleotides that are highly conserved in the CTCF binding motif, form multiple hydrogen bonds with the ZF4 domain via R368, K365 and E362. The polar contacts between guanine and arginine are a common mechanism for DNA recognition by protein factors [76, 77]. Other conserved nucleotides A3, A6, G8 interact with R448, Q418, R396 and K393 respectively. For nucleotides that are not conserved in the CTCF binding motif, either the closest amino acid capable of forming a polar interaction is too far away (like in case of C14 and T15) or the interactions are mediated by water; a situation observed for example for C2 or G9. The second available structure of CTCF, for which a structure containing the ZF3–11 segments was observed, was solved together with the CTCF sequence from the HS5-1a enhancer of the *Pcdh α* cluster and it provides complementary results (a summary of contacts is provided in Fig. 3A). The contacts between the DNA fragment and ZF4–7 are virtually identical to those observed in the complex with the CORE sequence. This suggests that CTCF adopts a very similar conformation irrespective of the DNA fragment it binds to.

In the core motif, cytosines at position 2 and 12 are usually followed by guanine or adenine forming a CpG or CpA dinucleotide; typical substrates for methylation enzymes [7, 78]. In a work by Wang et al., 41% of the variability in the CTCF occupancy in 13 cell types was linked to methylation of the core motif. However, only methylation of C2 has a substantial impact on the CTCF binding, while methylation of C12 remains mostly neutral [70, 75]. The structural analysis revealed that C2 interacts with D451 (two polar contacts, distance 2.6 and 2.2 Å) from ZF7. An additional methyl group attached to C3 would sterically obstruct this residue explaining why methylation of this nucleotide prevents CTCF binding. The C12 forms a single polar contact with E362 and a methyl group could easily be placed between the C γ -atom of E36 and the side chain of Y343; thus, no adjustment of the ZF4 position is required upon C12 methylation.

The ChIP-seq data suggest the existence of an additional motif located upstream of the core sequence. To address the question how the CTCF structure is modified upon binding to both the core and upstream motifs, a structure of CTCF containing the ZF6–11 segments was solved together with chimeric DNA containing the core motif from the *Pcdh α* HS5–1a enhancer and the upstream motif from the *Pcdh γ 7* promoter (5'-TTGCAGTAC-3'). Only the ZF9–11 recognize the upstream motif, but contrary to recognition of the core motif by ZF3–7, each domain interacts with at most 2 nucleotides. The most conserved in the upstream motif G3 interacts with R566 of ZF11, while interactions with G6 and G9' (from the complementary strand) are mediated by Q534 of ZF10 and R508 of ZF9. The ZF8 serves as a linker bridging the domains interacting with the M1 and M2 motifs although the presence of ZF8 increases nonspecific binding of a control sequence [75]. When no

upstream motif was present, the electron densities for the zinc finger domains after ZF8 could not be clearly seen, suggesting an auxiliary role of this part of a protein.

Convergent loops are the most common structure

The striking observation from the 3C data is that the orientation of the CTCF motif (found at the peak loci or anchors of the CTCF ChIA-PET) is not completely random. In 92% of the peak-loci and for 64.5 % of the ChIA -PET anchors, the corresponding core motifs have a convergent (i.e., facing each other) orientation [42, 43]. In the remaining regions, a tandem orientation of the CTCF motifs was identified and only in isolated cases was a reverse-orientated motif found. Since the motif orientation determines how the CTCF interacts with DNA, only two CTCFs in a very specific orientation would be found at the base of a given loop. This observation also inclines one to conclude that the loop formation cannot be explained so well by an equilibrium model, since the entropic cost of bending DNA over a length scale of hundreds of kilobases should be small and the initial orientation of motifs should not play any factor in this process [79].

The paramount importance of the convergent orientation of CTCF binding motifs was explored in several analyses that disrupted this pattern in well-studied loci. In a work by Gao et al. [15], the authors analyzed the protocadherin gene (*Pcdh*) clusters. The region contains two clusters of the gene called *Pcdh α* and *Pcdh $\beta\gamma$* , each having a similar structure. For the *Pcdh α* cluster, 12 “alternately expressed” exons, which correspond to the extracellular part of the protocadherin protein, are located at the 5' end of the cluster. A unique promoter is localized immediately upstream of each exon. Downstream of the “alternately expressed” exons, two “ubiquitously expressed” exons are localized, encoding the transmembrane part of the protocadherin proteins. These exons are followed by a regulatory region with two enhancers (named HS7 and HS5-1). The *Pcdh $\beta\gamma$* clusters has a similar structure with two instead of one group of “alternately expressed” exons (β 1-22, γ a1-12 and γ b1-8) followed by 3 “ubiquitously expressed” exons (γ c3-5) and a regulatory region with multiple enhancers. Most promoters and enhancers in a given cluster have the CTCF binding motif and these motifs are in a convergent orientation. Previous experiments showed substantial looping occurring within these regions that are essential for transcription [80].

To test if the convergent orientation of motifs might have a biological implication. The authors used the CRISPR/Cas9 technique to reverse the orientation of the CTCF motif localized in the HS5-1 enhancer of the *Pcdh α* cluster. The 4C experiment showed that upon this event, the number of contacts with exons from the *Pcdh α* cluster, previously bridged by loops formed by the convergently oriented CTCF motifs, dropped from 72% to 21% and, simultaneously, more contacts were observed between enhancers from the *Pcdh α* cluster and promoters of alternative exons from the *Pcdh β* cluster. Upon editing with the CRISPR/Cas9 system, these originally divergently oriented CTCF motifs assumed the tandem orientation. At the same time, the ChIP-qPCR analysis of CTCF showed no substantial change in its binding pattern, while the binding of cohesin was significantly reduced. The RNA-seq analysis showed that alternatively expressed exons in the *Pcdh α* cluster were significantly under-expressed upon reorientation of the motif near HS5-1 enhancer. Interestingly, despite the new loops being formed between the enhancer from the *Pcdh α* cluster and promoters of

the alternative expressed exons from the *Pcdh β* cluster, the expression of these exons was also disrupted (Fig 4B). It is important to note that gene expression is, however, not universally affected by changes in chromatin 3D structure. In an analysis of 3 loci *Malt1*, *Sox2* and *Fbn2*, in which the CTCF binding motif was either removed or reversed, although the pattern of chromatin contacts was altered, as confirmed by 4C, this change in the chromatin structure had only a limited effect on gene expression. For the *Malt1* and *Sox2* loci, the modification of the CTCF binding site did not change the expression levels and in case of the *Fbn2* locus the *Fbn* gene expression was increased [16]. For the mouse α globin cluster changes in the expression of genes neighboring were observed after the CTCF binding site located at the upstream edge of this cluster was removed. At the edge of this cluster, 3 CTCF-cohesin binding sites were identified and one, called HS-38, carries the core sequence in a convergent orientation with respect to the upstream-located CTCF core sequences. After HS-38 and a nearby HS-39 sites were removed in erythroid cells, the enhancers located inside the α -globin cluster interacted with the upstream region of the chromatin containing 3 genes. This could be interpreted as expanding the α globin sub-TAD with these upstream elements. The RNA-seq analysis also confirmed the overexpression of these newly incorporated genes, in case of the *Rhbdf1* gene 600-fold higher than in wild-type cells [18].

