Studying the Oncogenic and Epigenetic Impact of CTCF Loss of Heterozygosity and Zinc-Finger 1 Mutation in Breast Cancer

Benjamin Lebeau Division of Experimental Medicine
McGill University, Montréal November 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Ph.D. in Experimental Medicine

Contents

Abstract

English:

While epigenetic processes are important drivers of tumor progression, the contribution of deregulated chromatin architecture, including topologically associated domains (TADs), to cancer progression remains ambiguous. CTCF is a central regulator of higher-order chromatin structure that undergoes copy number loss in over half of all breast cancers. Also, mutations of CTCF ZF1 are exclusive to breast cancer and are associated with metastasis and therapeutic resistance. The impact of these CTCF defects on epigenetic programming, chromatin architecture and cancer progression remain unclear. We find that under physiological conditions, CTCF organizes subTADs to limit the expression of oncogenic pathways, including PI3K and cell adhesion networks. Loss of a single CTCF allele potentiates cell invasion through compromised chromatin insulation, a reorganization of chromatin architecture and histone programming that facilitates de novo promoter-enhancer contacts within TADs. However, this change in the higher-order chromatin landscape leads to a vulnerability to inhibitors of mTOR. Next, we developed and employed a novel motif analysis software, MoMotif, to define the previously uncharacterized recognition motif of CTCF zinc-finger 1 (ZF1), and to characterize the impact of CTCF ZF1 mutation on its association with chromatin. Using MoMotif, we identified an extension of the CTCF core binding motif that is recognized by a functional CTCF ZF1. Using a combination of ChIP-Seq and RNA-Seq, we discover that the inability to bind this extended motif drives an altered transcriptional program, here again enriched within TADs, that mimics the harmful oncogenic phenotypes observed clinically. These data support a model whereby subTAD reorganization drives both the modification of histones at de novo enhancer promoter-contacts and transcriptional upregulation of oncogenic transcriptional networks.

Français:

L’impact de la dérégulation de l'organisation architecturale de la chromatine, y compris les domaines d'association topologique (TAD) sur la progression du cancer reste ambiguë. CTCF est un régulateur essentiel de la structure de la chromatine qui voit son expression réduite dans plus de la moitié de tous les cancers du sein. De plus, les mutations du doigt de zinc 1 (ZF1) de CTCF sont exclusives au cancer du sein et sont associées au statut métastatique et à la résistance thérapeutique. Cependant, l'impact de ces altérations de CTCF sur la programmation épigénétique, l'architecture de la chromatine et la progression du cancer reste incertain. Nous constatons que dans des conditions physiologiques, CTCF organise les subTADs pour limiter l'expression de réseaux oncogènes, y compris les réseaux de PI3K et d'adhésion cellulaire. La perte d'un seul allèle CTCF permet l'invasion cellulaire en compromettant l’isolation de la chromatine, menant à une réorganisation de l'architecture de la chromatine et une reprogrammation des marques des histones, qui facilite de nouveaux contacts promoteur-amplificateur à l’intérieur des TADs. Ces changements épigénétiques créent une vulnérabilité aux inhibiteurs de mTOR. Aussi, nous avons développé et utilisé un nouveau logiciel d'analyse de motifs, MoMotif, pour définir le motif de reconnaissance non caractérisé de CTCF ZF1 et pour étudier l'impact de la mutation de CTCF ZF1 sur son association avec la chromatine. À l'aide de MoMotif, nous avons identifié une nouvelle extension du motif caractéristique de CTCF, qui nécessite un ZF1 fonctionnel pour s’y lier de manière appropriée. En utilisant une combinaison de ChIP-Seq et d'ARN-Seq, nous découvrons que l'incapacité à lier cette extension de motif altère les programmes transcriptionnels, ici encore, principalement à l’intérieur des TADs et d’une façon qui imite les phénotypes oncogènes nocifs observés cliniquement. En conclusion, ces données soutiennent un modèle dans lequel la réorganisation à l’intérieur des

TADs entraîne à la fois la modification des histones au niveau de nouveaux contacts promoteur-amplificateur, facilitant la transcription des réseaux transcriptionnels oncogènes.

Acknowledgements

This research was supported by the Canadian Institutes of Health Research (CIHR) (159759 and 159663 to M.W. An Oncopole EMC2 grant supported M.W, M.B. and M.P. B.L. is supported by a Lady Davis Institute TD Bank Studentship Award and M.W. is supported by a Fonds de Recherche du Québec Santé (FRQ-S), Senior Chercheur Boursier award. I further thank Kathlein Klein, of the Lady Davis bioinformatics core facility, for her valuable input.

Contribution to Original Knowledge

- Define the epigenetic and biological impact of CTCF Loss of Heterozygosity (LOH) in human mammary epithelial models. In short, the partial loss of CTCF leads to compromised subTAD insulation which permits the overexpression and over-activation of key effectors of the PI3K pathway, including the potent oncogene SNAI1. These aberrant epigenetic events drive the invasiveness of CTCF +/- epithelial cells, contributing to cancer progression.

- Demonstrate the biological and epigenetic difference between two levels of topological organization of the chromatin, namely TAD and subTADs. More specifically, that subTAD organization is more prone to changes following reduced CTCF levels and more profoundly impacts the transcription of oncogenic pathways, such as the PI3K signaling pathway.

- Propose a potential therapeutic avenue targeting cancer progression by which the epigenetic and biological consequences of CTCF LOH predict a vulnerability to mTOR or histone acetyltransferase inhibitors.

- As a collaboration, we developed a new tool, MoMotif, to investigate and quantify DNA binding motifs that discriminates between 2 conditions and outperforms currently available bioinformatic software.

- Identify an extension of the CTCF consensus motif recognized by its zinc finger 1, expanding our understanding of the role of this zinc-finger in CTCF-DNA recognition and the impact of its mutation in cancer.

- Associated the inability to bind the downstream extension of the CTCF consensus motif to transcriptional changes consistent with clinical phenotypes displayed by breast cancer carrying CTCF zinc-finger 1 mutations.

Contribution of Authors

Benjamin Lebeau: All bioinformatic analysis. Unless otherwise stated, all laboratory experiments, including RNA-Seq on CTCF H284N cells, all Hi-C experiments, ChIP-Seq on H3K27ac and invasion assays. Development and troubleshooting of MoMotif. Graphical representation of all data within the figures.

Co-authors in Chapter 1:

Cheng Kit Wong: Assistance with ChIP-Seq on CTCF.

Nolan Wong: ChIP-qPCR validation in supplementary figures.

Eduardo Cepeda Cañedo, Geneviève Deblois: Sample preparation and Western Blot of in vivo PDX models.

Adriana Aguilar-Mahecha, Catherine Chabot, Marguerite Buchanan, Morag Park, Mark Basik: Established the PDX cell lines and cell culture protocol, genotyping of the PDX cell lines

Rachel Catterall, Luke McCaffrey: Quantification of the fluorescence of the outer layer of the mammosphere

Co-authors in Chapter 2:

Kaiqiong Zhao, Celia M.T. Greenwood: Development and troubleshooting of MoMotif. Maria Guerra: Quantification of the consistency of TAD callers.

Co-authors in Chapter 1 & 2:

Maika Jangal: ChIP-Seq on CTCF and H3K4me3 experiments. RNA-Seq in CTCF +/- cells. Mammosphere experiments and immunofluorescence.

Tiejun Zhao: Assistance with western blot, invasion assay and mammosphere assay. Michael Witcher: Supervision and insight on all aspects of both chapters.

Introduction

Hierarchical nuclear organization of chromatin plays essential roles during development and cell specification (1,2). As such, mapping and understanding the functionality of three-dimensional (3D) chromatin structure is now at the forefront of epigenetics research. Based on the advent of Hi-C sequencing technology (3), we know that the entire genome is partitioned into an assembly of Topologically Associated Domains (TADs). TADs comprise 100kb to 1Mb regions of chromatin defined as a contiguous region enriched for DNA-DNA contacts between loci within the TAD, with few interactions outside of the TAD (4). TADs are commonly anchored by CTCF together with the cohesin complex, establishing a stable chromatin domain (5-9). Within TADs, smaller regions of self-interaction, called subTADs, add an additional layer of complexity to 3D chromatin architecture (10).

TADs and subTADs regulate gene transcription in mechanistically similar ways. By confining chromatin interactions in cis, to regions within a defined genomic neighborhood, they promote local interactions between cis-regulatory elements, such as enhancer-promoter interactions, while insulating from outside cis-regulatory elements. This allows for the specific pairings of promoters and enhancers required for proper temporal regulation of gene expression (11).

Organization of chromatin into subTADs facilitates a more precise and dynamic local regulation of transcription than TADs alone would allow. Indeed, dynamic changes in subTAD organization drive transcriptional events of differentiation and cell identity, while TAD boundaries are mostly stable during these processes (1).

Proper TAD/subTAD organization is essential for temporal control of gene expression during development (12,13). While aberrant activation of developmental programs appears to play an important role in tumor progression, it is unclear that widespread reorganization of

chromatindomains is involved in this process. Despite evidence that altered TAD or subTAD organization locally at specific oncogenic loci may promote tumor initiation via aberrant changes to gene transcription (14), genome-wide analysis of chromatin contacts using relevant models of tumor initiation and progression are clearly needed to provide further insights into a potential role of TAD reorganization in these processes.

Considering the central role of CCCTC-binding factor (CTCF) in maintenance of genomic TADs (15), it is not surprising that CTCF knockout leads to lethality at very early stages of embryonic development (16,17). Although not lethal, the loss of heterozygosity at the CTCF locus is also detrimental to cellular homeostasis (18) and CTCF appears to act as an haploinsufficent tumor suppressor gene (19,20), with its loss impacting hematopoietic tumor initiation in CTCF hemizygous mice (19). In humans, Down Syndrome related Acute MegaKaryoblastic Leukemia (DS-AMKL) carries CTCF deletions or mutations in 20% of all cases (21). Here, the loss of CTCF is thought to be important for clonal evolution to more aggressive phenotypes following GATA1 mutations (21). Despite such clear evidence for a tumor suppressive role for CTCF in hematopoietic tissue, the importance of physiological levels and functionality of CTCF for the prevention of solid tumors remains ambiguous. Consistent with a putative tumor suppressor role, data from the Cancer Genome Atlas reveals that sixty-three percent of all breast tumors harbor CTCF copy number loss (CNL) (22). While it has been hypothesized that fluctuations in CTCF levels may impact chromatin looping, this has not been formally examined (23). Thus, it remains unclear whether transcriptional networks and topological features may be deregulated in breast epithelium undergoing CTCF CNL.

In the first chapter, I explain how CTCF CNL in mammary epithelial cells potentiates subTAD reorganization and cell invasion. I also described that restructuring of chromatin architecture,

especially at the subTAD level, drives activation of the phosphatidylinositol 3-kinase (PI3K) pathway and overexpression of the classical oncogene SNAI1. These changes are associated with epigenetic reprogramming of H3K27ac and H3K4me3 at regulatory regions. These altered transcriptional events may predict sensitivity to mTor inhibitors (24), potently repressing the invasive capacity of cells carrying a single functional CTCF allele.

Furthermore, aberrant Transcription Factor (TF) activities or non-coding mutations located at promoters, enhancers or chromatin domain boundaries drive diverse pathologies, including a range of cancers (25-28). Biological investigation into the pathology of such events necessitates high-throughput sequencing based epigenomic approaches such as ChIP-Seq (29), and Hi-C (3). These epigenomic endeavors are expensive and require substantial quantities of biological samples (30). However, the development and fine-tuning of complementary bioinformatic analyses allow us to infer biological impact and subsequently predict sensitivity to personalized therapies. In particular, identifying context-dependent modifications of DNA-binding motifs specific to TFs is important for our understanding of cancer biology as motifs are frequently mutated, and mutated TFs may recognize altered motifs.

For instance, the biological impact of the mutation of the first zinc finger (ZF1) of the epigenetic regulatory protein CTCF, such as the H284N mutation, exclusive to breast cancer and prevalent in hormone resistant breast tumors (31), has remained elusive. In contrast to oncogenic mutations located within CTCF ZF3-7 (32), involved in CTCF's ability to bind its core motif (33,34) present in ~90% of CTCF binding sites (CBS) (35-37), CTCF ZF1 remains uncharacterized because its crystal structure has not been obtained (38). Although the truncation of CTCF ZF1 was shown to alter RNA dependent binding of CTCF to specific sites, the H284N mutation did not display such function (39). Also, CTCF ZF1 displays the weakest affinity for DNA of all

CTCF zinc fingers and is not required for the binding of CTCF to its core binding motif (33,38). It is known that bases outside the core binding motif modulate CTCF binding (33,40), but it remains unknown whether CTCF ZF1 mutations (ZF1M) regulate binding to an extended motif, or alternatively influence CTCF binding affinity through impeding its interaction with non-coding RNAs (39). However, computational tools designed to directly compare motifs between discriminative conditions are lacking. Therefore, we would expect current bioinformatic approaches to fall short in identifying possible motifs variations associated to differential binding of ZF1 mutated CTCF, because subtle changes would be "drowned" by the highly conserved elements of CTCF core binding motif. As such, new tools are required to predict the pathogenic mechanism of mutated DNA binding proteins, such as CTCF ZF1 mutations in breast cancer.

In the second chapter, I will describe a new R pipeline, developed in collaboration with Celia Greenwood's group, in which we designed a new tool, MoMotif (Modification of Motif analysis at single base-pair resolution). Our R pipeline incorporates, and builds upon, the three central analysis steps to mine ChIP-Seq data for DNA-binding motifs that discriminate between biological conditions. I profiled the potential of MoMotif by identifying the protein-DNA affinity changes conferred by the CTCF H284 mutation and different genomic locations. Further, I explained that the loss of binding, driven by mutant CTCF ZF1 causes changes in gene expression characteristic of the clinical phenotypes of CTCF mutated breast tumors.

Literature Review

Breast Cancer

Breast cancer is one of the 3 most common cancers and the most common malignancy in women (41). According to the Canadian Cancer Society, in 2017, a quarter of all new cancer cases in women are breast cancers, while 13% of all cancer related deaths in women were caused by breast cancer. Worldwide, around half a million people died from breast cancer in 2012 (41). The vast majority (~90%) die from metastatic disease. As a major burden on people's health worldwide, breast cancer rapidly became one of the most studied diseases on the planet. The hard work of these countless researchers provided major improvements to our understanding of the disease, leading to more efficient and less toxic treatments (42-44), which will be expanded upon later in this section. However, breast cancer is a heterogeneous and plastic disease. Therefore, researching mechanisms of its evolution and progression is important to offer improved treatments, both in terms of survival and toxicity, to millions of women.

Surgery

The earliest known mention of cancer dates back to about 3000 BC. Although not termed "cancer" back then, cases of breast cancer were described as untreatable, in an ancient Egyptian papyrus on trauma surgery, and needed to be removed with cautery, knives and salt (45). Despite the approximately five thousand years gap with modern medicine, surgery remains at the center of cancer care. Indeed, if a tumor is removed at an early, non-invasive stage, further consequence can frequently be avoided. However, due to the harm of over-diagnosis, the global health benefit of screening and removal of early tumors, such as during schedule mammography, are debatable (46). Additionally, not all tumors are easily detectable at early stages. For instance, early primary growth of pancreatic and ovarian cancer is often mostly symptomless and barely palpable through the skin,

greatly reducing the effectiveness of surgery and highly increasing their lethality (47,48). Further, tumors growing on essential organs or socially valued body parts, such as many head and neck cancers, may be unresectable or increase the morbidity of surgery (49). At later stages, tumors will invade the surrounding tissue, in a way that reminded Hippocrates of a crab, the Latin word for which being "cancer". These crustacean protrusions not only increase the area that needs to be surgically removed, leading to more morbidity, but also reduce the chances that the totality of the tumor mass will be removed by surgery. Therefore, as the millennia of oncology have taught us, surgery alone cannot treat cancer and therapeutic agents are required to prolong the life of the victims of this unrelenting disease.

Chemotherapy

"Oncology" comes from the Greek word "*oncos*", meaning swelling or growth. This ancient view of cancer, as an uncontrolled growth, remains the basis of most therapeutic approaches used today. Indeed, most chemotherapy harms, indiscriminately, fast dividing cells. Whether they result in massive amounts of DNA damage, such as cisplatin (50), or affect crucial elements of mitosis, such as paclitaxel (51), chemotherapies wreak havoc in rapidly dividing healthy or cancerous cells alike. This leads to numerous short-term toxicities, such as nausea, increased risk of infections, hair loss and neurotoxicity, which could result in permanent damage (51). In the long-term, chemotherapy is associated with an increased risk of developing another cancer, such as therapy related MDS or AML (52). Although effective against most tumor cells at first, breast cancer's ability to adapt and develop resistance to chemotherapies, coupled with an intratumor heterogenicity that can include slow dividing cells, result in about 40% of stage I to III tumors relapsing following standard therapy (53). The feared perspective of cancer and these toxic, yet flawed, treatments led many women, with family history of breast cancer, to undergo bilateral risk-reducing mastectomy (54). Hopefully, the advent of personalized target-based therapy is slowly

shifting this somber perception and benefiting the outcome of and patient experience during breast cancer treatments.

Luminal Breast Cancer

Around seventy percent of breast tumors overexpressed the estrogen receptor (ER) and/or Progesterone Receptor (PR). They are termed ER+/PR+ or Luminal A (41). This classification arises from the mechanistic importance of hormone receptors in cancer progression and, most importantly, their clinical relevance as predictive biomarkers.

When bound by estrogen, ER will dimerize and bind estrogen response elements on the genome to promote the expression of its target genes, such as GREB1 (55,56). As breast tissue is responsive to hormones, tumors can stem from an oncogenic highjack of this process to develop a dependence to estrogen-related pathways for their proliferation and survival (57). Therefore, ER positive tumors, contrarily to the majority of healthy cells of the human body, are highly sensitive to repression of estrogen signaling. By using estrogen receptor antagonist, blocking estrogen metabolism, and targeting estrogen receptor for degradation, modern targeted therapies evolved around this specific weakness of luminal breast cancer (Figure LR1) (58). Termed hormonal therapy, this pioneering approach to cancer care reduced recurrence

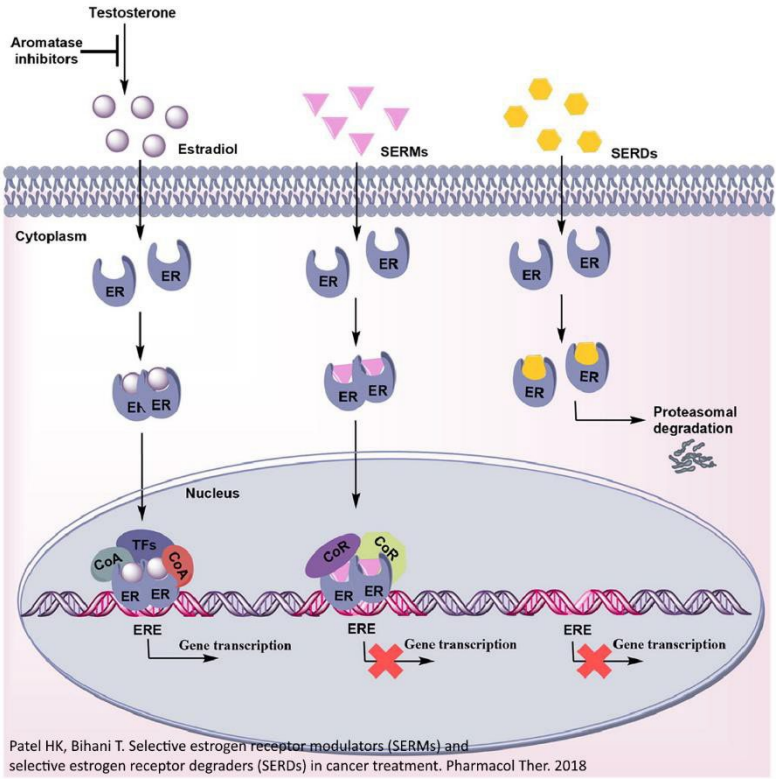


Figure LR1: Distinct therapeutic avenues to target ER.

and increased survival (58). The marked benefits of targeted therapies drove the discovery of new therapeutic targets.

HER2+ Breast Cancer

One such important target is the receptor tyrosine kinase, Her2, which is overexpressed in approximately twenty percent of breast tumors (59). Her2 overexpression promotes its dimerization with other receptor tyrosine kinase (RTK) of its family, such as HER3 or EGFR, resulting in their autophosphorylation and activation of downstream signaling cascades (59). Although signaling cascades branch out in various interconnected pathways of effectors, the activation of the PI3K signaling pathway is a common consequence of the overexpression of Her2 and is known to promote oncogenesis (60). The classical chain of reaction in this signaling pathway starts with the RTK's activity promoting PI3K phosphorylation of phosphatidylinositol-4,5-bisphosphate (PIP2) into phosphatidylinositol-3,4,5-triphosphate (PIP3) on the plasma membrane. Next, PIP3 will bind and permit the phosphorylation and activation of AKT by PDK1 and mTORC2. Then, AKT will phosphorylate the TSC1/TSC2 complex, blocking its inhibition of the mTORC1 kinase complex. mTORC1 will then phosphorylate 4EBP1 and S6K1, promoting translation and transcription of genes involved in cell survival and proliferation (Figure LR2) (60).

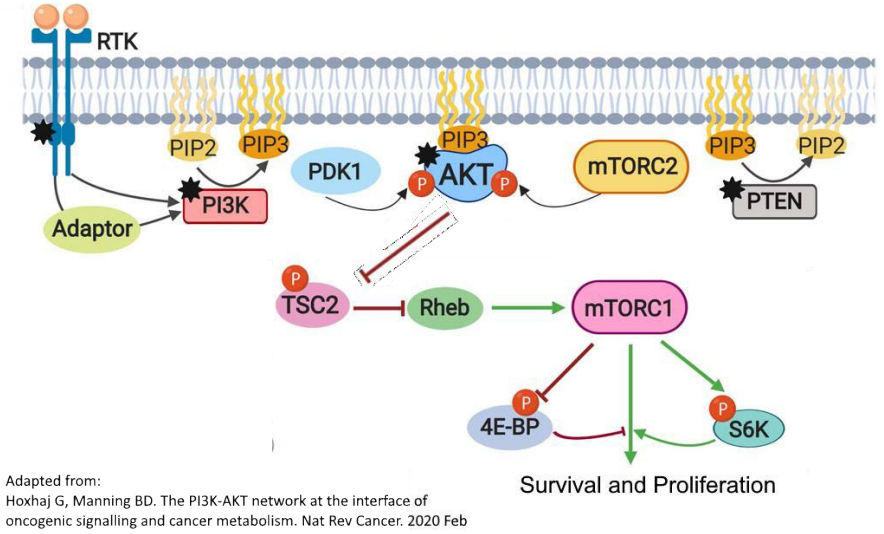


Figure LR2: Classical chain of reaction in the PI3K signaling pathway.

Similarly to ER+ breast cancer, Her2+ tumors will hijack and depend on Her2-related signaling pathways, such as the PI3K-pathway, for their oncogenesis. Therefore, targeting Her2 therapeutically will be more damaging for Her2+ tumors cells, than for most healthy cells. Now standard-of-care, monoclonal antibody targeting Her2+, such as trastuzumab, are used to inhibit dimerization with other RTKs, trigger antibody-dependent cellular cytotoxicity or to specifically deliver cytotoxic agent to Her2 expressing cells (59). Here again, a mechanism driving a tumor's growth and progression was turned into an exploitable weakness. However, actionable targets for cancer therapy are not always directly driving oncogenesis.

PARP Inhibitors and Synthetic Lethality

Synthetic lethality arises between two genes when the perturbation of either one is viable, but a simultaneous dysfunction of both leads to reduced viability (61). Such a relationship is observed between BRCA1/2 and the PARP1. BRCA1, named after the strong association between its heritable mutations and family history of breast cancer (62), is an essential player in Homologous Recombination (HR) double stranded-DNA repair pathways (63). BRCA1 promotes end resection and subsequent recruitment of BRCA2, which also depends on PARylated CTCF at the site of damage (64), for the proper deposition of RAD51, necessary for the following steps of HR (65). Similarly, PARP1 is involved in single-stranded DNA damage repair pathways, such as Base

Excision Repair (BER)(66). PARP1 activity leads to the recruitment of repair protein at the sites of damage. In the absence of functional PARP1, single stranded breaks are left unrepaired and become double-stranded breaks at the replication fork, which then requires HR for proper repair (67). In HR competent cells, this issue is usually resolved appropriately. But, in BRCA1/2 mutated or HR incompetent cells, the absence of PARP1 activity leads to a cytotoxic accumulation of DNA damage. Due to this synthetic lethality, PARP inhibitors lead to better progression-free survival and fewer toxicities than standard therapy for Her2- metastatic breast cancer with germline BRCA mutations (68). PARP inhibitors are the first approved cancer therapies based on synthetic lethality, but CRISPR-screen technologies are widely used today to find similarly actionable targets (61). Indeed, despite the existence of multiple targeted therapies for specific subtypes of breast cancer, a significant proportion of highly aggressive breast tumors are still lacking an actionable biomarker. This is especially true for Triple Negative Breast Cancer (TNBC).

Triple-Negative Breast Cancer

About fifteen to twenty percent of breast tumors are defined as TNBC due to their lack of ER, PR or Her2 expression. TNBC mostly occurs in younger woman and is more aggressive and prone to relapse, predicting a poorer prognosis than non-TNBC breast cancers. Without clear oncogenic dependence or synthetic lethality, the standard-of-care for TNBC remains highly toxic chemotherapies (69). Therefore, new actionable biomarkers are needed to provide better treatment to TNBC patients. Beside TNBC, relapsing primary or metastatic tumors that developed resistance to their initial therapy would also benefits from a wider variety of potential targets arising from a deeper understanding of cancer progression, both in term of biological process, but also genetic and epigenetic regulation and evolution.

Historical Perspective on Epigenetics and Chromatin Conformation

The field of epigenetics has been fast evolving in the last century. From the discovery of DNA methylation and its impact on gene expression, to the ever-increasing usage of RNA-Seq to study global gene expression and ChIP-Seq to map chromatin states and potential effectors of transcription, new transcriptional modulators are discovered continuously. In the last decade, the advent of chromatin conformation captures techniques revealed one such discovery: spatial organization of the DNA.

Deoxyribonucleic Acid and the Central Dogma of Molecular Biology

In 1871, Swiss physician and biochemist Friedrich Miescher isolated a novel organic substance in leukocyte nuclei purified from the pus of surgical bandage of the local hospital, which he termed nuclein (70). In 1944, this molecule, later named deoxyribonucleic acid (DNA), would be identified as the carrier of heritability within chromosomes (71). In 1953, Watson and Crick described the now iconic double-helix (72) to explain how DNA responds to the four requirements of any genetic material: replication, specificity, information content and ability to change (70). In agreement with Chargaff's rule (73), adenine (A) base-pairs with thymine (T) and cytosine (C) with guanine (G) to form the DNA double-helix constituting each chromosome. In brief, DNA allows for the genetic information to be carried as highly specifically ordered longitudinal sequences of bases. Following these discoveries, Francis Crick would enunciate and then specify the central dogma of molecular biology (74). In short, DNA is transcribed as messenger RNAs (mRNAs) which are then translated to proteins that will carry out their biological function in the cells.

In early studies of transcription, DNA was reduced to its coding sequence, a two-dimensional sequence of ATCG meant to be read by the transcription machinery in a, now widely criticized, 3-steps process. First, during initiation, an RNA Polymerase II (Pol II) is recruited to a promoter

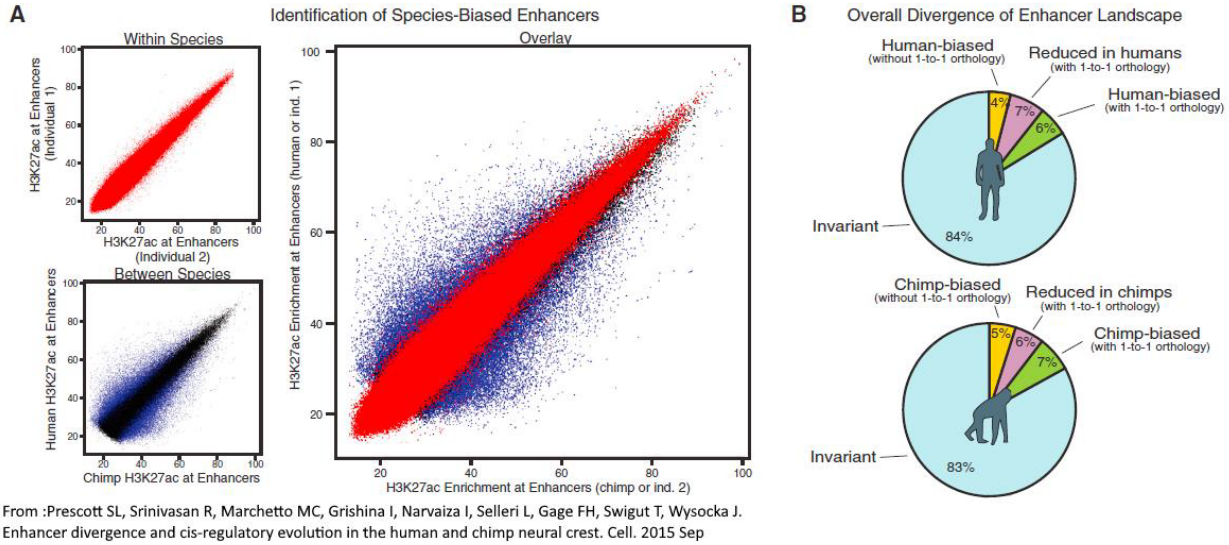
sequence slightly upstream of the coding region (75). Second, during elongation, Pol II travels along the DNA while writing a complementary strand of RNA from which non-coding introns within the gene are spliced out, directly ligating the protein coding exons. Third, Pol II will detach from the DNA and free the newly transcribed RNA, capped and polyA-tailed, to later be translated (76).

As the process of transcription became increasingly studied, the biochemical role of DNA and its modifications, such as the classical epigenetic modification, methylation of cytosine, were shown to be essential for the regulation of transcription, not simply as a coding element. Indeed, islands of dinucleotide CpG are present on approximately 70% of annotated promoters (77). The methylation of cytosine in these islands is associated with the silencing of gene expression (78). Further, the essential nature of proper DNA methylation programming was shown by the embryonic lethality caused by the knock-out or overexpression of DNMT1, necessary for the maintenance of DNA methylation (79,80). Interestingly, DNA methylation is also present on parts of the genome that cannot be transcribed as a mRNA, defined as non-coding DNA. Of all the base-pairs of the human genome, about ~99% are non-coding (81). This vast majority of our genome was previously termed "junk DNA", as it was thought to be intrinsically functionless (82,83). This belief has since been proven exquisitely incorrect.

Non-coding DNA and Chromatin States

Non-coding regions of the genome are associated with a diverse interrelated network of biological functions, essential for evolution, development, and homeostasis. A wide array of non-coding genetic mutations is associated with cancer (84) and other diseases (82). Non-coding changes between species, such as humans and chimps, are at the heart of characteristic phenotypic differences (85) (Figure LR3). Similarly, highly conserved non-coding elements between vertebrates are also critical for development (86). Additionally, proper expression of

non-coding RNA transcripts also plays a critical role in development (87,88) and to prevent cancer or other diseases (89,90). As this multitude of built-upon studies made clear the essentiality of "junk DNA", parallel investigations of the distinct chromatin state found across the genome helped to distinguish and understand the function of the different types of non-coding elements.

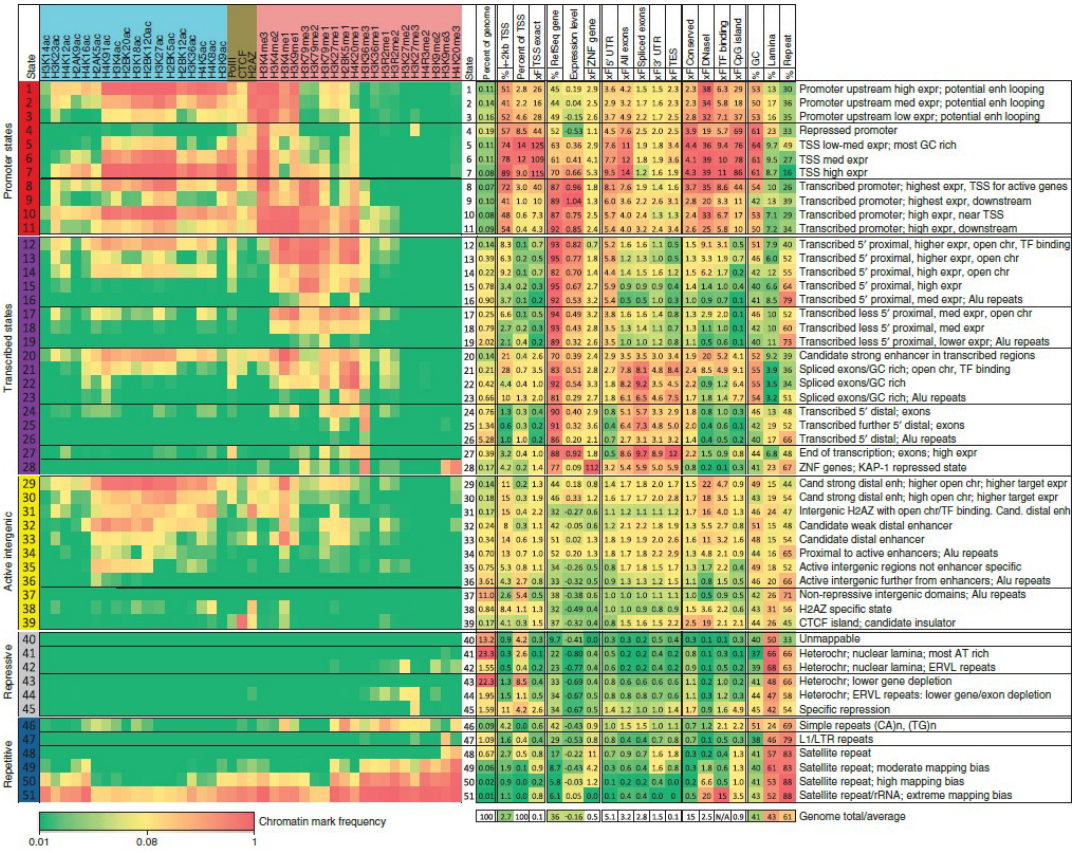


From :Prescott SL, Srinivasan R, Marchetto MC, Grishina I, Narvaiza I, Selleri L, Gage FH, Swigut T, Wysocka J. Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. Cell. 2015 Sep

Figure LR3: (A) Divergence between species of enhancer-associated histone mark H3K27ac. (B) Divergence of enhancer landscape between human and chimp.

One such type of non-coding element is the enhancer. Enhancers are regions of the genome that, when active, promote the activity of surrounding genes. Active enhancers are identifiable by the specific characteristics of the chromatin at their loci. A nucleosome is defined as a complex of 8 core histone proteins encircled by ~150bp of DNA (91). Nucleosome density is one indicator of enhancer activity. Low nucleosome density, identifiable genome-wide by an increased chromatin accessibility using ATAC-Seq (92) or DNAse-Seq (93), is associated to active or actively transcribed regions, while compact nucleosome placement is associated with inactivation and silencing of gene expression or enhancer activity (94). These "open" enhancers are thought to facilitate transcription factor recruitment.

Besides the placement and density of nucleosomes, the post-translational modification of their histone tails is another indicator of activity and function. Indeed, acetylation of the histone tails, such as acetylation of histone 3 lysine 27 (H3K27ac) or H3K9ac, is associated with active enhancers and promoters. Alternatively, specific methylation status can distinguish enhancer from promoters. H3K4me1 is usually a mark of poised enhancers, while H3K4me3 is found on active promoters (95). The range of histone posttranscriptional modifications is varied and specific combination of which are unique to distinct non-coding or coding region of the genome, depending on their activity (95) (Figure LR4). Interestingly, following the identification of enhancers, came the discovery of insulator elements, capable of blocking the effect of an active enhancer on adjacent promoters.



From: Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010 Aug

Figure LR4: Association of different combinations of chromatin marks and states to distinct characteristic or functions of the loci.

The classical insulator example is the Imprinting Control Region (ICR) around the H19 and IGF2 loci. Biallelic expression or repression of any of these genes results in severe developmental diseases, such as the Beckwith-Wiedemann syndrome (96). Imprinting refers to the maintenance of DNA methylation at distinct regions in the paternal and the maternal allele. The ICR on the paternal allele is methylated, inactivating its insulating potential, and enabling the downstream enhancer to interact and activate the upstream IGF2 genes. However, on the maternal allele, the ICR is unmethylated, allowing it to block the enhancer from interacting with the upstream IGF2 and limiting its activity to the downstream H19 genes (97). Interestingly, the mechanism of how DNA methylation blocks the insulating potential of the ICR is through the blockage of CCCTC-binding factor (CTCF) binding (98), a protein first discovered as a negative regulator of chicken c-myc expression (99). The mechanism by which non-coding loci, such as enhancers, and CTCF binding could influence the expression of gene thousands or millions of nucleotides away would later be explained through chromatin conformation.

In 2002, Dekker et al. published a study in which they developed 3C or Chromatin Conformation Capture (100). In this assay, chromatin is crosslinked and sparsely digested by a restriction enzyme to create a multitude of floating ends that are then stochastically ligated together. These manipulations produce fragments of DNA with a 3' end mapping anywhere on the genome and a 5' end mapping on a locus that was in physically proximal to it (100). The following years would see the rise of numerous variants of 3C, the most popular of which being Hi-C, introduced by Liebermann-Aiden et al. in 2009 (3) (Figure LR5), used to map DNA-DNA contact across the genome. Expanding on this newfound ability to precisely study genome-wide three-dimensional DNA organization led to numerous studies that, together, discovered that our genome is not just a two-dimensional sequence containing a parsimony of coding sequences hidden within an ocean of

random inconsequential assortments of bases, but a complex, hierarchically organized framework, programmed for precise control of transcription, development, and homeostasis.

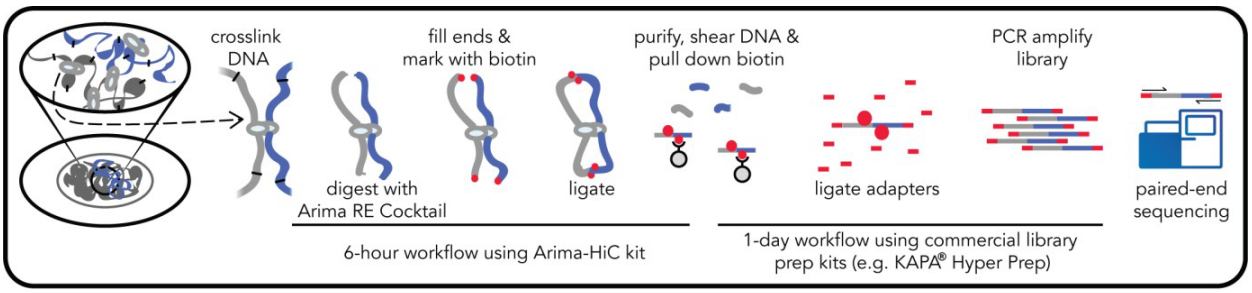


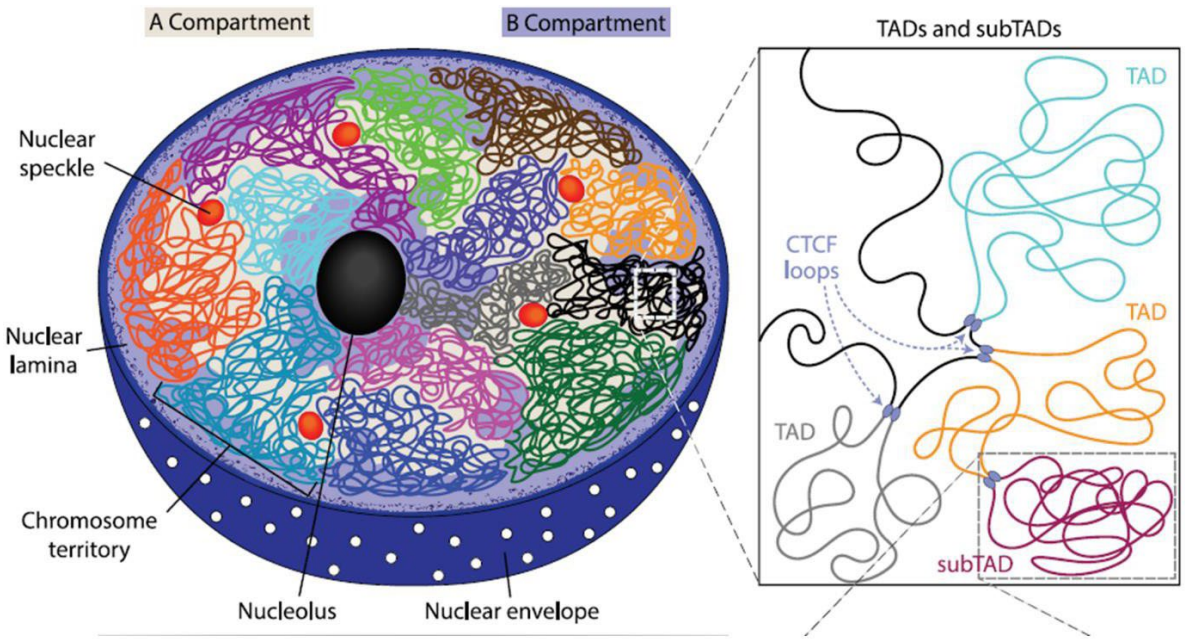
Figure LR5: Example of Hi-C experiment workflow.

Distinct Hierarchical Levels of Three-Dimensional Genomic Organization

The Passive Formation of Compartments

Within the nucleus, DNA is hierarchically organized on multiple levels. At the base of the hierarchy are the chromosome territories, in which each chromosome will occupy a specific section of the nucleus (101). Within these territories, active euchromatin will segregate from silent heterochromatin to form A and B compartments, respectively (3).

More precisely, compartments are the preferential clustering of continuous and non-continuous segment of chromatin of similar states. In the classical definition of compartment, these interactions range in the millions of base pairs (3). Compartments are thought to be formed passively by the opposing biophysical properties of active and inactive chromatin (102), such as phase separation (103). In addition to compartments, chromatin is also organized in Topologically Associated Domains (TADs). Contrarily to compartments, TADs are continuous, actively maintained, regions of increased short-range interactions, on average, ranging from one million base pairs to one hundred thousand base pairs (4) (Figure LR6).

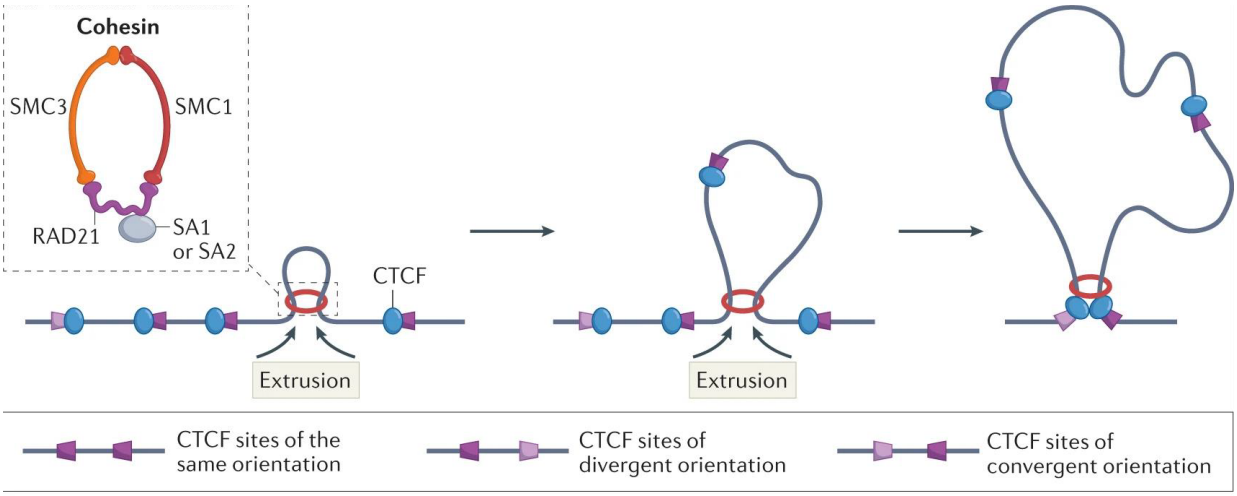


From : Goel VY, Hansen AS. The macro and micro of chromosome conformation capture. Wiley Interdiscip Rev Dev Biol. 2021 Nov

Figure LR6: Representation of the hierarchical organization of the chromatin in the nucleus.

The Active Extrusion of Topologically Associated Domains (TADs)

TADs are formed by loop extrusion, an active multistep process (104). First, the cohesin complex is loaded on DNA and handcuffs it in an ATP-dependent manner (105). Then, the strands of DNA are extruded out of the ring-like complex. This step is mediated by ATP-dependent cohesin activity. However, the precise mechanism for this proteomic "reel" is still ambiguous (5,106,107) and could also be promoted by transcription-induced supercoiling (107,108). Finally, extrusion is stopped once a second cohesin complex is encountered (109) or, more commonly, when both sides of newly formed loop reach convergent sites bound by the insulator protein CTCF (104,110) (Figure LR7). As such, loop extrusion allows for the formation of a continuous region of DNA for which internal interactions are promoted, while outside interactions are insulated.



From : Zhang Y, Zhang X, Dai HQ, Hu H, Alt FW. The role of chromatin loop extrusion in antibody diversification. Nat Rev Immunol. 2022 Sep

Figure LR7: Representation of the cohesin complex and loop extrusion

Loop extrusion also forms smaller domains within TADs, termed subTADs, that are also insulated from adjacent genomic loci. Although subTADs may be formed through the same process and play the same structural function as TADs, their smaller size allows for a more direct and precise regulation of coding and non-coding sequences within them. For example, most TAD boundaries are often bound by multiple redundant CTCF binding sites and colocalize with tRNA or housekeeping genes that are constitutively active (111,112). Unlike TADS, which are generally consistent across tissue types, subTADs reorganize themselves to allow for dynamic and precise transcriptional control of genomic loci (113). Consistent with this concept, long range chromatin interactions re-organized during serum starvation are dependent upon interactions between CTCF and binding partners (114) and a gain of CTCF-mediated interactions at the subTAD level have been correlated with gene expression (115). Additionally, the essentiality of CTCF and cohesin in embryonic development (17,116,117) further supports TADs and subTADs organization as essential for proper gene expression.

Smaller loops may also be observed within TADs and subTADs that are less “structural” in nature as they do not form insulated domains. These chromatin loops often represent enhancer-promoter

contacts, which can be formed by multiple mechanisms, some independent of cohesin loop extrusion (7).

Interplay between Topological Domains and Enhancers

By promoting internal chromatin contacts and insulating from external interactions, TAD and subTADs fine tune the precision and potency of regulatory non-coding elements, such as enhancers. These enhancers can be distributed sparsely around their target genes or clustered together in super enhancers (SE) (118). The clustering of enhancers and their additivity on a single promoter will usually result in a stronger upregulation of gene transcription (118,119). Indeed, active enhancers are bound by transcription factors, co-activators, such as BRD4 and Mediator (118,120) and epigenetic regulators, such as histone acetyltransferases (121), with critical roles in all aspects of transcription. By bringing these essential transcriptional actors in close proximity to a promoter, enhancers create an environment prone to transcriptional activity at that particular locus (119).

However, how enhancers find and interact with their promoters is still debated. ENCODE database counts around 668,000 candidate enhancer-like sequences (122). Therefore, a single gene is often regulated by a panel of enhancers, with distinct enhancers being specifically involved in distinct cell types or diseases (118,121,123). For instance, when deleting enhancers specific to microglia, which are enriched for risk variants for Alzheimer Disease, the expression of their target genes, such as BIN1, was uniquely reduced in microglia. However, in neurons and astrocytes, the gene expression of BIN1 was unchanged, as its expression is reliant on distinct enhancers in these cells (123). As such, understanding the epigenetic mechanisms driving and regulating enhancer-promoter interactions (E-P) is critical to gain pertinent insight on the development and disease progression.

