INVESTIGATING THE ROLE OF THE METHYL-CpG BINDING PROTEIN ZBTB33 IN CANCER

by

Tommy W. Terooatea

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Chemistry

The University of Utah

May 2017

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Tommy W. Terooatea**

has been approved by the following supervisory committee members:

Bethany A. Buck-Koehntop	, Chair	**07/29/16** Date Approved
Peter Francis Flynn	, Member	**07/29/16** Date Approved
Peter J. Stang	, Member	**07/29/16** Date Approved
Matthew T. Kieber-Emmons	, Member	**07/29/16** Date Approved
Markus Babst	, Member	**07/29/16** Date Approved

and by **Henry S. White** , Chair/Dean of

the Department/College/School of **Chemistry**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

ZBTB33 is the first identified zinc finger containing methyl-CpG binding protein (MBP), the cellular function of which is still unclear in both normal and disease state. Gene regulation by ZBTB33 is mediated through recognition of both methylated as well as sequence specific nonmethylated deoxyribonucleic acid (DNA) sites. DNA recognition is followed by recruitment of corepressors resulting in transcription changes. While evidence suggests that transcriptional regulation by ZBTB33 plays a role in cancer, an extensive analysis of the gene targets and signaling pathways regulated by this protein have yet to be investigated. The goal of this research is to begin to determine how ZBTB33 directs disease relevant transcription. To investigate this, a combination of techniques to characterize *in cell* global gene occupation, DNA methylation status, and transcriptome changes were utilized. Further, the effect of bimodal DNA recognition exhibited by ZBTB33 on transcriptional regulation and biological pathways in cancer were investigated.

Powerful techniques have emerged for investigating genome-wide protein-DNA interactions. One such technique incorporates chromatin immunoprecipitation followed by exonuclease digestion and next generation sequencing (ChIP-exo). While this technique maps global gene occupations of transcription factors at high resolution, it also harbors technique specific issues, including the significant generation of artifact peaks, limiting its application. Thus, a method for generating a background control that

substantially reduces these artifacts and improves the confidence level in peak identification for ChIP-exo data sets was developed. Application of our improved ChIP-exo method to ZBTB33 revealed additional issues with ChIP-based techniques coupled with next generation sequencing that can affect a number of systems. Thus, efforts to identify and resolve these issues are discussed.

Along with ChIP-based methods, whole genome shotgun bisulfite sequencing (WGSBS) and RNA sequencing (RNA-seq) are other powerful methodologies informing on the global DNA methylation status and transcriptome of a cell, respectively. By combining these techniques with ZBTB33 depletion/overexpression studies and phenotypic assays, we found mechanistic details for how ZBTB33 mediates cell-specific cell cycle regulation in a cancerous cell line. Specifically, we identified a mechanism by which ZBTB33 mediates the cyclin D1/cyclin E1/RB1/E2F pathway, controlling passage through the G1-restriction point and accelerating cancer cell proliferation.

TABLE OF CONTENTS

LIST OF FIGURES

LISTS OF TABLES

ABBREVIATIONS

ZBTB33: Zinc finger and BTB domain containing 33

KBS: Kaiso Binding Site

RB1: Retinoblastoma 1

siRNA: small interfering RNA

Scr: scrambled siRNA

IPA: Ingenuity pathway analysis

GSEA: Gene set enrichments analysis

TSS: Transcription Start Site

WGSBS: Whole genome shotgun bisulfite sequencing

CGI: CpG Island

MBD: Methyl-CpG-binding domain

UHRF1: Ubiquitin-like, containing PHD and RING finger domains

MBP: Methyl-CpG binding protein

RNA-seq: RNA sequencing

ChIP-seq: Chromatin immunoprecipitation coupled with high-throughput sequencing

ChIP-exo: Chromatin immunoprecipitation exonuclease

5-mC: 5-methylcytosine

ssRNA: single-stranded RNA

dsDNA: double-stranded DNA

DNA: deoxyribonucleic acid

PBS: Phosphate buffered saline

EDTA: ethylenediamminetetraacetic acid

ACKNOWLEDGMENTS

First, I would like to thank Professor Bethany Buck-Koehntop for offering me the opportunity to carry out this interesting Ph.D. project under her supervision. It has been challenging yet full of rewarding experiences that helped me grow to become a better scientist and a better person. I sincerely appreciate her patience, guidance, and full support as always.