CTCF and cohesin shape the genome 3D structure

The role of CTCF as an insulator blocking the contacts between enhancers and promoters and thus influencing gene expression has been known for years [81–83]. This role was linked to methylation of specific DNA regions that either promoted or hindered CTCF binding to DNA. Consequently, the expression pattern of nearby genes was also altered. This situation was observed for example at the *H19/IGF2* locus. Subsequent studies revealed that CTCF is accompanied by *SCC1* [84], a protein that is a part of the cohesin complex [14], and that both proteins are involved in establishing insulation. The role of CTCF and cohesin as an insulator was for example tested with the luciferase enzyme [84]. First a gene encoding the luciferase enzyme was inserted between two Imprinting Control Regions (ICRs) from the *H19/IGF2* locus, each carrying the CTCF binding site. An enhancer was placed downstream of the region flanked by ICRs. By depletion of either CTCF or *SCC1* protein, a significant increase in luciferase levels were observed, most likely due to the lack of insulation between the enhancer and promoter. Thus, it was concluded that CTCF and cohesin exert their effects cooperatively. The insulation can be a consequence of loop formation as equilibrium simulations of a confined polymer chain suggest that elements on the loops are less frequently in contact with elements located outside it [57] and in neighboring loops. Methylation of the CTCF binding motif would disrupt the formation of the enhancer-promoter loops and, as a result, alter gene expression.

The role of cohesin in loop formation was highlighted by several studies as it colocalized with CTCF at the borders of the loops and at the loop anchors [42, 43]. The cohesin is a complex of several proteins responsible for establishing cohesion of sister chromatids (part of the name) during interphase; however, it persists on the chromatin also during other phases of cell cycle. This complex is built of several proteins, with the core protein – *SCC1/RAD21*, *SMC1* and *SMC3* – forming an unusual ring that encompasses the dsDNA. Other

proteins like STAG1, STAG2, PDS5A or PDS5B serve auxiliary roles; e.g., as a docking platform for other proteins [14]. It was shown that the cohesin complex can shape the chromatin 3D structure without changes in the CTCF distribution. For example, Tedeschi et al. [85] observed that the depletion of cells of the cohesin release factor (WAPL), a protein involved in cohesin dissociation from DNA, resulted in a 20 fold increase of residence time of GFP-tagged cohesin from around 25 to 540 min. Moreover, cohesin became enriched in axial structures, which the authors referred to as vermicelli. substantial effect on chromatin structure was also observed as the chromosomes appeared to be more condensed despite no changes in the distribution of heterochromatin or H3 K9 trimethylation (H3K9me3). All these effects could be partially reduced if the Scc1 subunit was depleted with RNAi (RNA interference), preventing the formation of a stable cohesin ring in the first place. Interestingly, the vermicelli structures were found to be enriched with CTCF. Experiments suggest that depletion of CTCF and cohesin affect the chromatin structure in a different manner [86]. The 4C analysis of the H19/IGF2 locus showed that opening the cohesin ring by cleavage of Rad21 results in a global loss of interactions within this locus. In a subsequent Hi- experiment, Zuin and colleagues confirmed this local effect on a global scale with interactions occurring at a distance of 100 to 200 kb being the most affected. However, the removal had a limited effect on the observed patterns of TADs, which did not change their positions. The reduction of CTCF levels disrupted mostly short-range interaction that occurred at a distance shorter than 100kb. Interestingly, at the same time, an increase in contacts between TADs was observed; suggesting that CTCF is important for maintaining TAD boundaries. A more recent study involving a conditional degradation strategy in mouse embryonic stem cells suggests a far more profound change to chromatin structure upon a more complete CTCF depletion [87]. A loss of insulation was observed for more than 80% of the boundaries between the TADs; however, no changes in the segregation of active and inactive chromosome domains into A and B compartments were observed. This suggests a CTCF-independent mechanism behind the establishment of the compartments. Interestingly, the effects of the CTCF depletion are “dose-dependent”, which may explain why residual levels of CTCF, of around 10–15% after treatment with RNAi, showed less-impact on the TADs boundaries in the work by Zuin and colleagues.

The cohesin position on DNA can change over time

Analyses of cohesin ChIP-seq data show that there is a discrepancy between the cohesin position and the localization of cohesin loading sites that suggest that the cohesin position can change over time [88]. Two possible mechanism of cohesin translocation can be devised; passive diffusion or transcription mediated translocation. Recent experiments show that cohesin can indeed move passively along DNA in yeast; however, evidence of active translocation was found as well [13, 89]. An experiment on mouse embryonic fibroblasts [13] suggests that the transcription machinery might be responsible for cohesion translocation. In the article [13], the authors observed that upon CTCF depletion, cohesin localization changes dramatically and accumulation in vermicelli, a situation reminiscent of what was observed in the Wapl-depleted cells. Beside vermicelli, the cohesin complex was also positioned downstream to actively transcribed genes and near transcription starting sites (TSS). This last localization might, however, only reflect the region where cohesin is normally loaded onto chromatin, as indicated by ChIP-seq on the NIPBL protein, a part of

the cohesin loading complex. A double Wapl-ctcf knockout results in the formation of ‘cohesin islands’ which are around 70kb in length and are predominantly located downstream of actively transcribed genes (Figure 5). Busslinger and colleagues also noticed a clear correlation between the shape of the cohesin islands and the level of transcription. For convergently transcribed genes that have similar gene expression level, the density of cohesin in the cohesin islands is symmetrical. Distortions in this symmetrical shape of the cohesin distribution can be directly correlated with the expression levels of genes located next to a given cohesin island. Thus, it seems that cohesin is translocated actively by the transcription machinery. Other experiments confirm that fact as cohesin was also translocated by other DNA processing enzymes like in case of budding yeast transfected with the T7 bacteriophage RNA polymerase [89]. One might speculate that cohesin is translocated until a region with a CTCF (or CTCF dimer) bound to DNA is found.

Why CTCF serves as a main boundary for cohesin translocation is not well understood. The inner diameter of the cohesin ring is roughly ~25–50 nm [90, 91] and experiments show that cohesin was able to pass by the nucleosomes (diameter ~11nm) but proteins with diameter ~21nm provide a sufficient boundary. The CTCF diameter is probably smaller and thus should not be considered a sufficient obstacle; however, the specific interactions that are formed between cohesin and CTCF might be sufficient to block cohesin passage. A different possibility is that, CTCF might *in vivo* dimerize or even multimerize thus creating a complex that would serve as a mere steric obstacle for the cohesin ring [92, 93]. Of note, the CTCF dimerization was not observed during crystallography, which might be partially accounted for by the lack of the N- and C-terminal parts of the protein. A recent study also showed that upon embracing DNA, cohesin adopts a “rod-shaped”, closed conformation with a smaller ring diameter [94].