To better understand enhancer-promoter interactions, multiple models have been studied and are

likely to be involved in this process, be it in parallel or in complement of each other. Enhancers can find and start interacting with their targets during the loop extrusion process, in a model termed enhancer scanning and critical for V(D)J recombination (124,125). Additionally, E-Ps could be analog to micro-compartments: being driven by the physical properties of their chromatin and surrounding effectors, independently of loop extrusion. Indeed, E-Ps can still be formed when essential members of the cohesin complex are knocked out (7). However, independently of which recent model of enhancer-promoter interaction is used, subTADs and TADs do play an essential role in regulating and confining their interactions.

By definition, TADs and subTADs promote interactions within them, while insulating interactions from outside. Although enhancer-promoter interactions can happen independently of topological domains, TADs and subTADs guide the stochastic nature of E-P interactions by strongly promoting intradomain interactions (7,126). Alternatively, in the enhancer scanning model, TFs also guide the recruitment of the cohesin to start the loop extrusion process at a specific enhancer. As the chromatin is reeled through the cohesin ring, the loaded enhancer scans through it and retains contact with intradomain promoters and enhancers to cluster them together until extrusion is stopped by a convergent CTCF sites (124). In both models, the formation of TAD and subTADs and the proper definition of their boundaries, usually defined by CTCF binding, are necessary to limit what is included and excluded from each domain, therefore guiding the specificity of E-P interactions.

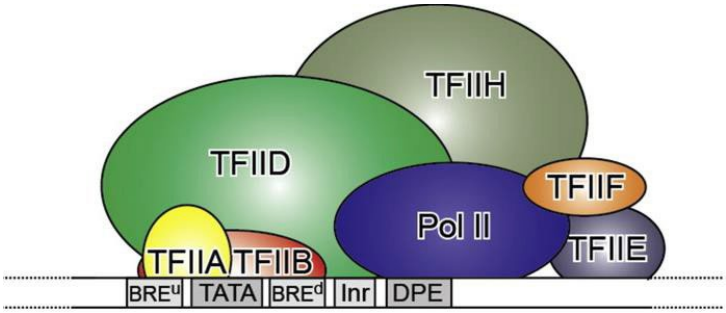
Current Model of Transcription and Epigenetic Regulation

Although there is still much to learn about transcription, the precision and complexity of our understanding of this mechanism evolved markedly since the early days of the central dogma. Indeed, most transcriptional activities are now modeled as a 6-steps process, each with multiple

layers of epigenetic regulation built, in part, upon the concepts described above.

Preinitiation, Initiation and Pausing

First, the RNA polymerase II will be recruited to the promoter in the Preinitiation Complex (PIC) (127). Preinitiation starts with the recruitment of TFIID, a General Transcription Factor (GTF) multiprotein complex (128). TFIID recruitment is strongly influenced by epigenetic events. H3K4me3 will guide TFIID on promoters (129), while enhancer-promoter interactions will promote its recruitment through the Mediator complex (130,131) and histone acetylation (132). Once recruited, TFIID will help the recruitment of other GTF and Pol II itself, readying the machinery to move on to initiation (127) (Figure LR8). Interestingly, other transcription factors are also able to mediate the recruitment of GTF, such as CTCF promoting the recruitment of TFII-I at transcriptional start sites (133).



From: Robinson PJ, Trnka MJ, Bushnell DA, Davis RE, Mattei PJ, Burlingame AL, Kornberg RD. Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex. Cell. 2016 Sep

Figure LR8: Simplified representation of the Preinitiation Complex.

Second and third are the initiation and pausing steps. To kickstart initiation, CDK7, a subunit of TFIIH, will phosphorylate serine 5 and 7 in the 52 YSPTSPS repeat of the C-terminal of Pol II (134,135). Serine 5 phosphorylation (Ser5-P) will destabilize the interaction between Poll II and the Mediator complex, facilitating its escape from the promoter (136). Ser5-P Pol II will advance on the gene and start synthesizing the mRNA, that will be capped co-transcriptionally while the, approximately, first thirty nucleotides are assembled (137,138). About twenty to sixty base pairs downstream of the initiation site, the transcriptional complex will be paused by DSIF, NELF and

PAF1 complex (139,140). Pausing is essential for transcription to continue on properly and its disruption will, among other things, influence the positioning of nucleosome and the presence of histone marks associated to elongation, such as H3K36me3 (140). As a majority of genes outside of heterochromatin are bound by paused Pol II downstream of their promoters (141), transcriptional pausing is also an additional opportunity for regulating and fine-tuning the expression of a gene. Indeed, further interactions with active enhancers will promote the transition from pausing to elongation, as BRD4 plays an important part in the recruitment of P-TEFB and the Super Elongation Complex (142).

Elongation, Reinitation and Termination

To push Pol II toward the fourth step, elongation, P-TEFB, a protein complex containing the kinase CDK9, will phosphorylate NELF and DSIF, displacing them from the transcriptional machinery (143,144). P-TEFB will also phosphorylate Ser2 on Pol II CTD, which is required for proper elongation (145). Similar to the previous steps of transcription, epigenetic regulation of Pol II speed of transcription is also critical, from the start of elongation to polyadenylation and termination (146). For example, mRNAs are spliced simultaneously with elongation and the presence of an epigenetic events along the transcript, such as CTCF and chromatin loops, can slow down the RNA polymerase and influence which exons will be integrated in the mRNA (147,148). Since the positioning and binding of CTCF is influenced, in part, by DNA methylation and nucleosome placement (149), the epigenetic landscape and chromatin structure along a gene, as it was on its promoter, is essential to the regulation of its expression, be it the first wave of transcription or the multiple potential subsequent ones.

Before the end of the sixth and final step of transcription, the termination, the transcription machinery will, most of the time, be recycled on the same gene for multiple rounds of transcription in a process termed reinitiation (150). Here again, chromatin conformation can

promote this process by bringing in close proximity to the transcriptional start and end sites (9). Although this short section only skimmed the surface of how transcription is regulated at each step, it is clear that chromatin conformation and epigenetic landscape play distinct, complementary, and essential roles in proper gene expression. Therefore, it is not surprising that a protein, such as CTCF, directly guiding some of these epigenetic events can be so deeply intertwined in the regulation of transcription and thus, cellular health and homeostasis.

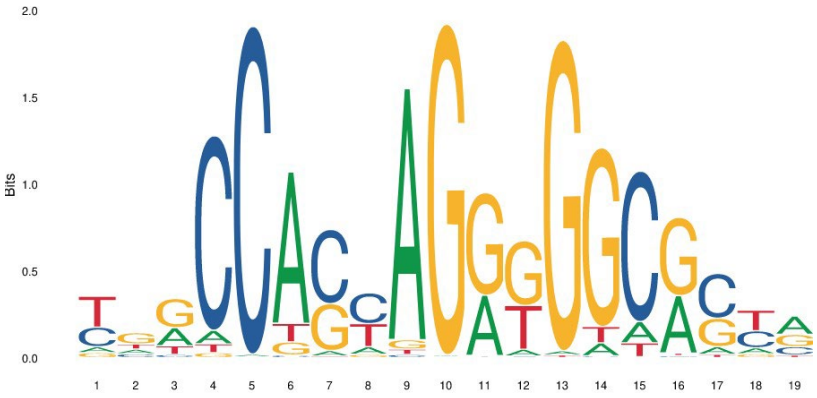
Understanding CTCF in health and diseases

Throughout this literature review, the multifunctional epigenetic regulator CTCF has already been mentioned multiple times: as an actor in double-stranded beak DNA damage repair (64); as an insulator of enhancer-promoter contact (98); as the factor that blocks cohesin loop extrusion to define the boundaries of topologically associated domains (124); as a transcription factor helping the recruitment of TFII-I (133); as a barrier for transcriptional elongation mediating alternative splicing (147,148). Yet, it was not aforementioned that the loss of a single allele of CTCF or its mutation may promote tumor onset and progression, which will be discussed later. Due to the polyvalence of CTCF, the biological and epigenetic mechanisms behind the potential oncogenic impact of its deficiency are still unresolved.

CTCF Binding to DNA

The multiple roles of CTCF within the nucleus are primarily dependent on its binding to DNA, for which the affinity, and specificity, is regulated on multiple levels. First, a major aspect of CTCF binding to DNA is its direct protein-DNA interaction. CTCF interacts directly with DNA with its complex 11 zinc-finger domain, located between its mostly unstructured N and C terminals. Further, not every individual zinc finger (ZF) of CTCF is binding to DNA simultaneously. Indeed, studies of the crystal structure of CTCF revealed that ZF3 to ZF7 are required for proper binding to the CTCF core binding motif of the consensus DNA recognition element (33,34,38). The core

binding motif of CTCF is present in approximately ninety percent of CTCF binding sites (35-37) and is represented in Figure LR9. Therefore, when deleted, ZF3 to ZF7 have the biggest influence on CTCF binding to DNA, compared to other ZFs (33,38). Although not binding to the core binding motif, ZF10 and ZF11 are thought to bind an upstream extension of CTCF motif stabilizing it on the DNA (40). Despite not being shown to directly bind DNA, ZF8 and ZF9 act as a spacer to ensure the proper positioning of the neighboring zinc fingers (38). Interestingly, the crystal structure of ZF1 with DNA could not be obtained (38). The specific deletion of ZF1 was shown to be less impactful on DNA binding than other CTCF zinc fingers. It is currently unknown whether ZF1 directly interacts with DNA by binding to conserved nucleotides beyond the core motif, such as the one identified by Rhee and Pugh (40), or whether its influence is mediated by its affinity for RNA.



From : JASPAR 2022 (MA0139.1)

Figure LR9: Classical CTCF binding motif

Similarly to how they bind DNA, zinc-fingers are also capable of binding RNA. CTCF's zinc-fingers are no exception. Indeed, RNA binding activity has been detected for ZF1 and ZF10, although no RNA sequence specificity was investigated (39). Further, RNA-CTCF interactions have been shown to mediate CTCF spatial recruitment to the DNA (151) or binding to specific

sites on the genome (39). Interactions with HOTTIP or MYCNOS guided or strengthened CTCF interaction at distinct sites for each ncRNA (152,153). Oppositely, the interaction of ncRNA JPX with CTCF removed it from key sites relevant to X chromosome inactivation (154). But, as mentioned before, non-coding RNAs are not the only epigenetic events that can mediate CTCF binding to the DNA.

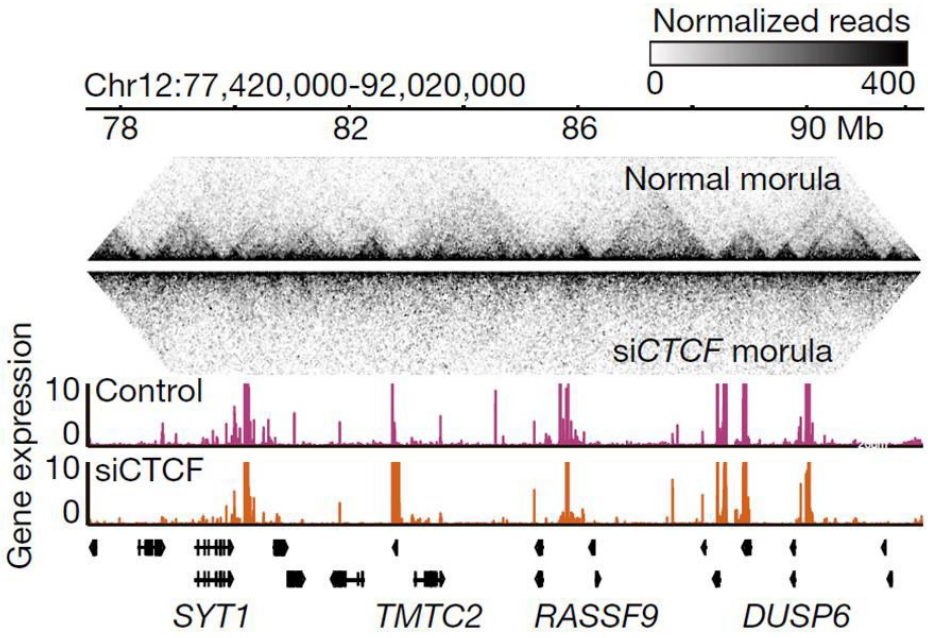
As shown at the IGF2/H19 locus, DNA methylation can block the binding of CTCF to the DNA (98). Specifically, methylation at the second cytosine of CTCF core binding motif was shown to display the most significant ability to block CTCF binding (38). Beside DNA methylation, the positioning and state of the nucleosomes also impacts CTCF binding to DNA. When nucleosomes are densely packed, they will block the binding of CTCF, as is the case for most transcription factors. Alternatively, CTCF influences nucleosome positioning and composition, as neighboring nucleosomes are spaced consistently around CTCF binding sites and are enriched for H2A.Z sub-unit (155). Further, CTCF residence time on the chromatin seems to be increased when it is at the boundary of a cohesin mediated loop (156).

Overall, there are a wealth of factors involved in orchestrating the binding of CTCF to DNA, and this coordination is essential for homeostasis and development.

CTCF and Chromatin Conformation in Normal Homeostasis and Development

Due to the essentiality of proper organization of chromatin during development (126), complete loss of CTCF is embryonic lethal (16,17). Indeed, as pluripotent stem cells advance toward differentiation, they gain a stricter, or less dynamic, topological organization of their chromatin to regulate their cell-type specific identity and expression. However, the absence of CTCF hinders this process (12) (Figure LR10). CTCF related development defects also arise when CTCF binding sites are mutated or when CTCF carries certain point mutation. For example, acheiropodia, a genetic limb disorder, can be caused by a heritable deletion of CTCF binding

sites mediating enhancer-promoter interaction at the SHH locus (157). Additionally, genetic defects of CTCF, from partial deletions to missense mutations, were associated to a wide spectrum of neurodevelopmental disorders in human patients (158). In mice, the effects of CTCF dysfunctions on neurological development were associated to its impact on differentiation ratio and increased apoptosis due to overexpression of the p53 target gene PUMA (159). As multiple recent studies showed the importance of CTCF in memory formation (160,161), the notion that proper CTCF function is critical for diverse neurological processes is strengthening. As is the importance of CTCF dysfunction in cancer progression.



From: Chen X, Ke Y, Wu K, Zhao H, Sun Y, Gao L, Liu Z, Zhang J, Tao W, Hou Z, Liu H, Liu J, Chen ZJ. Key role for CTCF in establishing chromatin structure in human embryos. Nature. 2019 Dec

Figure LR10: CTCF knockdown disrupts the establishment of TADs and subTADs in early embryo.

The Present, yet Unexplained, Role of CTCF in Cancer Initiation and Progression

In mice hemizygous for CTCF, the mutated animals displayed increased risk of hematological tumors (19). This observation has been found relevant to humans. In Down Syndrome related Acute MegaKaryoblastic Leukemia (DS-AMKL), CTCF deletions or mutations are found in

20% of all cases (21). Here, the loss of CTCF is thought to be important for clonal evolution to more aggressive phenotypes following GATA1 mutations (21). Investigations of the effect of reduced CTCF levels, through auxin-induced degradation, in the B-Cell Acute Lymphoblastic Leukemia (B-ALL) SEM cell line, revealed that chromatin accessibility was increased, while intra-TAD interactions were disrupted (162,163). However, the resulting changes in RNA expression were described in a fairly superficial manner and explained by the increased activity of other transcription factors, such as MYC (162,163). This model has the disadvantage that key chromatin rearrangements necessary for tumor development have already taken place in the B-ALL line. The causal relationship by which the repressed function of CTCF affects differentiation and therefore promotes hematopoietic cancer, is a potential model to explain the results of the B-ALL studies and supported by the study in DS-AMKL (21). If it is the case, then the mildness of CTCF degradation in B-ALL would be expected, as the impact of low CTCF on hematopoietic cancer would be the earliest steps of its initiation and not in its progression. However, this hypothesis has yet to be validated, as are most models of the impact of CTCF LOH or point mutations in epithelial tumors.

According to TCGA 2018 dataset, CTCF LOH is present in about sixty-three percent of all breast tumors and about thirty percent of endometrial cancer, compared to less than ten percent for leukemia or lymphoma (22). Although a significant fraction of CTCF deletion or mutations are identified in breast cancer, there is still no clear epigenetic or biological information regarding how this defect impacts cancer initiation or progression. Further, while it has been shown in other models that fluctuations in CTCF levels may impact chromatin looping (162), this has not been formally examined in human breast models (23). Thus, it remains unclear whether transcriptional networks and topological features may be deregulated in breast

epithelium undergoing CTCF LOH.

Similar questions are also still unanswered about the mechanistic link between CTCF point mutations and cancer (Figure LR11). For instance, the biological impact of the mutation of the first zinc finger (ZF1) of the epigenetic regulatory protein CTCF, such as the H284N mutation, exclusive to breast cancer and prevalent in hormone resistant breast tumors (31), has remained elusive.

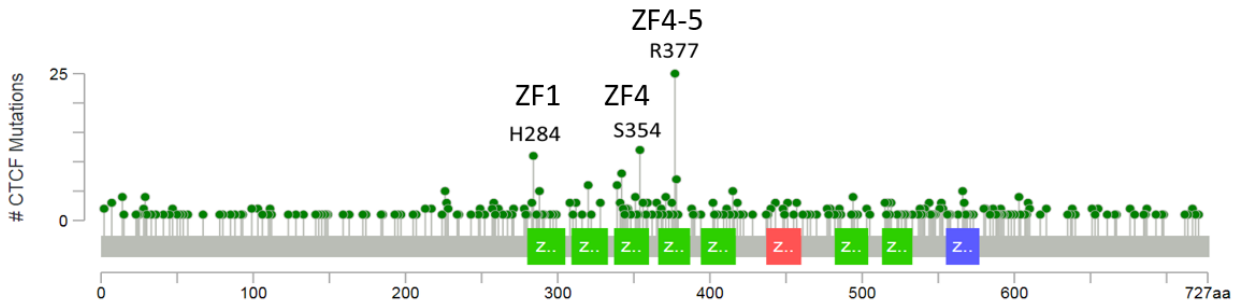


Figure LR11: cBioPortal Lollipop Plot of the most common CTCF point mutations in cancer.

Interestingly, CTCF mutations are among the most enriched in metastatic breast tumors compared to primary tumors, behind only ESR1 mutations (164). In contrast to oncogenic mutations located within CTCF ZF3-7, which have been partially studied in cellular model of endometrial cancer (32), CTCF ZF1 remains uncharacterized, and its biological or epigenetic impact remains unknown. As is the influence of CTCF ZF1 mutation (ZF1M) on CTCF ability to bind DNA.

Although the truncation of CTCF ZF1 was shown to alter RNA dependent binding of CTCF to specific sites, the H284N mutation did not display such function (39). It is known that bases outside the core binding motif modulate CTCF binding (33,40), but it remains unknown whether CTCF ZF1 mutations regulate binding to an extended motif, or alternatively influence CTCF binding affinity through impeding its interaction with non-coding RNAs (39). However, computational tools required to answer this question are lacking, as current tools do not allow for

the surveying of long, complex motifs, or direct quantification of motifs that discriminate subtle differences between motifs (discriminative motif comparison).

Motif Analysis

Beside a better understanding of the biological and epigenetic mechanisms underlying cancer progression, advancements in bioinformatics also promote innovation in the field of oncology. Biological investigation into pathologies driven by aberrant Transcription Factor (TF) activities or non-coding mutations located at promoters, enhancers, or chromatin domain boundaries (25-28,165) might benefit substantially from computationally driven motif analysis.

For the task of identifying DNA motifs, motif discovery tools, such as GADEM (166) or MEME (167), coupled with DNA motif databases for TFs, such as JASPAR (168), CisBP (169) and UniPROBE (170), are widely used. By comparing the immediate DNA sequence surrounding an oncogenic, non-coding, somatic mutation to an online TF motif database, one can predict which TF or family of TFs is likely to experience hindered DNA binding at this locus. From this prediction, mechanisms of oncogenic progression may be surmised. For example, multiple oncogenic non-coding variants were identified to colocalize with the core recognition motif of CTCF (171-173). When coupled with available Hi-C datasets, motif driven hypotheses provide mechanistic insights into the role of these non-coding variants through altered chromatin looping at key, actionable, oncogenes (174). Although current tools are well-suited to detect the presence of a known binding motif in the examples above, their intrinsic limitations hinder their predictive abilities when subtle modifications or extensions of known binding motifs are involved.

Motif discovery is studied in diverse biochemical environments, each with their pros and cons. *In Silico* DNA motif discovery tools can identify binding motifs by computing a position weight matrix (PWM), derived from normalized relative frequencies of each nucleic acid base, within the

aligned TF binding sites identified by experiments such as ChIP-seq (175). Compared to *in vitro* techniques, such as protein binding microarrays (PBM), which test for motifs directly involved in TF-DNA interaction (176), motifs discovered by *in-silico* analysis of ChIP-Seq datasets are influenced by cellular conditions. For instance, both methods will identify a similar motif for a TF whose binding is primarily driven by direct DNA-protein interactions. Alternatively, if a TF's motif is acutely influenced by chromatin state (177), cofactor interaction (178) or recruitment by another TF (179), the identified motif will be markedly different if discovered from a ChIP-Seq or a PBM experiment. As such, these two complementary approaches of motif discovery are competent to predict the primary recognition motif of a given TF, be it a direct or indirect DNA interaction. However, both fail to identify underrepresented motif variability. Subtle changes or extensions of the core motif are statistically overlooked by the strict thresholding required for motif discovery from ChIP-Seq. Further, condition-dependent motif alterations cannot be detected *in-vitro*, as current tools are programmed to identify motifs within a given group of sequences, compared to background or a complementary set of sequences, but not to compare the motifs, and surrounding nucleotides, themselves.

Due to these limitations, defining the impact on DNA recognition of mutated TFs, such as the aforementioned CTCF H284N mutation, or mutated co-factors, remains a challenge.

Body

Chapter 1: Mechanistic Investigation of CTCF Loss of Heterozygosity in Cancer (180)

Low CTCF expression promotes invasiveness in diverse breast cancer models

CTCF single allele deletions are prevalent in a majority of breast tumors (22). To better understand the consequences of this genetic aberration, I first surveyed conditionally reprogrammed cell lines from Patient Derived Xenografts from triple-negative breast cancer patients, termed PDX, harboring loss of one allele of CTCF (Supplementary Figure 1.1A) and with CTCF expression equivalent, or lower, than in our previously established MCF10A CTCF+/- mammary epithelial cell line, harboring a knockout of one CTCF allele (64) (Figure 1.1A). The low levels of CTCF in my cell lines tightly mimic those observed *in vivo* from a panel of tumor xenografts derived from a distinct set of triple-negative breast cancer patient's tumors (Supplementary Figure 1.1B). Thus, these data support these *in vitro* models as relevant systems to study the effects of low CTCF on oncogenic phenotypes.

Multiple clinical reports link 16q22.1 deletion, where CTCF resides, with metastasis (181-183). Therefore, I investigated the impact of altered CTCF levels on the invasive capacity of cells, a critical step in cancer progression. For this, I employed matrigel transwell invasion assays and CTCF addback to the PDX cell lines carrying low levels of CTCF. Following lentiviral addback of HA-CTCF (Figure 1.1B), the increased CTCF expression led to reduced invasiveness of all three PDX cell lines tested ($p = 0.0031$, 0.0084 and 0.0015 for HA-CTCF #1-3 against their respective Control) (Figure 1.1C), despite the varied mutational background of each cell line.

Next, I validated the impact of CTCF levels on invasiveness by comparing the effects of shRNAs against CTCF (shCTCFs) or scrambled shRNA (shCTL) in MCF7 cells, a widely used, CTCF WT, breast cancer cell line. Consistent with the inhibition of invasiveness brought about by

CTCF addbacks, the reduction in CTCF levels resultant from the shCTCFs (Supplementary Figure 1.1C) led to a significant increase in MCF7 invasiveness compared to shCTL ($p = 0.021$ and 0.0051 for shCTCF #1 and #2) (Supplementary Figure 1.1C). Altogether, these studies confirm a relationship between low CTCF expression and increased invasiveness in distinct breast cancer models.

CTCF Single Allele Knockout induces oncogenic phenotypes in mammary epithelial cells

Next, to study the effect of altered CTCF expression in mammary tissue independently of the heavy mutational burden of breast cancer models, I investigated the effect of low CTCF expression in the non-transformed breast epithelial cell line MCF10A carrying a single allele knockout of CTCF (MCF10A-CTCF+/-), previously edited via CRISPR-Cas9 (64) (Figure 1.1D).

Using this model, I screened for several classical oncogenic phenotypes including cell invasion, altered morphology, increased proliferation, and deregulated mammosphere growth. Similar to what I observed in breast cancer models, the loss of one copy of CTCF increased the capacity of MCF10A to invade through a matrigel matrix. CTCF+/- cells readily invaded through matrigel at a rate significantly higher than CTCF+/+ control (CTL) MCF10A cells ($p = 0.0066$ and $p < 0.0001$ for CTCF+/- #1 and #2) (Figure 1.1D). To support a direct link between the loss of CTCF and the acquired invasiveness, I carried out lentiviral-mediated addback of HA-CTCF within our MCF10A models. As in the PDX cell lines, the restoration of CTCF levels was able to significantly reduce the invasiveness of CTCF+/- cells (Figure 1.1E, Supplementary Figure 1.1D). Phenotypically, the morphology of the CTCF +/- cells in two-dimensional culture was markedly similar (Supplementary Figure 1.1E), while their proliferation rate was slightly reduced compared to CTL cells (Supplementary Figure 1.1F). MCF10As spontaneously form organized hollow ductal acinar-like structures in three-dimensional (3D) culture (184),

presenting the opportunity to study the impact of lowered CTCF levels on the regulation of unanchored growth. Strikingly, MCF10A CTCF+/- acini form significantly larger, less hollow, ($p < 0.0001$) and structurally deformed (Supplementary Figure 1.1G/F) mammospheres compared to CTL counterparts.

Together, these results point towards a potentially important role for the loss of heterozygosity of CTCF in cancer progression, as it promotes disorganized 3D growth and invasiveness, two strongly linked oncogenic abilities critical for tumors to progress from benign, to advanced stages of cancer.

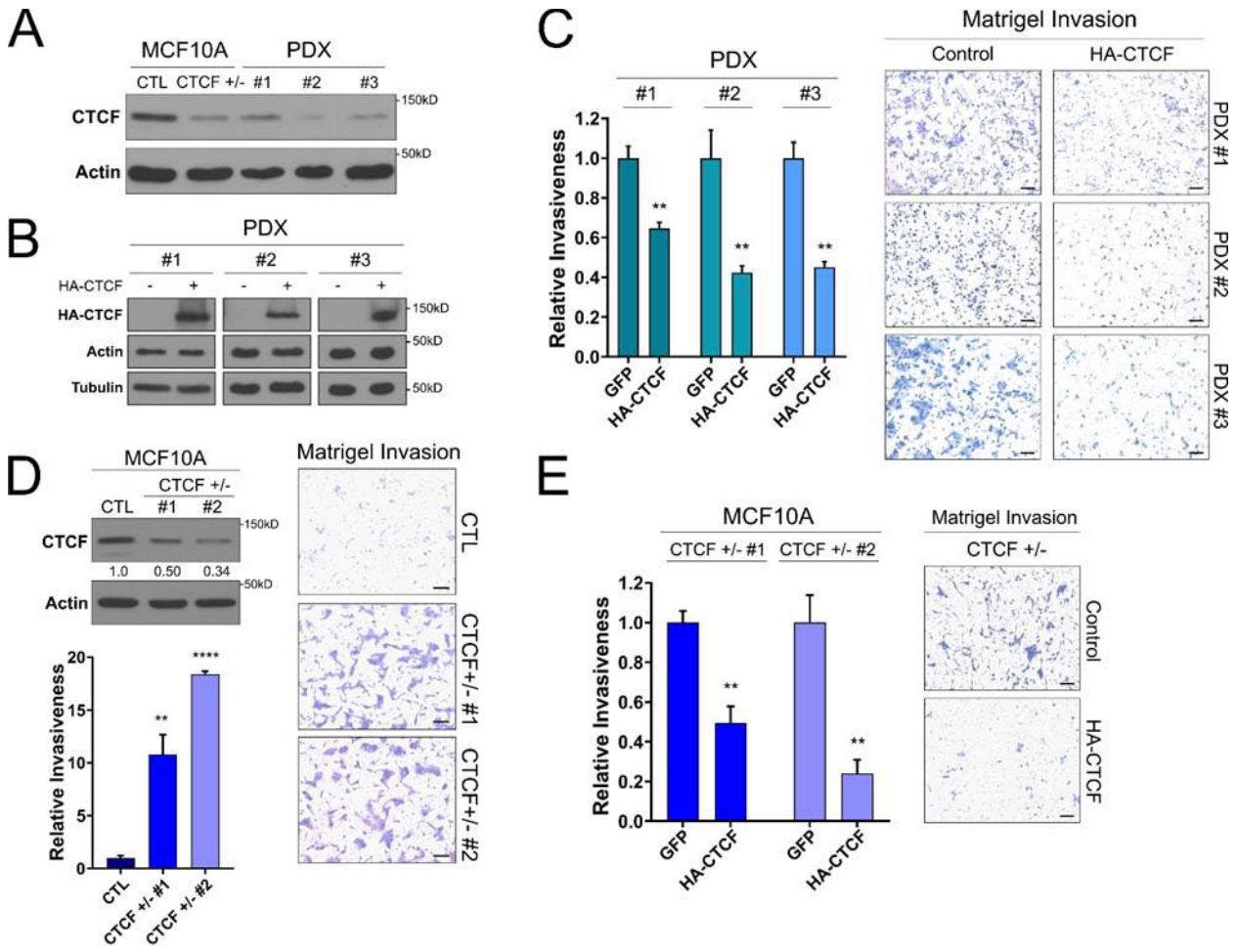


Figure 1.1. CTCF loss of heterozygosity promotes invasiveness and unorganized growth in distinct breast epithelial models. (**A**) Western Blot showing low levels of CTCF, similar to the CTCF+/- MCF10A, in the PDX cells. Loading control : Actin. (**B**) Western Blot of ectopic HA-CTCF expression in PDX cell lines. Loading control : Actin and Tubulin. (**C**) Decrease in relative invasiveness of HA-CTCF PDXs to respective GFP controls (mean ± SEM). p = 0.0031, 0.0084 and 0.0015 for PDX #1, 2 and 3. (**D**) Western Blot of low CTCF levels in the CTCF+/- compared to CTL MCF10A. Quantification of relative CTCF band intensity in CTCF +/- to CTL. Loading control: Actin. Bar Chart of the increased relative invasiveness of CTCF+/- to CTL (mean ± SEM). p = 0.0066 and p< 0.0001 for CTCF+/- #1 and #2. (**E**) Decreased relative invasiveness of CTCF+/- MCF10A with HA-CTCF addback relative to their respective GFP control (mean ± SEM). p-value = 0.0013 for both CTCF+/- #1 and #2.

Reprogramming of transcriptional networks leads to activation of oncogenic signaling in CTCF+/- cells

To gain insight into the mechanism whereby CTCF+/- cells acquire the oncogenic phenotypes described above, we carried out an RNA-Seq to compare global gene expression profiles of MCF10A CTL and MCF10A CTCF+/- cells. Using DESEQ2 (185), I detected 2976 and 2893 genes that were significantly transcriptionally altered in the CTCF+/- #1 and #2, respectively, compared to CTL (Basemean > 100, abs (log2FC) >1, adjusted p-value < 0.05). The transcriptional changes were highly reproducible as the respective changes in gene expression in each CTCF +/- clone compared to CTL correlated very strongly ($r = 0.9937$, $p < 0.0001$) (Supplementary Figure 1.2A). Of the 2765 genes commonly altered in both CTCF+/- clones, a slight majority of 1503 genes were upregulated (54%), compared 1261 genes that were downregulated (46%) (Figure 1.2A). These results demonstrate that a specific subset of genes is consistently sensitive to CTCF depletion in breast epithelial cells.

GSEA analysis of the RNA-Seq data revealed that multiple gene sets related to the phosphoinositide 3-kinase (PI3K) and epithelial to mesenchyme transition (EMT) pathways were strongly upregulated in the CTCF+/- cells (Figure 1.2B). Indeed, Gene Ontology "Positive Regulation of Phosphatidylinositol-3-Kinase Signaling" and " Epithelial-to-Mesenchymal Transition" were in the top 5% and 7% by Normalized Enrichment Score in our GSEA Analysis, respectively, out of 839 significantly upregulated pathways. Similarly, gene sets related to these pathways were consistently among the top 10 enriched pathways using KEGG, Reactome or PANTHER pathway analysis tools and distinct ranking methods (Supplementary Figure 1.2B).

The PI3K pathway is a classical oncogenic pathway aberrantly activated in diverse cancers that drives both invasiveness and altered mammosphere morphology (184). Among the top upregulated genes in our CTCF+/- clones were ERBB3 and FGFR1, well-characterized receptor

tyrosine kinases and oncogenes that activate the phosphorylation cascade of the PI3K pathway (186,187). The PI3K signaling feeds into the EMT pathway, partially through translational upregulation of classical oncogenes, such as SNAI1, which itself promotes invasion (188). Interestingly, CTCF sites surrounding SNAI1 are enriched for non-coding mutations in cancer (189) and SNAI1 was among the top hits within the EMT pathway based on our RNA-seq data (Figure 1.2B). Consistent with an upregulation of EMT related genes, such as SNAI1 (190), and our invasive phenotype, downregulated genes were enriched for those involved in the promotion of cell-to-cell contact, such as cellular adhesion pathways (Supplementary Figure 1.2C). I validated the marked upregulation and downregulation of top hits using qPCR (Supplementary Figure 1.2D). Considering that CTCF+/- cells do not undergo obvious changes in morphology, it is likely that these cells undergo a partial EMT that is reversible upon re-expression of CTCF (Figure 1.1E).

Next, I investigated RNA-Seq data from the TCGA database of breast cancer patients. To detect groups of genes whose expression may be altered by varying CTCF levels, I computed the statistical correlation between each gene and CTCF expression in each patient. We then carried out gene ranking based on Spearman Test p-values coupled with GSEA PreRank pathways analysis. Similar to what I observed in MCF10A cells, pathways involved in PI3K signaling were overrepresented in the Top10 altered pathways by Normalized Enrichment Score (Supplementary Figure 1.2E) and p-values (Figure 1.2C). Interestingly, the association between high SNAI1 expression and low CTCF is also observed clinically, as I found a significant correlation between low CTCF and high SNAI1 expression in patients' breast tumors (Figure 1.2D). Overall, these results imply a role for CTCF in the regulation of genes involved in the PI3K signaling pathway gene and important regulator of EMT, such as SNAI1.

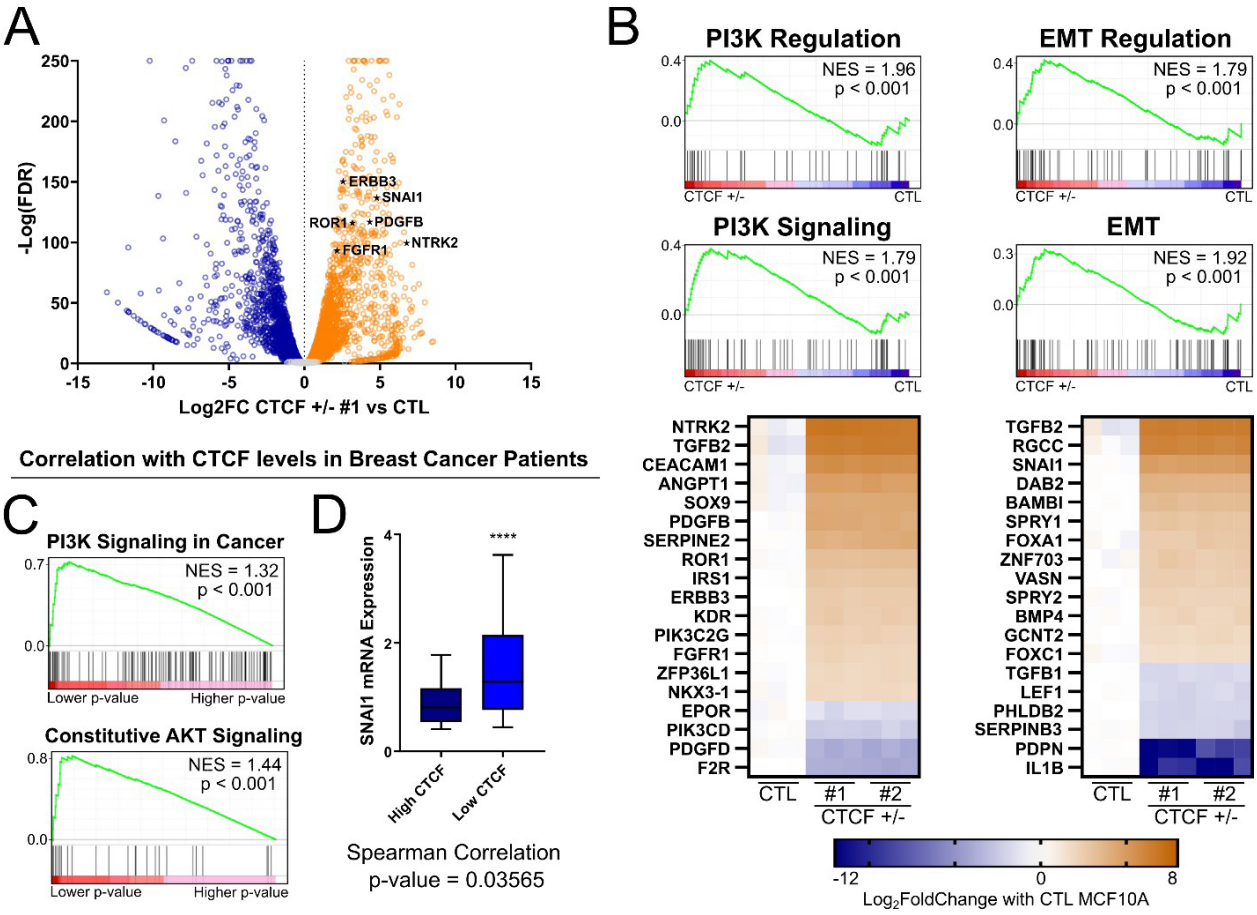


Figure 1.2. RNA-Seq reveals oncogenic expression underlying the invasive phenotypes. (A) Volcano plot of transcriptomic changes between CTCF+/- #1 and CTL MCF10A. Genes of the PI3K and EMT pathways, marked as black stars, are among the top upregulated genes. (B) GSEA of Gene Ontology PI3K Pathways and EMT Pathways upregulated in CTCF+/- MCF10A compared to CTL. Heatmaps of the top twenty most up or downregulated genes, ranked by abs(Log2FC), in PI3K and EMT Regulation pathway presented above. (C) Most enriched pathway by p-value (PI3K Signaling) and NES (AKT signaling) following GSEA Pre-rank analysis of genes significantly correlating with CTCF expression in TCGA breast cancer patient RNA-Seq data. Genes were ranked by -log(Spearman test p-value). (D) Box plot (10-90 percentile) of higher SNAI1 expression levels in CTCF low breast tumors ($p < 0.0001$, one-tailed Student's T Test) detected by RNA-Seq in TCGA Breast Cancer Patients for tumor in the top 20% low CTCF expression compared to top 20% high CTCF expression. p-value for the Spearman correlation test is also noted below.

Activation of the PI3K pathway following CTCF Copy Number Loss

Now guided by the RNA-seq data, I examined whether the PI3K pathway was hyper-activated in the MCF10A CTCF+/- cells, to determine whether this signaling contributes to the invasive phenotype. First, I screened for increased activation of key downstream effectors of PI3K signaling, including phosphorylation of 4EBP1 (serine 65) and S6K1 (threonine 389), direct targets of the mTORC1 complex (191). Under conditions of serum starvation, where phosphorylation of S6K1 and 4EBP1 were weakly detected in the CTL cells, a strong phosphorylation signal was detected in the CTCF+/- cells (Figure 1.3A). As the upregulation of the PI3K pathway can alter the morphology of mammospheres (184), my colleagues tested whether the elevated activation of PI3K might be observed under 3D culture conditions.

Similarly, they detected a pronounced phosphorylation of S6, the direct target of S6K1, in both CTCF+/- mammosphere populations, while it was absent in the CTL acini (Figure 1.3B). Since the outer region of mammospheres is expected to be the primary proliferative zone, due to the accessibility of nutrients and oxygen, they developed a custom script to visually isolate and quantify the fluorescence of this region for individual mammospheres. They detected that phosphorylation of S6 was 3.1 and 2.7 times higher in CTCF+/- #1 and #2 than in the CTL acini ($p < 0.0001$). Thus, under conditions of both 2D and 3D culture, reduced pools of CTCF leads to transcriptional reprogramming that activates the PI3K pathway.

Based on the aberrant activation of PI3K signaling in CTCF+/- cells, we surmised that their invasivity may be vulnerable to inhibitors of this pathway. I targeted mTORC1/2 because these kinase complexes assimilate the signals from diverse branches of the PI3K signaling cascade (191). I carried out matrigel transwell invasion assays following a 48h mTORC1/2 inhibition, using the second generation mTor inhibitor Torin1 (192). Since the CTL MCF10A are mostly non-invasive, I compared the changes in invasiveness of our CTCF+/- MCF10A cells and PDX

cell lines to the well-characterized breast cancer cell lines: MDA-MB-231, MCF7 and SKBR3. Following Torin1 treatment at 25nM, both MCF10A CTCF +/- and CTCF-low PDX lines, were markedly sensitive to these low concentrations, being significantly more repressed in their ability to invade than the trio of breast cancer cell lines carrying higher CTCF levels (Figure 1.3C). These data indicate that the PI3K pathway plays a central role in driving the invasion of normal epithelial cells with reduced CTCF levels, while late stage TNBC lines, such as MDA-MB-231 may utilize multiple, or alternative, pathways to achieve this phenotype.

I also validated that low concentrations of Torin1 treatment efficiently inhibited mTORC1. Concentrations as low as 5nM strongly abrogate the phosphorylation of 4EBP1 in the CTCF+/- cells under starved condition (Figure 1.3D). As the PI3K pathway has also been shown to control the protein expression of SNAI1 through translational upregulation (188), I investigated the impact of Torin1 treatment on SNAI1 expression. I detected a marked, dose-dependent drop in SNAI1 protein levels following 24h of Torin1 exposure (Figure 1.3D). Since SNAI1 overexpression promotes invasiveness in multiple models (193,194) and it is strongly overexpressed at the mRNA and protein level in our CTCF+/- MCF10A (Figure 1.3E), we decided to investigate whether it is an important downstream target of PI3K and playing a role in the invasiveness of the CTCF+/- cells. To do so, my colleague used lentiviral-mediated shRNA knockdown of SNAI1 and surveyed the changes to cell invasion. The downregulation of SNAI1 led to both a significant reduction of SNAI1 protein levels and of CTCF+/- invasiveness (Figure 1.3F). Overall, these results highlight the importance of the upregulation of the PI3K pathway, and its downstream effector, SNAI1, for the oncogenicity of the CTCF+/- cells. These indicate that the invasion of tumors harboring CTCF copy number loss, coupled with elevated SNAI1, may be susceptible to therapeutic intervention with inhibitors of mTORC1/2.

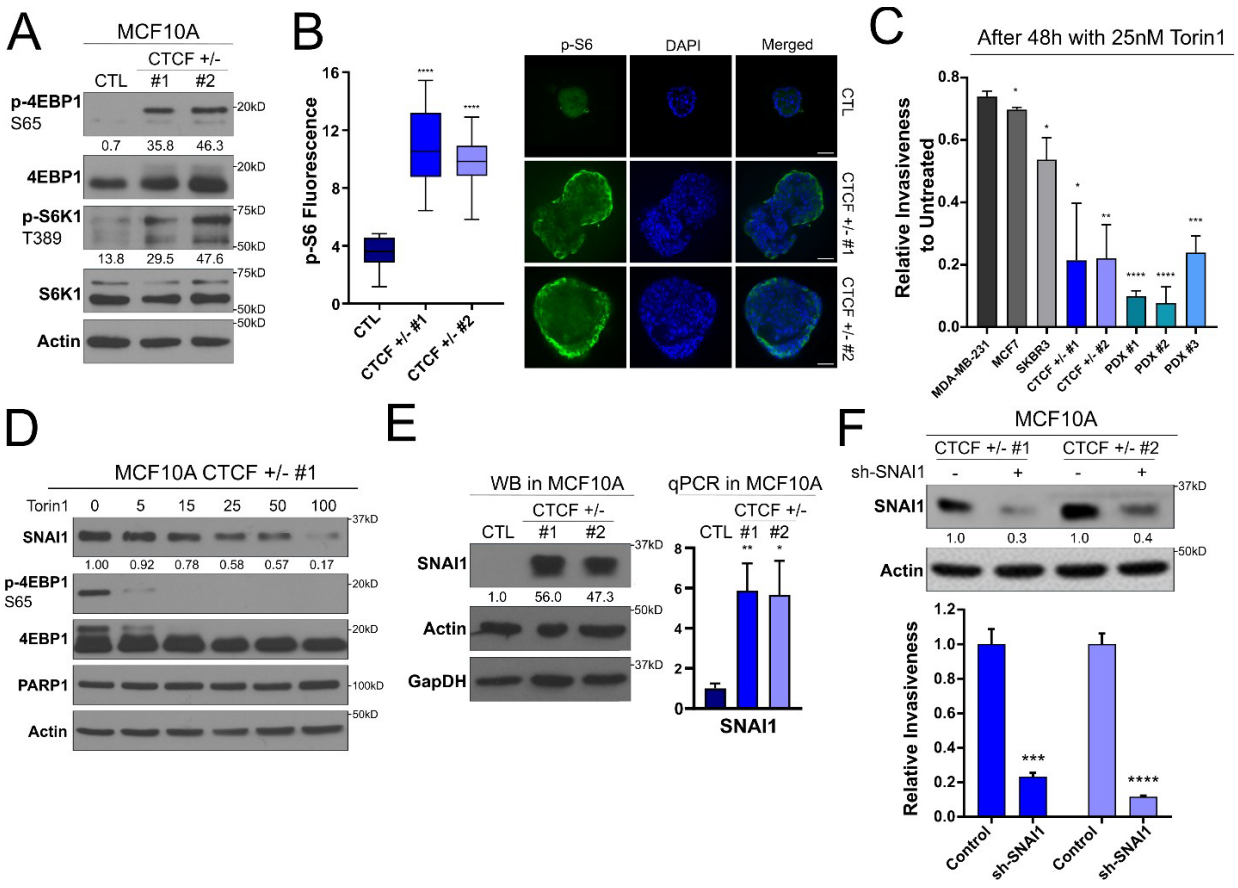


Figure 1.3. The PI3K pathway and SNAI1 are central for the oncogenic properties of CTCF+/- cells (A) Western Blot showing maintained phosphorylation of mTORC1 targets under serum-free conditions in CTCF+/- cells. Quantification represents the band intensity normalized on background. Loading control: Actin. (B) Mammosphere immunofluorescence and quantification of increased S6 fluorescence of the outer layer of the mammosphere in CTCF+/- compared to CTL MCF10A. Represented as a Min to Max Box Plot. $p < 0.0001$ for CTCF+/- #1 and #2 compared to CTL. (C) Relative invasiveness following 48h 25nM Torin1 treatment, normalized relative to the untreated invasiveness of each cell line (mean ± SEM). p-values comparing each cell line to the relative invasiveness of MDA-MB-231: MCF7 = 0.031508, SKBR3 = 0.019340, CTCF+/- #1 = 0.018925 and #2 = 0.003089, PDX #1 = 0.000002, #2 = 0.000076, #3 = 0.000262. (D) Western Blot, using serum-free conditions, for SNAI1 levels and 4EBP1 phosphorylation following 24h Torin1 treatment. Quantification of relative SNAI1 band intensity relative to untreated levels is shown below each blot. Loading control: PARP1 and Actin. (E) Western Blot of SNAI1 levels in CTCF+/- MCF10A cells. Quantification of relative SNAI1 band intensity to CTL is shown below the top-most blot. Loading control: Actin and GapDH. Bar chart (mean ± SEM) of qPCR validation of SNAI1 overexpression at the mRNA levels. (F) Western Blot of SNAI1 levels following sh-SNAI1 treatment. Below blot, quantification of relative SNAI1 band intensity to shCTL. Loading control: Actin. Bar chart of decreased relative invasiveness of the shSNAI1 treated CTCF+/- MCF10A compared to shCTL treated (mean ± SEM). p-values = 0.00103 and 0.000141 for #1 and #2. All p-values were calculated using Student's T Test.