Second, I would like to thank all past and current research group members, Professors on my committee and the University of Utah Genomic and Bioinformatics core facility directors for their kind help and mentorship. More specifically, I would like to express my gratitude to the following individuals: Dr. David Nix, Dr. Brian Dailey, Brett Milash, and Tim Mosbruger for their patience in teaching everything I know about next-generation sequencing and bioinformatics; Dr. Amir Pozner for his guidance with cell biology techniques and valuable discussions; Dr. Marcus Babst for his expertise in all the IP techniques we used in the lab; Dr. Peter Flynn for improving the quality of this dissertation with his incredible patience and critical comments; Dr. Kenneth Woycechowsky for his guidance and excellent advice; Dr. Michael Bartl and his previous group for their kindness and great scientific examples; for all other colleagues from various institutions that I came to know through conferences and seminars for their contribution to my success with their kind words and motivations.

Third, I would like to thank my wife for putting up with me during these long

years, for her patience, love and constant encouragement, I would never have done this without her; for my mother and brothers who always believed in me; for the Polynesian community in Salt Lake City notably, the Raihauti, Munanui, Temauri, Jones, Hosea, Mahuru and the Tupua families for helping me remain true to my culture through all the luaus, traditional chants and dances.

CHAPTER 1

INTRODUCTION

Deoxyribonucleic Acid (DNA) methylation in normal cell function and disease

It has been known since 1940 that DNA CpG sites in vertebrates can be methylated at the cytosine base, yet it took over 40 years[1] to realize the critical role of this epigenetic mark in gene regulation.[2] Today, it is well established that DNA methylation in the context of CpG dinucleotides is present in all mammals and is fundamental for controlling gene expression, cell development, cell identity, and differentiation. Tight regulation of DNA methylation patterns and propagation to newly formed cells must be achieved, as any alterations are often deleterious leading to the onset of various diseases.[3] In cells, cytosine methylation occurs through the transfer of a methyl group from *S*-adenosyl methionine (SAM) to the five position of the cytosine base by DNA methyltransferases (DNMTs) (Figure 1.1).[4] DNMT1 functions by ensuring that methylation patterns are propagated to nascent oligonucleotides formed during DNA replication,[4] while DNMT3a and DNMT3b, known as the de novo DNMTs, are responsible for establishing new methylation patterns on unmodified DNA during embryogenesis and differentiation.[5]

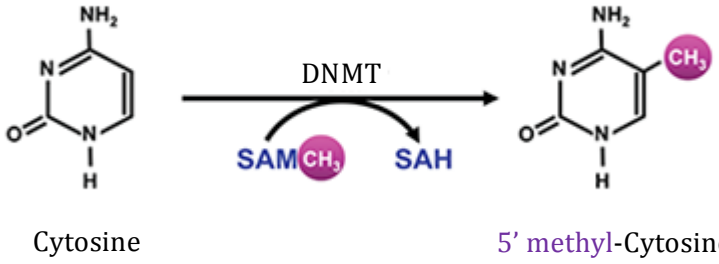


Figure 1.1. Representation of enzymatic DNA methylation by a DNMT. DNMT catalyzes the methylation of a cytosine base that precedes a guanosine at carbon 5 by transferring a methyl group from SAM.

In mammals, many methylated DNA sites are found in GC-rich genomic regions localized within promoters and gene bodies, referred to as CpG islands (CGIs). In addition, high levels of DNA methylation are found at highly repetitive sequences such as centromeric regions. In humans, more than 50% of all CpGs are coincident with heterochromatin structure while euchromatin regions contain mainly unmethylated CGIs.[6] Specific global genomic methylation patterns give rise to differential cellular phenotypes and are essential for regulating gene expression and maintaining genome stability. Open chromatin at unmethylated promoter regions can be targeted by transcription factors leading to gene activation, while compacted methylated promoters reduce transcriptional accessibility leading to gene silencing.[6,7]

Three proposed models explain the inhibitory effect of promoter DNA methylation on gene expression (Figure 1.2). The first involves the inability of transcription factors to bind their cognate sites due to the presence of CpG methylation. In the second model, specialized proteins with high-affinity for recognizing methyl-CpG sites, termed methyl-CpG binding proteins (MBPs), hinder promoter binding of transcriptional activators. Finally, methylated promoter binding by MBPs recruits repressive chromatin modifiers such as histone deacetylases (HDACs) and histone methyltransfereases (HMTs) to the target site leading to heterochromatin formation and gene repression.[8] Interestingly, recent evidence suggests that DNA methylation within gene bodies is associated with a higher transcription level in normally proliferating cells even when occupied by MBPs.[9-12] However, in slowly proliferating cells such as brain cells, gene body methylation is associated with a decrease in gene expression.[13,14] Additionally, in the murine frontal cortex, it has been shown that methylation of