Cohesin is not the only protein that traverse along DNA, suggesting that this mechanism is evolutionary ancient and can serve other functions. In *Bacillus subtilis* SMC-condensin complex, which structurally closely resembles cohesin, was shown to slide along DNA with a rate of 50 kbp per minute [95]. Because condensin movement along both DNA tracks was not constant, the authors suggested that the ‘handcuff’ model is a likely explanation of this behavior. This model assumes that instead of a single ring-like structure that traverse DNA, two ring-like molecules embraced together are loaded near the loading site (parS in the case of *Bacillus subtilis*). Next, each ring would traverse in the corresponding direction independently while still being connected to the other ring. Whether the same situation occurs for cohesin is not known. Nevertheless, considering the structural similarity between condensin and cohesion, it seems likely. Initially it was assumed that the complex formed between CTCF and cohesin is rather stable; however, recent experiments suggest it to be a rather dynamic entity [72].

Loops extrusion model

How the loops are initiated and established is an ongoing issue with several proposed models. In one model, MacPherson and colleagues showed that CTCF alone is capable of bending DNA to form a “loop-like” structures” [64]. Another option is the “hijacking” of small loops established during transcription initiation by the mediator complex that bridges

transcription factors bound to enhancers and promoters. This small loop can be next stabilized and enlarge by cohesin which is recruited to DNA by the mediator-associated protein NIPBL. Several theoretical models have been developed to explain the abovementioned experimental results. The one that is currently most widely accepted is the loop extrusion model (LE). The model was introduced with the notion that it behaves conceptually like the long-distance intracellular delivery systems such as the kinesin and dynein motor proteins that transport their cargoes along microtubule tracks [96, 97]. In the LE model, originally proposed by Marko and coworkers [98] and further developed by several other groups [99, 100], the chromatin is represented as a polymer consisting of beads separated by some specific distance; e.g., 1 kbp, 5 kbp, etc. loop extruding factor (LEF) initially binds to a region on the chromatin comprising neighboring beads or, possibly, a small loop formed by a few beads. Although the role of LEF was not assigned to any particular protein, it is now often assumed that this role is realized by cohesin [56]. Recent work also suggests that the cohesin loading complex localizes in transcription initiation sites, thus it is plausible that the extrusion process begins in the vicinity of actively transcribed genes [13]. The LEF then begins to traverse along the chromatin and the process stops when the cohesin reaches the insulator (also called a boundary element) or another cohesion ring traversing in the opposite direction. Considering the mounting evidence, CTCF is now widely accepted as the boundary element.

CTCF involvement in diseases

Undoubtedly the changes to chromatin structure due to various alterations in the CTCF-mediated loop formation can lead to various pathologies. For example, polymorphisms modifying the CTCF binding motif can alter the expression of human leukocyte antigens (HLA) and increase susceptibility to autoimmune diseases such as vitiligo [101], multiple sclerosis [102] and systemic lupus erythematosus [103]. Another example, SNP rs34481144, was found to influence the severity of an influenza virus infection [104]. This single nucleotide polymorphism targets the CTCF binding motif in Chr11 resulting in the substitution of C with T and is associated with an increased binding of CTCF to the promoter of Interferon Induced Transmembrane Protein 3, which leads to decreased expression of this gene [104]. Epigenetic changes that regulate CTCF occupancy and interactions with the cohesin complex also influence the expression of the *HTT* and *CFTR* genes related to well-known genetic disorders such as Huntington disease or cystic fibrosis, respectively [105, 106]. The mutations in the CTCF sequence, resulting in truncated protein, have proven to play a role in the progression of head and neck cancer [107], while changes to the CTCF binding motif are observed in colorectal cancer [108]. The CTCF mediated high-order spatial chromatin organization may contribute to the somatic co-mutations of certain cancer genes [109]. CTCF upregulation has been shown to decrease disease-free survival of patients affected with hepatocellular carcinoma by enhancing the expression of forkhead box protein M1 and telomerase reverse transcriptase [110].

CTCF was also found to bind to insulators located in the intergenic region between the gene translin-associated factor X (TSNAX) and the downstream gene disrupted-in-schizophrenia 1 (DISC1). A long intergenic non-coding RNA (lncRNA-NR_034037) that is also located in between these two genes, but on the opposite strand, can extract CTCF from the insulators

and promotes the expression of a chimeric transcript TSNAX–DISC1 in endometrial carcinoma [111]. The aggressiveness and progression of neuroblastoma was recently linked to interactions between CTCF and a long non-coding RNA (lncRNA). RNA pull-down and *in vitro* binding assay indicate that lncRNA called MYCNOS, which is located in the surrounding of the MYCN promoter region, induces binding of CTCF to this promoter. This results in chromatin remodeling and enhanced MYCN levels [112].

It has been reported that CF during lytic infection interacts with several sites in the *Herpes Simplex Virus 1* sequence and its knockdown resulted in a lowered number of viral genome copies and a decreased transcription of viral proteins. A postulated mechanism is that the CTCF knockdown promotes histones methylation (H3K27me3, H3K9me3) leading to chromatin condensation [113]. Satou et al. [114] have shown that CTCF binds to the *Human T-lymphotropic virus type 1* (HTLV-1) forming loops between the provirus sequence and the host genome at a border of epigenetic modifications in the pX region. CTCF bound to HTLV-1 acts as an enhancer blocker, regulates HTLV-1 mRNA splicing, and forms long-distance interactions with flanking host chromatin.

Maintaining the CTCF-mediated TAD boundaries is crucial for the proper development of organisms. In a recent work Lupianez and colleagues [115] noticed that several conditions involving limb malformation can be linked to disruption in the TAD boundaries in the vicinity of the EPHA4 gene. This gene is involved in proper limb innervation but itself does not directly influence the formation of the limb skeleton. The vicinity of the EPHA4 gene can be divided into 3 TADs: (A) a gene rich TAD with the WNT6 and IHH genes (B) the second TAD that encompasses only a single gene – EPHA4 and finally (C) the TAD that includes the PAX3 gene. Lupianez noted that structural variants in the vicinity of these three TADs can be identified in families affected with brachydactyly, F-syndrome, or polysyndactyly with craniofacial abnormalities. These variants resulted in (I) deletion of the long DNA fragment that includes both the EPHA4 gene and the border between the EPHA4 TAD and the PAX3 TAD, (II) inversion of the fragment encompassing both the WNT6 TAD and the EPHA4 TAD, and finally (III) duplication of the region at the border of the WNT6 TAD and the EPHA4 TAD. To analyze if these variants are actually responsible for the phenotypes observed in humans, the CRISPR/Cas9 system was harnessed to test the effects of the same changes in mice. The results are transferable between these organisms as both humans and mice have an identical structural arrangement of TADs (and genes) in this region. The analysis of heterozygous mice revealed that the structural variants were in fact responsible for the observed phenotype, moreover, 4C experiments confirmed that the TAD borders indeed shifted due to changes occurring at the boundaries of the originally observed TADs. The modification of the 3D landscape of this region influenced the expression of several genes. For example, in case (II) the IHH gene becomes a part of the EPHA4 TAD and its expression pattern mimics that of the EPHA4 gene. Interestingly, the pathological conditions could not be attributed to a mere decreased distance between the genes and enhancers. Introducing the same structural variants but retaining the TAD boundaries, resulted in no changes in the phenotypes. The CTCF ChIP-seq data confirmed that this protein binds at the borders of the EPHA4 TAD but not inside this TAD.