Low CTCF expression alters its binding to DNA surrounding oncogenes

It is reasonable to expect that the reduced nuclear pool of CTCF would compromise the number of occupied CTCF sites on the chromatin. To gain mechanistic insight into the altered oncogenic transcriptional networks of MCF10A CTCF +/- cells, we carried out a ChIP-Seq to map CTCF binding across the genome (Figure 1.4A). I identified that a majority of CTCF binding sites, 38 775 out of the 44 802 peaks called by MACS2 (195), were left unchanged between the MCF10A CTL and MCF10A CTCF+/- cells. Considering that CTCF levels were reduced by ~50-60% in these cells, it is clear that the nuclear pool of CTCF is in excess of that required for genomic regulation, consistent with previous reports showing a significant fraction of CTCF is unbound within interphase cell populations (156). This excess is likely a safeguard against genomic instability and protection of the transcriptome that might stem from fluctuating CTCF levels.

However, as expected, a subset of 5313 sites displayed reduced or lost CTCF binding in the CTCF+/- MCF10As compared to the CTL (FDR < 0.01, LogFC < -1). Surprisingly, a small cluster of 714 sites displayed a gain of CTCF binding (FDR < 0.01, LogFC > 1) (Supplementary Figure 1.3A).

Next, I investigated the differences in binding strength and distribution between the sites lost and constant in CTCF+/- cells. The average read density was lower for sites within the lost cluster compared to the constant cluster in the CTL MCF10As (Figure 1.4B). The genomic distributions of lost and gained sites were also unique compared to constant sites. 29% of CTCF lost sites compared to 19% of constant sites, were found on promoters, such as the promoter of SNAI1 (Supplementary Figure 1.3B/C). While 37% of lost sites, including a ERBB3 downstream site (Supplementary Figure 1.3D) and 41% of constant sites were located in distal intergenic regions, compared to 53% of gained sites (Supplementary Figure 1.3B).

Consistent with our RNA-Seq, KEGG pathway analysis of the CTCF binding sites within the

lost clusters were strongly enriched surrounding genes involved in the PI3K-Akt signaling pathway (ranked 2nd by relative gene count). Multiple pathways related to cell mobility, such as ECM-receptor activation (ranked 4th), were also observed, which display significant enrichment when compared to an equinumerous set of constant CTCF sites (Figure 1.4C). These results suggest that a subset of weakly binding CTCF sites, showing enrichment around promoters and genes involved in the PI3K pathway and cell invasion, such as SNAI1, are more sensitive to CTCF depletion in mammary epithelial cells.

CTCF Lost Sites are frequently proximal to deregulated genes

CTCF may impact gene transcription through binding proximal or distal to transcription start sites (TSS) through multiple mechanisms. To gain further insights into the mechanisms whereby altered CTCF binding might impact transcriptional events in MCF10A cells carrying a single functional CTCF allele, I mapped lost sites to determine their proximity to TSS. About half of all CTCF lost sites (2408 out of the 5313) were found with proximity (+/- 3kb) to significantly altered genes (Basemean > 100, adjusted p-value < 0.05). Interestingly, a significant fraction of lost sites was found around both strongly upregulated (Log2FC > 1, 530 sites) and strongly downregulated (Log2FC < -1, 716 sites) genes, in pathways consistent with the RNA-Seq (Figure 1.4C/D). This highlights the complexity of gene regulation mediated by CTCF, but underscores that loss of CTCF binding frequently impacts the transcription of proximal loci. The intensity of the loss of CTCF significantly, but weakly, correlated with both upregulation ($r = -0.1$, $p = 0.0056$) and downregulation ($r = 0.2$, $p < 0.0001$) of gene expression. Similarly, both upregulated and downregulated genes displayed a slightly lower average CTCF ChIP-Seq read density on their TSS (Figure 1.4E).

Regarding gained CTCF sites, of 714 such sites, 334 are proximal to 267 unique altered genes (adjusted p-value < 0.05). Of these genes, 192 (72%) are significantly upregulated, but only 107

genes reach a Log2FC >= 1. Pathway analysis of the 107 upregulated genes or 267 significantly altered genes revealed no clear enrichment, as no pathways were significant by FDR. These results are expected when the number of genes is low and the distribution is primarily stochastic. While proximal loss of CTCF is associated with both upregulation and downregulation of gene expression, it only mildly correlated with the intensity of altered gene expression. This hints that in many cases, the changes in transcription observed in CTCF heterozygous cells are likely driven through indirect, or downstream, mechanisms including changes in chromatin conformation or epigenetic reprogramming that are potentiated by CTCF loss, but likely not due to a loss of CTCF interaction with the core transcription machinery (196,197).

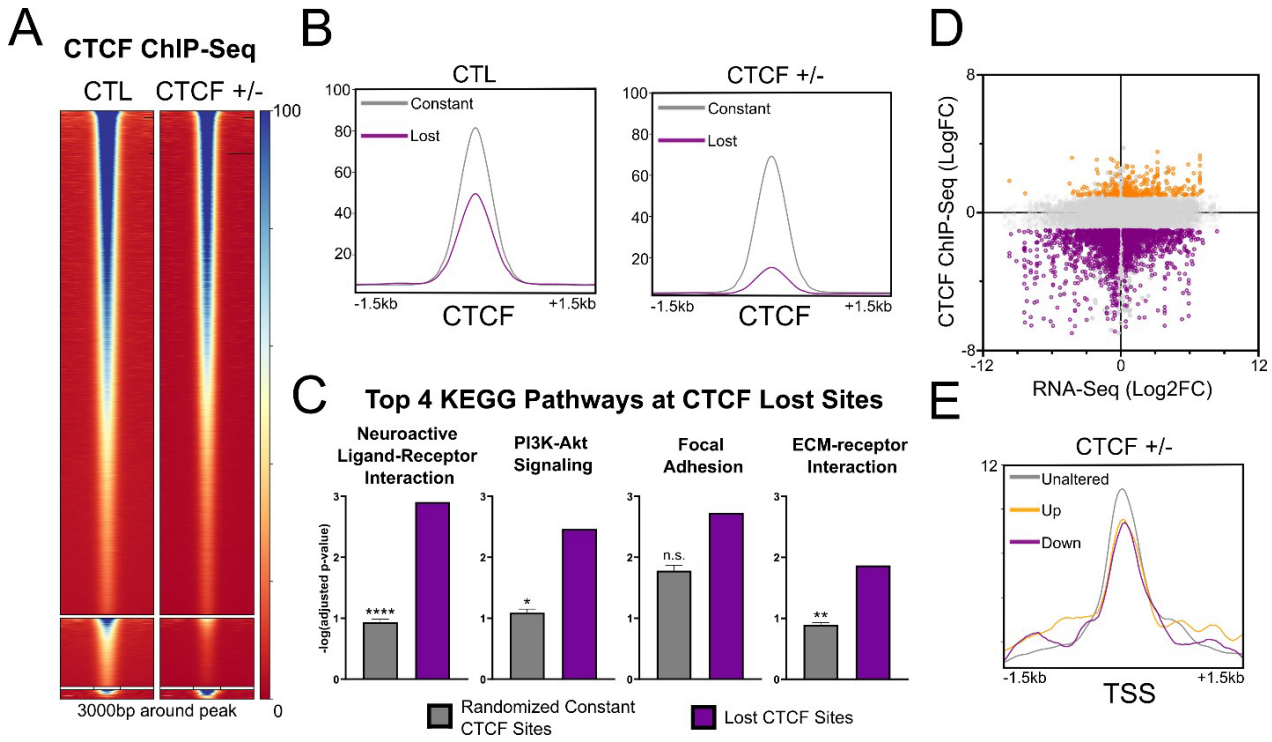


Figure 1.4. CTCF depletion alters CTCF DNA binding pattern (A) CTCF ChIP-Seq Heatmap of constant, lost and gained sites (from top to bottom). (B) Reduced average CTCF ChIP-Seq read density of lost sites compared to constant sites in the MCF10A CTL and CTCF+/-. (C) Differential enrichment of the top 4 KEGG Pathways, dominated by PI3K and ECM related pathways (ranked by geneRatio), at lost sites of CTCF compared to 100 equinumerous subsets of constant sites (mean ± SEM). (D) Dot plot of gene expression (Log2FC) and CTCF binding (LogFC) changes between CTCF+/- and CTL MCF10A for binding sites colocalizing (+/- 3kb) with expressed genes. Lost sites (purple) are found in proximity to both up and downregulated genes. Gained sites (orange) are differentially found in proximity to upregulated genes. (E) Decreased average CTCF ChIP-Seq read density in CTCF+/- MCF10A at the TSS of all upregulated genes (adjusted p-value < 0.05, Log2FC > 1) and downregulated genes (in purple, adjusted p-value < 0.05, Log2FC < -1) compared to unaltered genes (adjusted p-value > 0.05, abs(Log2FC) < 0.5).

CTCF loss potentiates epigenetic reprogramming at transcriptionally altered genes

Destabilization of CTCF binding has been linked to numerous epigenetic changes (198-200). Thus, I investigated whether the changes to gene expression and CTCF binding were associated with changes to chromatin marks. First, we screened for multiple activating and silencing histone marks on representative altered genes using ChIP-qPCR. Although I detected significant changes in H3K4me3 and H2K27ac associated with altered transcription (Supplementary Figure 1.4A/B), I did not detect strong changes with the repressive marks H3K27me3 and H3K9me3 (Supplementary Figure 1.4C/D). This is consistent with a previous study where changes to CTCF binding across multiple genomes were not strongly linked to differences in H3K27me3 (199).

Therefore, we proceeded to map H3K4me3 and H3K27ac genome-wide using ChIP-seq to compare CTL MCF10A with CTCF+/- cells (Figure 1.5A). A majority of H3K4me3 (~82%) and H3K27ac (~87%) peaks were conserved between CTCF+/- and the CTL MCF10A. However, both H3K4me3 and H3K27ac showed significant alterations upon loss of CTCF. H3K4me3 and H3K27ac gained enrichment at 2929 and 5188 loci respectively. Further, H3K4me3 was reduced at 1932 and H3K27ac at 2060 sites (abs(LogFC) >= 1, FDR <= 0.05) (Figure 1.5B). Overall, CTCF loss potentiated a gain of marks associated with gene activation. Then, I assessed whether these changes to histone marks correlated with altered gene expression. We observed a pronounced, and statistically significant gain of H3K27ac at upregulated genes ($r = 0.64$, $p < 0.0001$) (Figure 1.5C) including oncogenes such as ERBB3 and SNAI1 (Figure 1.5D), compared to a more modest correlation of H3K4me3 with upregulated genes ($r = 0.45$, $p < 0.0001$) (Figure 1.5C/D).

CTCF loss has been indirectly linked to deregulated DNA methylation (201,202) and it is possible that altered DNA methylation contributes to transcriptomic changes observed in CTCF+/- cells. I carried out bisulfite conversion and investigated the association between

genome wide changes in DNA methylation and transcriptomic changes, using Illumina EPIC methyl array with our RNA-Seq data. Contrary to the strong correlation detected between changes in activating marks and gene expression, the changes in DNA methylation pattern observed in the CTCF+/- did not correlate with changes in gene expression ($r = -0.04$, $p<0.0001$) (Figure 1.5E). These results indicate that under conditions of sub-physiological CTCF levels, changes in gene expression most specifically linked to a global reprogramming of H3K27ac.

To test for a role of gained H3K27ac in the promotion of cell invasion, I treated CTCF+/- cells with the Histone Acetyl-Transferase (HAT) inhibitor A485, that targets CBP (203). First, we validated the ability of A485 to inhibit the deposition of H3K27ac using western blotting (Supplementary Figure 1.4E). Linking acetylation to the transcriptomic profiles defined on our MCF10A CTCF +/- cells, under serum starved conditions, A485 treatment efficiently resolves the hyperactivation of the PI3K/mTor pathway, as indicated by a dose-dependent reduction of 4EBP1 phosphorylation (Figure 1.5F). Similarly, CBP inhibition blocked SNAI1 expression, linking the gain of H3K27ac to its upregulation (Figure 1.5F). As inhibition of both the PI3K pathway and SNAI1 expression reduced the invasiveness of the CTCF+/- cells, I tested their invasivity after exposure to A485 treatment and further, compared the effects with those observed in MDA-MB-231 cells. Similar to mTor inhibition and SNAI1 knockdown, A485 treatment significantly reduced the invasiveness of the CTCF+/- cells (Figure 1.5G) further supporting the hypothesis that the increased deposition of H3K27ac plays a key role in the oncogenic phenotypes caused by the loss of CTCF. Interestingly, MDA-MB-231 cells were noticeably sensitive to this treatment (Figure 1.5G) as well. These results highlight a general dependency on increased histone acetylation during the invasion process of aggressive epithelial cancer cells, regardless of CTCF status, and supports an essential role of epigenetic

reprogramming during cancer progression.

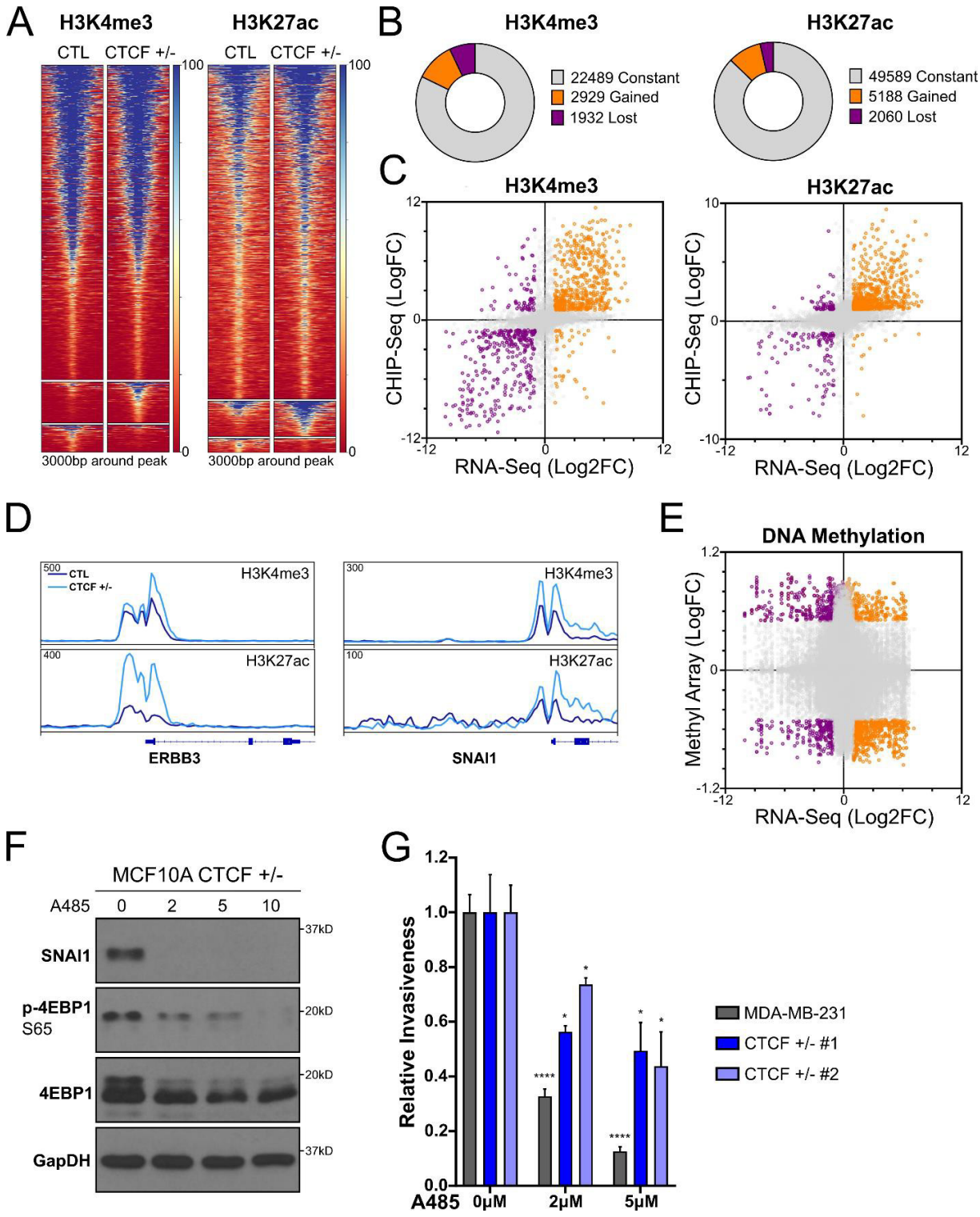


Figure 1.5. Epigenetic reprogramming of activating histone marks drives changes in gene expression (A) H3K4me3 and H3K27ac ChIP-Seq heatmaps for constant, gained and lost sites (from top to bottom). (B) Partitioning of constant, gain and lost clusters from Figure5A. (C) Dot plot of highly correlating gene expression (Log2FC) and H3K4me3 or H3K27ac (LogFC) changes between CTCF+/- and CTL MCF10A for binding sites colocalizing (+/- 3kb) with expressed genes. (D) ChIP-Seq track of the normalized read density for H3K27ac or H3K4me3 surrounding ERBB3 and SNAI1. (E) Dot plot of gene expression (Log2FC) and Methyl EPIC Array (LogFC) changes between CTCF+/- and CTL MCF10A for binding sites colocalizing (+/- 3kb) with expressed genes. (F) Western Blot, under starved conditions, for 4EBP1 phosphorylation and SNAI1 levels following 48h treatment with HATi A485 treatment (μM). (G) Relative invasiveness of A485 treated CTCF+/- MCF10A and CTL MDA-MB-231 (mean ± SEM) . P-values of treated compared to untreated cells; 2μM: MDA < 0.0001, #1 = 0.0444 and #2 = 0.0232; 5μM: MDA < 0.0001, #1 = 0.0284 and #2 = 0.0252. See also Supplementary Figure 4.

Reduced CTCF Levels lead to loss of insulation of subTAD structures

Following our ChIP-Seq experiments, we posited that the loss of CTCF binding and the relative increase in open chromatin at activated genes may stem from a loss of insulation. Therefore, I investigated changes in 3D chromatin architecture using Hi-C. I generated 600 million reads per condition with biological replicates of the CTL, and replicates of CTCF+/- #1 and #2 were merged for high resolution analysis. This sequencing depth allowed us to reach a complete genomic coverage at 5kb resolution, consistent with previous high-resolution Hi-C data (204-206). Statistical analysis of the correlation between the contact matrices of each cell line revealed a marked difference between the CTL and both CTCF +/- clones at 5kb resolution (Figure 1.6A), while all three groups were more homogenous when the resolution was moved to 500kb or 1Mb (Supplementary Figure 1.5A). Consistent with this analysis, at the megabase scale, I also did not detect notable genome-wide changes in chromosome organization between CTCF+/- and CTL cells (Supplementary Figure 1.5B), which were strikingly consistent with previously published data (207).

Next, I queried whether more local changes in chromatin architecture may underlie the RNA profiles resulting from CTCF CNL. First, I used a hierarchical TAD caller, hiTAD (208), to call TAD boundaries and domain boundaries within TADs (termed subTADs) at a 10kb resolution. I then compared the colocalization of called boundaries (+/- 10kb) between the CTL and two CTCF+/- clones. Of the 11,580 TAD boundaries called, 10% were lost in both CTCF+/- lines compared to the CTL. These changes were more pronounced when looking at subTAD boundaries, where 17% of the total number was lost (Figure 1.6B). The loss of these boundaries might potentiate de novo contacts between DNA elements due to loss of insulation. Indeed, CTCF+/- cells gained 810 new TAD boundaries (7% gain) and 606 subTAD boundaries (11% gain), which are enriched next to lost boundaries (Supplementary Figure 1.5C), indicating a re-

organization of sub-genomic regions. Altered boundaries frequently colocalized with altered sites of CTCF binding (+/- 10kb), with lost boundaries showing a marked enrichment for lost CTCF elements, while gained boundaries are generally CTCF-null (Figure 1.6C). These de novo TAD/subTAD interactions, demarcated by gained boundaries, are likely generated from a loss of insulator activity that limit long range DNA contacts, so it is logical that these regions would be devoid of CTCF. This novel mechanism is supported by a recent study demonstrating that CTCF-independent enhancer looping is potentiated by the loss of proximal CTCF binding (209).

To validate that the loss of CTCF binding leads to local loss of DNA insulation, I imaged the average local interaction centered around lost sites of CTCF (+/-200kb, Figure 1.6D). In agreement with our hypothesis, I detected a marked reduction of boundary strength, represented by a decreased interaction intensity at CTCF sites delimiting two domains. DNA insulation was also clearly compromised as represented by an increased interaction intensity between the domains spanning the lost sites of CTCF (Figure 1.6D).

Subsequently, I asked whether loss of insulation was equally compromised at TAD and subTAD boundaries. To answer this question, I subdivided the lost CTCF sites into lost sites colocalizing with TAD boundaries or located within TADs. For each subset of the lost sites, I plotted local interactions centered around the lost sites of CTCF (+/-200kb) and measured the average insulation score of these regions using FAN-C (210). I detected a slight loss of boundary strength and insulation at lost sites of CTCF colocalizing with TAD boundaries. These results were expected since TAD boundaries are often bound by redundant CTCF sites and recent evidence indicates that many TADs insulate themselves from their neighbors independently of CTCF (109,211). However, lost sites of CTCF within TADs resulted in a nearly complete loss of boundary strength and insulation, allowing inter-domain DNA interactions (Figure 1.6E). These

observations validate, in a quantifiable manner, the prominent loss of insulation at subTADs under conditions of low CTCF expression. Since subTADs are localized within TADs, in a chromatin environment that promotes interactions, the resulting loss in insulation is more permissive to the formation of new, potentially oncogenic, contacts.

Changes in subTAD organization drives epigenetic reprogramming and changes in gene expression

I continued the Hi-C analysis to investigate whether the changes in subTAD interactions are connected to the changes in gene expression. First, I measured the average gene expression changes at altered subTAD and TAD boundaries. Genes colocalizing with the gained subTAD boundaries were the most significantly upregulated ($p < 0.0001$, Figure 1.6F) compared to all genes. As expected, altered TAD boundaries were not significantly associated to transcriptional changes (Figure 1.6F).

I found that altered subTAD interactions and changes to activating marks are both associated with changes in gene expression, so I investigated whether colocalization of H3K27ac or H3K4me3 was observed at domain boundaries. I detected a strong enrichment of gained sites of H3K27ac and H3K4me3 with gained subTAD boundaries and vice-versa with lost subTADs boundaries (Figure 1.6G, Supplementary Figure 1.5D). Altered TAD boundaries were not enriched for changes in either mark (Figure 1.6G, Supplementary Figure 1.5D), consistent with the lack of transcriptional changes in these regions. These results are validated by comparing the average changes in insulation at gained sites of H3K27ac colocalizing with TAD boundaries or within TADs (Figure 1.6H). Gain of H3K27ac at TAD boundaries was not predictive of altered insulation, while gain of H3K27ac within TADs led to a marked gain of insulation (Figure 1.6H), confirming the formation of de novo subTAD boundaries at these sites. Importantly, the genes found at gained H3K27ac within TADs were enriched for genes involved in mTor signaling

(Supplementary Figure 1.5E), as this pathway was among the ten most differentially enriched pathways in gained H3K27ac compared constant H3K27ac within TADs. Then, using pileup plots, I looked at the average density of interactions between regions of gained H3K27ac and all sites of either H3K27ac or H3K4me3 (Figure 1.6I). Considering all combinations, I detected a marked gain of interaction at loci where H3K27ac was gained in the CTCF+/- cells (Figure 1.6I). These results indicate that the reconfiguration of subTADs, specifically, allows for de novo interactions at regulatory regions enriched for gains of H3K27ac that drive the expression of oncogenic programs.

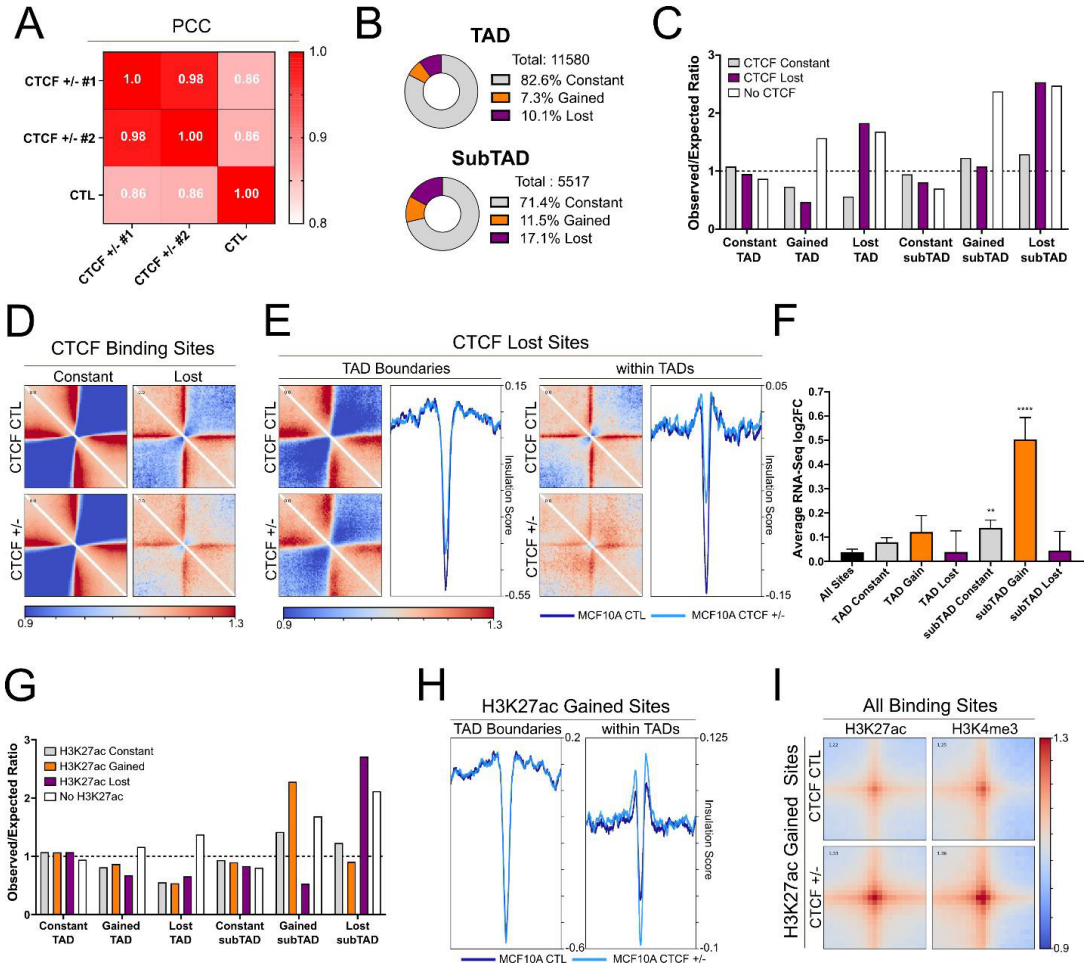


Figure 1.6. Loss of subTAD insulation drives gene expression changes. (A) Pearson Correlation Coefficient heatmaps showing diverging contact frequencies between CTCF +/- #1 and #2 and CTL. (B) Partitioning of constant, gained and lost TAD and subTAD boundaries (+/- 10kb). (C) Enrichment of CTCF sites at boundaries (O/E Ratio), showing an association between loss of CTCF and lost boundaries, and absence of CTCF and altered boundaries. These are both more pronounced for subTAD boundaries. (D) Pile-up plots showing local interaction, relative to randomize average genome-wide interaction, around constant and lost sites of CTCF (range: 200kb). CTCF lost sites show less insulation in CTL MCF10A, which is further reduced upon loss of CTCF. (E) Pile-up plots of local interactions at CTCF sites localizing at TAD boundaries or within TADs. Profile plot of average insulation score in each region quantifies the specific loss of insulation observed at lost sites of CTCF within TADs. (F) Average RNA-Seq log2FC between CTCF+/- and CTL of genes colocalizing with TAD and subTAD boundaries (+/- 10kb) (mean ± SEM) showing that gained subTAD boundaries are strongly associated with upregulation of gene expression. (G) Enrichment of altered H3K27ac sites at altered subTAD , but not TAD, boundaries (O/E Ratio). (H) Increased average insulation score at sites of gained H3K27ac within TAD, but not at TAD boundaries (colocalization : +/- 10kb, range: 200kb). (I) Pile-up plots of increased interaction between gained H3K27ac and all sites of H3K27ac and H3K4me3 (range: 50kb).

An excellent example of this mechanism may be observed at the SNAI1 locus. At the megabase scale, conformational changes are not obvious (Supplementary Figure 1.6A). Using HIFI (212) to facilitate Hi-C resolution at a sub-5kb scale, I detected a discrete, novel interaction between the SNAI1 gene and a downstream potential enhancer in CTCF+/- cells (Figure 1.7A). This interaction is positioned adjacent to the lost CTCF binding site within the SNAI1 promoter, and is embedded with a region of gained H3K27ac (Figure 1.7B). The downstream enhancer, connecting with the promoter, is likewise enriched for H3K27ac in the CTCF+/- cells (Figure 1.7B).

To validate that the loss of CTCF at SNAI1 may drive its overexpression, I directed a dCAS9 construct to the CTCF site at SNAI1 promoter (sgSNAI1), in MCF10A CTL cells, where CTCF binding is compromised in CTCF+/- cells. Using ChIP-qPCR, I validated the specific displacement of CTCF at the promoter proximal CBS (Supplementary Figure 1.6B). Disruption of CTCF at this site with exogenous dCAS9 would be expected to facilitate an increase of SNAI1 expression if our model is correct. As a control I used a sgRNA targeting a CTCF-unbound region at the SNAI1 locus (sgCTL). Compared to CTL cells infected with sgCTL, cells infected with sgSNAI1 displayed a significant upregulation of SNAI1 mRNA levels (2.1 fold increase, $p = 0.006$) (Figure 1.7C). As a further control, directing dCAS9 to this CTCF binding site in CTCF+/- cells, where CTCF binding is already compromised, did not result in an upregulation of SNAI1 (Figure 1.7C). These data validate that disruption of CTCF may play a key role in driving the upregulation of oncogenes, including SNAI1.

In summary, the loss boundaries at the subTAD level compromises insulation from de novo contacts. These de novo contacts, and the associated enrichment for H3K27ac at these regions, in turn, play a major role in driving the oncogenic networks observed in cells with CTCF CNL.

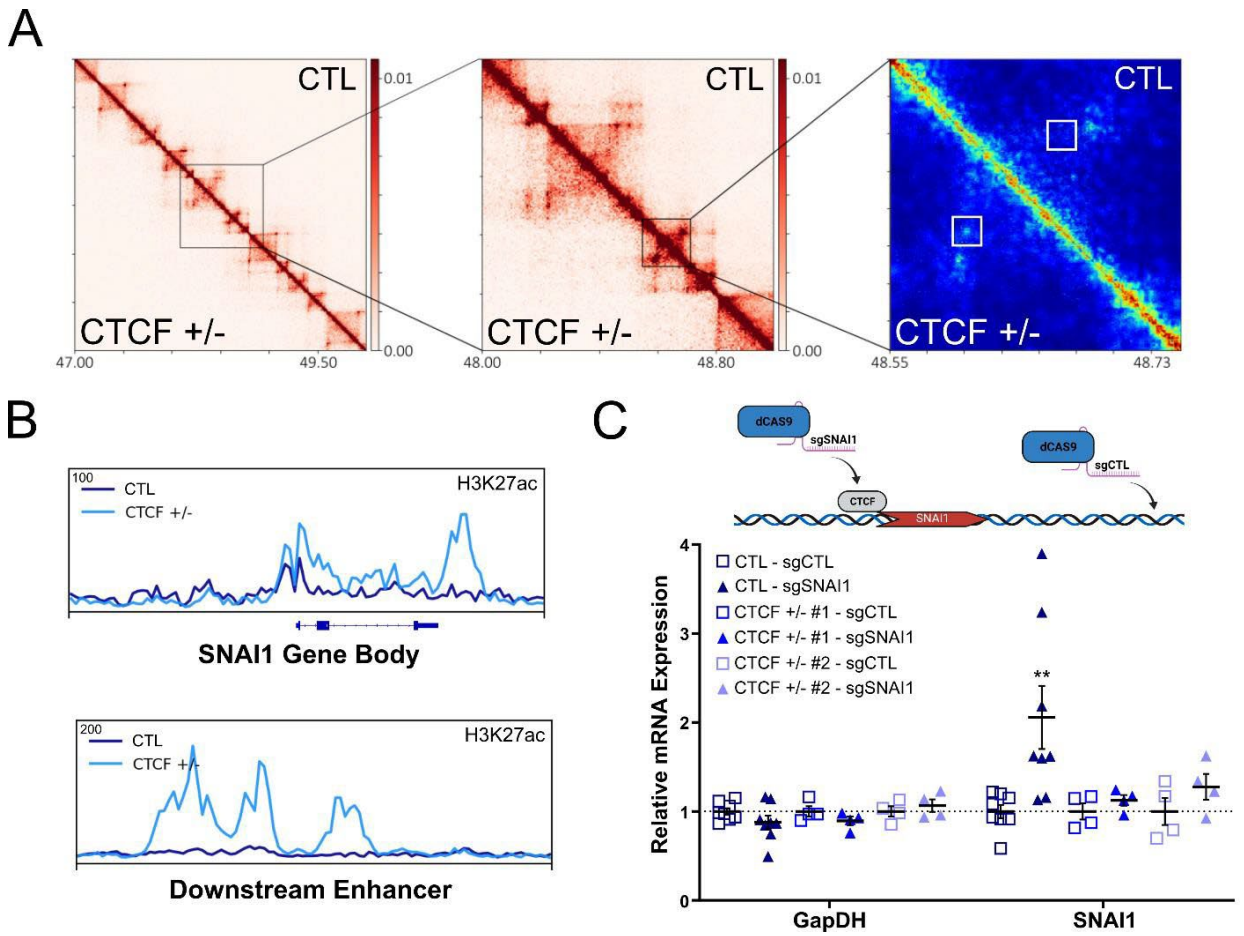


Figure 1.7. Loss of CTCF at SNAI1 drives reorganization of subTAD interactions. (A) Increasing zoom of 10kb and 5kb resolution HiC heatmap to HIFI high-resolution heatmap around SNAI1 loci (chr20, coordinates in Mb). Gain of enhancer-promoter interaction on SNAI1 body, specific to CTCF +/- cells, shown in the white boxes in the HIFI heatmap. (B) ChIP-Seq track of normalized read density of increased H3K27ac on SNAI1 gene body and the downstream enhancer which displayed a gain of interaction in Figure 7A. (C) Mean ± SEM and individual replicates mRNA expression, relative to sgCTL, of infected CTL and CTCF+/- MCF10A. SNAI1 mRNA levels ($p = 0.0057$) in CTL-sgSNAI1 compared to CTL-sgCTL. All other comparisons are non-significant. Schematic of the experimental conditions (made with Biorender) is depicted above.

Chapter 2: Identifying Altered DNA Recognition Motif Associated to Mutant CTCF (213)

CTCF ZF1M is associated with CTCF LOH in Breast Cancer

To gain insight into the biological importance of CTCF ZF1 mutation, I first sought to interrogate the clinical correlation between ZF1 mutation and CTCF Loss of Heterozygosity (LOH). CTCF LOH is observed in a majority of breast tumors and I investigated a potential association or exclusivity of CTCF ZF1M and CTCF LOH to identify the most common clinical genotypes of CTCF ZF1M in breast tumors. Using copy number variation data from cancer patients within the TCGA 2018 dataset, I detect a significant downregulation of copy number in patients with CTCF ZF1M, of which CTCF H284N was the most common, compared with patients with other CTCF mutations or with WT CTCF. (Figure 2.1A). Among CTCF mutations across tumor types, the association between ZF1 mutation and CTCF LOH is the most pronounced, especially in breast tumors (Figure 2.1A). Indeed, ~83% of breast tumors with CTCF ZF1M co-occur with CTCF LOH (Figure 2.1B). Comparatively, CTCF LOH is detected in ~52% of breast tumors and ~16% of other types of tumors when WT CTCF is expressed from the second allele. Therefore, we conclude that a significant co-occurrence of CTCF ZF1M and CTCF LOH is found within breast tumors.

In light of these observations, I decided to explore the biological impact of CTCF ZF1M in breast epithelium using two relevant models. First, the ZF1M/- model, in which the CTCF H284N mutation is inserted into one allele while the second allele of CTCF is knocked-out, similar to the most commonly observed genotype in the clinic. Second, the ZF1M/ZF1M model, in which a biallelic insertion of the CTCF H284N mutation results in the sole expression of the mutated form of CTCF at the same expression level as the control cell line, to account for any biological effects of the lower CTCF protein levels in the ZF1M/- cell line. Using CRISPR-Cas9, my colleague generated clonal lines for each of these genotypes, by combinations of knocking-in the CTCF H284N mutation and knocking-out CTCF in MCF10A cells (Supplementary Figure 2.1A/B).

MCF10A were chosen as they are immortalized, but not transformed, mammary epithelial cells, suitable to study the impact of the CTCF ZF1M in early events of breast cancer formation, without confounding effects of complex oncogenic mutations carried in breast cancer cell models.

CTCF H284N Mutation Leads to Altered DNA Binding

Next, to clarify the debated importance of ZF1 for coordinating CTCF-DNA interaction, I tested the hypothesis that the H284N mutation might alter CTCF binding to the DNA. Towards this goal, we carried out ChIP-Seq for CTCF using MCF10A CTL, ZF1M/ZF1M and ZF1M/-. 48 340 CTCF binding sites (CBS) were identified in CTCF CTL cells, consistent with other studies (202). Following csaw differential binding analysis, I identified 27997 constant CBS between all 3 conditions, 3812 gained CBS in both ZF1M/- and ZF1M/ZF1M and 6556 commonly lost CBS (FDR <= 0.05) (Figure 2.1C/D). Interestingly, the genomic distribution of the altered CBS was not prominently different from the constant CBS, beside a slight enrichment of altered CBS on distal intergenic elements (Supplementary Figure 2.1C). Overall, the changes in CTCF were consistent between the two mutant cell lines. CBS gained in the ZF1M/- cell line were also gained in ZF1M/ZF1M cells, however, without reaching a threshold for significance (Figure 2.1E). On the other hand, sites gained in the ZF1M/ZF1M cell line appeared as low signal CBS in CTL displaying a slightly increased read density in both mutants, but only reaching the significance threshold in ZF1M/ZF1M, likely due to the higher availability of CTCF in this cell line, compared to the ZF1M/- cells (Figure 1E). Similarly to gained CBS, lost CBS within the ZF1M/ZF1M cell line were likewise frequently lost in the ZF1M/- cells (Figure 2.1F), indicating a high degree of similarity between ZF1M/ZF1M and ZF1M/- cells. The only subset of altered CBS that did not display a strong similarity between the 2 mutant cell lines were the 1013 CBS uniquely lost in ZF1M/-, which are likely caused by the lower levels of CTCF (Figure 2.1F).

Independent ChIP analysis from previously published reports helps validate our findings. I compared the CTCF binding profile of our mutant cell lines with a CTCF WT ChIP-Seq dataset from an independent study (214). Here, the altered CBS called by csaw were markedly consistent (Supplementary Figure 2.1D), with our own MCF10A dataset, suggesting that changes in CTCF binding are intrinsic to the mutant clones. Further, the changes in CTCF binding in our mutant MCF10As were also consistent when the datasets were analyzed with a different pipeline, using MACS2 (195) for peak calling and DiffBind (215) for differential binding analysis (Supplementary Figure 2.1E). Therefore, these results indicate that the CTCF H284N, ZF1 mutation, likely induces a shift in the ability of CTCF to recognize or bind DNA. Also, due to the strong similarity between our models, the influence of the CTCF mutation on DNA binding seems to be largely independent of varying CTCF expression levels, hinting at a molecular mechanism underpinning the altered binding that does not include a stochastic loss in the general ability of CTCF to bind DNA.

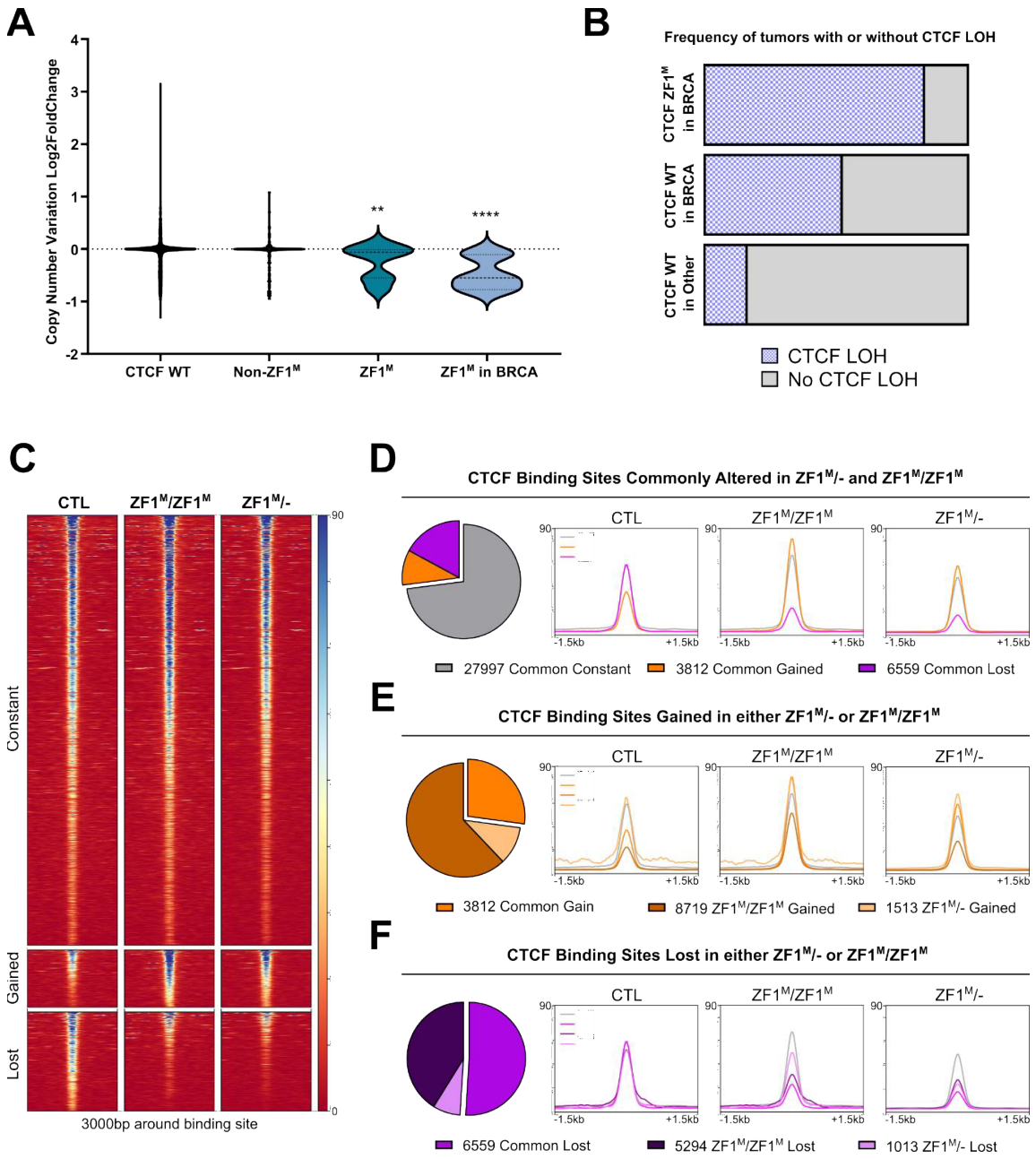


Figure 2.1. H284N mutation of CTCF ZF1 alters a subset of DNA binding sites. (**A**) Enrichment of copy number loss of CTCF in ZF1M in tumors of all origin (N = 13, p = 0.0018) and ZF1M in breast tumors (N = 6, p < 0.0001) compared to Non-WT Non-ZF1M CTCF tumors (N = 258). (**B**) Bar chart representation of the increased frequency of CTCF LOH in CTCF ZF1M in BRCA (N = 5) compared to CTCF WT BRCA (N =1045) and CTCF WT tumors of all cancer (N = 10607) (**C**) CTCF ChIP-Seq heatmaps of commonly constant, gained and lost CBS (csaw, FDR < 0.05). (**D**, **E**, **F**) Pie Charts of the number of CBS commonly altered or uniquely altered CBS in each clone, coupled with profile plot representation of read density at these specific sites. Beside the 1013 uniquely lost in ZF1M/-, all groups of altered CBS display nearly identical changes in read density in both mutant cell lines.

Classical Motif Enrichment Analysis

Following the identification of differentially bound sites, our next goal (and subsequent step of our motif discovery pipeline, represented in Figure 2.2) was the identification of enriched motifs. To do so, in collaboration with Kaiqiong Zhao from Dr. Celia Greenwood's laboratory, we first constructed representative subsets of each cluster by selecting the 1000 most significantly altered sites in the Gained and Lost clusters, based on the FDR-adjusted q-values. In contrast, to characterize the "Constant subset", of unchanged CBS, we selected the 1000 least significantly changed binding regions. Analyzing these subsets, as opposed to the entire cluster, focuses the analysis on the most relevant sites, thereby filtering out less significant differentially bound sites that might arise stochastically.

Once subsets were defined, we performed motif discovery analysis on these three clusters using rGADEM. We additionally compared the identified motif patterns to the JASPAR database and reported the significant matching motifs. Not surprisingly, rGADEM identified the CTCF motif as the most represented motif in all three clusters (Figure 2.3A). Indeed, the core CBS is found in 78% to 93% of all CTCF binding sites, depending on the cell line being probed (Figure 2.3B). However, as expected, the tools used for standard motif discovery analysis were unable to identify changes in motifs associated with altered binding affinity. This is expected since subtle changes would be drowned by the high representation of the CTCF core binding motif. Therefore, we continued our motif analysis using a new R-based software, developed in collaboration: MoMotif.