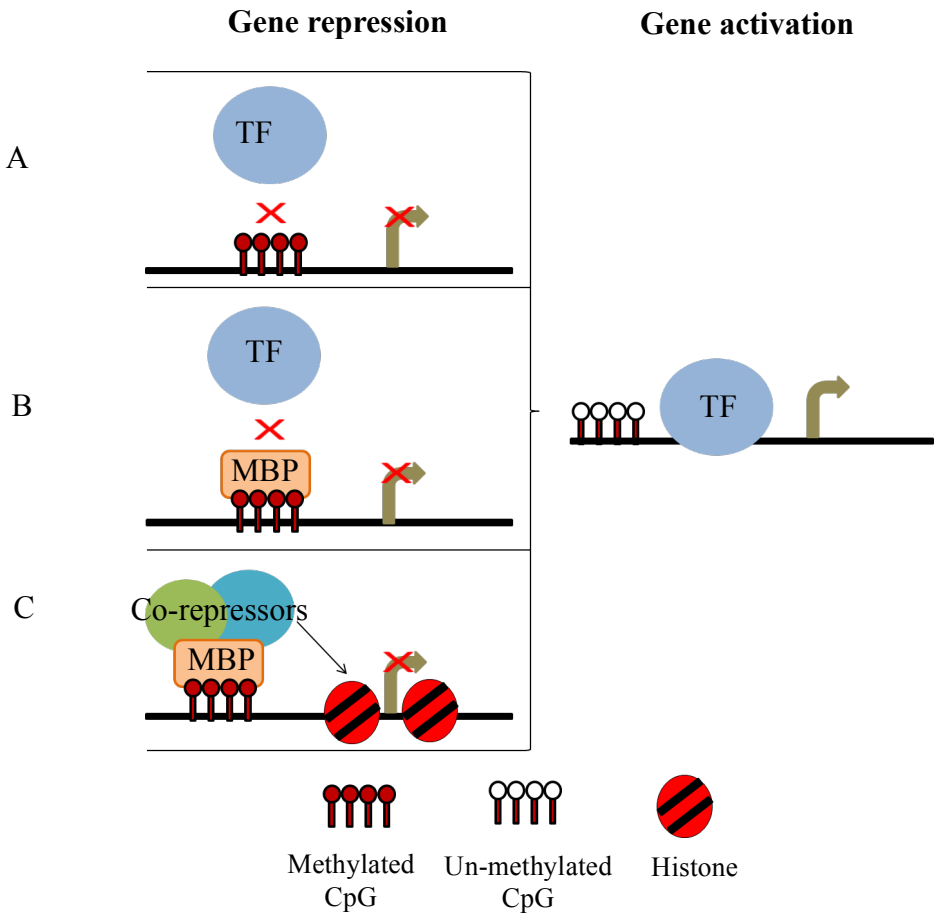


Figure 1.2. Mechanisms for gene repression induced via DNA methylation at gene promoters. (A) DNA methylation at gene promoters prohibits transcription factor (TF) recruitment, leading to gene inactivation. (B) Methyl-CpG binding proteins (MBPs) bind to methylated DNA at methylated gene promoters and prevent TF recruitment, resulting in gene inactivation. (C) MBPs bind to methylated DNA at methylated gene promoters, recruit co-repressor complexes, which induce chromatin compaction and gene inactivation.