CTCF involvement in imprinting-related diseases

Beside the abovementioned conditions, CTCF plays an important role in the developmental processes at the embryonic, prenatal and postnatal stages, where it is involved in maintaining proper genetic imprinting patterns. Genetic imprinting is a unique epigenetic mechanism; characteristic of placental mammals, marsupials and some flowering plants [116]. This phenomenon is characterized by expression of a single copy of a gene, either the one inherited from the mother or the father, where the other copy remains silenced [117]. Genetic imprinting is a complicated process in which the expression patterns can be gene- and tissue-specific and may depend on the current stage of development [118, 119]. Only specific genes undergo imprinting and typically they tend to form clusters controlled by Differentially Methylated Regions (DMR), which comprise CpG-rich regions called Imprinting Control Regions (ICRs) that are characterized by monoallelic DNA methylation and specific histone modifications. The CTCF binding motif has been identified in several ICRs and the ChIP-seq data confirmed that CTCF binds to ICRs and might be a causative agent behind genetic imprinting. Defects in ICR methylation can influence CTCF binding, which in turn can lead to harmful effects during fetal development as well as lead to several pathological conditions; e.g., transient neonatal diabetes mellitus, pseudohypoparathyroidism type 1, and various syndromes such as Angelman-Prader-Willy (APWS), Russel-Silver (RSS), and Beckwith-Wiedemann (BWS) [116].

The best-studied examples of CTCF involvement in genetic imprinting come from analyses of patients with RSS and BWS. RSS is characterized by growth retardation as defined by dwarfism, a low birth weight, triangular face and an asymmetric body [19]. BWS symptoms include overgrowth, which includes a higher birth weight, hemihypertrophy and abdominal wall defects. Some patients with BWS develop malignancies, like Wilms tumor [120]. The underlying molecular mechanism of these two syndromes involves a disruption in the genetic imprinting of two loci H19/IGF2 (location of ICR1) and CDKN1C/KCNQ1 (ICR2), both located on chromosome 11p15.5 Figure 6).

ICR1 is placed in an intergenic region between the H19 and IGF2 (Insulin-like growth factor 2) genes. DNA hypomethylation of this ICR results in binding CTCF to multiple target sites (TS) (between 6 and 7); hence it blocks the downstream enhancers from interacting with the IGF2 promoter and preventing gene expression. This situation normally occurs in the maternal allele, whereas for the paternal allele ICR1 is hypermethylated, thus preventing CTCF binding and allowing undisturbed expression of the IGF2 gene [116, 120, 121]. It is postulated that this mechanism works as a “cause-effect” loop meaning that due to CTCF interaction with the maternal ICR1, it is maintained in a methylation-free state. On the other hand, germline-derived DNA methylation of the parental chromosome prevents CTCF binding [121]. Deletions or uniparental disomy (UPD), might disrupt the ICR1 methylation pattern and, as a consequence, both IGF2 alleles are either unexpressed, resulting in RSS, or expressed, which leads to BWS. Since IGF2 is a crucial hormone player contributing to cell proliferation and growth promotion, its underexpression in RSS triggers developmental and growth retardation, whereas overexpression leads to prenatal and postnatal overgrowth [19, 116, 120]. The deletions observed in the ICR1 region seems to have various effects. The 2.2 kbp deletion (resulting in the removal of three TS) leads to limited methylation of the

maternal ICR1 and incomplete penetrance of BWS, whereas the 1.4 kbp and 1.8 kbp deletions (removing one or two TS) result in abundant methylation and severe BWS. Recently, it was shown that these microdeletions are specifically associated with changes in the distance between the CTCF TS and, as a consequence, a different pattern of TCF binding is observed. This suggests that the spatial organization of the CTCF TS, rather than their number, influences the CTCF binding. Furthermore, expressivity and penetrance of the BWS seem to be also dependent on the arrangement of the residual CTCF target sites on the alleles affected by different microdeletions [121].

Azzi et al. demonstrated that the hypomethylation does not spread equally throughout the entire /H19/IGF2 locus in patients with RSS. It was shown that the hypomethylation of specific TSs (TS2–4 and TS6) as well as DMR located in the H19 promoter (H19DMR), is observed more frequently than TS1, TS7 and IGF2DMR0 (DMR located in the IGF2 gene). Furthermore, 9% of RSS patients show a normal methylation levels of TS1 or TS7, whereas a subset of patients with BWS have normal levels of methylation at TS1, TS4, TS6 or TS7. This suggest that different mechanisms are responsible for protecting the unmethylated maternal allele from gaining methylation and the methylated paternal allele from losing methylation [122]. Beside cohesin and pluripotency factors, several other proteins have been reported to maintain of H19/IGF2 imprinting in cooperation with CTCF. These include the chromodomain helicase protein CDH8 [123] or the DEAD-box RNA-binding protein 68 [124]. It was also shown that the RNA-binding protein vigilin colocalize with CTCF at certain TSs regulating IGF2 expression [125]. Yu et al. have shown that vigilin interacts with the zinc-finger domains of CTCF through its seven K Homology domains. As this domain binds to RNA, further investigations indicated that the vigilin-CTCF interactions are RNA-dependent and mediated by the lncRNA H19. The H19 knock-down results in an alteration of the IGF2 imprinting pattern, supporting the idea that the CTCF-based imprinting also depends on the interactions between vigilin and H19 lncRNA [126].