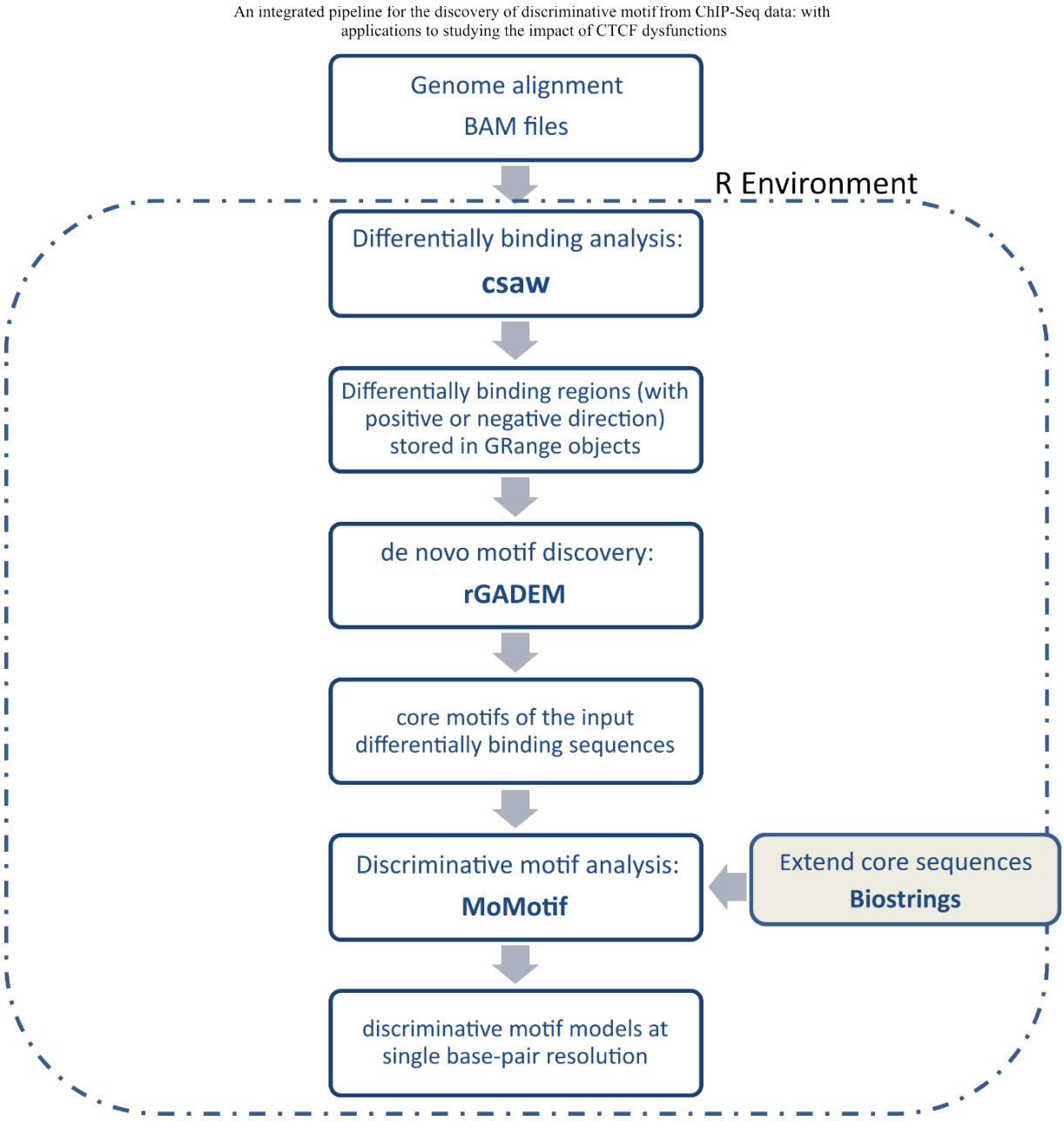


Figure 2.2 Flowchart representation of an R pipeline utilizing newly developed software MoMotif to identify complex DNA binding motifs based on ChIP-seq profiling.

Novel MoMotif Analysis

To detect single nucleotide changes in the binding sequences of Lost or Gained CTCF sites, we aligned and extended the CTCF-like motifs to a 61bp sequence centered at the mid-point of the canonical CTCF motif (represented by the purple dotted line in Figure 2.3C). The extension of the sequence allows us to focus on single nucleotide changes, within and outside of the classical ~15bp CTCF motif, that potentially influence CTCF binding affinity. Then, using MoMotif, we calculated frequency differences and p-values at each nucleotide within the extension, comparing the Lost and Gained subsets to the Constant subset, within the common altered sites (Figure 2.3C) and in each mutant cell line individually (Supplementary Figure 2.3A/B). We defined a section of the extended sequences containing every position reaching the required statistical threshold ($p < 1*10^{-10}$) and a frequency difference greater than 0.1, in the lost or gained sites. Specifically, from position 25 to 48, as indicated by the black dashed line in Figure 2.3C, which encompasses a downstream extended CTCF core binding motif. We therefore defined this subsection of the original 61bp sequence as our newly identified nucleotide region capable of influencing CTCF binding affinity in the context of the H284N mutation. Akin to the alteration of CTCF binding between our two mutants cell lines, the changes in nucleotides frequency were also markedly consistent (Supplementary Figure 2.3A/B).

By depicting these new motifs with the height of each nucleotide representing the Shannon Entropy of its occurrence frequency at each position (Figure 2.3D), we visually highlight the unique extended motif enriched at each position. This reveals an extended motif specific to the lost sites defined by an A at position 40, a G at position 43 and a C at position 46. Interestingly, the G at position 43 also displays the lowest p-value and highest frequency difference when comparing lost sites to constant sites in all conditions (Figure 2.3C, Supplementary Figure 2.3A/B). Furthermore, the extended motif identified with MoMotif is homologous to the

previously defined module 4 of CBS, carrying a very weak consensus, identified by ChIP-exo (40) (34). Although a mechanism explaining how CTCF recognizes this motif was not revealed in prior publications, module 4 of the CTCF binding motif has been associated with a stronger DNA-binding affinity of WT CTCF. This conclusion is supported by my observations (Supplementary Figure 2.4A) and these results, from independent studies, validate the predictive value of MoMotif.

The extended motif influencing the association of CTCF to DNA through ZF1 appears to be mediated by three nucleotides at position 40, 43, 46. The enrichment of the extended motif in the sites lost in cells carrying the H284N mutant becomes even more prominent when investigating the proportion of the 1000 sites that display a combination of 2 or 3 of these specific nucleotides. Indeed, 24% of common sites lost across both our mutant cell lines, co-localizing with a CTCF-like motif, displayed the 3 defining nucleotides of the extended sequence. In contrast, only 6% and 2% of Stable and Gained sites, respectively, carried this motif. Furthermore, the combination of at least 2 of these nucleotides was found in 66% of Lost sites, compared to 33% and 8% of Stable and Gained Sites, respectively. The exclusion of this extended sequence in the Gained Sites is also represented in the proportion of CTCF-like sites that do not include any of the three nucleotides, being 54% in the Gained Sites, compared to only 6% in the Lost Sites (Figure 2.3E).

As a comparison, I analyzed the lost sites using the classical motif enrichment tool SEA, from the MEME Suite. SEA identified the CTCF core binding motif as the most enriched motif in the lost sites (Supplementary Figure 2.4B), similar to earlier steps in our pipeline. When using the MEME suite software to carry out motif enrichment analysis comparing the lost CTCF sites with constant sites, SEA identified differentially enriched motifs in a small subset of lost sites, with low frequency of True Positives (TP) below 10% for each motif. These marginally differentially enriched motifs are also located in regions surrounding the center of the sequences, where a

consensus CTCF motif is located (Supplementary Figure 2.4C), inconsistent with a ZF1-specific effect. However, software from the MEME suite, such as SEA, does not identify unique motifs, differentially enriched between conditions, or motifs only partially present in both, a necessity to output a single nucleotide analysis of the modification of a specific motif between the conditions. Therefore, classical motif enrichment analysis is competent to identify TFs showing differential binding between conditions, but cannot precisely identify changes to a specific motif under variable conditions, as summarized in Supplementary Figure 2.4D.

In sum, MoMotif can be used to facilitate the discovery of subtle motif changes after the introduction of experimental variables. As will be detailed below, MoMotif may also be used to compare DNA motifs within subsets of single datasets, including ChIP-seq and Hi-C. Regarding CTCF, we used MoMotif to define a unique DNA motif that requires CTCF ZF1 for recognition. This motif is strongly associated with the sites lost upon ZF1 mutation and was ignored by classical motif analysis tools. These data suggest a model where the CTCF ZF1 mutation induces a loss of function rendering the mutant CTCF unable to bind, or recognize, the extended sequence, leading to its stochastic redistribution on CBS without this sequence, specifically those that do not require ZF1 to bind appropriately.

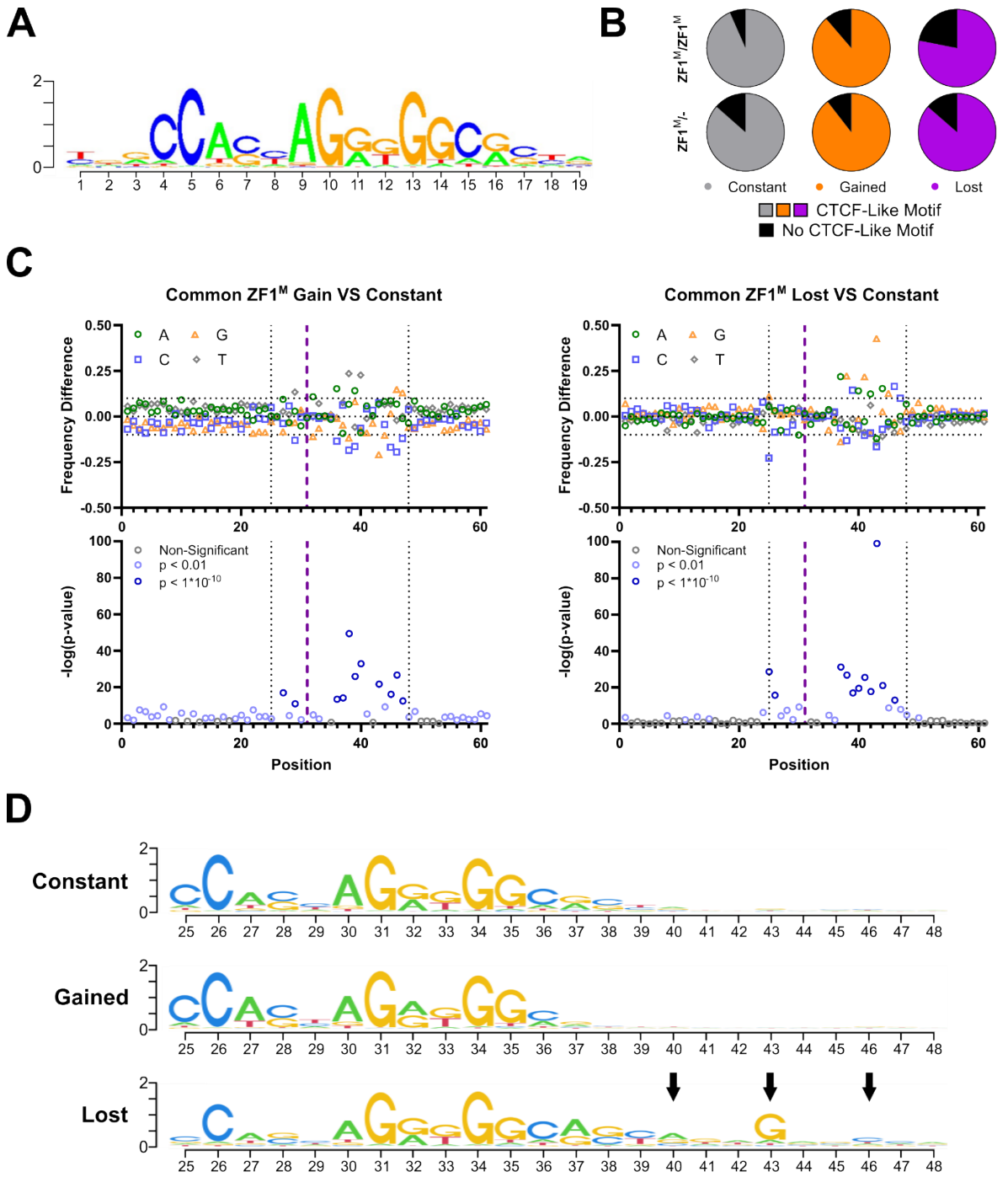

A
B
ZF1M/ZF1M
ZF1M/-
Constant
Gained
Lost
CTCF-Like Motif
No CTCF-Like Motif
C
Common ZF1M Gain VS Constant
Common ZF1M Lost VS Constant
Frequency Difference
A
G
C
T
-log(p-value)
Non-Significant
p < 0.01
p < 1*10-10
Position
D
Constant
Gained
Lost
E
Common Constant
Common Gained
Common Lost
Total=778
Total=790
Total=883
40A
43G
46C
40A-43G
40A-46C
43G-46C
40A-43G-46C
None

Figure 2.3.MoMotif identifies a unique motif enriched for CBS compromised upon mutation of ZF1. (**A**) Classical CTCF motif output by rGADEM. (**B**) Frequency of overlap with CTCF-Like motif in each 1000 sites subset. (**C**) MoMotif analysis of base frequency difference and p-value of bases distribution difference around CTCF-Like motif in common lost and gain CBS subsets compared to common constant subset. The purple line represents the middle of the CTCF Motif. The dotted line represented the selected region shown in D (**D**) MoMotif results depiction as the height of each nucleotide representing the Shannon Entropy of its occurrence frequency at each position in each subset. Highlighting the extended motif (40A, 43G, 46C) in the lost subset. (**E**) Bar chart representing the relative presence of each individual and combined element of the extended motif in each subset. Showing an enrichment of the partial or complete extended motif in the lost subset, while the complete or partial extended motif is absent from the gain sites. Highlighting a role for CTCF ZF1 in the recognition of this sequence.

Structural analysis of CTCF zinc finger-DNA contacts suggests conformation changes imparted by zinc finger 1 mutation

CTCF is known to use variable combinations of zinc fingers to flexibly bind diverse sites on the DNA (216). Therefore, I investigated whether the modified CTCF motif identified by MoMotif was recognized by a specific combination of CTCF zinc fingers requiring ZF1. I used per-domain predictions of CTCF ZFs DNA-binding specificity using the software and databases from Persikov et al. 2014 and 2015 (217,218), to identify 3bp sequences that are recognized by individual CTCF zinc finger (Figure 2.4A). CTCF ZF3 to ZF7 are known to mediate strong binding to the CTCF core binding motif (33,34). When aligning the ZF3-7 consensus motif with the motifs identified in the constant and gained clusters, a majority of the bases identified at each position match between motifs (92.8% and 85.7% against the constant and gain motifs respectively) (Figure 2.4B). These associations indicate that CTCF recognizes the motif identified within the constant and gained sites independently of CTCF ZF1 and is therefore not directly hindered by the mutation of ZF1. In contrast, the extended motif enriched in the lost cluster aligns with a different combination of ZFs. Indeed, although ZF7 to ZF4 match similarly to the first half of the extended motif (Figure 2.4C), the primary DNA base matches with ZF3 at the constant sites is replaced by secondary matches at lost sites. Further, a strong de novo primary match motif is observed at both ZF2 and ZF1 within the sites lost in ZF1M cells (Figure 2.4C). These results hint at an enrichment, at lost sites, of sequences that require the combination of ZF4-7 and ZF1-2, with a possible variation in ZF3 binding, to be appropriately recognized and bound by CTCF. As CTCF H284 is necessary for the coupling of the zinc ion, crucial to the ZF structure, it is expected that ZF1M structure would be aberrant and therefore, unable to carry out its function. In turn, blocking the ability of the ZF1-2 tandem to properly recognize the A and G of the extended motif, resulting in a dissociation specifically at these sites. However, the zinc finger structure alone cannot explain the presence

ofan extended motif from position 46 to 48, primarily defined by a C at position 46, hinting that a secondary binding mechanism of ZF1, beyond its binding of 3 core bp is at play, or alternatively, a protein, or RNA co-factor may influence the DNA recognition by CTCF ZF1. Overall, this analysis strongly supports our model that ZF1-mutation of CTCF is unable to bind an extended motif at a subset of CBS, and this pool of CTCF is then redistributed to gained sites stochastically, where ZF1 binding is not required.

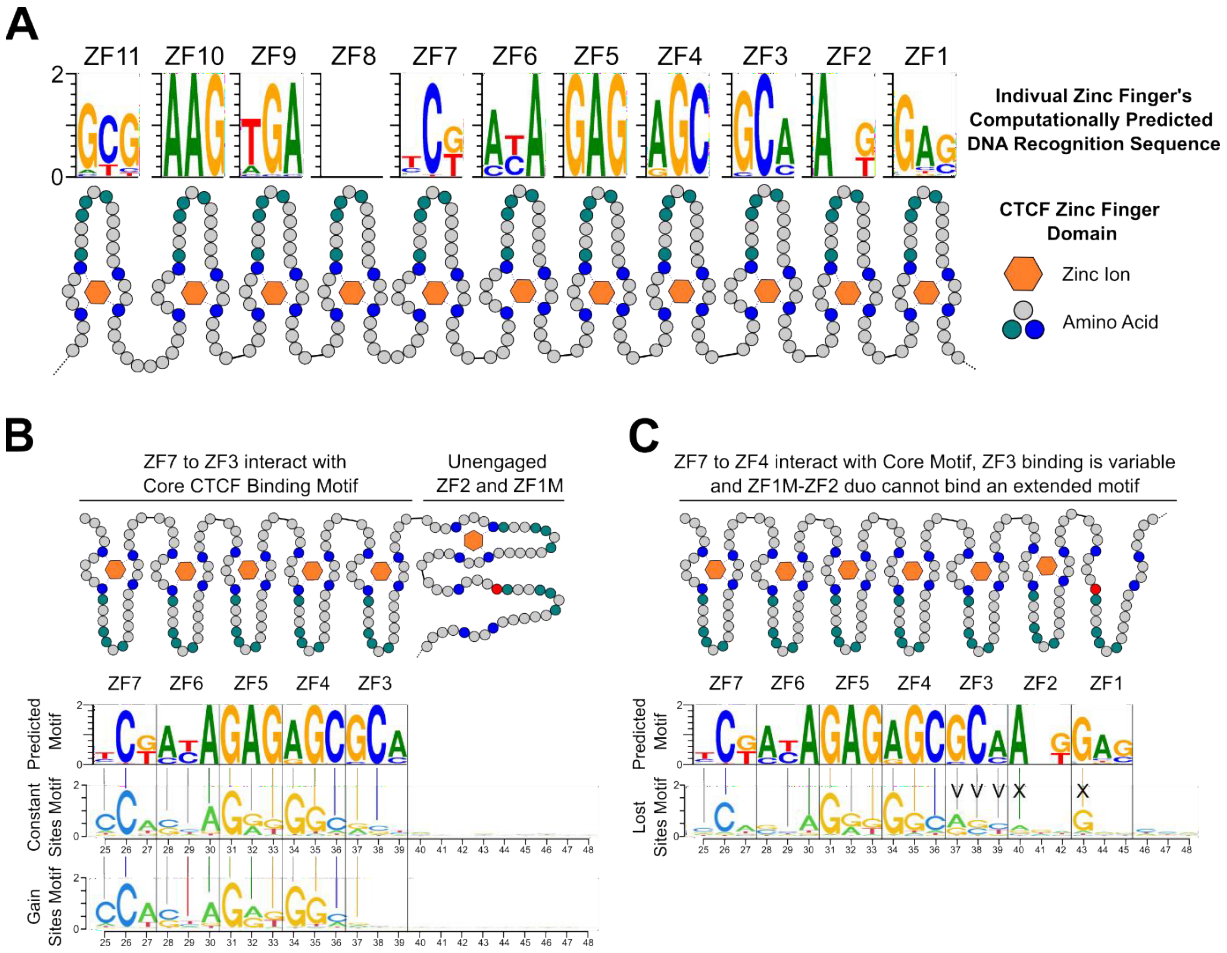


Figure 2.4. Extended Motif of CTCF is associated to an altered binding conformation (**A**) Predicted 3bp sequences recognized by each ZF of CTCF by Persikov et al. 2014 and 2015 (58,59). (**B**) Alignment of the predicted motif to the motif identified by MoMotif for Constant and Gain CTCF binding sites. (**C**) Alignment of the predicted motif to the extended motif identified by MoMotif for Lost CTCF binding sites. For B and C, colored vertical bars represent a match between the primary called base at each position and grey vertical bars represent a match between a secondary called base and a primary base.

MoMotif reveals increased stability of the core CTCF binding motif at domain boundaries

MoMotif is a versatile computational tool that may not only be used to compare DNA-binding motifs across ChIP-seq samples, but can also be used to compare complex DNA motifs present within subsets of a single ChIP-seq dataset.

CTCF plays an essential role in the organization of chromatin conformation, in part by defining the boundaries of Topologically Associated Domains (TADs). Therefore, we asked whether ZF1 mutation impacted differentially the binding of CTCF at sites maintaining 3D chromatin organization. Towards this goal, we used the genomic coordinates of TADs and subTADs (defined as self-associating domains within TADs), binned at 10kb, using our Hi-C datasets from CTL MCF10A to provide topological context to our CTCF ChIP-Seq. Here again, TAD and subTADs were called using the hierarchical TAD caller hiTAD (208), ranked best TAD caller in term of average concordance over normalizations and resolutions in Zufferey et al. 2018 (219). Further, as different TAD callers may output variable boundaries from the same sample, we also used SpectralTAD (220) to call and compare boundaries at 10kb resolution. Overall, ~97% of boundaries called by hiTAD in our CTL MCF10A were called in the same region (+/- ½ bin/5kb) by SpectralTAD (Supplementary Figure 2.5A), confirming the reproducibility of the topological context we provided.

Next, I categorized all CBS of the CTL MCF10A cells based on their co-localization with a TAD boundary, a subTAD boundary, or not on a domain boundary, independently of whether they are constant or lost in the ZF1M lines. Overall, 10276 and 4915 CBSs colocalized with a TAD or a subTAD boundary, respectively, compared to 36029 CBS that did not colocalized with any boundaries. These ratios are consistent with multiple previous investigations of CTCF and TAD colocalization (4,221).

Next, I used MoMotif to identify any discriminative modifications of the CTCF motif comparing

sites at subTAD boundaries, TAD boundaries, or not at boundaries (Supplementary Figure 2.5B/C). I found that the CTCF core binding motif is exquisitely consistent on subTAD and TAD boundaries (Supplementary Figure 2.5B/C). However, when comparing the CBS motif found at TAD boundaries to CBS outside TAD and subTAD boundaries, MoMotif detected an increased variability around ZF3 and ZF2 and to, a lesser extent, between ZF6 and ZF7. However, no specific enrichment for a particular base was observed at these positions. Instead, the bases recognized by these ZFs displayed a reduced Shannon Entropy, hinting at an increased motif disparity for CBS found within domains compared to CBS found at their boundaries, perhaps highlighting their diverse roles. Interestingly, the extended motif associated with lost CBS in CTCF ZF1M mutated cell lines is equally present on CBS colocalizing or not with a boundary. Supporting this conclusion, when comparing the genomic localization of constant and lost CBS between CTL and ZF1M MCF10A lines, the sites are distributed equally among domain boundaries or within domains (Supplementary Figure 2.5D). These results demonstrate, in a unique context, the sensitivity of MoMotif to identify precise regions of variability around a given motif, while showing that CTCF extended motif and its associated lost binding sites of CTCF ZF1M are not enriched in specific topological contexts.

Gene Expression Changes induced by CTCF ZF1M concur with observed clinical phenotypes

Next, I investigated whether the changes in CTCF binding might be associated with transcriptional changes that might underpin the clinical phenotypes observed in CTCF mutated breast tumors (31,164). First, I used RNA-Seq to define the differences in steady state RNA levels between MCF10A CTL, CTCF ZF1M/ZF1M and CTCF ZF1M/-. Overall, the changes in gene expression observed were highly conserved in both mutant cell lines, highlighting the impact of the H284N mutation on regulating gene expression. Indeed, when correlating the respective log2FC of both mutant lines with MCF10A CTL, the lines carrying the H284N mutation displayed a strong

correlation (r = 0.7811 and p-value < 0.0001) (Figure 2.5A). Approximately 95% of significantly altered genes (FDR <= 0.05) in ZF1M/ZF1M cells were altered in the same direction in ZF1M/-, while 69% and 76% of strongly up and downregulated genes (abs(log2FC) >= 1) in ZF1M/- were strongly altered in both cell lines. Similar to the ChIP-Seq distributions and MoMotif nucleotide frequency, the effect of the mutation appears to be dominant over any effects of the LOH.

Next, I used GSEA to run pathway analysis of altered genes in both mutant cell lines. Interestingly, pathways associated with drug metabolism were consistently among the top upregulated pathways (Figure 2.5B/C). My RNA-seq analysis also revealed that pathways involved in extracellular matrix (ECM) organization were among the top downregulated pathways (Figure 2.5B/C). Multiple genes involved in these pathways, such as ADAMST1 and SLC20A1, are proximal to lost sites of CTCF in ZF1M/ZF1M or ZF1M/- cell lines (Figure 2.5D). These genes are also within the majority of genes that were significantly altered in the same direction in our model and in patient's CTCF ZF1M breast tumors compared to CTCF WT breast tumors from TCGA datasets (Figure 2.5E). Consistent with our data, CTCF H284N mutations are frequently enriched in hormone resistant breast tumors (31). We propose that the upregulation of metabolic pathways that target xenobiotics may explain this phenomenon. We also propose that changes to the ECM may underlie the increased metastatic abilities of CTCF mutated breast tumors (164), also consistent with previous reports (222-225).

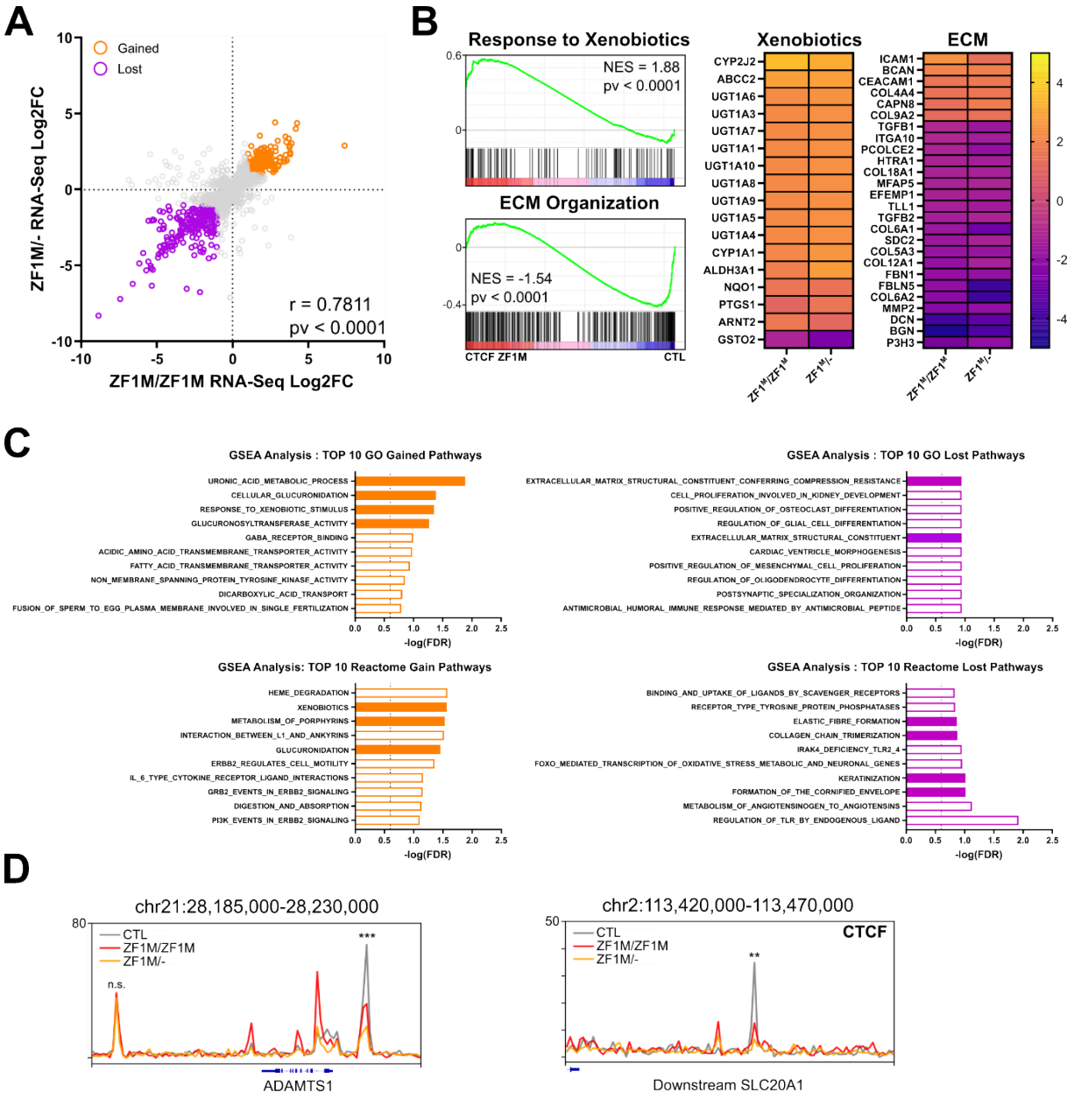


Figure 2.5. CTCF ZF1M drives oncogenic transcription profiles. (**A**) Dot plot representation of the RNA-Seq Log2FC of the individual mutant to control MCF10A on each axis. Showing a

strong correlation and reproducibility between the samples (with Pearson correlation and test p-value displayed). (**B**) GSEA enrichment representation of significantly upregulated and downregulated pathways. Heatmap of the Log2FC with control MCF10A of significantly altered genes in these pathways. Showing an upregulation of genes related to drug metabolism and downregulation of genes related to ECM. (**C**) Top 10 up and downregulated pathways (sorted by GSEA FDR) in Gene Ontology and Reactome Databases. Filled orange bars are linked to drug metabolism and filled purple bars are linked to ECM organization. Showing an over-representation of these pathways among the top altered pathways in diverse databases. (**D**) CTCF ChIP-Seq track around altered genes from the RNA-Seq in MCF10A CTCF ZF1M vs CTL and in TCGA Breast Tumor CTCF ZF1M vs CTCF WT related to Xenobiotic metabolism and extracellular matrix organization. Showing a significant loss of CTCF binding in proximity to ADAMTS1 promoter ($p=8.91*10-5$ and 0.003054 for ZF1M/ZF1M and ZF1M/- respectively) and within SLC20A1 ($p=7.28*10-5$ and 0.001778 for ZF1M/ZF1M and ZF1M/- respectively) (**E**) Pie chart showing a majority of genes significantly altered in the MCF10A models are also significantly altered in the same direction in breast tumors data from TCGA database when comparing changes in gene expression associated to CTCF ZF1M. Significance of the correlation between the alteration of gene expression of the two datasets is also shown.

Loss of CTCF binding within TADs is associated with the changes in gene expression

We next sought to determine the mechanisms underlying the transcriptional changes apparent in H284N-carrying cells. Because CTCF modulation of transcription may be highly dependent on the topological organization of the chromatin (226), an altered CBS could influence the expression of a gene thousands of kilobases away. Thus, I used the genomic coordinates of TAD and subTAD from our Hi-C datasets from CTL MCF10A to provide topological context to our RNA-Seq and ChIP-Seq results.

TAD boundaries are known to be highly conserved between cell types, often colocalizing with ubiquitously expressed genes, while CTCF-mediated interactions within TADs are prone to changes and less conserved between cell types (227). Therefore, we expect that genes most strongly deregulated by ZF1 mutated CTCF would likely be located within TADs and not at their boundaries, similarly to what I observed in the CTCF +/- model from Chapter 1. To study this hypothesis, I divided the TADs into 2 groups; TADs in which the TSS of all altered genes (FDR<= 0.05) are localized exclusively on their boundaries (+\- 1 resolution bin/10kb) (termed TAD- B) and TADs in which the TSS of all altered genes are found exclusively within the domains, and not at boundaries (TAD-I). Then, I computed and compared the distribution of strongly altered genes (abs(Log2FC) >= 1) in each condition. As predicted, the TAD-I group is significantly associated with strong alteration of gene expression, while the TAD-B group was not enriched for significant changes in gene expression (Figure 2.6A/B).

I then layered the CBS altered in ZF1-mutant cells onto my analysis to identify the cluster of CBS which was the most influential for altered gene expression. Within the TAD-I group, both the loss and the gain of CTCF within TADs was associated with RNA-Seq alterations. However, the association between lost CBS and changes to gene expression was markedly more significant than for the gained CBS (p-value = 0.0027 for CTCF Lost Sites, p-value = 0.028 for CTCF Gained

Sites) (Figure 2.6A/B). Supporting the validity of these findings, TAD-I in which no CBS displayed significantly less changes in transcription. Although the TAD-B group was not associated with significant changes in gene expression, loss of CTCF binding at the boundaries of these TADs still led to increased transcriptional variability (Figure 2.6A/B). In contrast, gain of CTCF at TAD boundaries, likely brought about through a stochastic redistribution of the mutant CTCF to strongly conserved CBS, was significantly associated to a conservation, instead of an alteration, of gene expression (Figure 2.6A).

The distribution of CTCF and gene expression changes at subTADs also supported a model where lost CTCF sites are driving gene expression changes. When investigating subTADs with TSS of altered genes colocalizing exclusively at their boundaries (subTAD-B), the only changes in CTCF binding promoting upregulation or downregulation of gene expression were lost CBS located at the boundaries of these subTADs (Supplementary Figure 2.6A). Overall, these results suggest that changes to CTCF binding within TADs predicts the altered gene expression through reorganization of intra-TAD interactions.

Supporting this theory, pathway analysis of altered genes proximal to a lost site of CTCF within a TAD (TAD-I) reproduces the top pathways I identified in the global RNA-Seq, being dominated by drug metabolism and ECM related pathways (Figure2.6C). Therefore, my contextual analysis of ChIP-Seq and RNA-Seq revealed that the loss of CTCF binding sites within TADs, including those sites at the boundaries of subTADs, are the main drivers of the changes in gene expression resultant from CTCF ZF1 mutation. This supports a model where the inability of CTCF to bind the extended recognition motif drives aberrant phenotypical changes.

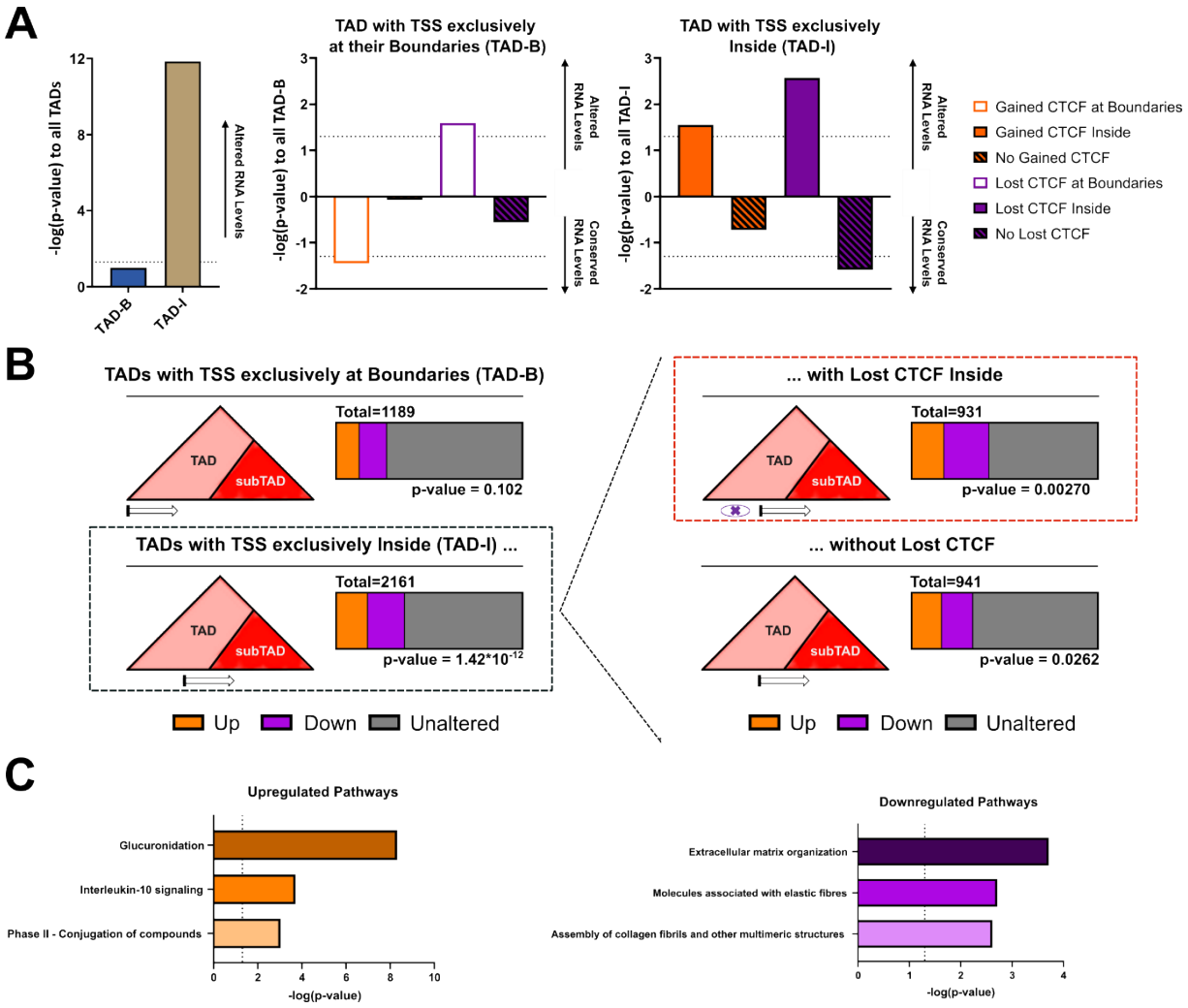


Figure 2.6. Loss of CTCF binding within TADs drives oncogenic transcription (**A**, **B**) Impact on the distribution of altered genes TSS (DESEQ2, FDR < 0.05) and altered CBS (csaw, FDR<0.05) in the context of TAD on the enrichment of strongly altered genes (ZF1M/ZF1M to CTL abs(Log2FC) >= 1). Showing the most significant impact of the loss of CTCF at TADs encompassing genes within them (TAD-I), compared to gain of CTCF or at TAD encompassing genes at their boundaries only (TAD-B) (p-value were generated from Chi-Square test on distribution of altered genes, -log(p-values) depicting significantly less strongly altered genes were turned negative in A to ease comprehensiveness of the graph). (**C**) Top 3 pathway, sorted by p-value, of Reactome Pathway Enrichment Analysis of strongly upregulated and downregulated genes from the distribution highlighted in red in B. Showing that loss of CTCF within TAD is driving the major changes in gene expression observed in global GSEA analysis of the RNA-Seq.

MoMotif identifies promoter proximal variability of TF recognition motif

Next, we wanted to validate the capacity of MoMotif to be used as a computational tool to compare DNA binding motifs across ChIP-Seq datasets incorporating independent experimental variables. To this end, I used previously published ChIP-Seq datasets from diverse transcription factors and compared their recognition motif among promoters (+/-3kb), non-coding intronic and distal intergenic regions.

First, I investigated ligand-dependent sites of Estrogen Receptor (ER) binding from Swinstead et al. (228). Of the 8173 ligand-dependent sites identified by csaw, 988 colocalized with promoter, while 6737 were found on non-coding regions. The ER recognition motif (shown from JASPAR database in Supplementary Figure 2.7A) was present in 48% of promoter proximal and 60% of non-coding binding sites (Supplementary Figure 2.7B). Interestingly, bases within the core recognition motif were slightly differently enriched following rGADEM motif discovery (Supplementary Figure 2.7C). These changes were validated and quantified by MoMotif, which also reveal that differential motif recognition at promoter and non-coding regions are limited within the core recognition motif of ER, as no extensions were detected, and the only noticeable change involves a background enrichment of C within the spacing region of the motif (Supplementary Figure 2.7D/E). This data supports MoMotif as being amenable to motif discovery using diverse datasets, and also indicate that changes in DNA-binding motifs are not invariably identified, highlighting the robustness of both the tool and our CTCF ZF1 mutation data.

Secondly, I probed ZNF263 ChIP-seq data (229) using MoMotif, and again divided the called peaks by their proximity to promoters or non-coding regions. Of the 2202 ZNF263 binding regions common among the two published peaksets, 314 were promoter proximal, while 1729 were found in non-coding regions. ZNF263 recognizes a GA rich repetitive motif without a clear consensus

(229,230). Using Perkisov et al. software and databases (217,218), I validated the specificity of ZNF263 zinc-fingers for G and A enriched motifs, as each of its zinc fingers recognizes primarily these 2 bases (Supplementary Figure 2.7F). This hints that the repetitive motif is likely directly recognized by ZNF263 and not artificial. Following rGADEM analysis, both groups display a G and A rich motif, with a slightly longer motif being found in promoter proximal ZNF263 binding sites (Supplementary Figure 2.7G). Due to the repetitive nature of the motif, direct comparison of both motifs at this step is arduous, as it is unknown how the 2 motifs align together and whether the bases present in the longer, promoter proximal, motif are also present outside of the identified non-coding regions motif. However, using MoMotif sequence alignment, extension, quantification, and analysis, revealed the exclusivity of the motif extension at both end of the promoter proximal motif and the strong enrichment of A at two positions within the non-coding sites motif, while offering detailed quantification and statistical analysis of the changes in bases frequencies (Supplementary Figure 2.7H/I). These observations promote the concept that ZNF263 binding on promoters is dependent on a longer combination of its zinc fingers, or cofactors beyond its zinc-fingers, while binding at non-coding region might be facilitated by a fewer, but more specific, combination of zinc-fingers.

Overall, these analysis provide examples of the power of MoMotif to expand classical motif analysis with discovery, validation and quantification of motif variability between experimental conditions or functional regions.

Methods:

Cell Culture Details:

MCF10A cell lines were maintained in DMEM/F12 50/50 (Wisent, #319-085-CL) supplemented with EGF (100µg/ml, Wisent, #511-110-UM), Insulin (10mg/ml, Wisent, #H511-016-U6), Hydrocortisone (1mg/ml, Sigma, #H0888-1G), Horse Serum (2%, Wisent, #065150) and Choleratoxin (1mg/ml, Sigma, #C8052-2MG) in an incubator at 37C and 5% CO2.

MDA-MB-231, MCF7, SKBR3 and HEK293T cell lines were maintained in DMEM (Wisent, #319-005-CL) supplemented with 10% FBS (Gibco, #12483-020) in an incubator at 37C and 5% CO2.

Conditionally Reprogrammed Cells of Patients Derived Xenograft of Triple Negative Breast Cancer tumors, termed PDXs, were established following as in Liu et al. 2017 (231) and Sirois et al. 2019 (232) and given to us by Dr. Park's and Dr. Basik's Laboratories from McGill University. Low CTCF protein expression in these cell lines was confirmed by Western Blot. Loss of Heterozygosity was confirmed using the Chromosome Analysis Suite from ThermoFisher. They were maintained in DMEM (Wisent, #319-005-CL), with 25% Ham's F12 (Wisent, #312-250-CL), 8% FBS (Gibco, #12483-020), L-Glutamine (1.5mM, Wisent, #609-065-EL), EGF (50µg/ml, Wisent, #511-110-UM), Insulin (5mg/ml, Wisent, #H511-016-U6), Hydrocortisone (0.8mg/ml, Sigma, #H0888-1G), Choleratoxin (84µg/ml, Sigma, #C8052-2MG), RhoK Inhibitor (0.01mM, Y-27623, StemCell Technologies, #72304) in an incubator at 37C and 5% CO2.

CRISPR/Cas9 Editing

CTCF H284N knock-in was performed similarly to those we previously described in Hilmi et al. 2017 (64). sgRNA guides targeting the genomic region around the nucleotide triplet coding for CTCF H284 were inserted into the vector backbone pSpCas9(BB)-2A-GFP (PX458) (Addgene, #48138) (Supplementary Table 1). A 250 base pair DNA donor, homologous to the region, but

replacing the CAC, coding for H284, by AAC, coding for H284N, were also designed and ordered with IDT (Supplementary Table 1). Introduction of plasmids and donor to $1x10^6$ MCF10A cells was carried out in a 6-centimeter dish using Lipofectamine 3000 (Invitrogen, # L3000001), 6µg of pCas9+guide and 12µl of 10mM DNA Donor. Two days later, GFP-positive cells were selected by fluorescence-activated cell sorting of individual cells into 96-well plates. To screen for CTCF H284N mutant cell clones, we isolated genomic DNA of each clone and amplified proximal sequences surrounding the Cas9 targets by polymerase chain reaction. Positive clones were first identified using the SURVEYOR Assay Kit (IDT, #706020). Then, individual alleles of positive clones were validated by Sanger Sequencing (GenomeQuebec) following PCR amplification and Zero Blunt TOPO PCR insertion and Cloning (Invitrogen, #45-0245) . Genomic DNA sequences were also compared to CTCF coding sequence using BlastX (blast.ncbi.nlm.nih.gov), to validate the presence of a mutation at the H284 position (Supplementary Figure 2.1A).

Western Blot Protocol:

Western blots were carried out as previously described (64). For Western Blot conducted on PDX tumors, tissue was harvested as in Savage et al. 2020 (233). For all other Western Blots, cells were harvested by scrapping. Then, cells are lysed in whole-cell lysis buffer [20 mM tris (pH 7.5), 420 mM NaCl, 2 mM MgCl2, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.5% Triton X-100, supplemented with fresh 1 mM dithiothreitol, phenylmethylsulfonyl fluoride, protease inhibitor cocktail (Roche) and phosphatase inhibitors, bis-glycerol phosphate, and NaF] for 15min, then spined at 13000rpm at 4°C for 15min to pellet cellular debris. Then, the protein concentration of the supernatant is assessed using a Bradford assay (Fisher, #1856209). 40µg of proteins are loaded on an 8 to 12% acrylamide gel and electrophoresed at 120V for 1h. Then, proteins on the gel are transferred on nitrocellulose membrane (Pall, #66485) at 4°C, 34V overnight for 8% gel and 100V for 1 hour for 12% gel. The membrane is then blocked with 5%

milk in TBST [20 mM Tris base, 137 mM NaCl, and 0.1% Tween 20] for 3h at 4°C. The membrane is then incubated with primary antibodies (see Supplementary Table 2) overnight at 4°C. Membranes are rinsed and washed for 10 minutes twice with TBST prior to secondary antibody incubation with goat anti-rabbit (SeraCare, #5220-0458) or anti-mouse (SeraCare, #5450-0011) dilute 1/10000 or 1/20000 in 5% milk in TBST. Membranes are washed again 3 times for 10 minutes in TBST, then revealed using ECL (Bio-Rad, #170-5061). Band intensity quantification and normalization on background is performed using ImageJ software.

Growth Curve Details:

15000 cells were plated in one well of a 12 wells petri dish (Fisher Scientific, #3513). Cells were fixed at day 1, day 3, and day 5 with 4% formaldehyde and stored at 4°C. Once all the conditions are fixed, the cells are stained with 1ml of a crystal violet solution [1% crystal violet, 10% EtOH] and dried. The stained cells are then diluted in 10% acetic acid. The growth ration of cells is calculated by reading and the comparing the DO at 595nm, the intensity of the violet dye, using PerkinElmer Multimode Plate Reader.

Lentiviral Infection for CTCF Addback and SNAI1 Knockdown Details:

CTCF addback (Genecopoeia, #EX-Z8806-Lv120), SNAI1 shRNA knockdown (Sigma, #NM_005985, target sequence: GCAAATACTGCAACAAGGAAT) and GFP control vector (Genecopoeia, #EX-EGFP-Lv120) were done using lentiviral vectors packages in HEK293T cells, as described previously (64). HEK293T cells were transfected with 7µg of the required lentiviral vector combined with 5µg of packaging vector MD2G and 2µg of envelope vector Pax2 using polyethylenimine (1 mg/ml). 24h after transfection, media was changed for the culture media used for maintenance. Viruses were collected at 48h and 72h after transfection and passed through a 0.45-µm filter. For the infection, MCF10A and PDXs cells were infected in six-well dishes and incubated with 1ml of viral supernatant along with 1ml of their respective culture media and

8µg/ml of hexadimethrine bromide (Polybrene) for MCF10A and 60µg/ml for PDXs. 24h after infection, culture media is changed and puromycin selection starts.

MCF10A cells were selected with 1 µg/ml of puromycin for the first two days following infection, followed by 0.25µg/ml of puromycin for 2 to 3 more days. Culture media is changed for puromycin free media 24h before the starvation period preceding the invasion assay. For the PDX cells, 1µg/ml of puromycin is used on the first day of selection, 0.5µg/ml of puromycin is used for the second day, culture media is changed for puromycin-free media on the third day and the preliminary starvation starts on the fourth day.

Transfection for shRNA CTCF Knockdown:

Transfection is done following the Lipofectamine™ 3000 Transfection Reagent protocol (Invitrogen, # L3000001), using 5ug of shCTL or shCTCF plasmid from Origene (Origene, # TL313675, Locus ID 10664). 30min before transfection, culture medium of 50% confluent cells is changed to Opti-MEM (Gibco, #11058-021). 6h after transfection, culture medium is changed for normal culture medium, with 1ug/ml puromycin. 24h before starvation, culture medium is changed for puromycin-free medium.