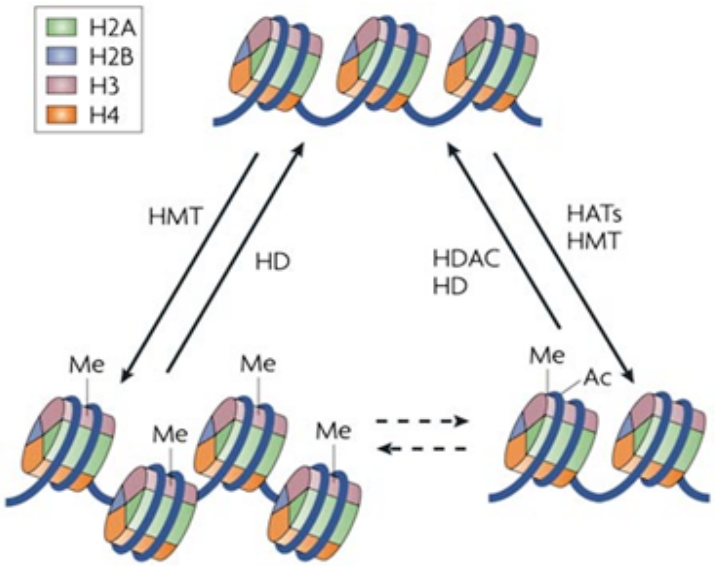
nonCpG sites in gene bodies is negatively correlated with gene expression.[14] Thus, it is increasingly evident that the mechanisms by which DNA methylation regulates gene expression are complex.

Histone modifications

Acetylation and methylation of lysines within histone polypeptide tails are prevalent posttranslational epigenetic associated modifications.[15] These modifications are critical for controlling chromatin structure, which can in turn affect various cellular processes such as DNA repair, replication, transcription, and recombination. In contrast to DNA methylation, histone modifications are more dynamic.[16,17]

Histone acetyltransferases (HATs) acetylate histones by transferring an acetyl group from acetyl coenzyme A to specific lysine residues.[15] The reverse process is mediated by histone deacetylaces (HDACs), which enzymatically remove the acetyl groups of histone lysine residues[18] (Figure 1.3). Lysine acetylation results in increased local negative charges, reducing the affinity between histones and DNA and inducing the open chromatin state. Thus, histone acetylation, particularly at H3K9ac, H4K5ac, H4K8ac and H2AK5ac, is considered a reliable epigenetic indicator for regions of transcriptional activation[19] (Figure 1.3). Global changes to histone acetylation patterns in humans has been associated with tumorigenesis and cancer progression.[20]

Histone methyltransferases (HMTs) catalyze the transfer of a methyl group from SAM onto lysines within the polypeptide tails of histones H3 and H4 and in contrast to histone acetylation, which primarily leads to gene activation, histone methylation can lead to either gene activation or silencing.[21] Thus, the degree of methylation on a given



Examples of histone modifications

Heterochromatin	**Euchromatin**	
H3K9me2	H3K4me2	H3K9ac
H3K9me3	H3K4me3	H4K5ac
H4K20me3	H3K36me1	H4K8ac
H3K27me3	H3K79me1	H2AK5ac

Figure 1.3. Histone tail posttranslational modifications. Illustration of histone tail methylation and acetylation patterns on chromatin structure. Histones can be reversibly acetylated and methylated at particular residues. Specific methylation and acetylation patterns induce the formation of either heterochromatin or euchromatin structures ultimately leading to gene silencing or activation, respectively (adapted from (22)).

lysine residue determines the downstream transcriptional outcome. Generally, enriched histone methylation at lysine residues can acquire between 1-3 methyl groups, H3K4me2, H3K4me3, H3K36me1, or H3K79me1, are indicative of gene activation, while enriched histone methylation at H3K9me2, H3K9me3, H4K20me3, or H3K27me3 correlates with gene silencing.[23]

Methyl-CpG binding proteins (MBPs) – interpreters of DNA methylation

Reading and interpreting the DNA methylation status of genes has been the focus of many important studies. Three main families of transcription factor known as the methyl-CpG binding proteins (MBPs) have been identified to bind methylated DNA in mammalian cells (Figure 1.4).

In 1989, Dr. Adrian Bird identified and named the first MBP, MeCP2, after observing it binding to methylated CpG DNA sites.[24] This early result prompted further characterization of the MeCP2 protein domain responsible for methylated DNA recognition, which was identified to encompass amino acids 78 to 162 and was termed the methyl-CpG binding domain (MBD)[25-27] (Figure 1.4A). Once identified, homology searching with the MBD of MeCP2 led to the identification of multiple other MBD family proteins including MBD1 and MBD2 (Figure 1.4A).[28-30] In addition to the MBD family of MBPs, two other main families have been identified including the structurally unrelated SET and RING associated domain (SRA) domain–containing proteins (UHRF1 and UHRF2) and the zinc finger (ZF) family (ZBTB33, ZBTB4, and ZBTB38) (Figure 1.4B and C).