ICR2, which regulates the CDKN1C/KCNQ1 (cyclin-dependent kinase inhibitor 1C/potassium voltage-gated channel subfamily Q member 1) cluster, is located in the promoter region of the KCNQ1OT1 gene (KCNQ –overlapping transcript 1). When ICR2 is unmethylated, KCNQ1OT1 expresses lncRNA that functions as a cis element antisense to the KCNQ1 gene silencing it and other genes within this cluster. This situation is normally present in the paternal allele. In the maternal allele ICR2 methylation result in KCNQ1OT1 silencing and the expression of genes in the CDKN1C/KCNQ1 cluster. Biallelic ICR2 methylation or maternal UPD results in overexpression of the genes in the CDKN1C/KCNQ1 locus, which leads to R S, while biallelic unmethylation or paternal UPD results in BWS. As CDKN1C acts not only as a negative regulator of cell proliferation but also as a tumour-suppressing agent, underexpression of this protein may lead to the cooccurrence of paediatric tumours with BWS [116]

Recently it has been shown that imprinting of the CDKN1C/KCNQ1 locus can be differentially affected by the 4.5 kbp region located in the second intron of the KCNQ1 gene, nearly 170 kbp upstream of ICR2 [127]. Additionally, this region lies in the middle of the 50 kbp region previously defined as cis-duplicated in the familial case of BWS with complete loss of ICR2 methylation [128]. Two distinct haplotypes associated with BWS,

called “protective” and “risk”, were identified using the linkage disequilibrium method. The “risk” haplotype, upon maternal transmission, leads to loss in ICR2 methylation and consequently to BWS. Two highly conserved CTCF binding sites were identified within this region. Demars et al. found that two SNPs (rs11823023 and rs179436), lying in close proximity to these CTCF binding sites, can affect CTCF occupancy. Both SNPs are associated with a loss of methylation but, according to EMSA experiments, the first SNP showed a decrease in the CTCF occupancy while the second SNP had the opposite effect. CTCF ChIA-PET data confirmed that both CTCF binding sites are involved in the formation of loops either with the 3'-end of the CDKN1C gene or with the region between the TRPM5 and KCNQ1 genes. Thus, the authors suggested that CTCF is involved in protecting the maternal ICR2 methylation by a mechanism involving the formation of a DNA loop between this 4.5 kbp region and the 3' end of the CDKN1C gene [127].

Summary

The field of chromatin 3D structure is prone to rapid changes. In this review we briefly described our current understanding how the genome 3D structure is organized first into 10-nm fiber followed by different types of DNA loops that constitute topologically associated domains and finally into interphase or mitotic chromosomes. We highlighted the role of CTCF in this process and described how mutations in the CTCF-binding sequence can lead to development of serious illnesses. Although CTCF is often described as the master weaver of the genome new experiments suggest that it is not the only agent involved in loop formation. Other proteins involved in this process are among others the mediator complex [129], which is used as a hub for multiple transcription factors. The primary role of the YY1 protein is associated with short-range promoter-enhancer contacts, while CTCFL (also called BORIS), a homologue of CTCF, recognizes the same DNA sequences but its expression is restricted to germ cells and a subset of cancer [12]. Cohesin, a complex consisting of several proteins, is now assumed to be an actual motor behind extrusion of loops, in a mechanism similar to that observed for condensin in some bacteria. Because the field of genome 3D structure is changing rapidly and new experimental data and information is advancing rapidly, to what extent the observation presented here will stand the test of time remains to be seen. But with knowledge that was gained in last two decades we are much closer to understanding the spatial organization of the genome.

Acknowledgements

This work has been supported by the Polish National Science Centre (2014/15/B/ST6/05082 and 2015/17/N/NZ2/01932), Foundation for Polish Science (TEAM to DP) and by the grant from the Polish Ministry of Science and Higher Education together with the Department of Science and Technology, India under Indo-Polish/Polish-Indo project No.: DST/INT/POL/P-36/2016. The work was co-supported by grant 1U54DK107967-01 “Nucleome Positioning System for Spatiotemporal Genome Organization and Regulation” within 4DNucleome NIH program. We thank Michal Kadlof, Lukasz Knizewski, Agnieszka Kraft and Przemyslaw Szalaj for careful reading of the manuscript.

References

1. Maeshima K, et al., Chromatin as dynamic 10-nm fibers. *Chromosoma*, 2014 123(3): p. 225–37. [PubMed: 24737122]

2. Cavalier-Smith T, Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Ann Bot*, 2005 95(1): p. 147–75. [PubMed: 15596464]
3. Webster M, Witkin KL, and Cohen-Fix O, Sizing up the nucleus: nuclear shape, size and nuclear-envelope assembly. *J Cell Sci*, 2009 122(Pt 10): p. 1477–86. [PubMed: 19420234]
4. Lieberman-Aiden E, et al., Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 2009 326(5950): p. 289–93. [PubMed: 19815776]
5. Flyamer IM, et al., Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, 2017 544(7648): p. 110–114. [PubMed: 28355183]
6. Barbieri M, et al., Complexity of chromatin folding is captured by the strings and binders switch model. *Proc Natl Acad Sci U S A*, 2012 109(40): p. 16173–8. [PubMed: 22988072]
7. Ong CT and Corces VG, CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet*, 2014 15(4): p. 234–46. [PubMed: 24614316]
8. Ruiz-Velasco M, et al., CTCF-Mediated Chromatin Loops between Promoter and Gene Body Regulate Alternative Splicing across Individuals. *Cell Syst*, 2017 5(6): p. 628–637.e6. [PubMed: 29199022]
9. Phillips JE and Corces VG, CTCF: master weaver of the genome. *ell*, 2009 137(7): p. 1194–211.
10. Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ, Cohen MA, Nabet B, Buckley DL, Guo YE, Hnisz D, Jaenisch R, Bradner JE, Gray NS and Young RA (2017). “YY1 Is a Structural Regulator of Enhancer-Promoter Loops.” *Cell* 171(7): 1573–1588.e1528. [PubMed: 29224777]
11. Bergmaier P, et al., Choice of binding sites for CTCFL compared to CTCF is driven by chromatin and by sequence preference. *Nucleic Acids Res*, 2018.
12. Pugacheva EM, Rivero-Hinojosa S, Espinoza CA, Mendez-Catala CF, Kang S, Suzuki T, Kosaka-Suzuki N, Robinson S, Nagarajan V, Ye Z, Boukaba A, Rasko JE, Strunnikov AV, Loukinov D, Ren B and Lobanenko VV (2015). “Comparative analyses of CTCF and BORIS occupancies uncover two distinct classes of CTCF binding genomic regions.” *Genome Biol* 16: 161. [PubMed: 26268681]
13. Busslinger GA, et al., Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. *Nature*, 2017 544(7651): p. 503–507. [PubMed: 28424523]
14. Peters JM, Tedeschi A, and Schmitz J, The cohesin complex and its roles in chromosome biology. *Genes Dev*, 2008 22(22): p. 3089–114. [PubMed: 19056890]
15. Guo Y, et al., CRISPRInversionofCTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell*, 2015 162(4): p. 900–10. [PubMed: 26276636]
16. de Wit E, et al., CTCF Binding Polarity Determines Chromatin Looping. *Mol Cell*, 2015 60(4) p. 676–84. [PubMed: 26527277]
17. Downen JM, et al., Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell*, 2014 159(2): p. 374–387. [PubMed: 25303531]
18. Hanssen LL, et al., Tissue-specific CTCF-cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nat Cell Biol*, 2017 19(8): p. 952–961. [PubMed: 28737770]
19. Spiteri BS, Stafrace Y, and Calleja-Agius J, Silver-Russell Syndrome: A Review. *Neonatal Netw*, 2017 36(4): p. 206–212. [PubMed: 28764823]
20. Robinson PJ and Rhodes D, Structure of the ‘30 nm’ chromatin fibre: a key role for the linker histone. *Curr Opin Struct Biol*, 2006 16(3): p. 336–43. [PubMed: 16714106]
21. Nishino Y, et al., Three-dimensional visualization of a human chromosome using coherent X-ray diffraction. *Phys Rev Lett*, 2009 102(1): p. 018101. [PubMed: 19257243]
22. Ou H., et al., ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science*, 2017 357(6349).
23. Finch JT and Klug A, Solenoidal model for superstructure in chromatin. *Proc Natl Acad Sci U S A*, 1976 73(6): p. 1897–901. [PubMed: 1064861]
24. Widom J and Klug A (1985). “Structure of the 300A chromatin filament: X-ray diffraction from oriented samples.” *Cell* 43(1): 207–213. [PubMed: 4075395]