Transwell Invasion Assay and Quantification Details:

Conditions for invasion were optimized for each cell lines as follows. 70% confluent cells were starved for 24h in non-supplemented DMEM/F12, for the MCF10A cells, or DMEM for MDA-MB-231, MCF7, SKBR3 and PDXs. 50 000 MCF10A cells or 100 000 MDA-MB-231 or SKBR3 cells or 200 000 PDX or MCF7 cells are seeded into an insert (Falcon, #353182) coated with 25µg/ml matrigel (Corning, #354230) for the MCF10A, MCF7 and MDA-MB-231 or 20µg/ml matrigel for the PDXs or SKBR3, diluted in in a 0.01M Tris and 0.7M NaCl solution. The cells are maintained in non-supplemented media in the insert. The inserts are then placed in companion plate chambers (Falcon, #353503) containing supplemented media used for cell

culture overnight for the MCF10A and MDA-MB-231 or for 24h for the PDX, MCF7 and SKBR3.

For Torin1 (Tocris, #4247) or A485 (Tocris, #6387) treated conditions, cells were treated with the indicated concentration of Torin1 (0nM and 25nM) or A485 (0µM, 2µM and 5µM), diluted in DMSO, for 24h before starvation, during starvation and during invasion, in both the insert media and the companion plate media.

Then, the inserts are washed in PBS, fixed in 5% glutaraldehyde for 10 minutes, stained with a crystal violet solution [1% crystal violet, 10% EtOH] for 30 minutes, rinsed in water and dried. For each biological replicates, 2 to 3 inserts are plated; for each insert, 5 pictures are taken, at 10X resolution. The total number of invading cells on each picture are then counted using ImageJ software and the average number of invasive cells per 5 pictures per inserts are averaged within each sample and compared between samples. Statistical test between samples is performed using Student t-test.

Lentiviral Infection for dCAS9 Details:

20k MCF10A cells are plated in each well of a 6 well-plate. To increase the rate of infection, cells are infected sequentially over a period of 4 days: 24h and 72h after seeding, cells are infected with dCAS9+blasticidin resistance lentiviral construct (Addgene, #85417), with a ratio of media to dCAS9 viral media of 1:1 and 30ug/ul of polybrene; 48h and 96h after seeding, cells are infected with guideRNAbackbone+puromycin resistance lentiviral construct (Addgene, #52963, in which gRNA sequence (CACCGGAGGACAGAGAGACAAGTGT) generated with CHOPCHOP(234) were cloned into by Norclone), with a ratio of media to gRNA viral media of 1:1 and 30ug/ul of polybrene. Culture media is changed every 24h during infection period and afterward, until cells are harvested. 2ug/ul blasticidin selection starts after 48h and is reduced to 1ug/ul after 72h, until cells are harvested. 1ug/ul puromycin selection starts after 72h and is

reduced to 0.5ug/ul after 96h, until cells are harvested. Cells are harvested for RNA extraction 3 days after the last infection (7 days after seeding).

Mammosphere assay and Quantification Details:

5000 cells were seeded on a 50µl matrigel cushion (10-12mg/ml, Corning, #354230) and maintained in supplemented DMEM/F12 containing 4% matrigel for 8 days. The media is carefully replaced every 3 days. Average mammosphere size was measured from brightfield microscopy images on ImageJ software. Statistical test between the average mammosphere size of each sample was performed using Student t-test.

Mammosphere Immunofluorescence and Quantification:

p-S6 immunofluorescence was performed using p-S6 S240/244 antibody from Cell Signaling (Rabbit, #2215S) and Goat Anti-Rabbit IgG with Alexa 488 fluorophore (Invitrogen, #A32731). DAPI was used for DNA fluorescence of the whole mammosphere used for normalization of p-S6 fluorescence quantification and mammosphere filling quantification.

Mammosphere filling was quantified from Z-stacks of DAPI stained mammosphere images using ImageJ software. The ratio between the area of the hollow cavity and the total area of the mammosphere was measured on each Z-stacks of each mammosphere of each sample and the Z-stack with the highest ratio was selected and quantified for each mammosphere of each sample. Statistical test between the filling ratio of samples was performed using Student t-test.

The quantification of p-S6 fluorescence was performed using a custom script in ImageJ developed by Dr. Luke McCaffrey's group. In brief, mammospheres were detected by thresholding the image (Mean method) to create a whole-organoid mask. This mask was duplicated and then iteratively eroded (13 times) to create an inner mask that excluded the outer layer of cells. A mask for the outer layer of cells was generated using an XOR gate applied to the whole organoid and inner mask. The mean pixel intensity (8-bit) was measured under the mask, for each whole organoid,

outer, and inner regions. The mean pixel intensity of each region was then compared between the samples. Statistical test between the p-S6 outer fluorescence of each sample was performed using Student t-test.

RNA-Seq Data Processing and Analysis Details:

The overall quality of reads and sequencing was assessed before and after trimming using the FastQC package (Babraham Bioinformatics). Prior to mapping, reads were trimmed with Trimmomatics (235) using the following condition: ILLUMINACLIP:$Adapters:2:30:10:8:true, HEADCROP:4, SLIDINGWINDOW:4:30, LEADING:3, TRAILING:3, MINLEN:30. Alignment on hg19 human genome was performed with STAR 2.5.4b (236) default parameters, and converted into bam format using Samtools 1.9 (237). Differential expression analysis was generated using FeatureCounts count matrix (238) followed by DESEQ2 analysis (185), using default parameters and prefiltering, for comparison across samples.

Downstream RNA-Seq Analysis in Chapter 1:

RNA-Seq Correlation and Volcano Plot: Correlation and Volcano plot representation of the RNA-Seq results was generated using the DESEQ2 calculated Log2FC and -log(adjusted p-value) of the respective MCF10A CTCF+/- compared to MCF10A CTL for every gene with a basemean > 100. Genes with p-value < 0.05 were represented in grey. Genes with Log2FC > 0 were represented in orange. Genes with Log2FC < 0 were represented in purple.

RNA-Seq GSEA Pathway Analysis: Pathway analysis was performed using GSEA tools default setting on the read count matrix of all significantly altered genes (basemean > 100, p.value<0.05)(239). All gene sets shown were significant for both p-value (< 0.001) and FDR (<0.25). Pathway names were shortened as follows, with the full name of each pathway being:

PI3K Signaling : GO_PHOSPHATIDYLINOSITOL_3_KINASE_SIGNALING,

EMT : GO_EPITHELIAL_TO_MESENCHYMAL_TRANSITION,

EMT Regulation : GO_POSITIVE_REGULATION_OF_EPITHELIAL_CELL_MIGRATION

PI3K Regulation: GO_POSITIVE_REGULATION_OF_PHOSPHATIDYLINOSITOL_3_KINASE_SIGNALING

Cell-Cell Adhesion:

GOBP_HETEROPHILIC_CELL_CELL_ADHESION_VIA_PLASMA_MEMBRANE_CELL_ADHESION_MOLECULES

PI3K Signaling in Cancer : REACTOME_PI3K_AKT_SIGNALING_IN_CANCER

Constitutive AKT Signaling: REACTOME_CONSTITUTIVE_SIGNALING_BY_AKT1_E17K_IN_CANCER

RNA-Seq KEGG and REACTOME Pathway Analysis: Pathway analysis for KEGG and PANTHER-Reactome was done using PANTHER webtool (http://www.pantherdb.org/) (240). Reactome pathway analysis was done using Reactome webtool (https://reactome.org/) (241).

RNA-Seq Heatmaps: Heatmaps were generated using the Log2FC to the average normalized read counts in MCF10A CTL of the 20 genes with the highest absolute Log2FC in the PI3K Regulation and EMT Regulation genesets.

Downstream RNA-Seq Analysis in Chapter 2:

RNA-Seq Dot Plot: Dot plot representation of the RNA-Seq results was generated using the DESEQ2 calculated Log2FC and -log(adjusted p-value) of the respective mutant MCF10A compared to CTL MCF10A for every gene with a basemean > 100. Genes with p-value < 0.05 were represented in grey. Genes with Log2FC > 1 were represented in orange. Genes with Log2FC< -1 were represented in purple.

RNA-Seq GSEA Pathway Analysis: Pathway analysis was performed using GSEA tools (239) default setting on the read count matrix of all genes (basemean > 10). All gene sets shown were significant for both p-value (< 0.001) and FDR (<0.25). Pathway names were shortened for esthetic purposes in the Figure 5B, with the full name of each pathway being written in Figure 5C.

RNA-Seq Heatmaps: Heatmaps were generated using the Log2FC with CTL MCF10A of genes with the highest absolute Log2FC from the following significantly altered pathways:

"GOBP_RESPONSE_TO_XENOBIOTIC_STIMULUS";

"REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION"

TCGA RNA-Seq Analysis: Average gene expression of each gene in breast cancer patient with CTCF ZF1M was compared to average gene expression in CTCF WT breast tumors. Log2FC of significantly altered genes in patients were then compared to Log2FC of significantly altered genes in MCF10A CTCF ZF1M/ZF1M and CTL.

RT-qPCR Protocol:

Total RNA was extracted according to Sigma RNA Extraction Kit (Sigma, #RTN350-1KT) protocol. RNA quantity and quality was measured using Nanodrop. 500ng of RNA are used as template for Reverse-Transcriptase PCR, following the manufacturer protocol (All-In-One RT MasterMix, ABM, #G490). The cDNA is diluted 1:10 and 2µl is used for qPCR amplification, following the manufacturer protocol (GoTaq qPCR MasterMix 2X, Promega, #A600A). Relative levels of cDNA are compared between the samples using the $2^{-\Delta\Delta CT}$ formula normalized on the average level of 3 housekeeping genes (GapDH, RPL4 and RPLPO). Statistical tests between normalized gene expression of each gene for each sample is performed using Student t-test. The sets of primers used for RT-qPCR are listed in Supplementary Table 3.

TCGA Data Analysis: For pathway analysis, Spearman correlation test was calculated between CTCF expression in patients and the individual expression of each gene surveyed in TCGA and with a baseMean > 100 in our RNA-Seq. Genes were then ranked by -log(p-value), in which p-values equal to zero were brought to the smallest non-zero p-value measured. The ranked list was then analyzed using GSEA PreRank analysis (239) and ranked by Normalized Enrichment Score (NES) or p-value. For the SNAI1 box plot, the dataset was separated into the high CTCF group, being the top 20% of patients in term of highest CTCF expression, and the low CTCF group, being the top 20% of patients in term of low CTCF expression. The two groups were compared for

SNAI1 RNA expression using a Student's T Test and a Spearman correlation test.

ChIP-Seq Sample Preparation Details:

70-80% confluent cells were fixed for 10 minutes in 4% formaldehyde and stored at -80C. The pellets were subsequently resuspended in 1ml of ChIP-buffer [0.25% NP-40, 0.25% Triton X-100, 0.25% Sodium Deoxycholate, 0.005% SDS, 50nM Tris (pH8), 100mM NaCl, 5mM EDTA, 1X PMSF, 2mM NaF, 1X P8340 Cocktail Inhibitor (Roche)] and sonicated with a probe sonicator (Fisher Scientific Sonic Dismembrator Model 500) using the following cycles: 5 cycles at 20% power, 5 cycles at 25% power, and 5 cycles at 30% power. Each cycle lasts 10 seconds, and the samples are kept on ice between each cycle to avoid overheating. Next, the samples are spun at high speed in a microcentrifuge for 30 minutes. Then, lysates are collected and protein concentration measured using the Bradford assay, as described above. Based on protein concentrations, samples are diluted to 2mg/ml proteins in ChIP-buffer and 50ul/ml of Protein G Plus-Agarose Suspension Beads (Calbiochem, IP04-1.5ML) are added for 3h to preclear. 2% of the sample is collected as input and kept at -20 °C until DNA purification. Immunoprecipitation is carried out at 4°C overnight with 1ml of sample, 60ul of beads and primary antibody (see Supplementary Table 2). The beads are then washed once with Wash1, Wash2, Wash3 [0.10% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris (pH 8), 150/200/500mM NaCl for Wash 1,2,3 respectively], Wash LiCl [0.25M LiCl, 1% NP-40, 1% Sodium Deoxycholate, 1mM EDTA, 10mM Tris (pH8)] and twice with TE buffer [10mM Tris (pH8), 1mM EDTA]. Then, beads are resuspended in elution buffer [1% SDS, 0.1M $NaHCO_3$]. The samples are decrosslinked overnight at 65 °C. 20µg of Proteinase K (Sigma, # 39450-01-6) is added for 1h at 42 °C. Then, DNA is purified using BioBasic DNA collection column (BioBasic, #SD5005). DNA concentration was assessed via Picogreen assay (Invitrogen, #P7589).

ChIP-Seq Data Processing:

Quality control of reads and sequencing was assessed before and after trimming by FastQC (Babraham Bioinformatics). Reads were trimmed with Trimmomatics (235) using the following parameters: ILLUMINACLIP:$Adapters:2:30:10, LEADING:30, TRAILING:30, SLIDINGWINDOW:4:30, MINLEN:30. Alignment on hg19 human genome was performed using BWA (237) default conditions. Sam files generated by BWA were converted to bam format using Samtools (237). Peak calling was performed with MACS2 (195) default condition and normalization on the respective Input dataset of each cell lines. Bigwig files used for visualization were generated from the fragment pileup bedGraph using the BedGraphToBigwig function.

Downstream ChIP-Seq Analysis in Chapter 1:

ChIP-Seq Differential Binding Analysis: Differentially binding region were quantified using DiffBind 3.0 (215). Bam and narrowPeak files for each samples and bam files of the corresponding input were used. Default normalization and analysis was performed for H3K4me3 and H3K27ac. CTCF normalization and analysis was performed with the following parameters: normalize = DBA_NORM_DEFAULT, library = DBA_LIBSIZE_PEAKREADS, background = F, bREtrieve = F. Threshold of significance were set at FDR <= 0.01 and abs(LogFC) >= 1 in all conditions. Consensus differential peaksets between replicates and conditions were used for further downstream analysis and converted to Grange format using GenomicRanges R packages (242). The number of sites in each peak set was used for quantification and the generation of pie chart.

ChIP-Seq Genomic Distribution and Sites Annotation: Genomic distribution and annotation were performed using clusterProfiler package (243) and ChIPSeeker package (244) on the differential binding sites identified by DiffBind 3.0, using the TxDb.Hsapiens.UCSC.hg19.knownGene as reference for gene location. TSS regions were defined with a +/- 3000bp overlap during peak annotation. The regions are annotated as : 5'UTR, Promoter (<=1kb), Promoter (1-2kb),

Promoter (2-3kb) are referred to as “Promoter (+/- 3kb)”; 1st Exon and Other Exon are referred to as “Exons”; 1st Intron and Other Intron are referred to as “Introns”; 3’UTR and Downstream are referred to as “Downstream” and Distal Intergenic is referred to as is.

ChIP-Seq Differential KEGG Pathway Analysis: Pathway analysis was performed using the annotation files from above and using the compareCluster function from the previously mentioned clusterProfiler package and the following parameters: geneCluster = genes, fun = "enrichKEGG", pvalueCutoff = 0.05, pAdjustMethod = "BH". -log(p.value) of enrichment significance was used for bar chart representation, where pathways were ranked according to geneRatio. The differential analysis was then performed by repeating these steps on 100 randomized subsets of constant CTCF sites equinumerous to the number of lost sites and on 100 randomized subsets of constant H3K27ac within TADs equinumerous to the number of gained H3K27ac within TADs. Then, the average p-value of the surveyed pathway within the 100 subsets was calculated and compared to the p-value in the lost CTCF sites or gained H3K27ac sites. P-values in subsets in which the surveyed pathways were not detected were overestimated at 0.25, as the most significant value for which pathways are not called by the program.

ChIP-Seq and RNA-Seq Dot Plots and Correlation: Dot plots were made by combining the RNA-Seq Log2FC between CTL and CTCF+/- #1 and #2 and the logFC from DiffBind of any called peak annotated on that gene (+/- 3kb) by clusterProfiler. Spearman correlation on the dot plot were performed using the ChIP-Seq logFC and RNA-Seq log2FC of every peak colocalization with a gene. Genes associated with a peak with a DiffBind or DESEQ2 adjusted p-value > 0.05 or DiffBind LogFC or DESEQ2 Log2FC < 1 are represented in grey, while those with a LogFC/ Log2FC >= 1 or <= -1 are represented in orange and purple, respectively.

ChIP-Seq Heatmaps, Profile Plot, Tracks: Heatmaps, profile plot and tracks were generate using

deepTools and samtools (237,245). Heatmaps and Profile plot were generated using 3kb regions centered around the differential peakset identified by DiffBind and bigwig from MACS2. Both the computeMatrix and plotHeatmaps were runned with default parameter; yMax, zMax and colors were adjusted in each condition to better represent the results. Tracks were generated as profile plot of the single genomic regions of interest with a gene annotation track from IGV(246) under each figure to represent the relative location of the gene of interest.

ChIP-Seq Colocalization Analysis: Analysis of colocalization, +/- 3kb, was performed using a genomic overlap algorithm between the position of differentially binding peak sets identified by DiffBind. Observed/Expected ratios shown for colocalization of ChIP-Seq peaks were calculated using the Chi-Square formula in Microsoft Excel.

MoMotif Analysis pipeline

The analysis sequence for the discovery of modification of motif is comprised of three principal steps: Step1: identification of sites of differential DNA binding; Step 2: discovery of motifs enriched within DNA binding sites that are either gained, lost or stable binding under experimental conditions; and Step 3: learning the discriminative motifs. These steps are conducted using three R packages: csaw, rGADEM and MoMotif, as illustrated in Figure 2. The first two packages have been widely utilized by the scientific community, but MoMotif, written in *R*, was developed specifically for this project.

Step 1: Differentially binding analysis: csaw: The first step involves quantifying binding intensity/counts from the aligned ChIP-Seq reads and *de novo* detection of differentially bound regions while controlling the genome-wide false discovery rates (FDR). For these processes, we rely on an existing R package, csaw (247). csaw uses a sliding window-based approach to summarize read counts across the genome. It examines the differential binding at the window level using quasi-likelihood F-tests with empirical Bayes-based dispersion estimations, which naturally

handle low, over dispersed counts with a limited number of replicates (248). csaw then aggregates adjacent windows into regions for output. The p-values for the aggregated regions are calculated using Simes' method (249), which correctly controls FDR at the region level. Our detailed steps for this differential binding analysis are summarized in Supplementary Figure 2A. We used a window of size of 10bp with spacing of 50bp to count the aligned reads. The differentially bound regions were detected using an FDR cut-off of 0.05. The outputs from this csaw pipeline are three sets of genomic regions (of varying lengths); experimentally induced 1) gain of binding 2) lost binding and 3) binding regions with no statistically significant differences between control and experimental conditions. Hereafter we refer to these three sets of genomic sequences as gained, lost and constant clusters.

Step 2: de novo motif discovery, rGADEM: Once lists of binding regions are returned by csaw, the next step of our new pipeline involves discovering enriched motif models. For this step, we rely on another existing R package rGADEM (250) (Droit A, et al. R package version 2.42.0), built upon the GADEM algorithm (166). GADEM is an efficient de novo motif discovery method that combines the two commonly used techniques for pattern matching; word enumeration and probabilistic local search. Enumerative methods identify motifs by counting all m-letter patterns, such as the method Drim (251). Probabilistic approaches model starting positions of motif patterns as latent variables and infer the final motif models using the Expectation-Maximization (EM) algorithm; such methods include MEME (252,253) and fdrMotif (254). Specifically, GADEM constructs spaced dyads by enumerating candidate words (4 to 6 nucleotides), and then uses them as starting positions to guide an EM algorithm for unbiased motif discovery.

We applied rGADEM to the three clusters of sequences obtained from the differential binding analysis step. To ease the computational burden and to focus on the most robust differentially

bound motifs, we performed the motif discovery analysis exclusively on the top 1000 regions in the gained and lost clusters, and the bottom 1000 regions (with the largest adjusted p-values) in the stable cluster, separately. The main outputs include the enriched motif models for each cluster, represented by either position weight matrices or consensus logos. Along with a specific motif, rGADEM also reports other helpful information, including all sequences in the input data incorporating this motif and the location of the identified motif patterns in the original sequence data. This information is subsequently employed as the input for the following discriminative motif analysis step.

Step 3: Discriminative motif analysis and result visualization, MoMotif: To detect small or subtle variations built upon a primary known motif, we have developed a new discriminative motif analysis tool, MoMotif, that represents the concluding step in our pipeline. This approach starts with the short core motif reported by rGADEM, which incorporates the core pattern of our primary known motif. We then retrieve and align all sequences carrying this core motif, referred to as core sequences, for each cluster. For a comprehensive characterization of subtle variability occurring within and around the core motif, we extend both ends of the core sequences by several base pairs (a user-chosen parameter permitting versatility). This strategy results in a set of adequately aligned long sequences of the same lengths, which allows us to compare the nucleotide distribution at each single base-pair to see which base pairs seem to distinguish clusters.

Next, we are able to compare the extended sequences in the lost or gained cluster to the stable cluster by assessing the statistical significance of differences in nucleotide frequency at each position. We used the Pearson's chi-square test to assess the statistical significance of the difference in nucleotide distribution at one position between two sets of aligned sequences (lost vs. stable or gained vs. stable). For a given position, let n_i^j be the number of sequences in Group

i that have nucleotide j at this position, where $i = 1, 2$ and $j = A, T, G$ and C. Let n_i be the total number of sequences in Group i, n^j be the number of sequences with nucleotide j at this position in both groups, and n be the total number of considered sequences, i.e. $n = n_1 + n_2 = n^A + n^T + n^G + n^C$. These notations are summarized in the following contingency table.

	A	T	G	C	TOTAL
Group 1	n_1^A	n_1^T	n_1^G	n_1^G	n_1
Group 2	n_2^A	n_2^T	n_2^C	n_2^G	n_2
TOTAL	n^A	n^T	n^G	n^C	n

Specifically, the chi-square test compares the observed frequencies in each subcategory with the frequencies one would expect if the two groups had the same nucleotide distribution. The expected frequencies, denoted as E_i^j, are of the form:

$$E_i^j = \frac{n_i \times n^j}{n} \quad for\ i = 1, 2\ and\ j = A, T, G, C.$$

Then the observed chi-squared test statistics can be calculated as

$$obs\_\chi^2 = \sum_{i \in \{1,2\}} \sum_{j \in \{A,T,G,C\}} \frac{\left(n_i^j - E_i^j\right)^2}{E_i^j}.$$

The p-value for the chi-squared test is thus defined as the right-tailed probability in a χ^2 distribution with degrees of freedom 3, i.e.

$$p - value = P(\chi_3^2 > obs\_\chi^2).$$

We repeated the test for all positions in the extended sequences and reported the p-values for each position. To control the family-wise error rate at a 5%, we suggest a stringent p-value threshold of $1*10^{-10}$ for declaring significance of a single position, which was derived from the approximate total number of 50M nucleotides in a small human chromosome. We also provided visualization to compare the significance level at each position relative to the overall significance

level in the extended region. Therefore, discriminative motif models are then identified as the smallest sub-region containing all sites reaching our stringent threshold of significance.

In addition, the MoMotif package contains functions for various output visualizations, including bar-plots showing the frequency for each nucleotide in a given set of sequences, sequence logo for the identified discriminative motif models and their position to the core motif of our interest. In our data analysis, we treated the 10th nucleotide in the canonical CTCF motif, shown in Figure 3A, as the center and extended by 30 bp in both directions.

MEME Suite - SEA (Simple Enrichment Analysis) : The same subset of the top 1000 constant and lost sites from CTCF ZF1M/ZF1M used for MoMotif analysis were used for SEA analysis. SEA was run on the MEME suite web tool (https://meme-suite.org/meme/tools/sea), using the option "Shuffled Input Sequences" for the motif enrichment in mutant cell line alone and "User-provided Sequences" for the comparative enrichment analysis of lost sites against constant sites.

ChIP-qPCR Protocol:

The Chromatin Immunoprecipitation was done following the ChIP-Seq protocol, however using only 1mg/ml of chromatin and 30ul of beads with the antibodies listed in Supplementary Table 2. Final ChIP-product is diluted in 60ul of DNAse-free water. qPCR was performed with the ChIP product following the manufacturer protocol (GoTaq qPCR MasterMix 2X, Promega, #A600A). 2-ΔΔCT formula was used for quantification, normalized on a 2% chromatin input of each sample and compared between sites and conditions. Primers used for ChIP DNA amplification are documented in Supplementary Table 4.

Methyl Array Protocol and Data Analysis:

Bisulfite conversion was performed using EZ DNA Methylation Kit (Zymo Research, #D5001). 500ng per sample of Bisulfite-converted DNA was sent at Princess Margaret Genomics Centre for quality control and detection of methylated bases using Illumina Human Methylation EPIC

Array. The ".idat" files outputted from the Illumina EPIC Array experiment were analyzed using Minfi package (83) for the comparison of individual red/green CpG probe intensity and genomic annotation, using Illumina Methylation EPIC reference: ilm10b4.hg19. Quality control of methylation pattern was performed using the Shinymethyl R package (84).

Methyl Array and RNA-Seq Dot Plot: Dot plot of methylation profile and RNA-Seq was generated using the Methylation LogFC between MCF10A CTCF+/- #2 and MCF10A CTL for all CpG colocalizing with a gene (+/- 2kb) with a DESEQ2 called basemean > 100. Spearman correlation was calculated using all points under these criteria.

Hi-C Data Processing and Analysis:

Quality control of reads and sequencing was assessed by FastQC (Babraham Bioinformatics). Raw sequencing read were mapped, filtered, and binned using the runHiC pipeline (208). Contact matrix were binned at 5kb and 10kb resolution and stored in ".cool" format.

Downstream Hi-C Analysis in Chapter 1:

Hi-C Data Processing for Pearson Correlation Analysis: Raw sequencing read were analyzed using cLoops2 pipeline (255) pre-processing program tracPre2.py, "cLoops2 pre" and "cLoops2 combine" functions, using default parameters. Pearson Correlation Coefficient was calculated using the pulled biological replicates of the pre-processing results using "cLoops2 estSim" function, using default parameters and bin size set to 5000, 10000, 500000 and 1000000.

Hierarchical TAD Calling: Hierarchical TAD calling was performed using the hiTAD function of the TADLib package (208), using the 10kb resolution contact matrix and default settings.

Domain Boundaries Colocalization Analysis: Colocalization of TAD boundaries and ChIP-Seq peak was determine as described earlier, using a simple genomic overlap algorithm between the called TAD boundaries and the differentially bound peak list generated with DiffBind, with an accepted overlap of +/- 10kb (+/- 1 contact matrix bin). Observed/Expected calculations were

performed as described earlier. The same algorithm and overlap were used across samples to determine altered boundaries. TAD and subTAD boundaries present only in the CTL, but in none of both the CTCF+/- clones were defined and quantified as lost boundaries. TAD and subTAD boundaries present in both CTCF+/- clones, but absent in the CTL were defined and quantified as gained boundaries. Boundaries found in the CTL and any of the CTCF+/- clones were defined as constant boundaries.

Constant or gained boundaries were defined as adjacent to lost boundaries is the next or second to next boundaries, in any direction, is lost. Significant enrichment of the gained boundaries next to lost boundaries, compared to constant boundaries, was performed using the Chi-Square formula in Microsoft Excel.

RNA-Seq Changes at Domain Boundaries: Constant and altered boundaries' genomic location (10kb each) were annotated as previously described in the ChIP-Seq Analysis section. The average RNA-Seq log2FC of the CTCF+/- #2 against the CTL of genes colocalizing with each type of boundaries was calculated and represented by a bar chart. Statistical test of the difference between the average RNA Log2FC at each boundary was performed using a Student's T Test. Pathway analysis of altered boundaries was performed as in the ChIP-Seq section.

Hi-C 2D Heatmaps: Genome wide and Chromosome 20 heatmaps were generated using Juicer (256) representation of observed interacting reads value. 5kb resolution heatmaps were generated from h5 converted cool files using the HiCExplorer packages (257) default settings, using the hicPlotMatrix function for interaction matrix and hicCompareMatrices for comparative interaction matrix and hicConvertFormat for the cool to h5 conversion. High resolution sub5kb HiC imaging was performed with the HIFI pipeline (212), using the Markovian Recombination Method on defined subsection of the genome (around SNAI1, chr20:48,550,000-48,750,000) and

default setting for this method.

Hi-C Pile-Up Plots: Pile-Up plots were generated using the cooltools package (Open Chromosome Collective), centered at the differential peak list from DiffBind (for CTCF and histone marks) or from the TADlib TAD caller (for TAD/subTAD boundaries), normalized on random background interaction and using default settings. Local interactions were map at +/- 200kb around the defined regions. Average interactions were mapped at +/-50kb around the defined region.

Insulation Score and Profile Plots: Insulation Score was calculated at 30kb resolution and output as a bigwig file using the FAN-C insulation command default settings (210) and the respective 10kb .cool matrix from MCF10A CTL and MCF10A CTCF+/- #2. The bigwig file was used to generate the profile plot using deepTools, previously explained in the ChIP-Seq section, at +/- 200kb around lost sites of CTCF or gained sites of H3K27ac colocalizing with TAD boundaries (+/- 10kb) or within TADs.

Downstream Hi-C Analysis in Chapter 2:

Hierarchical TAD Calling: Hierarchical TAD calling was performed using the hiTAD function of the TADLib package(208) and SpectralTAD package (220), using the 10kb resolution contact matrix and default settings.

Colocalization Analysis: TSS of altered genes (FDR < 0.05) and altered CTCF sites were mapped to TAD/subTAD boundaries (+/- 1 resolution bin/10kb) or within each TAD/subTAD. The distribution of strongly gained and lost gene (abs(LogFC) > 1) compared to all mapped genes was measured and compared using a ChiSQ test in each distribution of: TAD/subTAD, TSS location and CTCF status.

Individual Zinc Finger Motif Prediction

Human CTCF amino acid sequence from Ensembl (https://useast.ensembl.org) was inputted in

Perkov et al. 2014 and 2015 (217,218) webtool (http://zf.princeton.edu/b1h/index.html). 3bp predicted sequence from the F2 model were used for our analysis.

Quantification and Statistical Analysis:

Unless stated otherwise, all graphical representations display the mean and SEM of the sample's distribution. Unless stated otherwise, graphics and statistical tests were generated and performed using GraphPad Prism 9.1, GraphPad Software, San Diego, California USA, www.graphpad.com. Unless stated otherwise, all Student T-test are using the one-tailed method. Graphical models were created using Biorender. For each graph:

* : $p <= 0.05$

** : $p <= 0.01$

*** : $p <= 0.001$

**** : $p < 0.0001$

Discussion

The effect of CTCF on gene expression is highly dependent on topological context

The study of CTCF LOH and zinc finger 1 mutation in the context of breast cancer provides multiple insights into the importance of epigenetic events on oncogenic progression and the physiological role of CTCF in the regulation of transcriptional process. In CTCF +/- MCF10A cells, the loss of CTCF likely did not reprogram H3K27ac or H3K4me3 directly through recruitment of enzymes to CTCF lost sites. Similarly, CTCF lost sites did not frequently impact the transcription of the most proximal genes. However, we did link the loss of CTCF to altered subTAD organization. Such reorganizations strongly correlated with the accumulation of activating marks and gene expression changes when genes were found within at the boundary of or within an altered subTAD. Thus, CTCF appears to control gene expression changes, to a large extent, indirectly through controlling chromatin contacts at the subTAD level.

In MCF10A cells carrying CTCF H284N, the most significantly altered genes were associated with lost CBS within TADs, or at the boundaries of subTADs. This indicates that CTCF mutation to zinc finger 1 is responsible for the precise control of gene expression at these levels of organization. Moving forward, it would be interesting to investigate the impact of zinc finger 3,4 and 5 missense mutations, that are observed in endometrial cancer and DS-AMKL. These mutations would be expected to prevent CTCF binding to its core binding motif, which would result in a different epigenetic phenotypes by a potential direct impact on TAD boundaries.

Both models utilized by the work described herein, and analysis methods, demonstrate the mutual dependence of chromatin conformation on CTCF binding and transcriptional regulation on chromatin conformation. Also, my analysis of our diverse Hi-C, ChIP-Seq and RNA-Seq datasets highlighted another key aspect that is often left out of most modern analysis; the

importance of hierarchical organization.

The rational between the intrinsic dichotomy of TADs and subTADs

One the main takeaway from my investigation of CTCF in breast cancer is that TADs within TADs (subTADs) are directly linking altered CTCF binding to aberrant regulation of histone marks and gene expression. Oppositely, TADs themselves displayed significantly less reorganization and less association with changes in gene expression following the mono-allelic KO of CTCF. The functional divergence between these two structurally similar types of chromatin domains can be explained by simple, rational, mechanisms.

First, multiple independent studies highlighted mechanisms allowing TAD boundaries to remain stable across cell types or species (4,10,221) or following removal of crucial regulators of chromatin conformation at their boundaries, such as CTCF (211). For example, TAD boundaries are often bound by multiple CBS (221), meaning that the loss of a single CBS might be compensated for by the additional binding sites, leaving the boundary unaltered. Similarly, as TADs are mostly continuous along the genome and the loop extrusion process cannot overlap (109), the boundary of a TAD will not shift further than the next adjacent boundary. Additionally, in our CTCF +/- MCF10A, loss CBS displayed less Hi-C insulation and lower ChIP-Seq read density. By definition, TAD boundaries are better insulated than subTAD boundaries. Also, CTCF binding at TAD boundaries is more conserved (221) and in an environment with higher average interactions caused by actively transcribed housekeeping genes and tRNAs (111,112). Therefore, not only can the loss of a CBS be compensated at TAD boundaries, but loss of CTCF binding is also less likely at the boundary of a TAD, than within a TAD. These observations explain why subTAD interactions are more sensitive to altered CTCF binding. However, these conclusions do not clearly explain why changes in subTAD, and not TAD, are directly associated to aberrant

epigenetic regulation of gene expression.

Second, even when TAD organization shifts, it is inherently less closely associated to possible transcriptional changes than subTAD organization, simply due to subTADs being located within TADs. Few possible topological contexts would be permissive to altered regulatory interactions by displaced or fused TAD boundaries. For example, if a previously insulated cis-regulatory element, or promoter, is located within the region of shifting insulation and a compatible, and not further insulated, interactor is present within the shifted TAD, then a new regulatory interaction is made possible, while previous interactions could be lost (Figure D1).

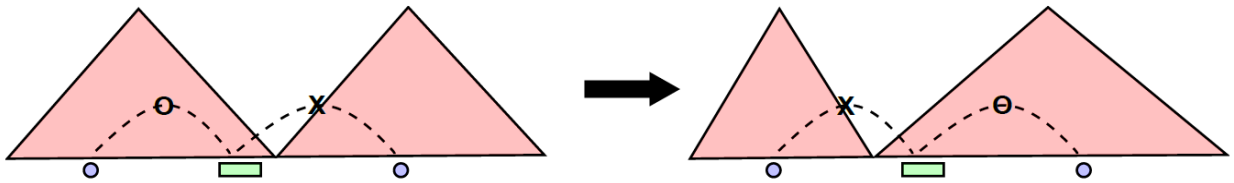


Figure D1: Example of a shift in TAD interaction permissive to transcriptional changes.

However, if the shifting TAD boundary does not encompass a regulatory element (Figure D2) or if the compatible interactors are insulated by subTADs (Figure D3/4), then changes in chromatin conformation at the TAD level are not permissive to transcriptional changes.

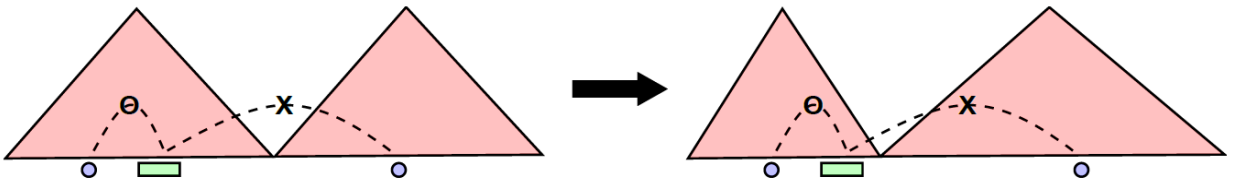


Figure D2: Example of a shift in TAD interaction not permissive to transcriptional changes since no regulatory element is encompassed within the shift.

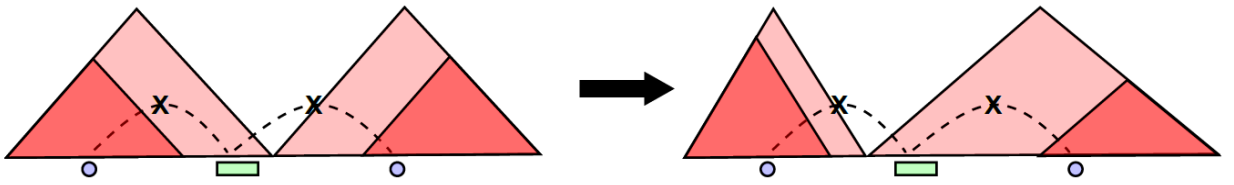


Figure D3 : Example of a shift in TAD interaction not permissive to transcriptional changes since the potential interactors are insulated within subTAD

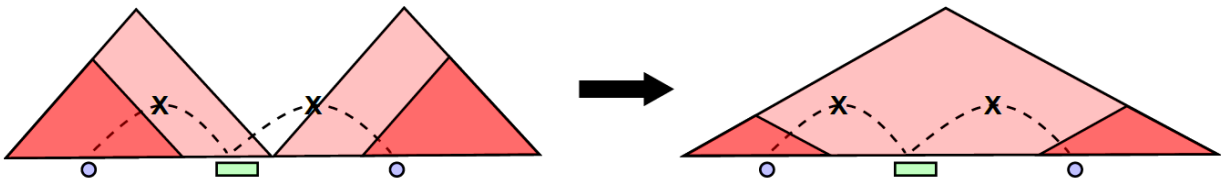


Figure D4 Example of a fusion of two TADs not permissive to transcriptional changes since the potential interactors are insulated within subTAD.

Shifts in subTAD boundaries follow the same rules. However, as CTCF binding at subTAD boundaries is less redundant and subTAD are not necessarily adjacent to another subTAD, shifts or loss of subTAD boundaries can encompass larger regulatory regions within the TAD, potentially permitting every regulatory elements within a single TAD to interact with each other. Therefore, the scenario presented in Figure D2 is less likely at the subTAD level. Further, the insulation of boundaries decreases from TAD to subTAD, and then subsequently from subTADs to interaction domains or loops within them. Therefore, regulatory elements encompassed by a shift in subTADs are more likely to be in an environment promoting their interaction with everything within the altered subTAD, making the scenario presented in Figure D3/4, here again, less likely.

In sum, these simple rational models demonstrate why, by definition, changes in subTAD organization are more likely to affect transcription than conformational shift at the TAD level. Additionally, the three-dimensional organization of the genome evolved together with the intrinsic biases of TAD and subTAD, enhancing the duality of stability at TAD boundaries versus adaptability at subTAD boundaries.

A hierarchy of stability and adaptability in chromatin conformation

The inherent differences created by the hierarchical organization of TAD and subTADs might have

been co-opted by the evolution of genome organization to promote homeostasis, while creating vulnerabilities that can be exploited by oncogenic events affecting chromatin conformation, such as CTCF loss or mutation.

As mentioned previously, TAD boundaries often colocalize with housekeeping or tRNA genes (111,112), many of which are essential transcripts for the survival of a cell. The stability of TADs, due to CTCF redundancy and conservation at TAD boundaries (221) and the exclusivity of chromatin extrusion by cohesin (109), is beneficial to the stable expression of these essential genes. Additionally, essential genes tend to cluster together (258). Therefore, an essential gene encompassed by a shifting TAD boundary could go from the transcriptionally active environment of a TAD boundary to the transcriptionally active environment of the neighboring one. This would result in a similar expression level, despite the shift in TAD organization. Coupled with the small range of possible shifts caused by the continuous presence of TADs along the genome, the enrichment of similarly express genes at TAD boundaries adds another layer to the stability they provide.

Genes and functions associated with subTADs benefit from the adaptability conferred by a hierarchically lower level of topological organization. subTADs are more dynamic than TAD, making them well suited to encompass and regulated groups of genes with condition dependent transcription. Already, multiple independent investigations highlighted the specific role of subTAD organization in the regulation of gene expression in response to nutrient availability or developmental signals, including PI3K signaling (259), circadian rhythm (260) and senescence and EMT pathways (261). To be competent in their ability to dynamically alter transcriptional states, subTADs need to be situated at a genomic locus where a small shift will insulate or incorporate cis-regulatory elements and promoters prone to altered transcriptional states. Without

that, dynamic subTADs organization could not, by itself, be conducive to swift changes in gene expression in response to diverse signals. However, this ability to adapt comes at the price of a reduced topological stability and higher risk of malignant transcriptional reprogramming caused by aberrant conformation in the presence of dysfunctional regulators, such as CTCF.

In sum, the specific sensitivity of disrupting subTADs, and acquiring new subTADs upon loss of CTCF can be explained by following a logical discourse. First, subTADs have, by definition, weaker insulation than TADs. Meaning that the binding of chromatin conformation regulators, such as CTCF, are more likely to be stochastically lost at their boundaries following reduced levels or loss of function. Second, subTAD boundaries evolved to be mobile. Therefore, the loss of an insulator is more likely to cause a shift in their boundaries, as we demonstrated in the CTCF LOH investigation. Third, subTADs are in an environment where small topological shifts are conducive to transcriptional changes. Thus, genes requiring condition-dependent regulations, such as genes of the PI3K pathway or EMT, in epithelial cells, are intrinsically more likely to be aberrantly transcribed following the mutation or loss of an insulator of chromatin conformation, such as CTCF. In other words, the altered DNA-DNA contact activating the genes of the PI3K and EMT pathway following the loss of CTCF are likely not unique to CTCF LOH cells, but could instead be interactions that are programmed to happen in specific cell states or conditions, such as changes to nutrient availability and growth factor signaling. However, the loss of insulation caused by lower CTCF levels permits these interactions, despite the cells not being in the required environment to inherently promote these contacts. In turn, this promotes oncogenic progression due to an aberrant transcriptional timing of key regulators of cellular functions or homeostasis. To validate this hypothesis, it would be interesting to carry out high-resolution Hi-C and RNA-Seq experiments under various physiological stress and growth conditions, such as starvation or

hypoxia. Then, the result could outline an enrichment of transcriptionally relevant subTAD changes compared to fewer relevant TAD changes. Additionally, analyzing which subTADs are altered in each conditions and how it affects gene expression would reveal if indeed, these sets of genes and subTADs are more sensitive to the loss of CTCF. This experiment would also expose whether the loss of CTCF epigenetically mimics any particular stresses, which could hint at potential therapeutic avenues.

Cell Type Specificity of Occurrence and Vulnerabilities

My investigations revealed that the transcription of genes of the PI3K and EMT pathways are especially sensitive to subTAD reorganization caused by CTCF LOH in mammary epithelial cells. The specificity of affected pathways and the mechanism of their altered transcription open potential targeted therapeutic avenues by inhibitors of the mTOR or histone acetylation. However, due to varied subTAD organizations between cell types and during differentiation, these sensitivities are likely to be cell type specific, as are the enrichments of CTCF deletions and mutations.

In terms of prevalence of CTCF genetic defects in diverse types of cancer, tumors of epithelial origin are dominant. Using cBioPortal analysis of TCGA Pan-Cancer datasets of 2018 (10,953 patients), the five types of cancer in which CTCF alterations are the most prevalent are all of epithelial origin (Endometrial, Bladder, Esophagogastric, Breast and Head and Neck). In these types of cancer, the sensitivities we highlighted in breast epithelial cells are likely to be reproducible. Indeed, most epithelial cells are competent to undergo EMT under specific conditions, such as wound healing (262). Further, epithelial cells of diverse origins can have adaptable survival or growth rate mediated, in part, by the PI3K pathway depending on nutrient availability (263) or tissue regeneration (264). Therefore, genes related to these pathways cannot

be stably silenced or constantly expressed in epithelial cells, as their expression needs to be adaptable. If their adaptability is reliant on proper subTAD organization or insulation (259-261), then the transcriptional phenotypes caused by lower levels of CTCF we discovered in breast epithelial cells are likely transposable to other CTCF low epithelial tumors. A similar combination of CRISPR-Cas9 editing, RNA-Seq, ChIP-Seq and Hi-C, as used in Chapter 1, to study the effect of CTCF LOH in diverse epithelial tissue and non-epithelial tissue could validate this hypothesis. For example, according to the results of my project, I would expect such experiments to outline an enrichment of altered genes related to PI3K and EMT pathways specifically in epithelial cells. However, as the epigenetic mechanism driven by the loss of CTCF is likely to be consistent across cell lines, the association with loss of insulation of subTADs, H3K27ac and changes in gene expression should be observed in most cell lines, independent of their origin. If proven correct, such investigation could be followed up with HAT inhibitor treatments, potentially detecting and expanding the sensitivity detected in CTCF+/- MCF10A to distinct independent models.

From a different perspective, other mechanisms may explain the high alteration rate of CTCF in specific subcategories of non-epithelial cancer types. For example, in humans, Down Syndrome related Acute MegaKaryoblastic Leukemia (DS-AMKL) carries CTCF deletions or mutations in 20% of all cases (21). Here, the loss of CTCF is thought to be important for clonal evolution to more aggressive phenotypes following GATA1 mutations (21), a crucial transcription factor regulating differentiation in erythropoiesis (265). DS-AMKL, as with most liquid cancers, are often caused by defects in differentiation (266). Interestingly, topological domains of the chromatin become gradually more insulated and defined to solidify cellular fate and identity throughout differentiation, in a CTCF dependent manner (12). Therefore, dysfunctions of CTCF could hinder differentiation by impeding the formation of properly insulated subTADs necessary

to define cell identity. Although this hypothesis has not been directly validated, it is supported by the synergy between CTCF and GATA1 dysfunction in DS-AMKL (21) and the early onset of hematological tumors in CTCF hemizygous mice (19). Investigating chromatin conformation throughout the timeline of erythropoietic differentiation in diverse models with altered CTCF expression could answer this hypothesis directly.

In sum, despite the major differences in cellular functions affected by altered CTCF, the epigenetic mechanism we highlighted in MCF10A cells is likely to be consistent across models. Indeed, in the diverse types of cancer mentioned above, loss of insulation of subTADs can explain the oncogenicity of CTCF dysfunctions, through transcriptional defect of adaptably expressed genes or by hindering the solidification of topological domains when establishing cellular identity. Additionally, strong oncogenic hits, such as GATA1 mutations (21), benefit from the decreased insulation caused by aberrant CTCF functions, which promotes tumor progression and not initiation, like in our MCF10A models. However, similarly to how altered subTAD insulation will promote cancer progression distinctly in specific cell types, each mutations or change in levels of CTCF will utilize this mechanism to lead to a unique phenotype.

Since my project focused on a deep and thorough investigation of a few inter-connected models, a pitfall of my research is that the loss of CTCF in different models was not directly investigated. Its implication could only be theorized and deduced through the logic of the mechanism we outlined. If direct answers were generated for the ideas and potential experiments mentioned in this section, such wider scope would have validated a universal sensitivity of subTAD organization to changing CTCF levels and its implication in cancer progression, outside of breast epithelial models.

Explaining the difference between CTCF mutation and loss of heterozygosity

Despite being both drivers of transcriptional changes within TADs, CTCF LOH and CTCF ZF1M lead to significantly different pathways being altered compared to CTL cells. Similar to how the different subTAD organization of diverse cell types utilized aberrant CTCF functions in unique ways to activate oncogenic pathways, different dysfunctions of CTCF may also affect subTAD topology in distinct manners.

First, the altered CBS are distinct in the CTCF +/- and CTCF ZF1M MCF10A. These differences stem from the biochemical mechanism explaining the changes in CTCF binding following these distinct modifications. Our MoMotif and ChIP-Seq analysis revealed that in CTCF +/- cells, CTCF is lost at CBS with lower read density and with a slightly more variable core motif. However, the downstream extended motif is not directly involved in the alteration of CTCF binding in CTCF +/- cells, despite being a key factor in CTCF ZF1M cells. Additionally, lost sites in CTCF ZF1M cells display a higher read density, which was also previously associated with the extended motif (40). Therefore, in CTCF +/- cells, CTCF is likely stochastically lost from less conserved CBS. While in CTCF ZF1M cells, changes in CTCF binding are driven by its inability to bind the extended motif, often enriched on more conserved CBS.