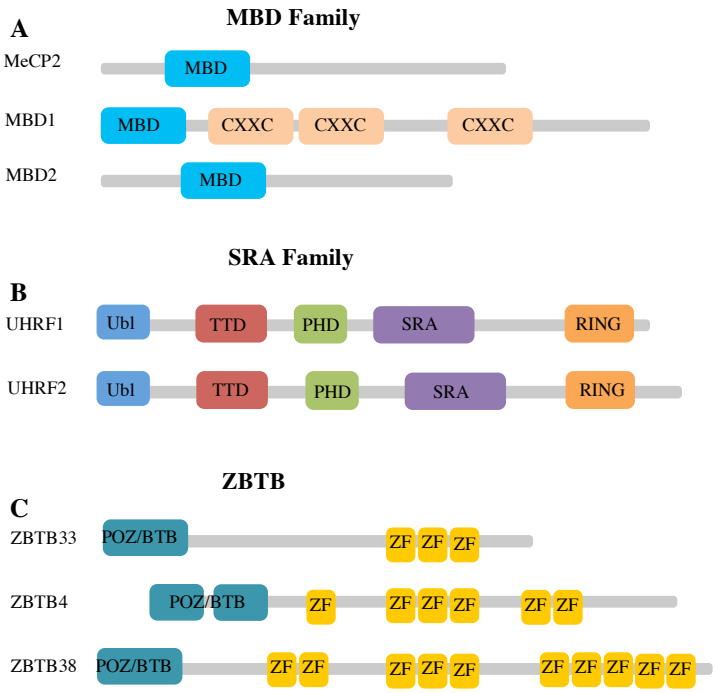


Figure 1.4. The three families of mammalian MBPs. (A) The MBD family: MeCP2, MBD1, and MBD2. MBD, methyl-CpG binding domain; CXXC ZF, unmethylated-CpG binding zinc finger. (B) SRA domain proteins: UHRF1 (Ubiquitin-like with PHD and RING finger domains 1) and UHRF2. Ubl, ubiquitin-like domain; TTD, tandem tudor domain; PHD, plant homeodomain finger domain; SRA, SET and RING associated domain; RING, Really Interesting New Gene finger domain. (C) ZBTB33 (Zinc finger and BTB domain containing 33) and ZBTB33-like proteins: ZBTB4 and ZBTB38. BTB/POZ (Broad complex, Tramtrack, Bric a brac/Pox virus, and Zinc finger) domain; ZF, zinc fingers.

The zinc finger family of MBPs

The Bric-a-brac, Tramtrack, and Broad complex/Poxvirus and Zinc finger (BTB/POZ) domain proteins constitute a large family of transcription factors that share a well-known protein-protein interaction domain.[31] ZBTB33 (also known as Kaiso) represents the founding member of the ZBTB MBP family. ZBTB33 was originally identified through its interaction with the Armadillo repeat containing protein p120-catenin ($p120^{ctn}$).[32] ZBTB33 contains an N-terminal BTB/POZ protein-protein interaction domain and three C-terminally localized Kruppel-like Cys_2His_2 zinc fingers (ZFs) responsible for DNA binding (Figure 1.4C). Homology searching with the Cys_2His_2 ZFs of ZBTB33 identified two other family members, ZBTB4 and ZBTB38, that have been confirmed to also be capable of binding methylated DNA (Figure 1.4C).[33] In addition to methylated DNA recognition, ZBTB33 is unique among MBPs in that it additionally recognizes a sequence specific nonmethylated DNA target termed the Kaiso Binding Site (KBS; TCCTGCNA).[34]

The role of ZBTB33 in cancer

A mounting number of studies indicate various roles of ZBTB33 transcriptional activities in cancer progression. Significantly increased survival rate was shown in ZBTB33 knockout mice prone to develop intestinal cancer, indicating an essential role for ZBTB33 in disease progression.[35] Further, in a colon cancer cell line, ZBTB33 was observed to induce cell cycle attenuation through methyl-dependent repression of CDKN2A.[36] Accordingly, a correlation between the increase of ZBTB33 cytoplasmic-to-nuclear shift and tumorigenesis was observed in both human prostate cancer (PCa) and

human breast cancer tissues.[37-39] Together, these results imply that ZBTB33 modulates various cancer promoting biological pathways specific to different cancer type. However, currently there are no comprehensive mechanistic studies that describe how ZBTB33 promotes and progresses the cancer state.