25. Song F, et al., Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. *Science*, 2014 344(6182): p. 376–80. [PubMed: 24763583]
26. Robinson PJ, et al., EM measurements define the dimensions of the “30-nm” chromatin fiber: evidence for a compact, interdigitated structure. *Proc Natl Acad Sci U S A*, 2006 103(17): p. 6506–11. [PubMed: 16617109]
27. Dubochet J, et al., Cryo-electron microscopy of vitrified specimens. *Q Rev Biophys*, 1988 21(2): p. 129–228. [PubMed: 3043536]
28. Dubochet J and Sartori Blanc N., The cell in absence of aggregation artifacts. *Micron*, 2001 32(1): p. 91–9. [PubMed: 10900384]
29. Joti Y, et al., Chromosomes without a 30-nm chromatin fiber. *Nucleus*, 2012 3(5): p. 404–10. [PubMed: 22825571]
30. Nishino Y, et al., Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure. *Embo j*, 2012 31(7): p. 1644–53. [PubMed: 22343941]
31. Maeshima K, et al., Nucleosomal arrays self-assemble into supramolecular globular structures lacking 30-nm fibers. *Embo j*, 2016 35(10): p. 1115–32. [PubMed: 27072995]
32. Bilokapic S, Strauss M, and Halic M, Cryo-EM of nucleosome core particle interactions in trans. *Sci Rep*, 2018 8(1): p. 7046. [PubMed: 29728587]
33. Dekker J, et al., Capturing chromosome conformation. *Science*, 2002 295(5558): p. 1306–11. [PubMed: 11847345]
34. Simonis M, et al., Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet*, 2006 38(11): p. 1348–54. [PubMed: 17033623]
35. Dostie J and Dekker J, Mapping networks of physical interactions between genomic elements using 5C technology. *Nat Protoc*, 2007 2(4): p. 988–1002. [PubMed: 17446898]
36. Fullwood MJ, et al., An oestrogen-receptor- α -bound human chromatin interactome. *Nature*, 2009 462(7269): p. 58–64. [PubMed: 19890323]
37. Lajoie BR, Dekker J, and Kaplan N, The Hitchhiker’s guide to Hi- analysis: practical guidelines. *Methods*, 2015 72: p. 65–75. [PubMed: 25448293]
38. de Wit E and de Laat W, A decade of 3C technologies: insights into nuclear organization. *Genes Dev*, 2012 26(1): p. 11–24. [PubMed: 22215806]
39. Denker A and de Laat W, The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev*, 2016 30(12): p. 1357–82. [PubMed: 27340173]
40. Fullwood MJ and Ruan Y, ChIP-based methods for the identification of long-range chromatin interactions. *J Cell Biochem*, 2009 107(1): p. 30–9. [PubMed: 19247990]
41. Han J, Zhang Z, and Wang K, 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. *Mol Cytogenet*, 2018 11: p. 21. [PubMed: 29541161]
42. Rao SS, et al., A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 2014 159(7): p. 1665–80. [PubMed: 25497547]
43. Tang Z, et al., CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell*, 2015 163(7): p. 1611–27. [PubMed: 26686651]
44. Bolzer A, et al., hree-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol*, 2005 3(5): p. e157. [PubMed: 15839726]
45. Fraser P and Bickmore W, Nuclear organization of the genome and the potential for gene regulation. *Nature*, 2007 447(7143): p. 413–7. [PubMed: 17522674]
46. Parada L and Misteli T, Chromosome positioning in the interphase nucleus. *Trends Cell Biol*, 2002 12(9): p. 425–32. [PubMed: 12220863]
47. Cremer T . and . Cremer, Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet*, 2001 2(4): p. 292–301. [PubMed: 11283701]
48. Yanez-Cuna JO and van Steensel B, Genome-nuclear lamina interactions: from cell populations to single cells. *Curr Opin Genet Dev*, 2017 43: p. 67–72. [PubMed: 28107752]
49. Peric-Hupkes D, et al., Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell*, 2010 38(4): p. 603–13. [PubMed: 20513434]

50. Nora EP, et al., Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 2012 485(7398): p. 381–5. [PubMed: 22495304]
51. Dixon JR, et al., Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 2012 485(7398): p. 376–80. [PubMed: 22495300]
52. Vietri Rudan M., et al., Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep*, 2015 10(8): p. 1297–309. [PubMed: 25732821]
53. Szalaj P and Plewczynski D, Three-dimensional organization and dynamics of the genome. *Cell Biol Toxicol*, 2018.
54. Nagano T, et al., Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 2017 547(7661): p. 61–67. [PubMed: 28682332]
55. Forcato M, et al., Comparison of computational methods for Hi-C data analysis. *Nat Methods*, 2017 14(7): p. 679–685. [PubMed: 28604721]
56. Ji X, et al., 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell Stem Cell*, 2016 18(2): p. 262–75. [PubMed: 26686465]
57. Doyle B, et al., Chromatin loops as allosteric modulators of enhancer-promoter interactions. *PLoS Comput Biol*, 2014 10(10): p. e1003867. [PubMed: 25340767]
58. Soutourina J, Transcription regulation by the Mediator complex. *Nat Rev Mol Cell Biol*, 2018 19(4): p. 262–274. [PubMed: 29209056]
59. Kagey MH, et al., Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 2010 467(7314): p. 430–5. [PubMed: 20720539]
60. Javierre BM, et al., Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, 2016 167(5): p. 1369–1384 e19. [PubMed: 27863249]
61. O’Sullivan JM, et al., Gene loops juxtapose promoters and terminators in yeast. *Nat Genet*, 2004 36(9): p. 1014–8. [PubMed: 15314641]
62. Eagen KP, Aiden EL, and Kornberg RD, Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proc Natl Acad Sci S A*, 2017 114(33): p. 8764–8769.
63. Schoenfelder S, et al., Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat Genet*, 2015 47(10): p. 1179–1186. [PubMed: 26323060]
64. MacPherson MJ and Sadowski PD, The CTCF insulator protein forms an unusual DNA structure. *BMC Mol Biol*, 2010 11: p. 101. [PubMed: 21176138]
65. Heger P, et al., The chromatin insulator CTCF and the emergence of metazoan diversity. *Proc Natl Acad Sci U S A*, 2012 109(43): p. 17507–12. [PubMed: 23045651]
66. Kim TH, et al., Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, 2007 128(6): p. 1231–45. [PubMed: 17382889]
67. Heath H, et al., CTCF regulates cell cycle progression of alphabeta T cells in the thymus. *EMBO J*, 2008 27(21): p. 2839–50. [PubMed: 18923423]
68. Yin M, et al., Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites. *Cell Res*, 2017 27(11): p. 1365–1377. [PubMed: 29076501]
69. Nakahashi H, et al., A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep*, 2013 3(5): p. 1678–1689. [PubMed: 23707059]
70. Wang H, et al., Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res*, 2012 22(9): p. 1680–8. [PubMed: 22955980]
71. Schmidt D, ., et al., Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, 2012 148(1–2): p. 335–48. [PubMed: 22244452]
72. Hansen S, et al., CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *Elife*, 2017. 6.
73. Kadota M, et al., CTCF binding landscape in jawless fish with reference to Hox cluster evolution. *Sci Rep*, 2017 7(1): p. 4957. [PubMed: 28694486]
74. Cassandri M, et al., Zinc-finger proteins in health and disease. *Cell Death Discov*, 2017 3: p. 17071. [PubMed: 29152378]