Second, if changes in CTCF binding are the initial driver of altered gene expression, then we should expect distinct transcriptional changes for unique CTCF binding changes. This was observed when comparing RNA-Seq data of CTCF +/- and CTCF ZF1M compared to CTL MCF10A. Besides strengthening our models, this observation also hints at the diversity of CBS across the genome. Indeed, an enrichment of lowly conserved CBS at the boundaries of subTADs regulating genes of the PI3K pathway in mammary epithelial cell would explain the results obtained in CTCF +/- MCF10A. Similarly, an enrichment of the extended CTCF motif at

boundaries of subTAD regulating genes involved in xenobiotic metabolism would explain the changes detected in CTCF ZF1M MCF10A. However, a global enrichment of genes involved in EMT and extracellular matrix organization in subTADs, compared to genes of other pathways, could be behind their enrichment in both models, as they would be more likely to be contained within a subTAD with a less conserved or motif extended CBS.

Third, the extent to which CTCF ZF1M mutation directly hinders CTCF binding to DNA was not fully uncovered with the computational model used, as extra bases outside of the predicted region of binding were enriched in the extended motif. This hints that a potential co-factor could also play a role in the alteration of CTCF H284N binding. CTCF ZF1 is known to contribute to RNA binding (39). As such, its mutation could hinder the interaction with RNA-dependent co-factors (267) necessary for recognition of or recruitment at this extended motif. Future "enhanced CrossLink and ImmunoPrecipitation and Sequencing" (eCLIP-Seq) (268,269) investigations of the altered RNA-binding properties of mutant CTCF could provide further insight into this relationship. Additionally, if such potential co-factors are regulated by specific pathways or signals, the alteration of CBS in CTCF ZF1M cells would be dependent on them, further dividing the phenotypes observed in ZF1 mutated CTCF versus lower expression of CTCF. Performing CTCF ChIP-Seq and RNA-Seq in different environments or stress conditions in CTCF ZF1M, CTCF +/- and CTCF WT cells could expose such dependence and narrow the range of potential effectors within known CTCF or CTCF ZF1M interactors.

In opposition to ZF1 mutations, other common mutations of CTCF might result in a similar phenotype to what we observed in CTCF +/- cells, although with some important nuances. Other common mutations of CTCF are located around ZF4-5, which are known to bind to the core binding motif of CTCF (33,34). Therefore, less conserved CBS harboring its consensus binding

motif, which represents the majority of loss sites in CTCF +/-, are expected to be the most sensitive to these mutations. However, the minority of loss sites in CTCF +/- that do not colocalize with CTCF core binding motif would not be expected to be sensitive to mutations of CTCF around ZF4-5, assuming these mutations do not affect the overall protein level of CTCF or its ability to DNA outside of CTCF consensus motif. To validate this hypothesis, it would be interesting to carry out a MoMotif analysis of altered CBS following CTCF ZF4-5M. A relative enrichment of the section of the core CTCF binding motif recognized by these zinc finger or the totality of the consensus motif would confirm a direct impact of these mutations on CTCF-DNA binding through motif recognition. Oppositely, in the unlikely event that no such enrichment is detected, it would hint at a mechanism of CTCF-DNA binding in which other zinc-fingers of CTCF can compensate for these mutations, leading to a relative enrichment of different sections of the CTCF motif in the constant sites. Similarly, if the mutations affect the interaction of CTCF with a co-factor, then MoMotif would detect an enrichment of any surrounding motif recognized by the co-factor or the altered CBS would be found in specific chromatin states influencing the co-factors interactions with CTCF and DNA, independently of DNA motifs. These studies would solidify our understanding of CTCF-DNA binding, while exploring the mostly unexplored biochemical and epigenetic impact of CTCF mutations. Additionally, as I demonstrated at the end of Chapter 2, MoMotif potential applications are wider than studying mutations of CTCF or other transcription factors.

The potential of MoMotif

Similar to our analysis of mutated CTCF or motif variations at promoter proximal regions, MoMotif analysis of available genomics data on DNA binding proteins and their co-factors in varied conditions has the potential to identify diverse modified motifs. These include motifs

specific to a context-dependent binding of transcription factors (TF), between mutated versions of proteins, in the presence of various co-factors and under various environmental conditions. In addition, differences in TF recognition motifs when adjacent to TAD boundaries, transcription start sites, enhancer elements or across tissue types may be explored. Further, MoMotif can investigate whether mutated TFs, similarly to CTCF, may harbor unknown context specific binding motifs. Other factors impacting both wild-type or mutated TF binding motifs may include proximity to regulatory elements, proximity of co-factor binding sites, chromatin states at binding sites or post-translational modifications. Mining available genomic databases using the MoMotif pipeline will allow the identification and association of subtle motif disparities between various contexts, greatly extending our compendium of knowledge regarding biological influencers of DNA binding. In turn, this knowledge may be helpful in identifying therapeutic vulnerabilities from diverse clinical datasets, including non-coding mutations identified by whole genome sequencing or altered chromatin states detected by ATAC-Seq.

Association between H3K27ac and topological changes

The histone modification H3K27ac is commonly used to demarcate contact between transcriptionally active regions of the genome (270,271). However, it remains unclear how chromatin contacts and histone modifications influence each other. Our data provides insight into the relationship between altered subTAD distribution and epigenetic changes. Since sites of compromised CTCF binding are generally not proximal to sites of gained H3K27ac or H3K4me3, it is unlikely that the loss of CTCF drives the gain of activating marks through direct recruitment of histone writers to specific loci. For example, it is unlikely that compromised CTCF binding would lead to the loss of recruitment of antagonistic epigenetic writers, such as EZH2 (272) at sites many kilobases away where H3K27ac is subsequently accumulated.

Therefore, we can infer that the reshuffling of subTADs is driving the redistribution of H3K27ac and H3K4me3 more so than altered CTCF itself. Our model predicts that de novo chromatin contacts between genomic regions may promote the recruitment of activating chromatin writers (273,274), leading to reprogrammed epigenetic landscape and transcriptional changes.

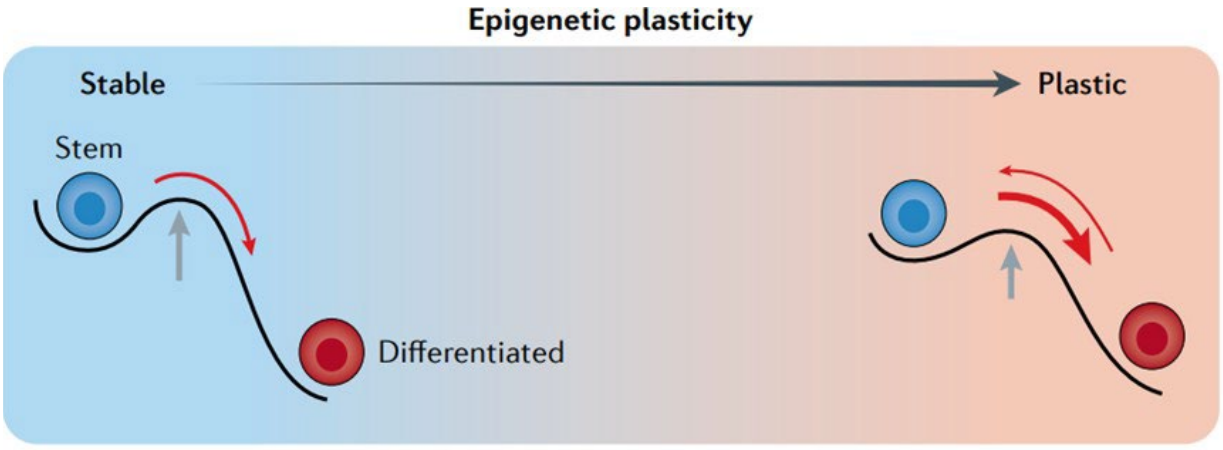
Modeling Epigenetic Plasticity and Evolution in the context of oncogenesis

During my investigations of CTCF LOH, cell models were cultured in constant culture conditions and kept at low passages. Despite being optimal for reproducible and reliable *in cellulo* data, this modus operandi cannot model the ever-present environment and evolution of *in vivo* oncogenesis. Indeed, animal models are needed to accurately study the impact of CTCF LOH in the context of cancer evolution.

It is well characterized that as most tumors progress, they will accumulate genetic mutations. Multiple rounds of mutations and selections will then lead to highly heterogeneous and aggressive tumors. The partial loss of CTCF could enhance this evolutive capacity of tumors, in ways that are predicted, but not directly tested, in our previous investigations. My predecessors and colleagues, in Hilmi et al. 2017, showed that CTCF was important for homologous recombination double-stranded beak DNA damage repair by mediating the recruitment of BRCA2 (64). Interestingly, other members of this DNA damage repair pathway, such as BRCA1 and BRCA2, are common tumor suppressor genes associated with breast cancer (275). Indeed, defective DNA damage repair can be beneficial for cancer progression. Often associated to a "mutator phenotype" or increased genomic instability (275), it increases the rate at which the tumors accumulate genetic mutations, and therefore driver mutations, speeding up cancer evolution and progression. This could be one way by which CTCF LOH promotes cancer evolution. However, in light of my research, the altered epigenetic regulation resulting from the partial loss of CTCF might,

synergistically, promote it even further.

Although genetic mutations have been at the forefront oncology for many decades, the epigenetic aspect of cancer progression is rising up and was recently added to the hallmarks of cancer (276). Cancer goes through multiple stages, from initiation and growth to invasion and metastasis. As such, a mutation beneficial in the initiation phase might be detrimental to the tumor's metastatic abilities. For example, multiple investigations termed the TGFB pathways, or other genes and pathways central to EMT, as tumor suppressor since it hinders growth (277), while multiple others termed it an oncogenic pathway as it promotes invasion and metastasis (278,279) . Contrarily to genetic mutations, which are technically irreversible, epigenetic regulation is dynamic. Therefore, the expression of key genes can be promoted or inhibited epigenetically at phases of cancer progression benefiting from such transcriptional changes. For example, primary breast tumors accumulate multiple genetic mutations, promoting their growth and heterogeneity, but very few mutations are accumulated during the metastatic process (280). As reviewed in Nam et al. 2012 (Figure D5) (281), dysfunctions of epigenetic regulatory processes are competent to facilitate transitions between two different cell states. Therefore, transcriptional plasticity renders more efficient the switch between two phenotypically very different stages of cancer.



From: Nam AS, Chaligne R, Landau DA. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. Nat Rev Genet. 2021 Jan

Figure D5: Depiction of how epigenetic plasticity facilitates the switch between 2 cell states.

According to my results, CTCF LOH has strong potential to promote epigenetic plasticity. By hindering global subTAD insulation, it facilitates potential oncogenic DNA-DNA contacts within TADs, such as at between SNAI1 promoter and downstream enhancer in our MCF10A model, to happen dynamically in tumor cell and promote its progression. Such aberrant epigenetic events could happen transiently in tumor cells requiring invasivity to progress or be carried over by descending sub-population. Specific examples of loss of CTCF mediated novel oncogenic contact were already described (282) and support the hypothesis arising from my research.

Additionally, it is possible that the genetic and epigenetic plasticity promoted by CTCF might synergize. Indeed, accumulation of genetic mutation seems to be an important driver of tumor initiation and primary growth, while Chapter 1 showed that epigenetic plasticity could be a crucial factor of the following invasion and metastasis. Therefore, each stage of cancer progression might benefit from these evolutionary advantages. However, since my project did not include a model compatible with the study of the temporal and evolutive aspect of tumor progression, this hypothesis was not tested directly. Therefore, to test such hypothesis, we developed a mouse model of CTCF loss in breast cancer, by crossbreeding cell type specific deletion of CTCF with Her2+ background of

NIC mice, a slow initiating model of breast cancer in an immune competent background (283). This model will allow us to detect changes in penetrance, tumor initiation and tumor progression by measuring classical metrics of tumor growth and progression. Further, single- cell RNA-Seq and ATAC-Seq analysis of primary tumors, circulating tumor cell and metastatic tumor will allow for the detection of tumor transcriptional heterogeneity compared to CTCF WT NIC mice, as well as identifying which transcriptional subpopulations are more permissive to the transition between each stages of cancer. Overall, this model and experiments will not only directly reveal the possible evolutionary advantages conferred by CTCF LOH, but also identify essential drivers of breast tumor progression, independent of or synergistic with CTCF dysfunctions.

Final Conclusion and Summary

In conclusion, my investigations revealed the epigenetic impact of CTCF LOH in mammary epithelial, through loss of insulation of subTADs. This mechanisms explains why genes of the PI3K pathway and EMT are prone to be altered by lower levels of CTCF and are therefore potential therapeutic targets. Due to the intrinsic properties of subTAD interactions, we predicted sensitivity of our model to inhibitors of mTor or histone acetylation. Further, I employed, and played a central role in the development of, a new computational tool we named MoMotif. Using MoMotif in the context of CTCF H284N mutation, I was able to identify an extension of the core CTCF motif requiring its ZF1 to bind appropriately. I then associated the H284N mutation dependent loss of CTCF binding at sites harboring this extended motif within TAD or at the boundaries of subTAD to changes in gene expression reminiscent of the clinical phenotypes observed in CTCF mutated breast cancer. Overall, I showed the importance of epigenetic and chromatin conformation regulation in cancer progression and strengthened the role of CTCF as a crucial tumor suppressor gene.

References

1. Chen, C., Yu, W., Tober, J., Gao, P., He, B., Lee, K., Trieu, T., Blobel, G.A., Speck, N.A. and Tan, K. (2019) Spatial Genome Re-organization between Fetal and Adult Hematopoietic Stem Cells. *Cell Rep*, **29**, 4200-4211 e4207.
2. Sivakumar, A., de Las Heras, J.I. and Schirmer, E.C. (2019) Spatial Genome Organization: From Development to Disease. *Front Cell Dev Biol*, **7**, 18.
3. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289-293.
4. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376-380.
5. Davidson, I.F., Bauer, B., Goetz, D., Tang, W., Wutz, G. and Peters, J.M. (2019) DNA loop extrusion by human cohesin. *Science*, **366**, 1338-1345.
6. Kim, Y., Shi, Z., Zhang, H., Finkelstein, I.J. and Yu, H. (2019) Human cohesin compacts DNA by loop extrusion. *Science*, **366**, 1345-1349.
7. Rao, S.S.P., Huang, S.C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D. *et al.* (2017) Cohesin Loss Eliminates All Loop Domains. *Cell*, **171**, 305-320 e324.
8. Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y. *et al.* (2015) CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell*, **162**, 900-910.
9. Hsieh, T.S., Cattoglio, C., Slobodyanyuk, E., Hansen, A.S., Rando, O.J., Tjian, R. and Darzacq, X. (2020) Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. *Mol Cell*, **78**, 539-553 e538.
10. Phillips-Cremins, J.E., Sauria, M.E., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S., Ong, C.T., Hookway, T.A., Guo, C., Sun, Y. *et al.* (2013) Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, **153**, 1281-1295.
11. Tena, J.J. and Santos-Pereira, J.M. (2021) Topologically Associating Domains and Regulatory Landscapes in Development, Evolution and Disease. *Front Cell Dev Biol*, **9**, 702787.
12. Chen, X., Ke, Y., Wu, K., Zhao, H., Sun, Y., Gao, L., Liu, Z., Zhang, J., Tao, W., Hou, Z. *et al.* (2019) Key role for CTCF in establishing chromatin structure in human embryos. *Nature*, **576**, 306-310.
13. Justice, M., Carico, Z.M., Stefan, H.C. and Dowen, J.M. (2020) A WIZ/Cohesin/CTCF Complex Anchors DNA Loops to Define Gene Expression and Cell Identity. *Cell Rep*, **31**, 107503.
14. Luo, H., Zhu, G., Xu, J., Lai, Q., Yan, B., Guo, Y., Fung, T.K., Zeisig, B.B., Cui, Y., Zha, J. *et al.* (2019) HOTTIP lncRNA Promotes Hematopoietic Stem Cell Self-Renewal Leading to AML-like Disease in Mice. *Cancer Cell*, **36**, 645-659 e648.
15. Willemin, A., Lopez-Delisle, L., Bolt, C.C., Gadolini, M.L., Duboule, D. and Rodriguez-Carballo, E. (2021) Induction of a chromatin boundary in vivo upon insertion of a TAD border. *PLoS Genet*, **17**, e1009691.
16. Bailey, C.G., Metierre, C., Feng, Y., Baidya, K., Filippova, G.N., Loukinov, D.I.,

Lobanenkov, V.V., Semaan, C. and Rasko, J.E. (2018) CTCF Expression is Essential for Somatic Cell Viability and Protection Against Cancer. *Int J Mol Sci*, **19**.
17. Moore, J.M., Rabaia, N.A., Smith, L.E., Fagerlie, S., Gurley, K., Loukinov, D., Disteche, C.M., Collins, S.J., Kemp, C.J., Lobanenkov, V.V. *et al.* (2012) Loss of maternal CTCF is associated with peri-implantation lethality of Ctcf null embryos. *PLoS One*, **7**, e34915.
18. Alharbi, A.B., Schmitz, U., Marshall, A.D., Vanichkina, D., Nagarajah, R., Vellozzi, M., Wong, J.J., Bailey, C.G. and Rasko, J.E. (2021) Ctcf haploinsufficiency mediates intron retention in a tissue-specific manner. *RNA Biol*, **18**, 93-103.
19. Kemp, C.J., Moore, J.M., Moser, R., Bernard, B., Teater, M., Smith, L.E., Rabaia, N.A., Gurley, K.E., Guinney, J., Busch, S.E. *et al.* (2014) CTCF haploinsufficiency destabilizes DNA methylation and predisposes to cancer. *Cell Rep*, **7**, 1020-1029.
20. Aitken, S.J., Ibarra-Soria, X., Kentepozidou, E., Flicek, P., Feig, C., Marioni, J.C. and Odom, D.T. (2018) CTCF maintains regulatory homeostasis of cancer pathways. *Genome Biol*, **19**, 106.
21. Yoshida, K., Toki, T., Okuno, Y., Kanezaki, R., Shiraishi, Y., Sato-Otsubo, A., Sanada, M., Park, M.J., Terui, K., Suzuki, H. *et al.* (2013) The landscape of somatic mutations in Down syndrome-related myeloid disorders. *Nat Genet*, **45**, 1293-1299.
22. Damaschke, N.A., Gawdzik, J., Avilla, M., Yang, B., Svaren, J., Roopra, A., Luo, J.H., Yu, Y.P., Keles, S. and Jarrard, D.F. (2020) CTCF loss mediates unique DNA hypermethylation landscapes in human cancers. *Clin Epigenetics*, **12**, 80.
23. Cattoglio, C., Pustova, I., Walther, N., Ho, J.J., Hantsche-Grininger, M., Inouye, C.J., Hossain, M.J., Dailey, G.M., Ellenberg, J., Darzacq, X. *et al.* (2019) Determining cellular CTCF and cohesin abundances to constrain 3D genome models. *Elife*, **8**.
24. Tian, T., Li, X. and Zhang, J. (2019) mTOR Signaling in Cancer and mTOR Inhibitors in Solid Tumor Targeting Therapy. *Int J Mol Sci*, **20**.
25. Bushweller, J.H. (2019) Targeting transcription factors in cancer - from undruggable to reality. *Nat Rev Cancer*, **19**, 611-624.
26. Akdemir, K.C., Le, V.T., Chandran, S., Li, Y., Verhaak, R.G., Beroukhim, R., Campbell, P.J., Chin, L., Dixon, J.R., Futreal, P.A. *et al.* (2020) Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat Genet*, **52**, 294-305.
27. Rheinbay, E., Nielsen, M.M., Abascal, F., Wala, J.A., Shapira, O., Tiao, G., Hornshoj, H., Hess, J.M., Juul, R.I., Lin, Z. *et al.* (2020) Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, **578**, 102-111.
28. Lee, T.I. and Young, R.A. (2013) Transcriptional regulation and its misregulation in disease. *Cell*, **152**, 1237-1251.
29. Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, **4**, 651-657.
30. Yan, H., Tian, S., Slager, S.L. and Sun, Z. (2016) ChIP-seq in studying epigenetic mechanisms of disease and promoting precision medicine: progresses and future directions. *Epigenomics*, **8**, 1239-1258.
31. Razavi, P., Chang, M.T., Xu, G., Bandlamudi, C., Ross, D.S., Vasan, N., Cai, Y., Bielski, C.M., Donoghue, M.T.A., Jonsson, P. *et al.* (2018) The Genomic Landscape of Endocrine-Resistant Advanced Breast Cancers. *Cancer Cell*, **34**, 427-438 e426.
32. Marshall, A.D., Bailey, C.G., Champ, K., Vellozzi, M., O'Young, P., Metierre, C., Feng,

Y., Thoeng, A., Richards, A.M., Schmitz, U. *et al.* (2017) CTCF genetic alterations in endometrial carcinoma are pro-tumorigenic. *Oncogene*, **36**, 4100-4110.
33. Nakahashi, H., Kieffer Kwon, K.R., Resch, W., Vian, L., Dose, M., Stavreva, D., Hakim, O., Pruett, N., Nelson, S., Yamane, A. *et al.* (2013) A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep*, **3**, 1678-1689.
34. Yin, M., Wang, J., Wang, M., Li, X., Zhang, M., Wu, Q. and Wang, Y. (2017) Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites. *Cell research*, **27**, 1365-1377.
35. Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V. and Ren, B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, **128**, 1231-1245.
36. Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.Y., Cui, K. and Zhao, K. (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome research*, **19**, 24-32.
37. Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Goncalves, A., Kutter, C., Brown, G.D., Marshall, A., Flicek, P. and Odom, D.T. (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, **148**, 335-348.
38. Hashimoto, H., Wang, D., Horton, J.R., Zhang, X., Corces, V.G. and Cheng, X. (2017) Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. *Mol Cell*, **66**, 711-720 e713.
39. Saldana-Meyer, R., Rodriguez-Hernaez, J., Escobar, T., Nishana, M., Jacome-Lopez, K., Nora, E.P., Bruneau, B.G., Tsirigos, A., Furlan-Magaril, M., Skok, J. *et al.* (2019) RNA Interactions Are Essential for CTCF-Mediated Genome Organization. *Mol Cell*, **76**, 412-422 e415.
40. Rhee, H.S. and Pugh, B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408-1419.
41. Harbeck, N. and Gnant, M. (2017) Breast cancer. *Lancet*, **389**, 1134-1150.
42. Osborne, C.K. (1998) Tamoxifen in the treatment of breast cancer. *N Engl J Med*, **339**, 1609-1618.
43. Hudis, C.A. (2007) Trastuzumab--mechanism of action and use in clinical practice. *N Engl J Med*, **357**, 39-51.
44. Livraghi, L. and Garber, J.E. (2015) PARP inhibitors in the management of breast cancer: current data and future prospects. *BMC Med*, **13**, 188.
45. Hajdu, S.I. (2011) A note from history: landmarks in history of cancer, part 1. *Cancer*, **117**, 1097-1102.
46. Loberg, M., Lousdal, M.L., Bretthauer, M. and Kalager, M. (2015) Benefits and harms of mammography screening. *Breast Cancer Res*, **17**, 63.
47. Kamisawa, T., Wood, L.D., Itoi, T. and Takaori, K. (2016) Pancreatic cancer. *Lancet*, **388**, 73-85.
48. Doubeni, C.A., Doubeni, A.R. and Myers, A.E. (2016) Diagnosis and Management of Ovarian Cancer. *Am Fam Physician*, **93**, 937-944.
49. Zebolsky, A.L., Ochoa, E., Badran, K.W., Heaton, C., Park, A., Seth, R. and Knott, P.D. (2021) Appearance-Related Distress and Social Functioning after Head and Neck Microvascular Reconstruction. *Laryngoscope*, **131**, E2204-E2211.
50. Dasari, S. and Tchounwou, P.B. (2014) Cisplatin in cancer therapy: molecular mechanisms

of action. *Eur J Pharmacol*, **740**, 364-378.
51. Marupudi, N.I., Han, J.E., Li, K.W., Renard, V.M., Tyler, B.M. and Brem, H. (2007) Paclitaxel: a review of adverse toxicities and novel delivery strategies. *Expert Opin Drug Saf*, **6**, 609-621.
52. Desai, P. and Roboz, G.J. (2019) Clonal Hematopoiesis and therapy related MDS/AML. *Best Pract Res Clin Haematol*, **32**, 13-23.
53. Stewart, R.L., Updike, K.L., Factor, R.E., Henry, N.L., Boucher, K.M., Bernard, P.S. and Varley, K.E. (2019) A Multigene Assay Determines Risk of Recurrence in Patients with Triple-Negative Breast Cancer. *Cancer Res*, **79**, 3466-3478.
54. Thorat, M.A. and Balasubramanian, R. (2020) Breast cancer prevention in high-risk women. *Best Pract Res Clin Obstet Gynaecol*, **65**, 18-31.
55. Kumar, V. and Chambon, P. (1988) The estrogen receptor binds tightly to its responsive element as a ligand-induced homodimer. *Cell*, **55**, 145-156.
56. Rae, J.M., Johnson, M.D., Scheys, J.O., Cordero, K.E., Larios, J.M. and Lippman, M.E. (2005) GREB 1 is a critical regulator of hormone dependent breast cancer growth. *Breast Cancer Res Treat*, **92**, 141-149.
57. Yager, J.D. and Davidson, N.E. (2006) Estrogen carcinogenesis in breast cancer. *N Engl J Med*, **354**, 270-282.
58. Patel, H.K. and Bihani, T. (2018) Selective estrogen receptor modulators (SERMs) and selective estrogen receptor degraders (SERDs) in cancer treatment. *Pharmacol Ther*, **186**, 1-24.
59. Oh, D.Y. and Bang, Y.J. (2020) HER2-targeted therapies - a role beyond breast cancer. *Nat Rev Clin Oncol*, **17**, 33-48.
60. Hoxhaj, G. and Manning, B.D. (2020) The PI3K-AKT network at the interface of oncogenic signalling and cancer metabolism. *Nat Rev Cancer*, **20**, 74-88.
61. O'Neil, N.J., Bailey, M.L. and Hieter, P. (2017) Synthetic lethality and cancer. *Nat Rev Genet*, **18**, 613-623.
62. Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M., Ding, W. *et al.* (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, **266**, 66-71.
63. Moynahan, M.E., Chiu, J.W., Koller, B.H. and Jasin, M. (1999) Brca1 controls homology-directed DNA repair. *Mol Cell*, **4**, 511-518.
64. Hilmi, K., Jangal, M., Marques, M., Zhao, T., Saad, A., Zhang, C., Luo, V.M., Syme, A., Rejon, C., Yu, Z. *et al.* (2017) CTCF facilitates DNA double-strand break repair by enhancing homologous recombination repair. *Sci Adv*, **3**, e1601898.
65. Prakash, R., Zhang, Y., Feng, W. and Jasin, M. (2015) Homologous recombination and human health: the roles of BRCA1, BRCA2, and associated proteins. *Cold Spring Harb Perspect Biol*, **7**, a016600.
66. Dantzer, F., Schreiber, V., Niedergang, C., Trucco, C., Flatter, E., De La Rubia, G., Oliver, J., Rolli, V., Menissier-de Murcia, J. and de Murcia, G. (1999) Involvement of poly(ADP-ribose) polymerase in base excision repair. *Biochimie*, **81**, 69-75.
67. Plummer, R. (2011) Poly(ADP-ribose) polymerase inhibition: a new direction for BRCA and triple-negative breast cancer? *Breast Cancer Res*, **13**, 218.
68. Robson, M., Im, S.A., Senkus, E., Xu, B., Domchek, S.M., Masuda, N., Delaloge, S., Li, W., Tung, N., Armstrong, A. *et al.* (2017) Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation. *N Engl J Med*, **377**, 523-533.

69. Yin, L., Duan, J.J., Bian, X.W. and Yu, S.C. (2020) Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Res*, **22**, 61.
70. Portin, P. (2014) The birth and development of the DNA theory of inheritance: sixty years since the discovery of the structure of DNA. *J Genet*, **93**, 293-302.
71. Avery, O.T., MacLeod, C.M. and McCarty, M. (1995) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. 1944. *Mol Med*, **1**, 344-365.
72. Watson, J.D. and Crick, F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**, 737-738.
73. Chargaff, E. (1950) Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, **6**, 201-209.
74. Crick, F. (1970) Central dogma of molecular biology. *Nature*, **227**, 561-563.
75. Grosveld, G.C., de Boer, E., Shewmaker, C.K. and Flavell, R.A. (1982) DNA sequences necessary for transcription of the rabbit beta-globin gene in vivo. *Nature*, **295**, 120-126.
76. Nikolov, D.B. and Burley, S.K. (1997) RNA polymerase II transcription initiation: a structural view. *Proc Natl Acad Sci U S A*, **94**, 15-22.
77. Saxonov, S., Berg, P. and Brutlag, D.L. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A*, **103**, 1412-1417.
78. Deaton, A.M. and Bird, A. (2011) CpG islands and the regulation of transcription. *Genes Dev*, **25**, 1010-1022.
79. Liao, J., Karnik, R., Gu, H., Ziller, M.J., Clement, K., Tsankov, A.M., Akopian, V., Gifford, C.A., Donaghey, J., Galonska, C. *et al.* (2015) Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. *Nat Genet*, **47**, 469-478.
80. Biniszkiewicz, D., Gribnau, J., Ramsahoye, B., Gaudet, F., Eggan, K., Humpherys, D., Mastrangelo, M.A., Jun, Z., Walter, J. and Jaenisch, R. (2002) Dnmt1 overexpression causes genomic hypermethylation, loss of imprinting, and embryonic lethality. *Mol Cell Biol*, **22**, 2124-2135.
81. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.
82. Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M. and Gerstein, M.B. (2010) Annotating non-coding regions of the genome. *Nat Rev Genet*, **11**, 559-571.
83. Ohno, S. (1972) So much "junk" DNA in our genome. *Brookhaven Symp Biol*, **23**, 366-370.
84. Elliott, K. and Larsson, E. (2021) Non-coding driver mutations in human cancer. *Nat Rev Cancer*, **21**, 500-509.
85. Prescott, S.L., Srinivasan, R., Marchetto, M.C., Grishina, I., Narvaiza, I., Selleri, L., Gage, F.H., Swigut, T. and Wysocka, J. (2015) Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell*, **163**, 68-83.
86. Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K. *et al.* (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol*, **3**, e7.
87. Stefani, G. and Slack, F.J. (2008) Small non-coding RNAs in animal development. *Nat Rev Mol Cell Biol*, **9**, 219-230.

88. Fatica, A. and Bozzoni, I. (2014) Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet*, **15**, 7-21.
89. Esteller, M. (2011) Non-coding RNAs in human disease. *Nat Rev Genet*, **12**, 861-874.
90. Anastasiadou, E., Jacob, L.S. and Slack, F.J. (2018) Non-coding RNA networks in cancer. *Nat Rev Cancer*, **18**, 5-18.
91. Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F. and Richmond, T.J. (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, **389**, 251-260.
92. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. and Greenleaf, W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*, **10**, 1213-1218.
93. Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S. and Crawford, G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311-322.
94. Klemm, S.L., Shipony, Z. and Greenleaf, W.J. (2019) Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet*, **20**, 207-220.
95. Ernst, J. and Kellis, M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*, **28**, 817-825.
96. Weksberg, R., Shuman, C. and Smith, A.C. (2005) Beckwith-Wiedemann syndrome. *Am J Med Genet C Semin Med Genet*, **137C**, 12-23.
97. Sasaki, H., Ishihara, K. and Kato, R. (2000) Mechanisms of Igf2/H19 imprinting: DNA methylation, chromatin and long-distance gene regulation. *J Biochem*, **127**, 711-715.
98. Bell, A.C. and Felsenfeld, G. (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, **405**, 482-485.
99. Lobanenkov, V.V., Nicolas, R.H., Adler, V.V., Paterson, H., Klenova, E.M., Polotskaja, A.V. and Goodwin, G.H. (1990) A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene*, **5**, 1743-1753.
100. Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306-1311.
101. Lichter, P., Cremer, T., Borden, J., Manuelidis, L. and Ward, D.C. (1988) Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. *Hum Genet*, **80**, 224-234.
102. Goel, V.Y. and Hansen, A.S. (2021) The macro and micro of chromosome conformation capture. *Wiley Interdiscip Rev Dev Biol*, **10**, e395.
103. Erdel, F. and Rippe, K. (2018) Formation of Chromatin Subcompartments by Phase Separation. *Biophys J*, **114**, 2262-2270.
104. Sanborn, A.L., Rao, S.S., Huang, S.C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J. *et al.* (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A*, **112**, E6456-6465.
105. Ladurner, R., Bhaskara, V., Huis in 't Veld, P.J., Davidson, I.F., Kreidl, E., Petzold, G. and Peters, J.M. (2014) Cohesin's ATPase activity couples cohesin loading onto DNA with Smc3 acetylation. *Curr Biol*, **24**, 2228-2237.
106. Banigan, E.J. and Mirny, L.A. (2020) Loop extrusion: theory meets single-molecule experiments. *Curr Opin Cell Biol*, **64**, 124-138.
107. Benedetti, F., Dorier, J., Burnier, Y. and Stasiak, A. (2014) Models that include

supercoiling of topological domains reproduce several known features of interphase chromosomes. *Nucleic Acids Res*, **42**, 2848-2855.
108. Racko, D., Benedetti, F., Dorier, J. and Stasiak, A. (2018) Transcription-induced supercoiling as the driving force of chromatin loop extrusion during formation of TADs in interphase chromosomes. *Nucleic Acids Res*, **46**, 1648-1660.
109. Xi, W. and Beer, M.A. (2021) Loop competition and extrusion model predicts CTCF interaction specificity. *Nat Commun*, **12**, 1046.
110. Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665-1680.
111. Iwasaki, Y., Ikemura, T., Kurokawa, K. and Okada, N. (2020) Implication of a new function of human tDNAs in chromatin organization. *Sci Rep*, **10**, 17440.
112. Cubenas-Potts, C., Rowley, M.J., Lyu, X., Li, G., Lei, E.P. and Corces, V.G. (2017) Different enhancer classes in Drosophila bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. *Nucleic Acids Res*, **45**, 1714-1730.
113. Oh, S., Shao, J., Mitra, J., Xiong, F., D'Antonio, M., Wang, R., Garcia-Bassets, I., Ma, Q., Zhu, X., Lee, J.H. *et al.* (2021) Enhancer release and retargeting activates disease-susceptibility genes. *Nature*, **595**, 735-740.
114. Ferrari, R., de Llobet Cucalon, L.I., Di Vona, C., Le Dilly, F., Vidal, E., Lioutas, A., Oliete, J.Q., Jochem, L., Cutts, E., Dieci, G. *et al.* (2020) TFIIIC Binding to Alu Elements Controls Gene Expression via Chromatin Looping and Histone Acetylation. *Mol Cell*, **77**, 475-487 e411.
115. Hua, P., Badat, M., Hanssen, L.L.P., Hentges, L.D., Crump, N., Downes, D.J., Jeziorska, D.M., Oudelaar, A.M., Schwessinger, R., Taylor, S. *et al.* (2021) Defining genome architecture at base-pair resolution. *Nature*, **595**, 125-129.
116. Xu, H., Balakrishnan, K., Malaterre, J., Beasley, M., Yan, Y., Essers, J., Appeldoorn, E., Tomaszewski, J.M., Vazquez, M., Verschoor, S. *et al.* (2010) Rad21-cohesin haploinsufficiency impedes DNA repair and enhances gastrointestinal radiosensitivity in mice. *PLoS One*, **5**, e12112.
117. De Koninck, M., Lapi, E., Badia-Careaga, C., Cossio, I., Gimenez-Llorente, D., Rodriguez-Corsino, M., Andrada, E., Hidalgo, A., Manzanares, M., Real, F.X. *et al.* (2020) Essential Roles of Cohesin STAG2 in Mouse Embryonic Development and Adult Tissue Homeostasis. *Cell Rep*, **32**, 108014.
118. Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I. and Young, R.A. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, 307-319.
119. Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A.V., Manteiga, J.C. *et al.* (2018) Coactivator condensation at super-enhancers links phase separation and gene control. *Science*, **361**.
120. Loven, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I. and Young, R.A. (2013) Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, **153**, 320-334.
121. Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854-858.

122. Consortium, E.P., Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A. *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699-710.
123. Nott, A., Holtman, I.R., Coufal, N.G., Schlachetzki, J.C.M., Yu, M., Hu, R., Han, C.Z., Pena, M., Xiao, J., Wu, Y. *et al.* (2019) Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science*, **366**, 1134-1139.
124. Zhang, Y., Zhang, X., Ba, Z., Liang, Z., Dring, E.W., Hu, H., Lou, J., Kyritsis, N., Zurita, J., Shamim, M.S. *et al.* (2019) The fundamental role of chromatin loop extrusion in physiological V(D)J recombination. *Nature*, **573**, 600-604.
125. Dai, H.Q., Hu, H., Lou, J., Ye, A.Y., Ba, Z., Zhang, X., Zhang, Y., Zhao, L., Yoon, H.S., Chapdelaine-Williams, A.M. *et al.* (2021) Loop extrusion mediates physiological Igh locus contraction for RAG scanning. *Nature*, **590**, 338-343.
126. Lupianez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R. *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, **161**, 1012-1025.
127. Gupta, K., Sari-Ak, D., Haffke, M., Trowitzsch, S. and Berger, I. (2016) Zooming in on Transcription Preinitiation. *J Mol Biol*, **428**, 2581-2591.
128. Cianfrocco, M.A., Kassavetis, G.A., Grob, P., Fang, J., Juven-Gershon, T., Kadonaga, J.T. and Nogales, E. (2013) Human TFIID binds to core promoter DNA in a reorganized structural state. *Cell*, **152**, 120-131.
129. Lauberth, S.M., Nakayama, T., Wu, X., Ferris, A.L., Tang, Z., Hughes, S.H. and Roeder, R.G. (2013) H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell*, **152**, 1021-1036.
130. Johnson, K.M., Wang, J., Smallwood, A., Arayata, C. and Carey, M. (2002) TFIID and human mediator coactivator complexes assemble cooperatively on promoter DNA. *Genes Dev*, **16**, 1852-1863.
131. Robinson, P.J., Trnka, M.J., Bushnell, D.A., Davis, R.E., Mattei, P.J., Burlingame, A.L. and Kornberg, R.D. (2016) Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex. *Cell*, **166**, 1411-1422 e1416.
132. Bhuiyan, T. and Timmers, H.T.M. (2019) Promoter Recognition: Putting TFIID on the Spot. *Trends Cell Biol*, **29**, 752-763.
133. Marques, M., Hernandez, R.P. and Witcher, M. (2015) Analysis of changes to mRNA levels and CTCF occupancy upon TFII-I knockdown. *Genom Data*, **4**, 17-21.
134. Makela, T.P., Parvin, J.D., Kim, J., Huber, L.J., Sharp, P.A. and Weinberg, R.A. (1995) A kinase-deficient transcription factor TFIIH is functional in basal and activated transcription. *Proc Natl Acad Sci U S A*, **92**, 5174-5178.
135. Akhtar, M.S., Heidemann, M., Tietjen, J.R., Zhang, D.W., Chapman, R.D., Eick, D. and Ansari, A.Z. (2009) TFIIH kinase places bivalent marks on the carboxy-terminal domain of RNA polymerase II. *Mol Cell*, **34**, 387-393.
136. Wong, K.H., Jin, Y. and Struhl, K. (2014) TFIIH phosphorylation of the Pol II CTD stimulates mediator dissociation from the preinitiation complex and promoter escape. *Mol Cell*, **54**, 601-612.
137. Moteki, S. and Price, D. (2002) Functional coupling of capping and transcription of mRNA. *Mol Cell*, **10**, 599-609.
138. Ramanathan, A., Robb, G.B. and Chan, S.H. (2016) mRNA capping: biological functions and applications. *Nucleic Acids Res*, **44**, 7511-7526.

139. Core, L. and Adelman, K. (2019) Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. *Genes Dev*, **33**, 960-982.
140. Wang, Z., Song, A., Xu, H., Hu, S., Tao, B., Peng, L., Wang, J., Li, J., Yu, J., Wang, L. *et al.* (2022) Coordinated regulation of RNA polymerase II pausing and elongation progression by PAF1. *Sci Adv*, **8**, eabm5504.
141. Gilchrist, D.A., Dos Santos, G., Fargo, D.C., Xie, B., Gao, Y., Li, L. and Adelman, K. (2010) Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell*, **143**, 540-551.
142. Jang, M.K., Mochizuki, K., Zhou, M., Jeong, H.S., Brady, J.N. and Ozato, K. (2005) The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. *Mol Cell*, **19**, 523-534.
143. Wada, T., Takagi, T., Yamaguchi, Y., Watanabe, D. and Handa, H. (1998) Evidence that P-TEFb alleviates the negative effect of DSIF on RNA polymerase II-dependent transcription in vitro. *EMBO J*, **17**, 7395-7403.
144. Fujinaga, K., Irwin, D., Huang, Y., Taube, R., Kurosu, T. and Peterlin, B.M. (2004) Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. *Mol Cell Biol*, **24**, 787-795.
145. Marshall, N.F., Peng, J., Xie, Z. and Price, D.H. (1996) Control of RNA polymerase II elongation potential by a novel carboxyl-terminal domain kinase. *J Biol Chem*, **271**, 27176-27183.
146. Gromak, N., West, S. and Proudfoot, N.J. (2006) Pause sites promote transcriptional termination of mammalian RNA polymerase II. *Mol Cell Biol*, **26**, 3986-3996.
147. Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R. and Oberdoerffer, S. (2011) CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, **479**, 74-79.
148. Ruiz-Velasco, M., Kumar, M., Lai, M.C., Bhat, P., Solis-Pinson, A.B., Reyes, A., Kleinsorg, S., Noh, K.M., Gibson, T.J. and Zaugg, J.B. (2017) CTCF-Mediated Chromatin Loops between Promoter and Gene Body Regulate Alternative Splicing across Individuals. *Cell Syst*, **5**, 628-637 e626.
149. Alharbi, A.B., Schmitz, U., Bailey, C.G. and Rasko, J.E.J. (2021) CTCF as a regulator of alternative splicing: new tricks for an old player. *Nucleic Acids Res*, **49**, 7825-7838.
150. Kang, W., Ha, K.S., Uhm, H., Park, K., Lee, J.Y., Hohng, S. and Kang, C. (2020) Transcription reinitiation by recycling RNA polymerase that diffuses on DNA after releasing terminated RNA. *Nat Commun*, **11**, 450.
151. Hansen, A.S., Amitai, A., Cattoglio, C., Tjian, R. and Darzacq, X. (2020) Guided nuclear exploration increases CTCF target search efficiency. *Nat Chem Biol*, **16**, 257-266.
152. Wang, F., Tang, Z., Shao, H., Guo, J., Tan, T., Dong, Y. and Lin, L. (2018) Long noncoding RNA HOTTIP cooperates with CCCTC-binding factor to coordinate HOXA gene expression. *Biochem Biophys Res Commun*, **500**, 852-859.
153. Zhao, X., Li, D., Pu, J., Mei, H., Yang, D., Xiang, X., Qu, H., Huang, K., Zheng, L. and Tong, Q. (2016) CTCF cooperates with noncoding RNA MYCNOS to promote neuroblastoma progression through facilitating MYCN expression. *Oncogene*, **35**, 3565-3576.
154. Sun, S., Del Rosario, B.C., Szanto, A., Ogawa, Y., Jeon, Y. and Lee, J.T. (2013) Jpx RNA activates Xist by evicting CTCF. *Cell*, **153**, 1537-1551.

155. Fu, Y., Sinha, M., Peterson, C.L. and Weng, Z. (2008) The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet*, **4**, e1000138.
156. Hansen, A.S., Pustova, I., Cattoglio, C., Tjian, R. and Darzacq, X. (2017) CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *Elife*, **6**.
157. Ushiki, A., Zhang, Y., Xiong, C., Zhao, J., Georgakopoulos-Soares, I., Kane, L., Jamieson, K., Bamshad, M.J., Nickerson, D.A., University of Washington Center for Mendelian, G. *et al.* (2021) Deletion of CTCF sites in the SHH locus alters enhancer-promoter interactions and leads to acheiropodia. *Nat Commun*, **12**, 2282.
158. Konrad, E.D.H., Nardini, N., Caliebe, A., Nagel, I., Young, D., Horvath, G., Santoro, S.L., Shuss, C., Ziegler, A., Bonneau, D. *et al.* (2019) CTCF variants in 39 individuals with a variable neurodevelopmental disorder broaden the mutational and clinical spectrum. *Genet Med*, **21**, 2723-2733.
159. Watson, L.A., Wang, X., Elbert, A., Kernohan, K.D., Galjart, N. and Berube, N.G. (2014) Dual effect of CTCF loss on neuroprogenitor differentiation and survival. *J Neurosci*, **34**, 2860-2870.
160. Kim, S., Yu, N.K., Shim, K.W., Kim, J.I., Kim, H., Han, D.H., Choi, J.E., Lee, S.W., Choi, D.I., Kim, M.W. *et al.* (2018) Remote Memory and Cortical Synaptic Plasticity Require Neuronal CCCTC-Binding Factor (CTCF). *J Neurosci*, **38**, 5042-5052.
161. Sams, D.S., Nardone, S., Getselter, D., Raz, D., Tal, M., Rayi, P.R., Kaphzan, H., Hakim, O. and Elliott, E. (2016) Neuronal CTCF Is Necessary for Basal and Experience-Dependent Gene Regulation, Memory Formation, and Genomic Structure of BDNF and Arc. *Cell Rep*, **17**, 2418-2430.
162. Hyle, J., Zhang, Y., Wright, S., Xu, B., Shao, Y., Easton, J., Tian, L., Feng, R., Xu, P. and Li, C. (2019) Acute depletion of CTCF directly affects MYC regulation through loss of enhancer-promoter looping. *Nucleic Acids Res*, **47**, 6699-6713.
163. Xu, B., Wang, H., Wright, S., Hyle, J., Zhang, Y., Shao, Y., Niu, M., Fan, Y., Rosikiewicz, W., Djekidel, M.N. *et al.* (2021) Acute depletion of CTCF rewires genome-wide chromatin accessibility. *Genome Biol*, **22**, 244.
164. Rinaldi, J., Sokol, E.S., Hartmaier, R.J., Trabucco, S.E., Frampton, G.M., Goldberg, M.E., Albacker, L.A., Daemen, A. and Manning, G. (2020) The genomic landscape of metastatic breast cancer: Insights from 11,000 tumors. *PLoS One*, **15**, e0231999.
165. Consortium, I.T.P.-C.A.o.W.G. (2020) Pan-cancer analysis of whole genomes. *Nature*, **578**, 82-93.
166. Li, L. (2009) GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. *J Comput Biol*, **16**, 317-329.
167. Bailey, T.L., Johnson, J., Grant, C.E. and Noble, W.S. (2015) The MEME Suite. *Nucleic Acids Res*, **43**, W39-49.
168. Fornes, O., Castro-Mondragon, J.A., Khan, A., van der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranasic, D. *et al.* (2020) JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*, **48**, D87-D92.
169. Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431-1443.
170. Hume, M.A., Barrera, L.A., Gisselbrecht, S.S. and Bulyk, M.L. (2015) UniPROBE, update

2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res*, **43**, D117-122.
171. Dai, J., Zhu, M., Wang, C., Shen, W., Zhou, W., Sun, J., Liu, J., Jin, G., Ma, H., Hu, Z. *et al.* (2015) Systematical analyses of variants in CTCF-binding sites identified a novel lung cancer susceptibility locus among Chinese population. *Sci Rep*, **5**, 7833.
172. Liu, Y., Walavalkar, N.M., Dozmorov, M.G., Rich, S.S., Civelek, M. and Guertin, M.J. (2017) Identification of breast cancer associated variants that modulate transcription factor binding. *PLoS Genet*, **13**, e1006761.
173. Liu, E.M., Martinez-Fundichely, A., Diaz, B.J., Aronson, B., Cuykendall, T., MacKay, M., Dhingra, P., Wong, E.W.P., Chi, P., Apostolou, E. *et al.* (2019) Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole Genomes. *Cell Syst*, **8**, 446-455 e448.
174. Liu, Y., Li, C., Shen, S., Chen, X., Szlachta, K., Edmonson, M.N., Shao, Y., Ma, X., Hyle, J., Wright, S. *et al.* (2020) Discovery of regulatory noncoding variants in individual cancer genomes by using cis-X. *Nat Genet*, **52**, 811-818.
175. Stormo, G.D. (2015) DNA Motif Databases and Their Uses. *Curr Protoc Bioinformatics*, **51**, 2 15 11-12 15 16.
176. Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd and Bulyk, M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, **24**, 1429-1435.
177. Meers, M.P., Janssens, D.H. and Henikoff, S. (2019) Pioneer Factor-Nucleosome Binding Events during Differentiation Are Motif Encoded. *Mol Cell*, **75**, 562-575 e565.
178. Hansen, A.S., Hsieh, T.S., Cattoglio, C., Pustova, I., Saldana-Meyer, R., Reinberg, D., Darzacq, X. and Tjian, R. (2019) Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region in CTCF. *Mol Cell*, **76**, 395-411 e313.
179. Partridge, E.C., Chhetri, S.B., Prokop, J.W., Ramaker, R.C., Jansen, C.S., Goh, S.T., Mackiewicz, M., Newberry, K.M., Brandsmeier, L.A., Meadows, S.K. *et al.* (2020) Occupancy maps of 208 chromatin-associated proteins in one human cell type. *Nature*, **583**, 720-728.
180. Lebeau, B., Jangal, M., Zhao, T., Wong, C.K., Wong, N., Canedo, E.C., Hebert, S., Aguilar-Mahecha, A., Chabot, C., Buchanan, M. *et al.* (2022) 3D chromatin remodeling potentiates transcriptional programs driving cell invasion. *Proc Natl Acad Sci U S A*, **119**, e2203452119.
181. Driouch, K., Dorion-Bonnet, F., Briffod, M., Champeme, M.H., Longy, M. and Lidereau, R. (1997) Loss of heterozygosity on chromosome arm 16q in breast cancer metastases. *Genes Chromosomes Cancer*, **19**, 185-191.
182. Thomassen, M., Tan, Q. and Kruse, T.A. (2009) Gene expression meta-analysis identifies chromosomal regions and candidate genes involved in breast cancer metastasis. *Breast Cancer Res Treat*, **113**, 239-249.
183. Rakha, E.A., Pinder, S.E., Paish, C.E. and Ellis, I.O. (2004) Expression of the transcription factor CTCF in invasive breast cancer: a candidate gene located at 16q22.1. *Br J Cancer*, **91**, 1591-1596.
184. Lourenco, C., Kalkat, M., Houlahan, K.E., De Melo, J., Longo, J., Done, S.J., Boutros, P.C. and Penn, L.Z. (2019) Modelling the MYC-driven normal-to-tumour switch in breast cancer. *Dis Model Mech*, **12**.
185. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, **15**, 550.