Conclusion

While sufficient evidence suggests that the ZBTB33 protein plays a role in cancer, an extensive analysis of the gene targets and signaling pathway regulated by this protein has yet to be extensively investigated. Thus, we attempted to establish the first complete list of genes targeted by ZBTB33, the methylation status of these genes, and the transcriptional outcomes of ZBTB33 occupation at these gene sites in cancer cell lines. This data provides initial important insights into how the bimodal DNA recognition by ZBTB33 mediates disease relevant transcription.

With recent advancements in applied genomics and high-throughput screening methods, new powerful techniques have been developed for investigating global genomic protein-DNA interactions. One such technique couples chromatin immunoprecipitation with exonuclease digestion and next generation sequencing (ChIP-exo).[40] While this technique is capable of mapping global DNA binding sites at single base pair resolution it also harbors numerous false positives, which reduces the ability to identify real binding targets for low occupancy transcription factors or for proteins in which the consensus-binding site is not known.

Thus, we designed a complementary method to generate a ChIP-exo specific input that substantially reduces the number of false positives and drastically improves the

confidence level of real binding targets in these data sets.[41] We show that this method allows for more accurate de novo motif discovery searching, which is important for high confidence identification of unknown transcription factor genomic binding sites. We tested this method on a well-established protein CTCF in HeLa cells and demonstrated that its binding motif can be found unbiasedly throughout the genome with high confidence.

Application of our improved ChIP-exo method for ZBTB33 revealed additional issues with ChIP-based techniques coupled with next-generation sequencing. In ChIP-exo, we found that the lack of idealized lysine or cysteine cross-linking with guanosine bases limits the chances for stable protein DNA interactions between ZBTB33 and its cognate DNA sequence. Therefore, we suggest that the unsuccessful ZBTB33 ChIP-exo attempts were due to premature reverse cross-linking as a result of insufficient protein:DNA cross-linking strength and prolonged high temperature exposure during the ChIP-exo procedure.

While ChIP-seq does not include high temperature steps that would compromise the protein-DNA cross-links, it also does not identify ZBTB33-specific genomic occupation sites. While the KBS and mCpG ZBTB33 targets can be detected through ChIP-PCR,[42-45] ChIP-seq analysis appears to mainly enrich DNA motifs for other well-known transcription factors such as CTCF.[46] Thus, we hypothesize that ZBTB33 ChIP-seq may only report on indirect protein:protein:DNA interactions. Surprisingly, we also observed that ChIP-qPCR analysis of the ChIP-seq samples after library preparation showed no genomic occupation of the sites observed prior to preparation of the library. This finding is significant and suggests that the inability to identify ZBTB33 genomic

occupation sites may also result from a loss of this information during one or more of the library preparation steps.

Along with ChIP-based high-throughput sequencing methods, whole genome shotgun bisulfite sequencing and RNA sequencing (RNA-seq) are other powerful methodologies that inform on the global DNA methylation status and global transcriptome of a given cell, respectively. Combined, these techniques can be used to precisely determine the methylation status of genes and inform about transcriptional outcomes at a given gene target. By combining these methods we found mechanistic details for how ZBTB33 mediates cell-specific cell cycle regulation in a cancerous cell line. Specifically, it was determined that cell proliferation in HeLa cells is induced by direct activation of the cyclin D1 and cyclin E1 genes by ZBTB33. Activation of these genes leads to RB1 hyper-phosphorylation and increased E2F transcriptional activity, which facilitates G1- to S-phase transition. On the other hand in HEK293 cells ZBTB33 indirect deactivation of cyclin E results in both decreased RB1 phosphorylation and E2F activity thus decelerating G1 transition. Therefore, we identified a novel mechanism by which ZBTB33 accelerates cancer cell proliferation through the cyclin D1/cyclin E1/RB1/E2F pathway, controlling passage through the G1-restriction point.

References

(1) Fuks, F. DNA Methylation and Histone Modifications: Teaming Up to Silence Genes. *Curr. Opin. Genet. Dev.* **2005**, *15*, 490-495.

(2) Grippo, P.; Iaccarino, M.; Parisi, E.; Scarano, E. Methylation of DNA in developing Sea Urchin Embryos. *J. Mol. Biol.* **1968**, *36*, 195-208.

(3) Jones, P. A.; Takai, D. The Role of DNA Methylation in Mammalian Epigenetics. *Science.* **2001**, *293*, 1068-1070.