75. Hashimoto H, Wang D, Horton JR, Zhang X, Corces VG and Cheng X (2017). “Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA.” *Mol Cell* 66(5): 711–720 e713. [PubMed: 28529057]
76. Luscombe NM, Laskowski RA, and Thornton JM, Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res*, 2001 29(13): p. 2860–74. [PubMed: 11433033]
77. Segal DJ, et al., Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5'-GNN-3' DNA target sequences. *Proc Natl Acad Sci U S A*, 1999 96(6): p. 2758–63. [PubMed: 10077584]
78. Lister R, et al., Human DNA methylomes at base resolution show widespread epigenomic differences. *nature*, 2009 462(7271): p. 315. [PubMed: 19829295]
79. Ghirlando R and Felsenfeld G, CTCF: making the right connections. *Genes Dev*, 2016 30(8): p. 881–91. [PubMed: 27083996]
80. Guo Y, et al., CTCF/cohesin-mediated DNA looping is required for protocadherin alpha promoter choice. *Proc Natl Acad Sci U S A*, 2012 109(51): p. 21081–6. [PubMed: 23204437]
81. Hark AT, et al., CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature*, 2000 405(6785): p. 486. [PubMed: 10839547]
82. Filippova GN, et al., An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol Cell Biol*, 1996 16(6): p. 2802–13. [PubMed: 8649389]
83. Merkenschlager M and Odom DT, CTCF and cohesin: linking gene regulatory elements with their targets. *Cell*, 2013 152(6): p. 1285–97. [PubMed: 23498937]
84. Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T, Yahata K, Imamoto F, Aburatani H, Nakao M, Imamoto N, Maeshima K, Shirahige K and Peters JM (2008). “Cohesin mediates transcriptional insulation by CCCTC-binding factor.” *Nature* 451(7180): 796–801. [PubMed: 18235444]
85. Tedeschi A, et al., Wapl is an essential regulator of chromatin structure and chromosome segregation. *Nature*, 2013 501(7468): p. 564. [PubMed: 23975099]
86. Zuin J, et al., Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci U S A*, 2014 111(3): p. 996–1001. [PubMed: 24335803]
87. Nora EP, et al., Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell*, 2017 169(5): p. 930–944.e22. [PubMed: 28525758]
88. Zuin J, et al., A cohesin-independent role for NIPBL at promoters provides insights in CdLS. *PLoS Genet*, 2014 10(2): p. e1004153. [PubMed: 24550742]
89. Davidson IF, et al., Rapid movement and transcriptional re-localization of human cohesin on DNA. *EMBO J*, 2016 35(24): p. 2671–2685. [PubMed: 27799150]
90. Haering CH, et al., Molecular architecture of C proteins and the yeast cohesin complex. *Mol Cell*, 2002 9(4): p. 773–88. [PubMed: 11983169]
91. Huis in 't Veld P.J., et al., Characterization of a DNA exit gate in the human cohesin ring. *Science*, 2014 346(6212): p. 968–72. [PubMed: 25414306]
92. Yusufzai TM, et al., CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol Cell*, 2004 13(2): p. 291–8. [PubMed: 14759373]
93. Nichols MH and Corces VG, A CTCF Code for 3D Genome Architecture. *Cell*, 2015 162(4): p. 703–5. [PubMed: 26276625]
94. Hons MT, et al., Topology and structure of an engineered human cohesin complex bound to Pds5B. *Nat Commun*, 2016 7: p. 12523. [PubMed: 27549742]
95. Wang X, et al., Bacillus subtilis SMC complexes juxtapose chromosome arms as they travel from origin to terminus. *Science*, 2017 355(6324): p. 524–527. [PubMed: 28154080]
96. Hirokawa N and Takemura R, Molecular motors and mechanisms of directional transport in neurons. *Nat Rev Neurosci*, 2005 6(3): p. 201–14. [PubMed: 15711600]
97. Vallee RB, et al., Dynein: An ancient motor protein involved in multiple modes of transport. *J Neurobiol*, 2004 58(2): p. 189–200. [PubMed: 14704951]