186. Kiavue, N., Cabel, L., Melaabi, S., Bataillon, G., Callens, C., Lerebours, F., Pierga, J.Y. and Bidard, F.C. (2020) ERBB3 mutations in cancer: biological aspects, prevalence and therapeutics. *Oncogene*, **39**, 487-502.
187. Servetto, A., Formisano, L. and Arteaga, C.L. (2021) FGFR signaling and endocrine resistance in breast cancer: Challenges for the clinical development of FGFR inhibitors. *Biochim Biophys Acta Rev Cancer*, **1876**, 188595.
188. Robichaud, N., del Rincon, S.V., Huor, B., Alain, T., Petruccelli, L.A., Hearnden, J., Goncalves, C., Grotegut, S., Spruck, C.H., Furic, L. *et al.* (2015) Phosphorylation of eIF4E promotes EMT and metastasis via translational control of SNAIL and MMP-3. *Oncogene*, **34**, 2032-2042.
189. Umer, H.M., Cavalli, M., Dabrowski, M.J., Diamanti, K., Kruczyk, M., Pan, G., Komorowski, J. and Wadelius, C. (2016) A Significant Regulatory Mutation Burden at a High-Affinity Position of the CTCF Motif in Gastrointestinal Cancers. *Hum Mutat*, **37**, 904-913.
190. Haraguchi, M., Okubo, T., Miyashita, Y., Miyamoto, Y., Hayashi, M., Crotti, T.N., McHugh, K.P. and Ozawa, M. (2008) Snail regulates cell-matrix adhesion by regulation of the expression of integrins and basement membrane proteins. *J Biol Chem*, **283**, 23514-23523.
191. Dibble, C.C. and Cantley, L.C. (2015) Regulation of mTORC1 by PI3K signaling. *Trends Cell Biol*, **25**, 545-555.
192. Thoreen, C.C., Kang, S.A., Chang, J.W., Liu, Q., Zhang, J., Gao, Y., Reichling, L.J., Sim, T., Sabatini, D.M. and Gray, N.S. (2009) An ATP-competitive mammalian target of rapamycin inhibitor reveals rapamycin-resistant functions of mTORC1. *J Biol Chem*, **284**, 8023-8032.
193. Jung, H.Y., Fattet, L., Tsai, J.H., Kajimoto, T., Chang, Q., Newton, A.C. and Yang, J. (2019) Apical-basal polarity inhibits epithelial-mesenchymal transition and tumour metastasis by PAR-complex-mediated SNAI1 degradation. *Nat Cell Biol*, **21**, 359-371.
194. Li, C., Xia, M., Wang, H., Li, W., Peng, J. and Jiang, H. (2020) Propofol facilitates migration and invasion of oral squamous cell carcinoma cells by upregulating SNAI1 expression. *Life Sci*, **241**, 117143.
195. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, **9**, R137.
196. Liu, Z., Scannell, D.R., Eisen, M.B. and Tjian, R. (2011) Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. *Cell*, **146**, 720-731.
197. Pena-Hernandez, R., Marques, M., Hilmi, K., Zhao, T., Saad, A., Alaoui-Jamali, M.A., del Rincon, S.V., Ashworth, T., Roy, A.L., Emerson, B.M. *et al.* (2015) Genome-wide targeting of the epigenetic regulatory protein CTCF to gene promoters by the transcription factor TFII-I. *Proc Natl Acad Sci U S A*, **112**, E677-686.
198. Witcher, M. and Emerson, B.M. (2009) Epigenetic silencing of the p16(INK4a) tumor suppressor is associated with loss of CTCF binding and a chromatin boundary. *Mol Cell*, **34**, 271-284.
199. Fang, C., Wang, Z., Han, C., Safgren, S.L., Helmin, K.A., Adelman, E.R., Serafin, V., Basso, G., Eagen, K.P., Gaspar-Maia, A. *et al.* (2020) Cancer-specific CTCF binding facilitates oncogenic transcriptional dysregulation. *Genome Biol*, **21**, 247.
200. Narendra, V., Rocha, P.P., An, D., Raviram, R., Skok, J.A., Mazzoni, E.O. and Reinberg,

D. (2015) CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science*, **347**, 1017-1021.
201. Soochit, W., Sleutels, F., Stik, G., Bartkuhn, M., Basu, S., Hernandez, S.C., Merzouk, S., Vidal, E., Boers, R., Boers, J. *et al.* (2021) CTCF chromatin residence time controls three-dimensional genome organization, gene expression and DNA methylation in pluripotent cells. *Nat Cell Biol*, **23**, 881-893.
202. Wang, H., Maurano, M.T., Qu, H., Varley, K.E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R. *et al.* (2012) Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res*, **22**, 1680-1688.
203. Lasko, L.M., Jakob, C.G., Edalji, R.P., Qiu, W., Montgomery, D., Digiammarino, E.L., Hansen, T.M., Risi, R.M., Frey, R., Manaves, V. *et al.* (2017) Discovery of a selective catalytic p300/CBP inhibitor that targets lineage-specific tumours. *Nature*, **550**, 128-132.
204. Ren, B., Yang, J., Wang, C., Yang, G., Wang, H., Chen, Y., Xu, R., Fan, X., You, L., Zhang, T. *et al.* (2021) High-resolution Hi-C maps highlight multiscale 3D epigenome reprogramming during pancreatic cancer metastasis. *J Hematol Oncol*, **14**, 120.
205. Bauer, M., Vidal, E., Zorita, E., Uresin, N., Pinter, S.F., Filion, G.J. and Payer, B. (2021) Chromosome compartments on the inactive X guide TAD formation independently of transcription during X-reactivation. *Nat Commun*, **12**, 3499.
206. Wang, J., Huang, T.Y., Hou, Y., Bartom, E., Lu, X., Shilatifard, A., Yue, F. and Saratsis, A. (2021) Epigenomic landscape and 3D genome structure in pediatric high-grade glioma. *Sci Adv*, **7**.
207. Barutcu, A.R., Lajoie, B.R., McCord, R.P., Tye, C.E., Hong, D., Messier, T.L., Browne, G., van Wijnen, A.J., Lian, J.B., Stein, J.L. *et al.* (2015) Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biol*, **16**, 214.
208. Wang, X.T., Cui, W. and Peng, C. (2017) HiTAD: detecting the structural and functional hierarchies of topologically associating domains from chromatin interactions. *Nucleic Acids Res*, **45**, e163.
209. Vos, E.S.M., Valdes-Quezada, C., Huang, Y., Allahyar, A., Verstegen, M., Felder, A.K., van der Vegt, F., Uijttewaal, E.C.H., Krijger, P.H.L. and de Laat, W. (2021) Interplay between CTCF boundaries and a super enhancer controls cohesin extrusion trajectories and gene expression. *Mol Cell*, **81**, 3082-3095 e3086.
210. Kruse, K., Hug, C.B. and Vaquerizas, J.M. (2020) FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome Biol*, **21**, 303.
211. Barutcu, A.R., Maass, P.G., Lewandowski, J.P., Weiner, C.L. and Rinn, J.L. (2018) A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. *Nat Commun*, **9**, 1444.
212. Cameron, C.J., Dostie, J. and Blanchette, M. (2020) HIFI: estimating DNA-DNA interaction frequency from Hi-C data at restriction-fragment resolution. *Genome Biol*, **21**, 11.
213. Lebeau, B., Zhao, K., Jangal, M., Zhao, T., Guerra, M., Greenwood, C.M.T. and Witcher, M. (2022) Single base-pair resolution analysis of DNA binding motif with MoMotif reveals an oncogenic function of CTCF zinc-finger 1 mutation. *Nucleic Acids Res*.
214. Fritz, A.J., Ghule, P.N., Boyd, J.R., Tye, C.E., Page, N.A., Hong, D., Shirley, D.J., Weinheimer, A.S., Barutcu, A.R., Gerrard, D.L. *et al.* (2018) Intranuclear and higher-order chromatin organization of the major histone gene cluster in breast cancer. *J Cell Physiol*,

233, 1278-1290.
215. Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J., Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R. *et al.* (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389-393.
216. Filippova, G.N., Fagerlie, S., Klenova, E.M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P.E., Collins, S.J. and Lobanenkov, V.V. (1996) 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*, **16**, 2802-2813.
217. Persikov, A.V., Rowland, E.F., Oakes, B.L., Singh, M. and Noyes, M.B. (2014) Deep sequencing of large library selections allows computational discovery of diverse sets of zinc fingers that bind common targets. *Nucleic Acids Res*, **42**, 1497-1508.
218. Persikov, A.V., Wetzel, J.L., Rowland, E.F., Oakes, B.L., Xu, D.J., Singh, M. and Noyes, M.B. (2015) A systematic survey of the Cys2His2 zinc finger DNA-binding landscape. *Nucleic Acids Res*, **43**, 1965-1984.
219. Zufferey, M., Tavernari, D., Oricchio, E. and Ciriello, G. (2018) Comparison of computational methods for the identification of topologically associating domains. *Genome Biol*, **19**, 217.
220. Cresswell, K.G., Stansfield, J.C. and Dozmorov, M.G. (2020) SpectralTAD: an R package for defining a hierarchy of topologically associated domains using spectral clustering. *BMC Bioinformatics*, **21**, 319.
221. Kentepozidou, E., Aitken, S.J., Feig, C., Stefflova, K., Ibarra-Soria, X., Odom, D.T., Roller, M. and Flicek, P. (2020) Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biol*, **21**, 5.
222. Shah, R., Smith, P., Purdie, C., Quinlan, P., Baker, L., Aman, P., Thompson, A.M. and Crook, T. (2009) The prolyl 3-hydroxylases P3H2 and P3H3 are novel targets for epigenetic silencing in breast cancer. *Br J Cancer*, **100**, 1687-1696.
223. Loftus, P.G., Watson, L., Deedigan, L.M., Camarillo-Retamosa, E., Dwyer, R.M., O'Flynn, L., Alagesan, S., Griffin, M., O'Brien, T., Kerin, M.J. *et al.* (2021) Targeting stromal cell Syndecan-2 reduces breast tumour growth, metastasis and limits immune evasion. *Int J Cancer*, **148**, 1245-1259.
224. Odagiri, H., Kadomatsu, T., Endo, M., Masuda, T., Morioka, M.S., Fukuhara, S., Miyamoto, T., Kobayashi, E., Miyata, K., Aoi, J. *et al.* (2014) The secreted protein ANGPTL2 promotes metastasis of osteosarcoma cells through integrin alpha5beta1, p38 MAPK, and matrix metalloproteinases. *Sci Signal*, **7**, ra7.
225. Lehner, A., Magdolen, V., Schuster, T., Kotzsch, M., Kiechle, M., Meindl, A., Sweep, F.C., Span, P.N. and Gross, E. (2013) Downregulation of serine protease HTRA1 is associated with poor survival in breast cancer. *PLoS One*, **8**, e60359.
226. Braccioli, L. and de Wit, E. (2019) CTCF: a Swiss-army knife for genome organization and transcription regulation. *Essays Biochem*, **63**, 157-165.
227. Hanssen, L.L.P., Kassouf, M.T., Oudelaar, A.M., Biggs, D., Preece, C., Downes, D.J., Gosden, M., Sharpe, J.A., Sloane-Stanley, J.A., Hughes, J.R. *et al.* (2017) Tissue-specific CTCF-cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nature cell biology*, **19**, 952-961.
228. Swinstead, E.E., Miranda, T.B., Paakinaho, V., Baek, S., Goldstein, I., Hawkins, M., Karpova, T.S., Ball, D., Mazza, D., Lavis, L.D. *et al.* (2016) Steroid Receptors Reprogram

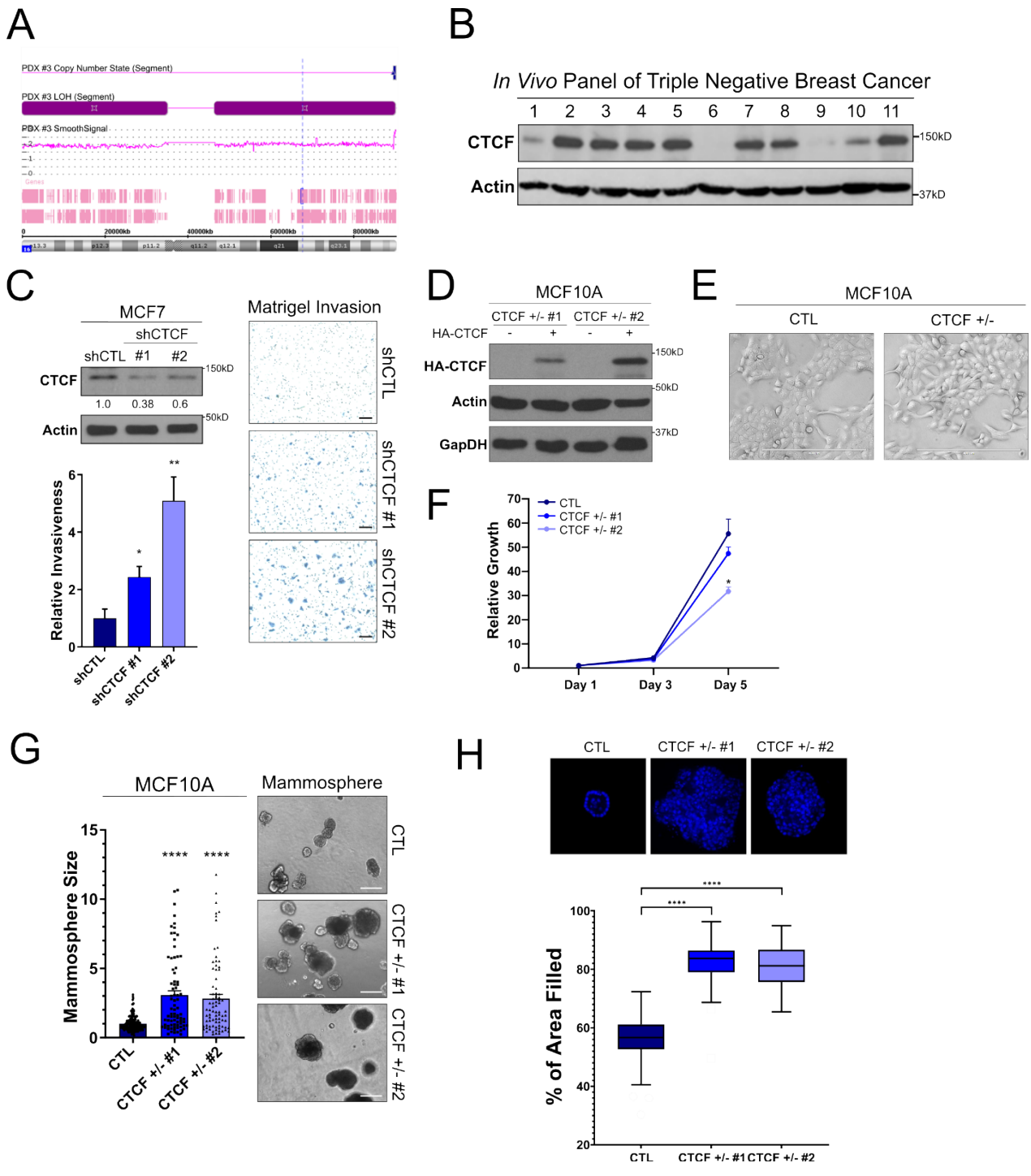
FoxA1 Occupancy through Dynamic Chromatin Transitions. *Cell*, **165**, 593-605.
229. Frietze, S., Lan, X., Jin, V.X. and Farnham, P.J. (2010) Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263. *J Biol Chem*, **285**, 1393-1403.
230. Kennedy, B.A., Lan, X., Huang, T.H., Farnham, P.J. and Jin, V.X. (2012) Using ChIPMotifs for de novo motif discovery of OCT4 and ZNF263 based on ChIP-based high-throughput experiments. *Methods Mol Biol*, **802**, 323-334.
231. Liu, X., Krawczyk, E., Suprynowicz, F.A., Palechor-Ceron, N., Yuan, H., Dakic, A., Simic, V., Zheng, Y.L., Sripadhan, P., Chen, C. *et al.* (2017) Conditional reprogramming and long-term expansion of normal and tumor cells from human biospecimens. *Nat Protoc*, **12**, 439-451.
232. Sirois, I., Aguilar-Mahecha, A., Lafleur, J., Fowler, E., Vu, V., Scriver, M., Buchanan, M., Chabot, C., Ramanathan, A., Balachandran, B. *et al.* (2019) A Unique Morphological Phenotype in Chemoresistant Triple-Negative Breast Cancer Reveals Metabolic Reprogramming and PLIN4 Expression as a Molecular Vulnerability. *Mol Cancer Res*, **17**, 2492-2507.
233. Savage, P., Pacis, A., Kuasne, H., Liu, L., Lai, D., Wan, A., Dankner, M., Martinez, C., Munoz-Ramos, V., Pilon, V. *et al.* (2020) Chemogenomic profiling of breast cancer patient-derived xenografts reveals targetable vulnerabilities for difficult-to-treat tumors. *Commun Biol*, **3**, 310.
234. Labun, K., Montague, T.G., Krause, M., Torres Cleuren, Y.N., Tjeldnes, H. and Valen, E. (2019) CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res*, **47**, W171-W174.
235. Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114-2120.
236. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15-21.
237. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.
238. Liao, Y., Smyth, G.K. and Shi, W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923-930.
239. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545-15550.
240. Mi, H. and Thomas, P. (2009) PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol Biol*, **563**, 123-140.
241. Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C. *et al.* (2022) The reactome pathway knowledgebase 2022. *Nucleic Acids Res*, **50**, D687-D692.
242. Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T. and Carey, V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput Biol*, **9**, e1003118.
243. Yu, G., Wang, L.G., Han, Y. and He, Q.Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**, 284-287.

244. Yu, G., Wang, L.G. and He, Q.Y. (2015) ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382-2383.
245. Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A. and Manke, T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res*, **42**, W187-191.
246. Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat Biotechnol*, **29**, 24-26.
247. Lun, A.T. and Smyth, G.K. (2016) csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Res*, **44**, e45.
248. Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139-140.
249. SIMES, R.J. (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751-754.
250. Jayaram, N., Usvyat, D. and AC, R.M. (2016) Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics*, **17**, 547.
251. Eden, E., Lipson, D., Yogev, S. and Yakhini, Z. (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol*, **3**, e39.
252. Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, **2**, 28-36.
253. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W. and Noble, W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*, **37**, W202-208.
254. Li, L., Bass, R.L. and Liang, Y. (2008) fdrMotif: identifying cis-elements by an EM algorithm coupled with false discovery rate control. *Bioinformatics*, **24**, 629-636.
255. Cao, Y., Liu, S., Ren, G., Tang, Q. and Zhao, K. (2022) cLoops2: a full-stack comprehensive analytical tool for chromatin interactions. *Nucleic Acids Res*, **50**, 57-71.
256. Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander, E.S. and Aiden, E.L. (2016) Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst*, **3**, 95-98.
257. Ramirez, F., Bhardwaj, V., Arrigoni, L., Lam, K.C., Gruning, B.A., Villaveces, J., Habermann, B., Akhtar, A. and Manke, T. (2018) High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun*, **9**, 189.
258. Batada, N.N. and Hurst, L.D. (2007) Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat Genet*, **39**, 945-949.
259. Luo, H., Yu, Q., Liu, Y., Tang, M., Liang, M., Zhang, D., Xiao, T.S., Wu, L., Tan, M., Ruan, Y. *et al.* (2020) LATS kinase-mediated CTCF phosphorylation and selective loss of genomic binding. *Sci Adv*, **6**, eaaw4651.
260. Kim, Y.H., Marhon, S.A., Zhang, Y., Steger, D.J., Won, K.J. and Lazar, M.A. (2018) Rev-erbalpha dynamically modulates chromatin looping to control circadian gene transcription. *Science*, **359**, 1274-1277.
261. Zampetidis, C.P., Galanos, P., Angelopoulou, A., Zhu, Y., Polyzou, A., Karamitros, T., Kotsinas, A., Lagopati, N., Mourkioti, I., Mirzazadeh, R. *et al.* (2021) A recurrent chromosomal inversion suffices for driving escape from oncogene-induced senescence via subTAD reorganization. *Mol Cell*, **81**, 4907-4923 e4908.
262. Nieto, M.A., Huang, R.Y., Jackson, R.A. and Thiery, J.P. (2016) Emt: 2016. *Cell*, **166**, 21-45.

263. Fingar, D.C. and Blenis, J. (2004) Target of rapamycin (TOR): an integrator of nutrient and growth factor signals and coordinator of cell growth and cell cycle progression. *Oncogene*, **23**, 3151-3171.
264. Chen, Y., Fan, Z., Wang, X., Mo, M., Zeng, S.B., Xu, R.H., Wang, X. and Wu, Y. (2020) PI3K/Akt signaling pathway is essential for de novo hair follicle regeneration. *Stem Cell Res Ther*, **11**, 144.
265. Gutierrez, L., Caballero, N., Fernandez-Calleja, L., Karkoulia, E. and Strouboulis, J. (2020) Regulation of GATA1 levels in erythropoiesis. *IUBMB Life*, **72**, 89-105.
266. Nowak, D., Stewart, D. and Koeffler, H.P. (2009) Differentiation therapy of leukemia: 3 decades of development. *Blood*, **113**, 3655-3665.
267. Kung, J.T., Kesner, B., An, J.Y., Ahn, J.Y., Cifuentes-Rojas, C., Colognori, D., Jeon, Y., Szanto, A., del Rosario, B.C., Pinter, S.F. *et al.* (2015) Locus-specific targeting to the X chromosome revealed by the RNA interactome of CTCF. *Mol Cell*, **57**, 361-375.
268. Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K. *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods*, **13**, 508-514.
269. Hollin, T., Abel, S. and Le Roch, K.G. (2021) Genome-Wide Analysis of RNA-Protein Interactions in Plasmodium falciparum Using eCLIP-Seq. *Methods Mol Biol*, **2369**, 139-164.
270. Mumbach, M.R., Satpathy, A.T., Boyle, E.A., Dai, C., Gowen, B.G., Cho, S.W., Nguyen, M.L., Rubin, A.J., Granja, J.M., Kazane, K.R. *et al.* (2017) Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat Genet*, **49**, 1602-1612.
271. Wang, C., Zhang, L., Ke, L., Ding, W., Jiang, S., Li, D., Narita, Y., Hou, I., Liang, J., Li, S. *et al.* (2020) Primary effusion lymphoma enhancer connectome links super-enhancers to dependency factors. *Nat Commun*, **11**, 6318.
272. Wei, L., Liu, Q., Huang, Y., Liu, Z., Zhao, R., Li, B., Zhang, J., Sun, C., Gao, B., Ding, X. *et al.* (2020) Knockdown of CTCF reduces the binding of EZH2 and affects the methylation of the SOCS3 promoter in hepatocellular carcinoma. *Int J Biochem Cell Biol*, **120**, 105685.
273. Kang, H., Shokhirev, M.N., Xu, Z., Chandran, S., Dixon, J.R. and Hetzer, M.W. (2020) Dynamic regulation of histone modifications and long-range chromosomal interactions during postmitotic transcriptional reactivation. *Genes Dev*, **34**, 913-930.
274. Greenwald, W.W., Li, H., Benaglio, P., Jakubosky, D., Matsui, H., Schmitt, A., Selvaraj, S., D'Antonio, M., D'Antonio-Chronowska, A., Smith, E.N. *et al.* (2019) Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nat Commun*, **10**, 1054.
275. Narod, S.A. and Foulkes, W.D. (2004) BRCA1 and BRCA2: 1994 and beyond. *Nat Rev Cancer*, **4**, 665-676.
276. Hanahan, D. (2022) Hallmarks of Cancer: New Dimensions. *Cancer Discov*, **12**, 31-46.
277. Laiho, M., DeCaprio, J.A., Ludlow, J.W., Livingston, D.M. and Massague, J. (1990) Growth inhibition by TGF-beta linked to suppression of retinoblastoma protein phosphorylation. *Cell*, **62**, 175-185.
278. Padua, D. and Massague, J. (2009) Roles of TGFbeta in metastasis. *Cell Res*, **19**, 89-102.
279. Leivonen, S.K. and Kahari, V.M. (2007) Transforming growth factor-beta signaling in cancer invasion and metastasis. *Int J Cancer*, **121**, 2119-2124.

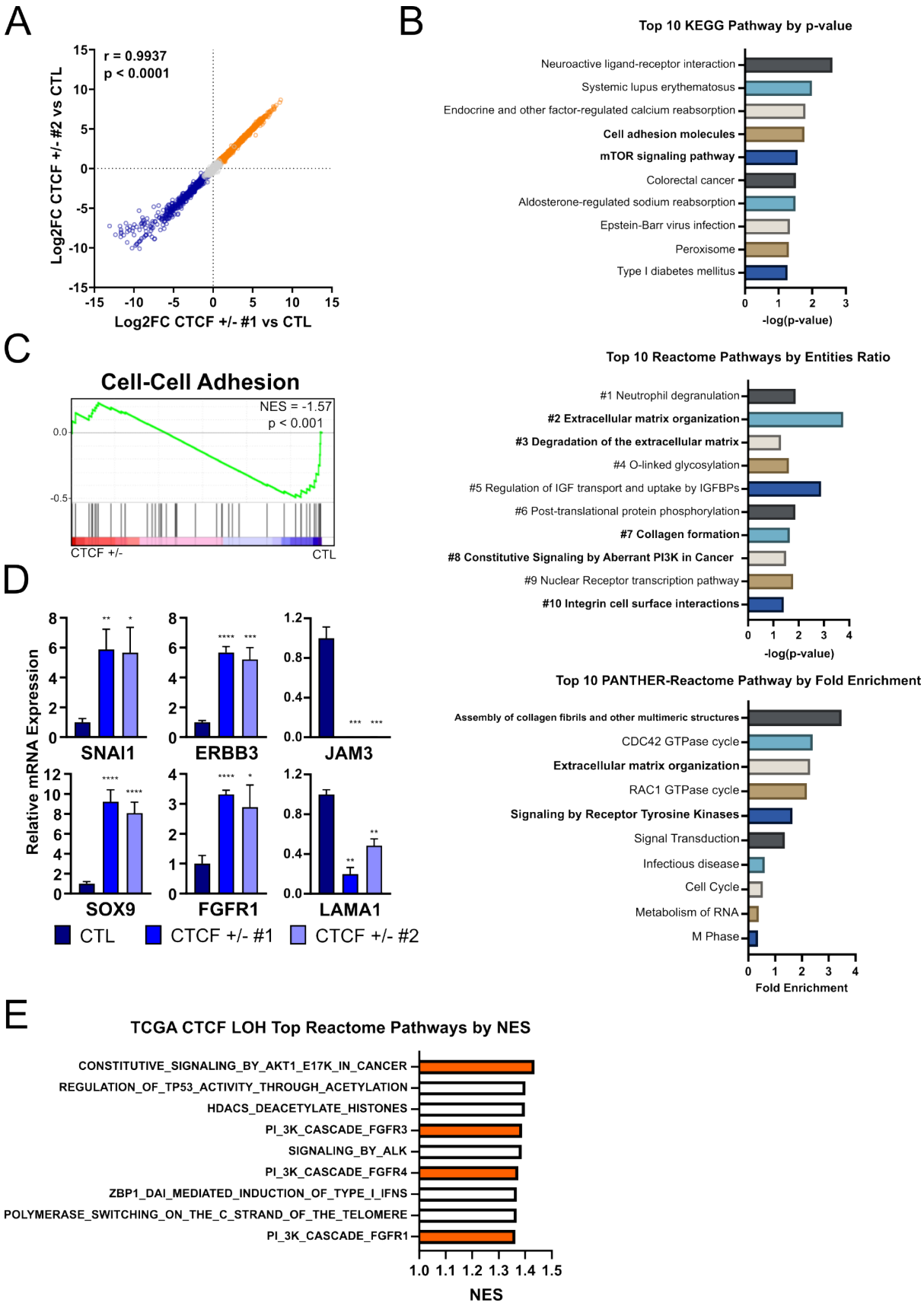
280. Casasent, A.K., Schalck, A., Gao, R., Sei, E., Long, A., Pangburn, W., Casasent, T., Meric-Bernstam, F., Edgerton, M.E. and Navin, N.E. (2018) Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell*, **172**, 205-217 e212.
281. Nam, A.S., Chaligne, R. and Landau, D.A. (2021) Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. *Nat Rev Genet*, **22**, 3-18.
282. Flavahan, W.A., Drier, Y., Liau, B.B., Gillespie, S.M., Venteicher, A.S., Stemmer-Rachamimov, A.O., Suva, M.L. and Bernstein, B.E. (2016) Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, **529**, 110-114.
283. Ursini-Siegel, J., Hardy, W.R., Zuo, D., Lam, S.H., Sanguin-Gendreau, V., Cardiff, R.D., Pawson, T. and Muller, W.J. (2008) ShcA signalling is essential for tumour progression in mouse models of human breast cancer. *EMBO J*, **27**, 910-920.

Supplementary Information

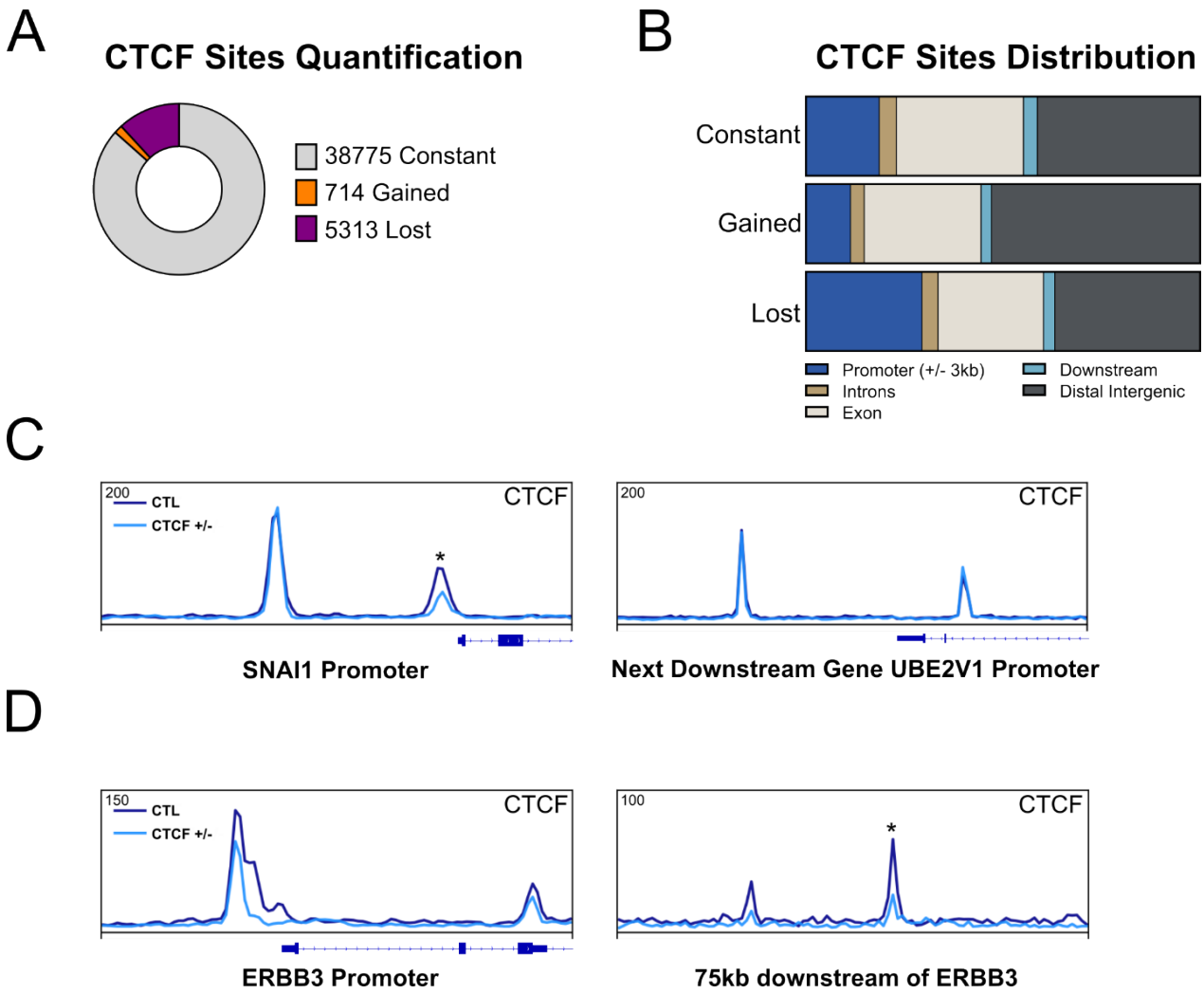


Supplementary Figure 1.1. (A) Copy number analysis of PDX #3 cell line using the Chromosome Analysis Suite from ThermoFisher showing CTCF Loss of Heterozygosity

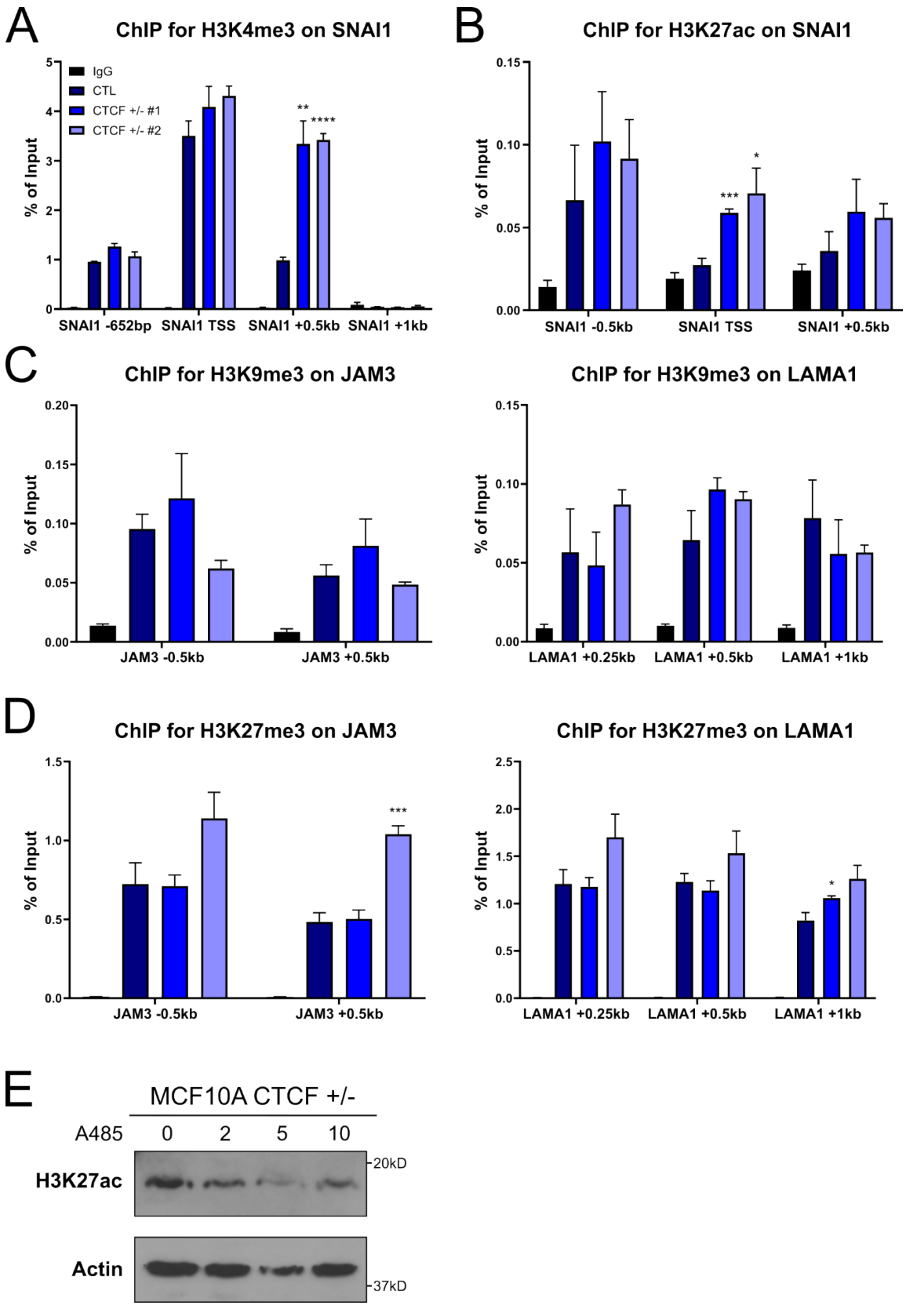
(demarcated by a purple box) at the CTCF loci (demarcated by the dotted vertical line). (B) Western Blot of tumors from triple negative breast cancer patient derived xenograft showing that physiological CTCF levels found in tumors 1, 9 and 10 are similar to the CTCF level of our cell line models. (C) Western Blot of CTCF levels following shCTCF expression in MCF7 with quantification of relative CTCF band intensity to shCTL. Loading control: Actin. Bar chart of relative invasiveness of MCF7 shCTCF #1 and #2 to shCTL (mean ± SEM). $p = 0.021$ and 0.0051 for shCTCF #1 and #2 compared to shCTL, showing an increased invasiveness following shCTCF treatment. (D) Western Blot showing ectopic expression of HA-CTCF in CTCF+/- cells. Actin and GapDH are used as loading control. (E) Brightfield microscopy picture of CTL and CTCF+/- MCF10A at similar confluence, showing no major morphological changes between the cell lines. (F) Line chart (mean ± SEM) of the growth of CTL, CTCF+/- #1 and CTCF+/- #2 MCF10A during a 5-day growth curve assay. CTCF+/- #2 proliferates at a modestly slower rate than CTL ($p = 0.02$ using a two-tailed Student's T Test). All other comparisons are non-significant. (G) Average and individual mammosphere size of CTCF+/- MCF10A relative to CTL (mean ± SEM) showing an increase in mammosphere size in CTCF +/- cells, p-values < 0.0001 for both CTCF+/- #1 and #2. All p-values were calculated using Student's T Test. (H) Mammosphere DAPI immunofluorescence of CTL and CTCF+/- MCF10A showing an increased filing of the CTCF+/- mammospheres. Tukey Box plot below images represents the distribution of mammosphere filling in each cell lines, showing a significant increase in cells filling the core of CTCF+/- MCF10A mammospheres compared to CTL. $p < 0.0001$ for CTCF+/- #1 and #2 using a two-tailed Student's T Test. (I) Western blot of CDH1 levels. GapDH was used as a loading control. Bar chart (mean ± SEM) below represents the relative invasiveness in a matrigel invasion assay of each cell line, normalized to CTL, showing no increase in invasiveness following CDH1 KO. $p = 0.0459$ and 0.129 for CDH1 KO #1 and #2 compared to CTL using a one-tailed Student's T Test. Pictures on the right show the inserts following invasion. The few invading cells are indicated with black arrows.


A
r = 0.9937
p < 0.0001
Log2FC CTCF +/- #2 vs CTL
Log2FC CTCF +/- #1 vs CTL
B
Top 10 KEGG Pathway by p-value
Neuroactive ligand-receptor interaction
Systemic lupus erythematosus
Endocrine and other factor-regulated calcium reabsorption
Cell adhesion molecules
mTOR signaling pathway
Colorectal cancer
Aldosterone-regulated sodium reabsorption
Epstein-Barr virus infection
Peroxisome
Type I diabetes mellitus
-log(p-value)
C
Cell-Cell Adhesion
NES = -1.57
p < 0.001
CTCF +/-
CTL
Top 10 Reactome Pathways by Entities Ratio
#1 Neutrophil degranulation
#2 Extracellular matrix organization
#3 Degradation of the extracellular matrix
#4 O-linked glycosylation
#5 Regulation of IGF transport and uptake by IGFBPs
#6 Post-translational protein phosphorylation
#7 Collagen formation
#8 Constitutive Signaling by Aberrant PI3K in Cancer
#9 Nuclear Receptor transcription pathway
#10 Integrin cell surface interactions
-log(p-value)
D
Relative mRNA Expression
SNAI1
ERBB3
JAM3
SOX9
FGFR1
LAMA1
CTL
CTCF +/- #1
CTCF +/- #2
Top 10 PANTHER-Reactome Pathway by Fold Enrichment
Assembly of collagen fibrils and other multimeric structures
CDC42 GTPase cycle
Extracellular matrix organization
RAC1 GTPase cycle
Signaling by Receptor Tyrosine Kinases
Signal Transduction
Infectious disease
Cell Cycle
Metabolism of RNA
M Phase
Fold Enrichment
E
TCGA CTCF LOH Top Reactome Pathways by NES
CONSTITUTIVE_SIGNALING_BY_AKT1_E17K_IN_CANCER
REGULATION_OF_TP53_ACTIVITY_THROUGH_ACETYLATION
HDACS_DEACETYLATE_HISTONES
PI_3K_CASCADE_FGFR3
SIGNALING_BY_ALK
PI_3K_CASCADE_FGFR4
ZBP1_DAI_MEDIATED_INDUCTION_OF_TYPE_I_IFNS
POLYMERASE_SWITCHING_ON_THE_C_STRAND_OF_THE_TELOMERE
PI_3K_CASCADE_FGFR1
NES

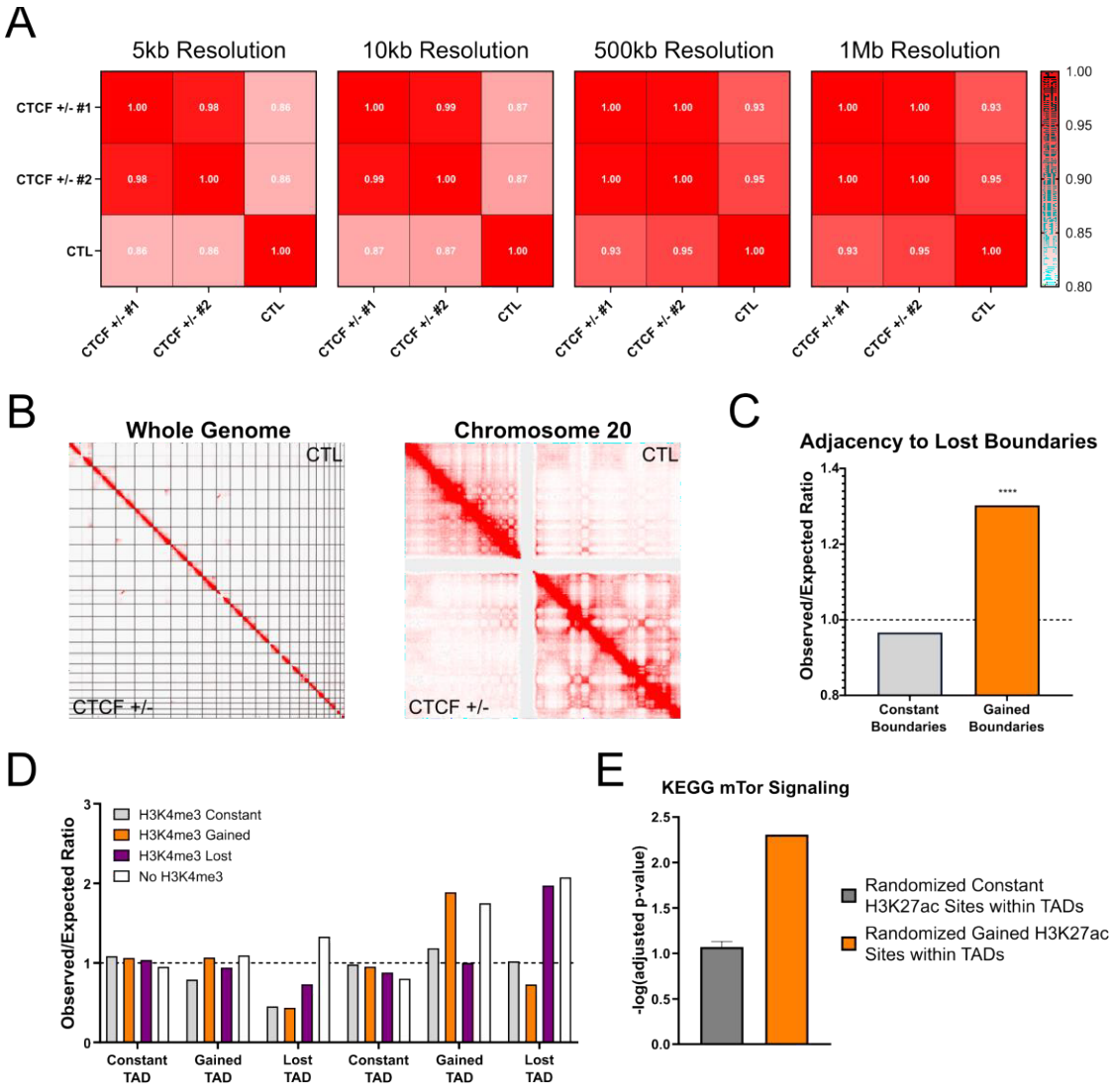
Supplementary Figure 1.2. (A) Correlation of changes in gene expression of the respective CTCF+/- clones to CTL (Pearson r and p-value), showing a strong reproducibility of gene expression changes in both cell lines. (B) Top 10 KEGG and Reactome pathway analysis by p-value and entities ratio, respectively, for upregulated genes (Log2FC >= 2, adjusted p-value <= 0.05). Top 10 Reactome pathways, analyzed with PANTHER and ranked by fold enrichment, of significantly up or downregulated genes (abs(Log2FC) >= 1, adjusted p-value <= 0.05). Gene sets related to PI3K signaling pathway or EMT are written in bold and are distinctly present in the top 10 of each analysis method. (C) GSEA Analysis highlighting the downregulation of cell-cell adhesion pathway from the Gene Ontology data set. (D) Bar chart (mean ± SEM) of the qPCR validation of the top hits in the RNA-Seq, showing the relative expression normalized on 3 housekeeping genes compared to CTL. p-values are listed below and were calculated using a two-tailed Student's T Test comparing CTCF+/- #1 and #2 to CTL: SNAI1 #1 = 0.0075 and #2 = 0.0175 ; ERBB3 #1 < 0.001 and #2 = 0.0004 ; SOX9 #1 and #2 < 0.0001 ; FGFR1 #1 < 0.0001 and #2 = 0.0379; JAM3 #1 and #2 = 0.0001 ; LAMA1 #1 = 0.0016 and #2 = 0.0081. (E) Top 10 Reactome Pathways ranked by NES of the GSEA Prerank analysis on TCGA dataset as described in Figure 2C. Pathways related to PI3K signaling are highlighted in orange and dominate the top 10.



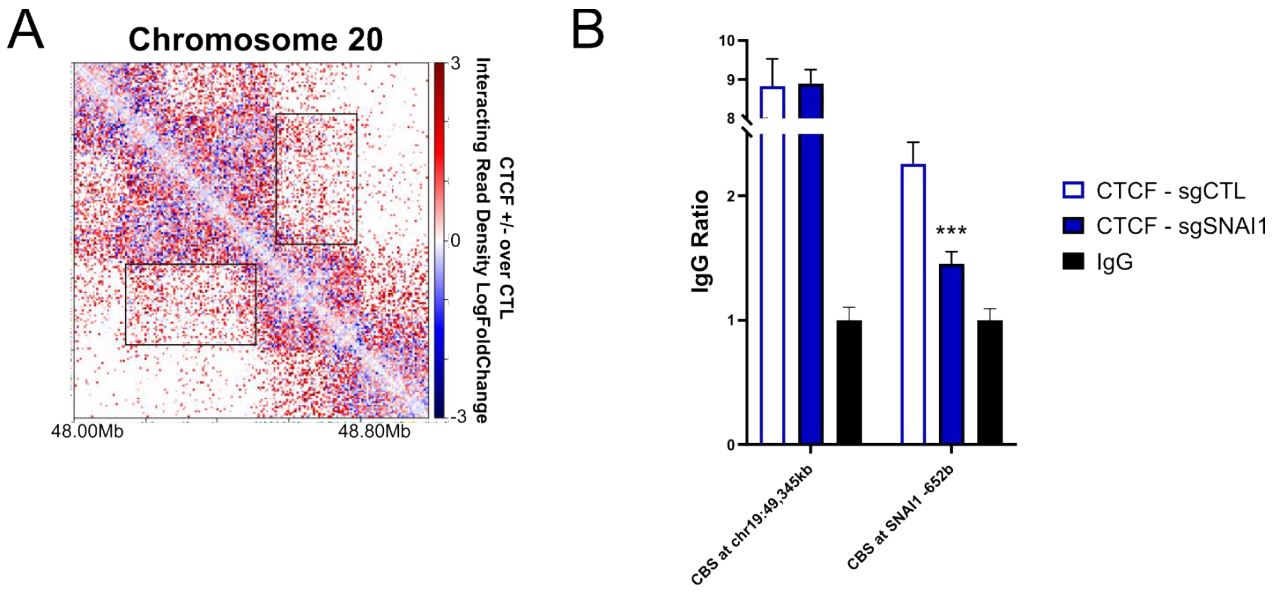
Supplementary Figure 1.3. (A) Partitioning of constant lost and gain sites from Figure 4A. (B) Genomic distribution of the different clusters shows a small enrichment of lost sites on promoters and gain sites on distal intergenic sites, compared to constant CTCF binding sites. (C/D) ChIP-Seq track of normalized read density showing the specific loss of CTCF binding at SNAI1 promoter and downstream of ERRB3 ($p <= 0.05$ calculated using DiffBind 3.0), as ChIP- Seq track of surrounding regions show no significant change in CTCF binding. The normalized read density of each track ranges from 0 to the number noted in the top left corner.


A
ChIP for H3K4me3 on SNAI1
IgG
CTL
CTCF +/- #1
CTCF +/- #2
% of Input
SNAI1 -652bp
SNAI1 TSS
SNAI1 +0.5kb
SNAI1 +1kb
B
ChIP for H3K27ac on SNAI1
% of Input
SNAI1 -0.5kb
SNAI1 TSS
SNAI1 +0.5kb
C
ChIP for H3K9me3 on JAM3
% of Input
JAM3 -0.5kb
JAM3 +0.5kb
ChIP for H3K9me3 on LAMA1
% of Input
LAMA1 +0.25kb
LAMA1 +0.5kb
LAMA1 +1kb
D
ChIP for H3K27me3 on JAM3
% of Input
JAM3 -0.5kb
JAM3 +0.5kb
ChIP for H3K27me3 on LAMA1
% of Input
LAMA1 +0.25kb
LAMA1 +0.5kb
LAMA1 +1kb
E
MCF10A CTCF +/-
A485
0
2
5
10
H3K27ac
20kD
Actin
37kD

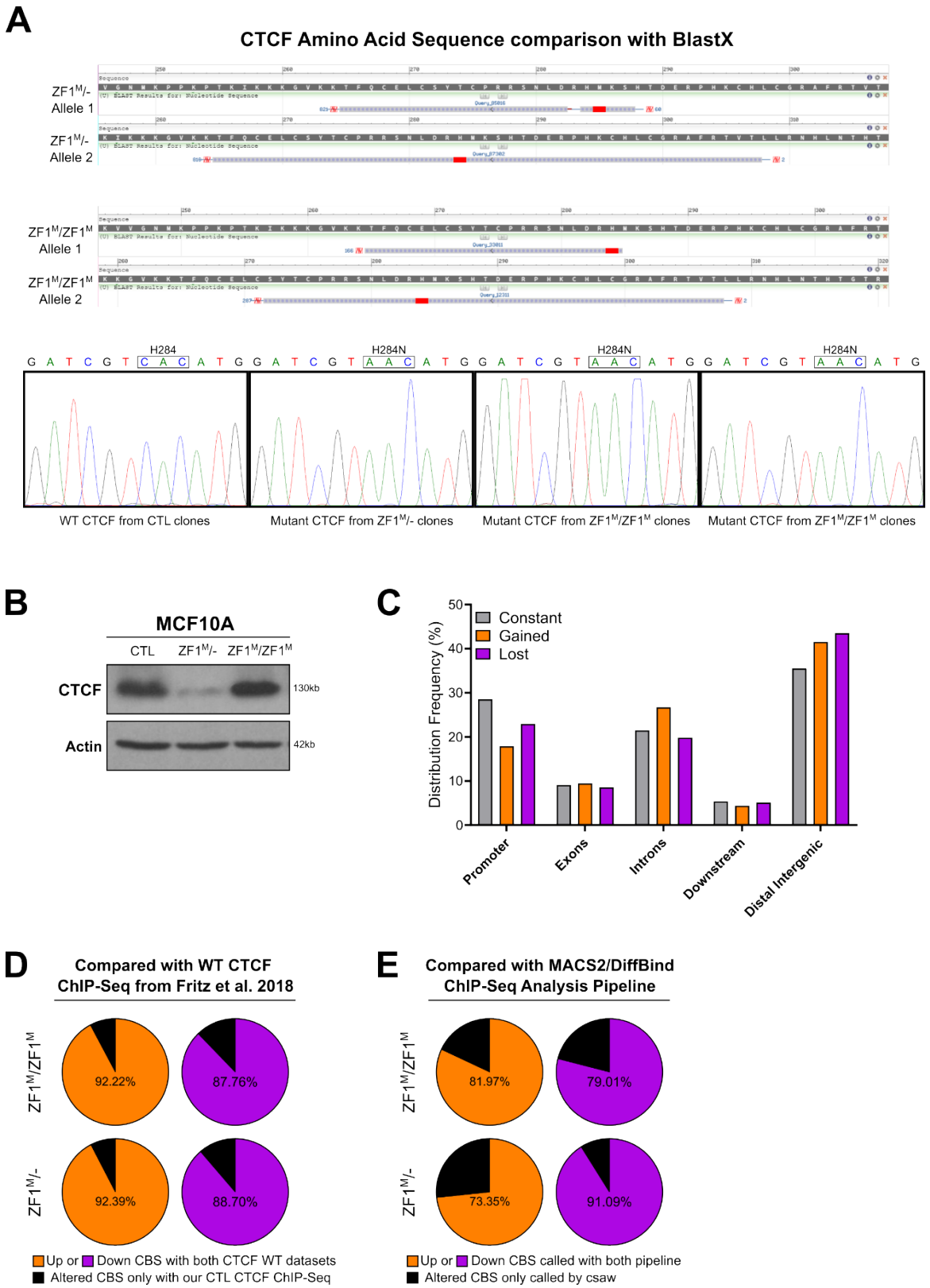
Supplementary Figure 1.4. ChIP-qPCR Screening of histone modifications around genes with altered expression in CTCF+/- MCF10A. All bar charts represent mean % of input ± SEM. The coordinates of each site is represented as distance to TSS. (A) ChIP-qPCR results showing the significant gain of H3K4me3 in CTCF+/- MCF10As relative to CTL at 500bp downstream of SNAI1. p= 0.00729 and 0.000068 in CTCF+/- #1 and #2, respectively. (B) ChIP-qPCR results showing the significant gain of H3K27ac in CTCF+/- MCF10As relative to CTL SNAI1 TSS. p
= 0.00169 and 0.0247 in CTCF+/- #1 and #2 respectively. (C) ChIP-qPCR results showing no significant changes in H3K9me3 binding between CTCF+/- and CTL MCF10As around JAM3 and LAMA1 TSS. (D) ChIP-qPCR results showing no consistent significant changes in H3K27me3 binding between CTCF+/- and CTL MCF10As around JAM3 and LAMA1 TSS. P- values at JAM3 TSS +0.5kb = 0.806 and 0.000170; p-values at LAMA1 TSS +1kb =0.0143 and 0.0708 in CTCF+/- #1 and #2, respectively. (E) Western Blot showing the decrease in H3K27ac levels following A485 treatments (in μM) in MCF10A CTCF+/- cell.



Supplementary Figure 1.5. (A) Pearson Correlation Coefficient heat maps comparing the contact frequencies between CTCF +/- #1 and #2 and CTL, showing a distinction between CTL and CTCF +/- at high resolution, where shorter interactions can be surveyed, while this distinction fades at lower resolution. (B) Juicebox heatmap of balanced interaction read count for the whole genome and chromosome 20, where SNAI1 loci is located. Interactions in CTL MCF10A are represented in the top/right half, while interaction in CTCF+/- are represented in the bottom/left half of each heat map. (C) Bar Chart (Observed/Expected Ratio) representing the enrichment of constant or gained boundaries adjacent to a lost boundaries (+/- 1 boundaries) showing that gained boundaries are significantly enriched around lost boundaries (Chi Square Test, $p < 0.0001$
). (D) Bar chart (O/E Ratio) showing the enrichment of gained H3K4me3 on gained subTAD and vice-versa for lost H3K4me3 and lost subTAD. (E) Differential Enrichment of mTor Signaling KEGG Pathway at gained H3K27ac sites within TADs compared to 100 equinumerous subsets of constant H3K27ac sites within TADs.



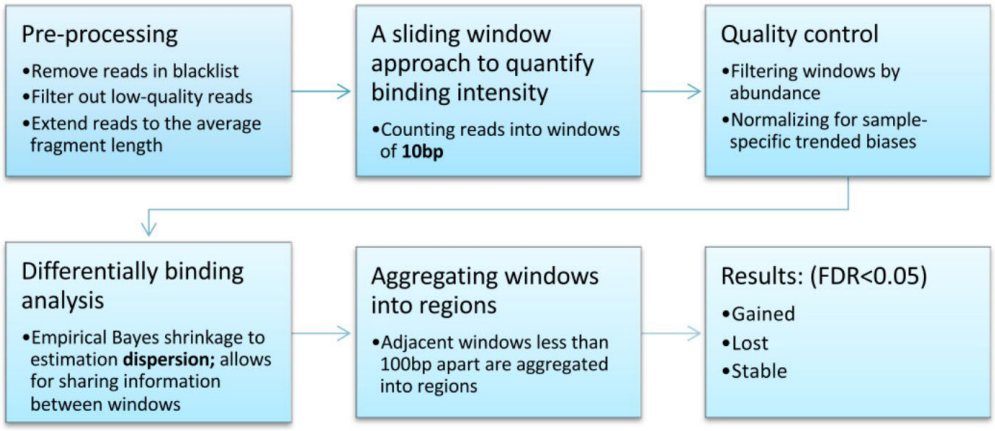
Supplementary Figure 1.6. (A) Symmetrical comparative heat map of the differential interaction (in logFC) of CTCF+/- compared to CTL MCF10A at the same genomic region as the second heatmap in Figure 7A. The black boxes highlight a zone of modestly increased interdomain interactions in the CTCF+/- compared to CTL MCF10A. (B) Bar chart (IgG Ratio
+/- SEM, n =4) of CTCF ChIP-qPCR results in CTL MCF10A following dCAS9 and sgCTL or sgSNAI1 expression showing a significant reduction of CTCF binding at the target site upstream of SNAI1 by sgSNAI1. As control, CBS at chr19:49,345kb is not target by sgSNAI1 and is therefore not hindered.



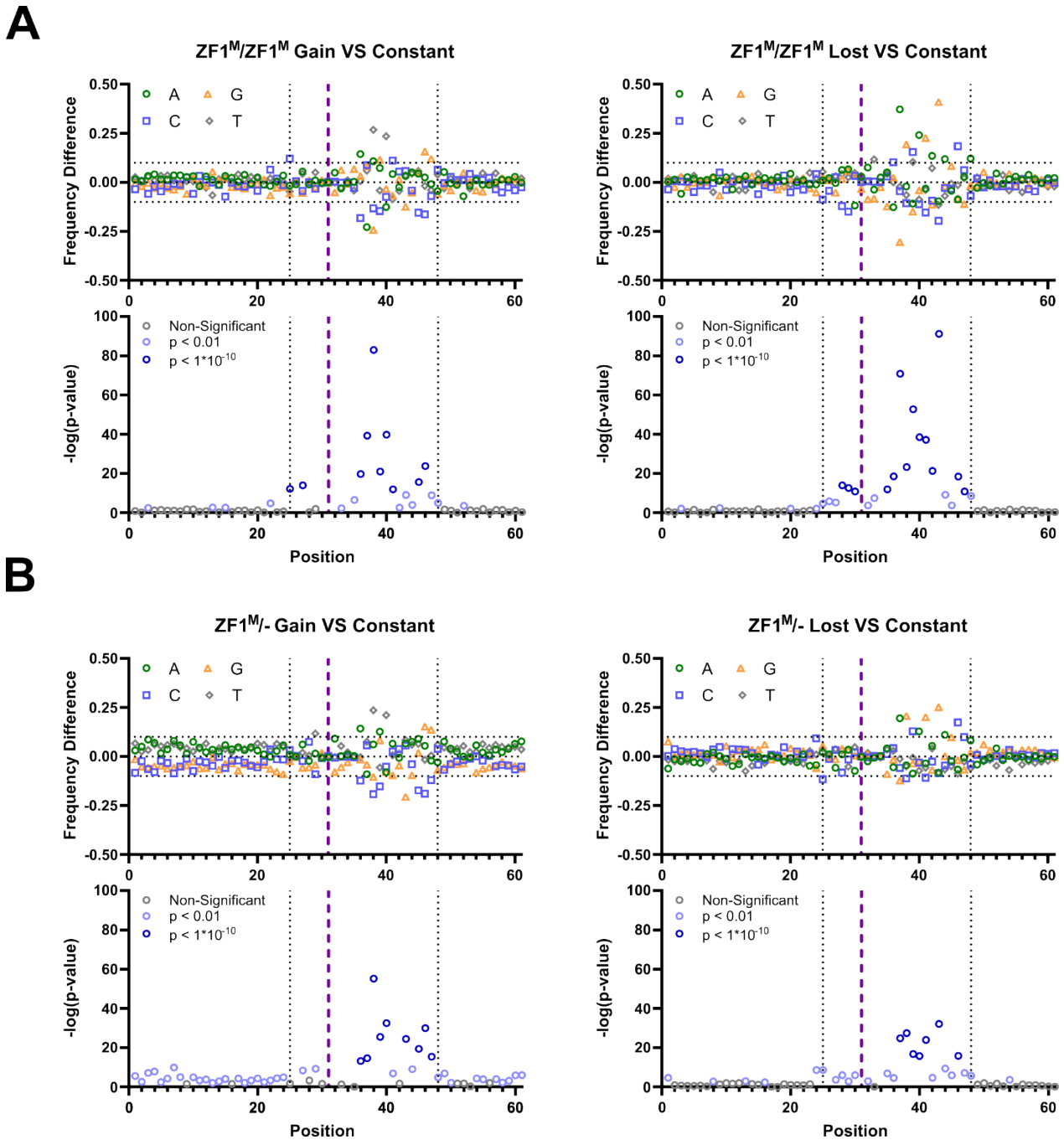
Supplementary Figure 2.1. (A) BlastX Results from single allele Sanger sequencing showing the presence of the H284N mutation and frameshift deletion in the respective alleles of ZF1M/ZF1M and ZF1M/- clones, respectively. Chromatograph results from one allele of the CTL clones and each

mutated CTCF alleles of ZF1M/- and ZF1M/ZF1M detailing the inserted mutation. (B) Western Blot of CTCF levels in the distinct CTCF ZF1M MCF10A clones. Actin used as loading controls. (C) Relative distribution of common CBS. Showing slight enrichment of altered CBS on distal intergenic elements and a slight enrichment of constant CBS on promoters. (D) Comparison between csaw called DB regions between ZF1M/ZF1M or ZF1M/- and our CTL MCF10A or CTCF WT MCF10A ChIP-Seq from Fritz et al. 2018, showing that DB regions are intrinsic to the mutant clones. (E) Comparison between csaw and MACS2/DiffBind to identify differentially binding regions between ZF1M/ZF1M or ZF1M/- and CTL MCF10A, showing that a majority of DB regions are called by both analysis methods.

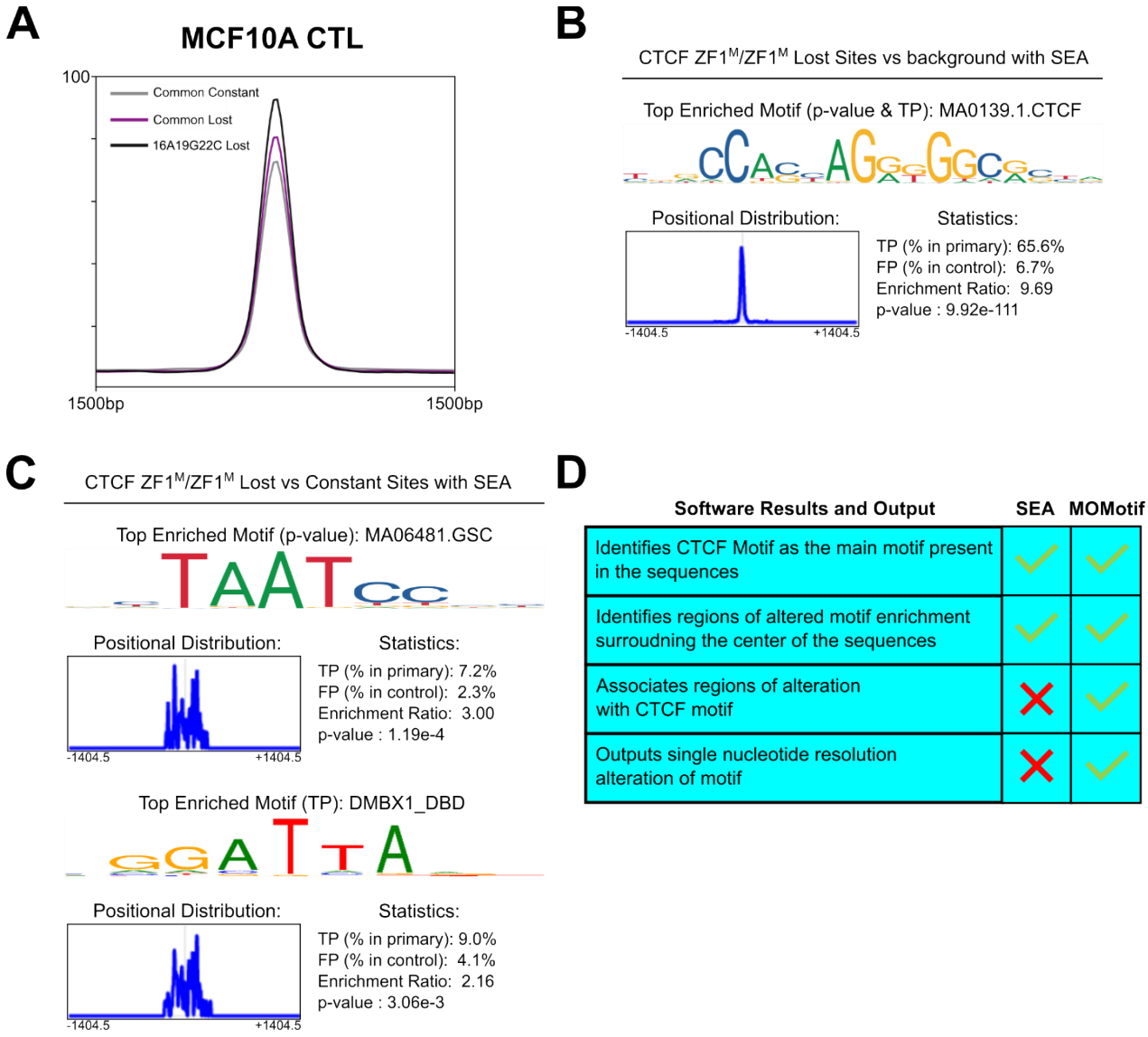
A



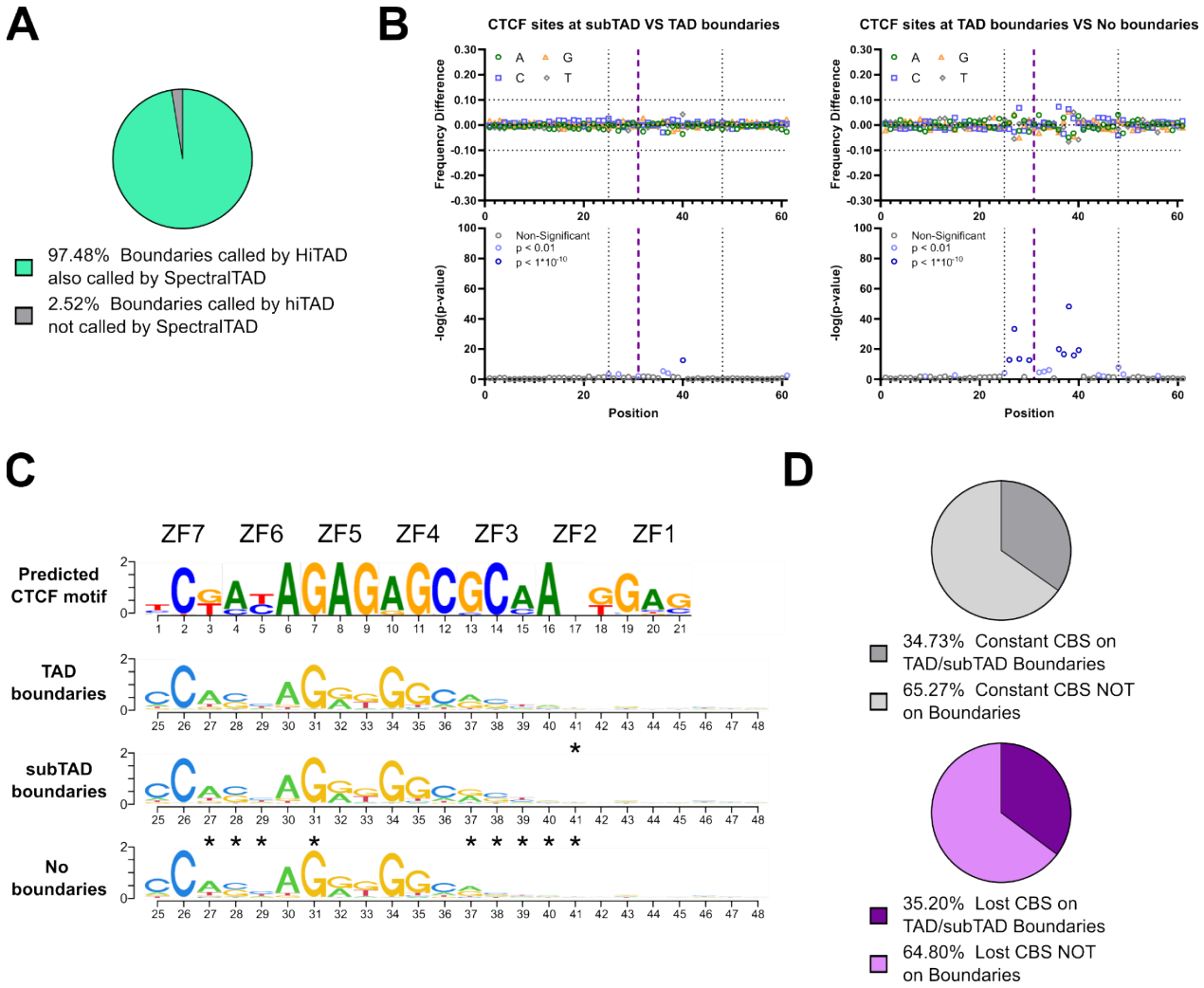
Supplementary Figure 2.2. (A) csaw flowcharts and specific settings used during this study.



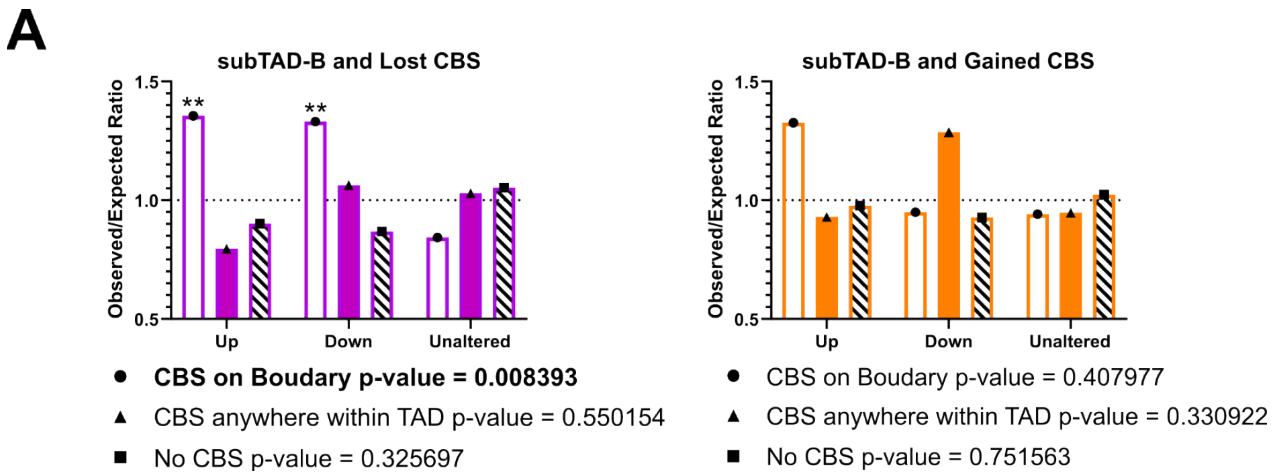
Supplementary Figure 2.3. (A) MoMotif analysis of base frequency difference and p-value of bases distribution difference around CTCF-Like motif in lost and gain CBS subsets compared to constant subset in CTCF ZF1M/ZF1M clones. (B) MoMotif analysis of base frequency difference and p-value of bases distribution difference around CTCF-Like motif in lost and gain CBS subsets compared to constant subset in CTCF ZF1M/- clones. The purple line represents the middle of the CTCF Motif. The dotted line represented the selected region.



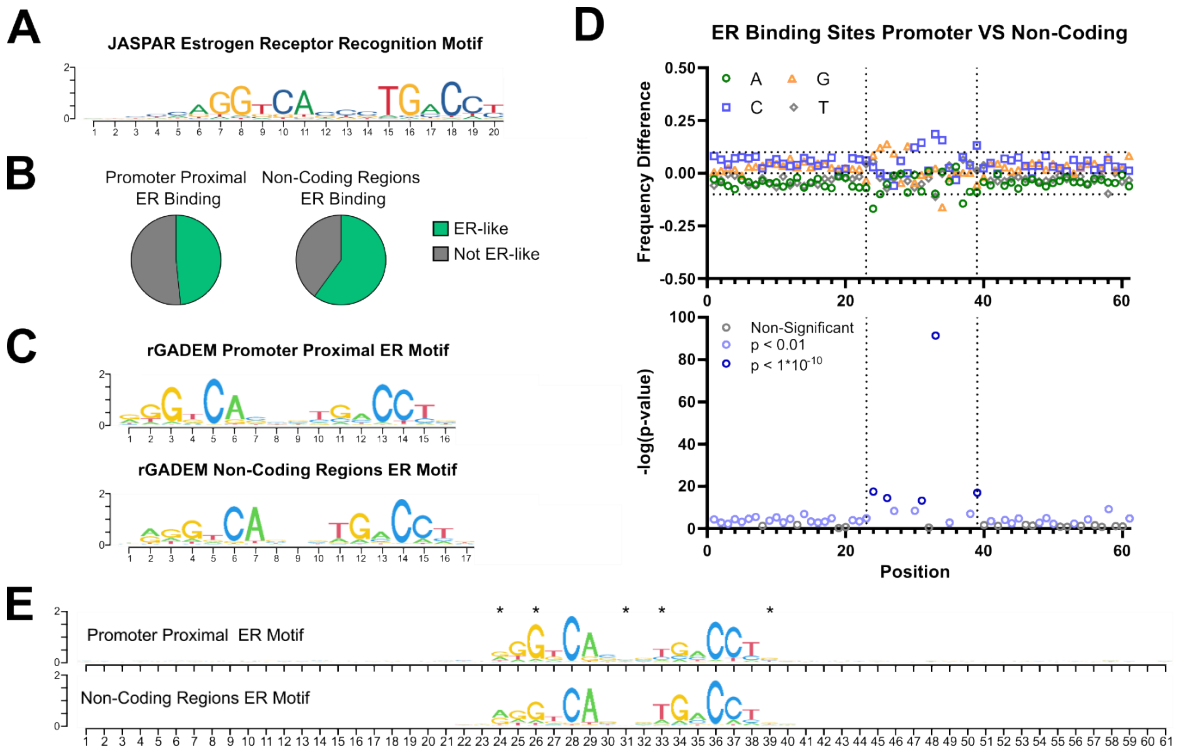
Supplementary Figure 2.4. (A) Profile plot of CTCF ChIP-Seq read density in CTL MCF10A at commonly constant, lost and lost with the full extended motif. Showing that sites harboring the full extended motif have higher affinity for WT CTCF. (B) MEME-Suite SEA analysis and output of CTCF ZF1M/ZF1M lost sites compared to background showing CTCF motif as the top hit for p-value and True Positive (TP) and centrally located on the sequences. (C) MEME-Suite SEA analysis and output of CTCF ZF1M/ZF1M lost sites compared to constant sites. Top enriched motifs by p-value and TP shows a less than 10% TP and are located adjacent to the middle of the sequences. (D) Summary of the functional differences between SEA and MoMotif.

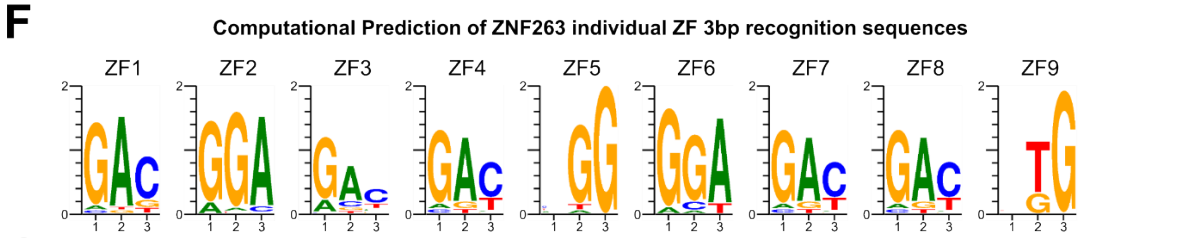


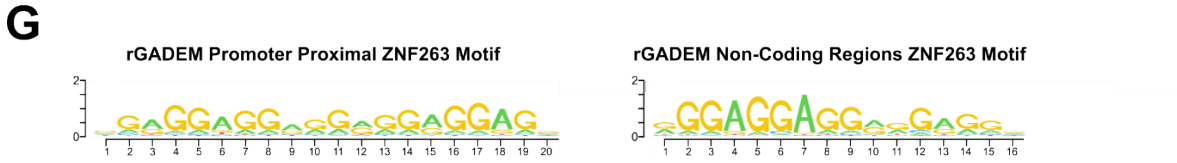
Supplementary Figure 2.5. (A) Pie chart comparing the reproducibility of called TAD and subTAD boundaries from HiTAD (used for this study) and SpectralTAD, an alternative hierarchical TAD caller. (B) MoMotif analysis of base frequency difference and p-value of bases distribution difference around CTCF-Like motif in CBS at subTAD boundaries compared to an equal size subset of CBS at TAD boundaries and the subset of CBS at TAD boundaries compared to an equal size subset of CBS located within domains, therefore not on boundaries (+/- ½ bin/5kb) (n = 4915). The purple line represents the middle of the CTCF Motif. The dotted line represented the selected region, which was kept the same as in Figure 3 to ease comparison between figures and because no significant changes were observed outside of this region. (C) MoMotif results depiction as the height of each nucleotide representing the Shannon Entropy of its occurrence frequency at each position in each group. Asterisk marks individual position with significantly altered base frequencies compared to TAD, highlighting the decreased enrichment of the called bases at these position in CBS not colocalizing with boundaries. (D) Pie charts of the frequency of CBS found on TAD/subTAD boundaries or not on boundaries in all CBS constant or lost between CTL and CTCF ZF1M mutated MCF10A. Showing no enrichment of lost CBS in these specific topological contexts.

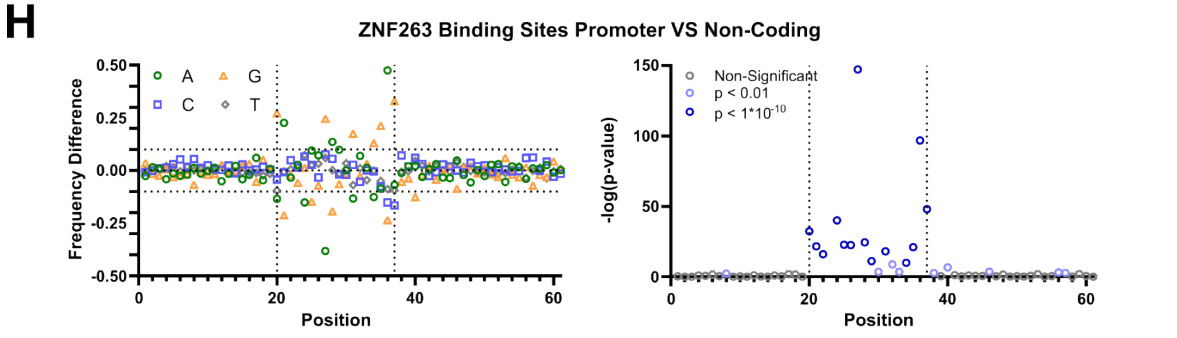


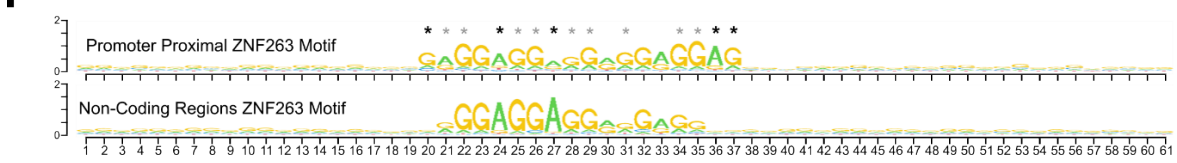
Supplementary Figure 2.6. (A) Enrichment of strongly up and downregulated genes for different distribution of subTAD, TSS and CBS. Showing that lost of CTCF near a gene at the boundaries of subTAD is significantly predictive of its up or downregulation. (p-values were generated from Chi-Square test on distribution of altered genes)


A
JASPAR Estrogen Receptor Recognition Motif
B
Promoter Proximal ER Binding
Non-Coding Regions ER Binding
ER-like
Not ER-like
C
rGADEM Promoter Proximal ER Motif
rGADEM Non-Coding Regions ER Motif
D
ER Binding Sites Promoter VS Non-Coding
A
G
C
T
Frequency Difference
Non-Significant
p < 0.01
p < 1*10-10
-log(p-value)
Position
E
Promoter Proximal ER Motif
Non-Coding Regions ER Motif


F
Computational Prediction of ZNF263 individual ZF 3bp recognition sequences
ZF1
ZF2
ZF3
ZF4
ZF5
ZF6
ZF7
ZF8
ZF9


G
rGADEM Promoter Proximal ZNF263 Motif
rGADEM Non-Coding Regions ZNF263 Motif


H
ZNF263 Binding Sites Promoter VS Non-Coding
A
G
C
T
Frequency Difference
Position
Non-Significant
p < 0.01
p < 1*10-10
-log(p-value)
Position


I
Promoter Proximal ZNF263 Motif
Non-Coding Regions ZNF263 Motif

Supplementary Figure 2.7. (A) Classical ER DNA binding (MA0112.2) motif from JASPAR. (B) Pie chart of the occurrence of the ER-like motif at promoter proximal or non-coding ligand- dependent ER binding sites. (C) ER-like motif identified by rGADEM and used as input for MoMotif analysis. (D) Single-nucleotide resolution base frequency difference and significance around the ER-like motif in promoter proximal versus non-coding sites. (E) MoMotif results depiction as the height of each nucleotide representing the Shannon Entropy of its occurrence frequency at each position for the regions analysed in D. Asterisks are placed above bases with significant difference between promoter proximal and non-coding ER binding sites. (F) Predicted 3bp sequences recognized by each ZF of ZNF263 by Persikov et al. 2014 and 2015. (G) ZNF263 motif identified by rGADEM and used as input for MoMotif analysis. (H) Single-nucleotide resolution base frequency difference and significance around the aligned ZNF263 GA rich motif in promoter proximal versus non-coding sites. (I) MoMotif results depiction as the height of each nucleotide representing the Shannon Entropy of its occurrence frequency at each position for the regions analysed in H. Dark Asterisks are placed above the five most significantly altered bases between promoter proximal and non-coding ER binding sites, showing an extension of the motif in promoter proximal binding sites and an enrichment intra-motif A in non-coding sites. Grey asterisks are placed above all significantly altered bases.

Supplementary Table 1: DNA donor to insert the H284N mutation coupled with the small guide RNAs targeting CTCF for the CRISPR-Cas9 experiment.

Name	Sequence
CTCF-H284N-Donor	ACATAGGTGTAAAGAAGACATTCCAGTGTGAGCTTTGCAGTTACACGTGTCCAC GGCGTTCAAATTTGGATCGTAACATGAAAAGCCACACTGATGAGAGACCACACA AGTGCCATCTCTGTGGCAGGGCATTCAGAACAGTCACCCTCC
CTCF-H284N-sgRNA F	CACCGCCACGGCGTTCAAATTTGGATCG
CTCF-H284N-sgRNA R	CGGTGCCGCAAGTTTAAACCTAGCCAAA

Supplementary Table 2: Antibodies

Antibodies	**Source**	**Catalog Number**
Rabbit Monoclonal anti- 4E-BP1	Cell Signaling	#9644
Rabbit Monoclonal anti-CDH1	Cell Signaling	#3195
Mouse Monoclonal anti-CTCF	BD	#612149
Mouse Monoclonal anti-GapDH	Origene	#TA802519
Rabbit Monoclonal anti- HA-Tag	Cell Signaling	#3724

Rabbit Polyclonal anti-phospho-4E-BP1 (Ser65)	Cell Signaling	#9451
Rabbit Polyclonal anti-PARP1	Cell Signaling	#9542
Mouse Monoclonal anti-phospho-p70S6K1 (Thr389)	Cell Signaling	#9206
Rabbit Monoclonal anti-p70S6K	Cell Signaling	#2708
Mouse Monoclonal anti- Snail	Cell Signaling	#3895
Mouse Monoclonal anti-α-Tubulin	Cell Signaling	#3873
Mouse Monoclonal anti-β-Actin	SigmaAldrich	#A5316
Rabbit Polyclonal anti- CTCF	EMD Millipore	#07-729
Rabbit Polyclonal anti-acetyl-Histone H3 (Lys27)	EMD Millipore	#07-360
Rabbit Polyclonal anti-trimethyl-Histone H3 (Lys4)	EMD Millipore	#07-473
Rabbit Polyclonal anti-trimethyl-Histone H3 (Lys27)	EMD Millipore	#07-449
Rabbit Polyclonal anti-trimethyl Histone H3 (Lys9)	Diagenode	#C15410056
Goat Polyclonal anti- Rabbit-IgG	SeraCare	#5220-0458
Goat Polyclonal anti- Mouse-IgG	SeraCare	#5450-0011
Rabbit Polyclonal anti-phospho-S6 Ribosomal Protein (Ser240/244)	Cell Signaling	#2215
Goat Polyclonal anti- Rabbit-IgG with Alexa 488 fluorophore	Invitrogen	#A32731

Supplementary Table 3: RT-qPCR Primers

Name	Sequence
GapDH F	CAGCCTCAAGATCATCAGCA
GapDH R	TGTGGTCATGAGTCCTTCCA
RPL4 F	GCTCTGGCCAGGGTGCTTTTG
RPL4 R	ATGGCGTATCGTTTTTGGGTTGT
RPLPO F	TTAAACCCTGCGTGGCAATCC
RPLPO R	CCACATTCCCCCGGATATGA
SOX9 F	AGCAAGACGCTGGGCAAG
SOX9 R	GTAATCCGGGTGGTCCTTCT
JAM3 F	CCAGGATCGAGTGGAAGAAA
JAM3 R	CAGGGATGTCTTCCCCAGT
ERBB3 F	AAAGGACCAGAGCTTCAAGA
ERBB3 R	CCAGCATCATGAAAATCACT
FGFR1 F	CCTCTTCAGAGGAGAAAGAAACA
FGFR1 R	TCTTTTCTGGGGATGTCCAA
LAMA1 F	GCAAAGGCAGAACAAAGGTC
LAMA1 R	GGCCGTCGACAGTTATGAAG
SNAI1 F	ACCTGTTTCCCGGGCAATTT
SNAI1 R	CTGGGAGACACATCGGTCAG

Supplementary Table 4: ChIP-qPCR Primer

Name	Sequence
JAM +0.5kb F	CAGTGCTGTGCTCTCCAGAA
JAM +0.5kb R	AGGGCTGTGACCAAGCAG
JAM -0.5kb F	GAAGGCGATAATGCTTCCAA
JAM -0.5kb R	CAGGTCGGAGAAGGAACACT
LAMA1 +0.25kb F	AAAGCCTAAGCCTGCAAAGA
LAMA1 +0.25kb R	ATCCTGATCCACCTCGGAGT
LAMA1 +0.5kb F	CTTTAACCTCCTCGGGCTTT
LAMA1 +0.5kb R	CAGCACTGCTCGCGTAGAT
LAMA1 +1kb F	TTTGTGACTGCCTAGCCAAC
LAMA1 +1kb R	TTTTGGGGGACAACCCTAGT
SNAI1 -0.5kb F	CGTAGACTGTCTGGGCCAAT
SNAI1 -0.5kb R	AGGCTTCCATCCTCCAACTT
SNAI1 TSS F	CCCTCCATTCTCATCAGCTC
SNAI1 TSS R	CCGATAAACTCCCTTGGACA
SNAI1 +0.5kb F	GCACACCTGACATGCTGACT
SNAI1 +0.5kb R	CCCTGACCATCACAGGCTAT
SNAI1 -652 F	CGGGAGAGGCTCTGAGTGTT
SNAI1 -652 R	CTAGCCAAGAGCACCCGTTC
SNAI1 +1kb F	GATGAGGACAGTGGGAAAGG
SNAI1 +1kb R	GCCTCCAAGGAAGAGACTGA
chr19:49,345kb F	AGTGGTCCTCACCCTCACAC
chr19:49,345kb R	GATGGCAGTAGCACACAGGA

Abbreviations:

- 3D: Three Dimensional
- Abs: Absolute
- AML:
- ATAC:
- B-ALL: B-Cell Acute Lymphoblastic Leukemia
- BER: Base-Excision Repair
- CBS: CTCF Binding Sites
- ChIP: Chromatin Immunoprecipitation
- CTCF: CCCTC-Binding Factor
- DS-AMKL: Down Syndrome related Acute MegaKaryoblastic Leukemia
- ECM: Extracellular Matrix
- EMT: Epithelial to Mesenchyme Transition
- E-P: Enhancer-Promoter Interactions
- ER: Estrogen Receptor
- FC: Fold Change
- FDR: False Discovery Rate
- GSEA: Gene Set Enrichment Analysis
- GTF: General Transcription Factor
- H3K27ac: Histone 3 Lysine 27 Acetylation
- H3K4me3: Histone 3 Lysine 4 Trimethylation
- HAT: Histone Acetyl-Transferase
- Hi-C: High-throughput Genome-Wide Chromosome Conformation Capture
- HR: Homologous Recombination
- ICR: Imprinting Control Region
- LOH: Loss of Heterozygosity
- MDS:
- mRNA: messenger RNA
- ncRNA: non-coding RNA
- NES: Normalized Enrichment Score
- PBM: Protein Binding Microarrays
- PDX: Patient Derived Xenograft
- PI3K: Phosphoinositide 3-Kinase
- PIC: Preinitiation Complex
- PIP2: phosphatidylinositol-4,5-bisphosphate
- PIP3: phosphatidylinositol-3,4,5-triphosphate
- Pol II: RNA Polymerase II
- PR: Progesterone Receptor
- PWN: Position Weight Matrix
- qPCR: quantitative Polymerase Chain Reaction
- RTK: Receptor Tyrosine Kinase
- RT-qPCR: Reverse-Transcriptase quantitative Polymerase Chain Reaction
- SEM: Standard Error of the Mean
- Seq: Sequencing
- Ser5-P: Serine 5 phosphorylation

- sgRNA: small guide RNA
- shRNA: short hairpin RNA
- TAD : Topologically Associated Domains
- TCGA: The Cancer Genome Atlas
- TF: Transcription Factor
- TNBC: Triple Negative Breast Cancer
- TP : True Positive
- WT : Wild Type
- ZF: Zinc Finger
- ZF1M: Zinc Finger 1 Mutation

List of Figures:

Figure 2.3.MoMotif identifies a unique motif enriched for CBS compromised upon mutation of ZF1.

Figure 2.4. Extended Motif of CTCF is associated to an altered binding conformation Figure 2.5. CTCF ZF1M drives oncogenic transcription profiles.

Figure 2.6. Loss of CTCF binding within TADs drives oncogenic transcription

Discussion:

Figure D1: Example of a shift in TAD interaction permissive to transcriptional changes.

Figure D2: Example of a shift in TAD interaction not permissive to transcriptional changes since no regulatory element is encompassed within the shift.

Figure D3 : Example of a shift in TAD interaction not permissive to transcriptional changes since the potential interactors are insulated within subTAD

Figure D4 Example of a fusion of two TADs not permissive to transcriptional changes since the potential interactors are insulated within subTAD.

Figure D5: Depiction of how epigenetic plasticity facilitates the switch between 2 cell states.