98. Alipour E, . and Marko JF, Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res*, 2012 40(22): p. 11202–12. [PubMed: 23074191]
99. Fudenberg G, et al., Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep*, 2016 15(9): p. 2038–49. [PubMed: 27210764]
100. Sanborn AL, et al., Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A*, 2015 112(47): p. E6456–65. [PubMed: 26499245]
101. Hayashi M, Jin Y, Yorgov D, Santorico SA, Hagman J, Ferrara TM, Jones KL, Cavalli G, Dinarello CA and Spritz RA (2016). “Autoimmune vitiligo is associated with gain-of-function by a transcriptional regulator that elevates expression of HLA-A*02:01 in vivo.” *Proc Natl Acad Sci U S A* 113(5): 1357–1362. [PubMed: 26787886]
102. Martin P, et al., Identifying Causal Genes at the Multiple Sclerosis Associated Region 6q23 Using Capture Hi-C. *PLoS One*, 2016 11(11): p. e0166923. [PubMed: 27861577]
103. Raj P, et al., Regulatory polymorphisms modulate the expression of HLA class II molecules and promote autoimmunity. *Elife*, 2016 5.
104. Allen EK, et al., SNP-mediated disruption of CTCF binding at the IFITM3 promoter is associated with risk of severe influenza in humans. *Nat Med*, 2017 23(8): p. 975–983. [PubMed: 28714988]
105. De Souza RA, et al., DNA methylation profiling in human Huntington’s disease brain. *Hum Mol Genet*, 2016 25(10): p. 2013–2030. [PubMed: 26953320]
106. Gosalia N and Harris A, Chromatin Dynamics in the Regulation of CFTR Expression. *Genes (Basel)*, 2015 6(3): p. 543–58. [PubMed: 26184320]
107. Bornstein S, Schmidt M, Choonoo G, Levin T, Gray J, Thomas CR, Jr., Wong M and McWeeney S (2016). “IL-10 and integrin signaling pathways are associated with head and neck cancer progression.” *BMC Genomics* 17: 38. [PubMed: 26747525]
108. Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE, Ristolainen H, Hanninen UA, Cajuso T, Kondelin J, Tanskanen T, Mecklin JP, Jarvinen H, Renkonen-Sinisalo L, Lepisto A, Kaasinen E, Kilpivaara O, Tuupanen S, Enge M, Taipale J and Aaltonen LA (2015). “CTCF/cohesin-binding sites are frequently mutated in cancer.” *Nat Genet* 47(7): 818–821. [PubMed: 26053496]
109. Shi Y, et al., Chromatin accessibility contributes to simultaneous mutations of cancer genes. *Sci Rep*, 2016 6: p. 35270. [PubMed: 27762310]
110. Zhang B, et al., The CCCTC-binding factor (CTCF)-forkhead box protein M1 axis regulates tumour growth and metastasis in hepatocellular carcinoma. *J Pathol*, 2017 243(4): p. 418–430. [PubMed: 28862757]
111. Li N, Zheng J, Li H, Deng J, Hu M, Wu H, Li W, Li F, Lan X, Lu J and Zhou Y (2014). “Identification of chimeric TSNAX-DISC1 resulting from intergenic splicing in endometrial carcinoma through high-throughput RNA sequencing.” *Carcinogenesis* 35(12): 2687–2697. [PubMed: 25239642]
112. Zhao X, et al., CTCF cooperates with noncoding RNA MYCNOS to promote neuroblastoma progression through facilitating MYCN expression. *Oncogene*, 2016 35(27): p. 3565–76 [PubMed: 26549029]
113. Lang F, et al., CTCF interacts with the lytic HSV-1 genome to promote viral transcription. *Sci Rep*, 2017 7: p. 39861...p..
114. Satou Y, et al., The retrovirus HTLV-1 inserts an ectopic CTCF-binding site into the human genome. *Proc Natl Acad Sci USA*, 2016 113(11): p. 3054–9.
115. Lupianez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, Santos-Simarro F, Gilbert-Dussardier B, Wittler L, Borschiwer M, Haas SA, Osterwalder M, Franke M, Timmermann B, Hecht J, Spielmann M, Visel A and Mundlos S (2015). “Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions.” *Cell* 161(5): 1012–1025. [PubMed: 25959774]
116. Elhamamsy AR, Role of DNA methylation in imprinting disorders: an updated review. *J Assist Reprod Genet*, 2017 34(5): p. 549–562. [PubMed: 28281142]

117. Renfree MB, Suzuki S, and Kaneko-Ishino T, The origin and evolution of genomic imprinting and viviparity in mammals. *Philos Trans R Soc Lond B Biol Sci*, 2013 368(1609): p. 20120151. [PubMed: 23166401]
118. Lalande M and Calciano M, Molecular epigenetics of Angelman syndrome. *Cell Mol Life Sci*, 2007 64(7–8): p. 947–60. [PubMed: 17347796]
119. Wilkinson LS, Davies W, and Isles AR, Genomic imprinting effects on brain development and function. *Nat Rev Neurosci*, 2007 8(11): p. 832–43. [PubMed: 17925812]
120. Brioude F, et al., Expert consensus document: Clinical and molecular diagnosis, screening and management of Beckwith-Wiedemann syndrome: an international consensus statement. *Nat Rev Endocrinol*, 2018 14(4): p. 229–249. [PubMed: 29377879]
121. Beygo J, et al., The molecular function and clinical phenotype of partial deletions of the IGF2/H19 imprinting control region depends on the spatial arrangement of the remaining CTCF-binding sites. *Hum Mol Genet*, 2013 22(3): p. 544–57. [PubMed: 23118352]
122. Azzi S, Steunou V, Tost J, Rossignol S, Thibaud N, Das Neves C, Le Jule M, Habib WA, Blaise A, Koudou Y, Busato F, Le Bouc Y and Netchine I (2015). “Exhaustive methylation analysis revealed uneven profiles of methylation at IGF2/ICR1/H19 11p15 loci in Russell Silver syndrome.” *J Med Genet* 52(1): 53–60. [PubMed: 25395389]
123. Ishihara K, Oshimura M, and Nakao M, CTCF-dependent chromatin insulator is linked to epigenetic remodeling. *Mol Cell*, 2006 23(5): p. 733–42. [PubMed: 16949368]
124. Yao H, et al., Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. *Genes Dev*, 2010 24(22): p. 2543–55. [PubMed: 20966046]
125. Liu Q, et al., Vigilin interacts with CCCTC-binding factor (CTCF) and is involved in CTCF-dependent regulation of the imprinted genes *Igf2* and *H19*. *FEBS J*, 2014 281(12): p. 2713–25. [PubMed: 24725430]
126. Yu X, et al., Vigilin interacts with CTCF and is involved in the maintenance of imprinting of *IGF2* through a novel RNA-mediated mechanism. *Int J Biol Macromol*, 2018 108: p. 515–522. [PubMed: 29157910]
127. Demars J, et al., Genetic variants within the second intron of the *KCNQ1* gene affect CTCF binding and confer a risk of Beckwith-Wiedemann syndrome upon maternal transmission. *J Med Genet*, 2014 51(8): p. 502–11. [PubMed: 24996904]
128. Demars J, et al., New insights into the pathogenesis of Beckwith-Wiedemann and Silver-Russell syndromes: contribution of small copy number variations to 11p15 imprinting defects. *Hum Mutat*, 2011 32(10): p. 1171–82. [PubMed: 21780245]
129. Allen BL and Taatjes DJ, The Mediator complex: a central integrator of transcription. *Nat Rev Mol Cell Biol*, 2015 16(3): p. 155–66. [PubMed: 25693131]
130. Woodcock CL, Frado LL and Rattner JB (1984). “The higher-order structure of chromatin: evidence for a helical ribbon arrangement.” *J Cell Biol* 99(1 Pt 1): 42–52. [PubMed: 6736132]
131. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A and Fraser P (2013). “Single-cell Hi-C reveals cell-to-cell variability in chromosome structure.” *Nature* 502(7469): 59–64. [PubMed: 24067610]

Fig 1.

Please refer to the Figure 1 of the manuscript “A 3D Map of the Human Genomeat Kilobase Resolution Reveals Principles of Chromatin Looping” by Rao et al.

Fig 2.

Please refer to the Figure 1 of the manuscript “CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription” by Tang et al.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig 3.

Please refer to the Figures 3 and 4 of the manuscript “Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites” by Yin et al.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig 4.

Please refer to the Figures 1 and 2 of the manuscript “CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function” by Guo et al

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig 5.

Please refer to the Figure 4 of the manuscript “Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl” by Busslinger et al.

Fig 6.

Please refer to the Figure 1 of the manuscript “Beckwith-Wiedemann syndrome” by Choufani et al.