(4) McGhee, J. D.; Ginder, G. D. Specific DNA Methylation Sites in the Vicinity of the Chicken Beta-Globin Genes. *Nature* **1979**, *280*, 419-420.

(5) Riggs, A. D. X Inactivation, Differentiation, and DNA Methylation. *Cytogenet. Cell Genet.* **1975**, *14*, 9-25.

(6) Blackledge, N. P.; Klose, R. J.; CpG island chromatin: a platform for gene regulation. *Epigenetics,* **2011,** *6*, 147–152 .

(7) Jaenisch, R.; Bird, A. Epigenetic Regulation of Gene Expression: How the Genome Integrates Intrinsic and Environmental Signals. *Nat. Genet.* **2003**, *33*, 245-254.

(8) Curradi, M.; Izzo, A.; Badaracco, G.; Landsberger, N. Molecular Mechanisms of Gene Silencing Mediated by DNA Methylation. *Mol. Cell Biol*. **2002**, *22*, 3157–317.

(9) Ball, M. P.; Li, J. B.; Gao, Y.; Lee, J. H.; LeProust, E. M.; Park, I. H.; Xie, B.; Daley, G. Q.; Church, G. M. Targeted and Genome-Scale Strategies Reveal Gene-Body Methylation Signatures in Human Cells. *Nat. Biotechnol.* **2009**, *27*, 361-368.

(10) Maunakea AK.; Nagarajan, R. P.; Bilenky, M.; Ballinger, T. J.; D'Souza, C.; Fouse, S. D.; Johnson, B. E.; Hong, C.; Nielsen, C.; Zhao, Y.; Turecki, G.; Delaney, A.; Varhol, R.; Thiessen, N.; Shchors, K.; Heine, V. M.; Rowitch, D. H.; Xing, X.; Fiore, C.; Schillebeeckx, M.; Jones, S. J.; Haussler, D.; Marra, M. A.; Hirst, M.; Wang, T.; Costello, J. F. Conserved Role of Intragenic DNA Methylaiton in Regulating Alternative Promoters. *Nature* **2010**, *466,* 253-257.

(11) Nguyen CT.; Gonzales, F. A..; Jones, P. A. Altered Chromatin Structure Associated with Methylation-Induced Gene Silencing in Cancer Cells: Correlation of Accessibility, Methylation, MeCP2 Binding and Acetylation. *Nucleic Acid Res.* **2001***, 22,* 4598-4606.

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

(13) Guo, J. U.; Ma, D. K.; Mo, H.; Ball, M. P.; Jang, M. H.; Bonaguidi, M. A.; Balazer, J. A.; Eaves, H. L.; Xie, B.; Ford, E.; Zhang, K.; Ming, G. L.; Gao, Y.; Song, H. Neuronal Activity Modifies the DNA Methylation Landscape in the Adult Brain. *Nat. Neurosci.* **2011**, *14,* 1345-1351.

(14) Xie, W.; Barr, C. L.; Kim, A.; Yue, F.; Lee, A. Y.; Eubanks, J.; Dempster, E. L.; Ren, B. Base-Resolution Analyses of Sequence and Parent-of-Origin Dependent DNA Methylation in the Mouse Genome. *Cell* **2012**, *148,* 816-831.

(15) Galdieri, L.; Vancura, A. Acetyl-CoA Carboxylase Regulates Global Histone Acetylation. *J. Biol. Chem.* **2012**, *287*, 23865–23876.

(16) Schultz, D. C.; Ayyanathan, K.; Negorev, D.; Maul, G. G.; Rauscher, F. J.; 3rd. SETDB1: a Novel KAP-1-Associated Histone H3, lysine 9-Specific Methyltransferase that Contributes to HP1-Mediated Silencing of Euchromatic Genes by KRAB Zinc-Finger Proteins. *Genes. Dev.* **2002**, *16*, 919-932.

(17) Wang, H.; An, W.; Cao, R.; Xia, L.; Erdjument-Bromage, H.; Chatton, B.; Tempst, P.; Roeder, R. G.; Zhang, Y. mAM Facilitates Conversion by ESET of Dimethyl to Trimethyl Lysine 9 of Histone H3 to Cause Transcriptional Repression. *Mol. Cell.* **2003**, *12*, 475-487.

(18) Sarkar, S.; Abujamra, A. L.; Loew, J. E.; Forman, L.W.; Perrine, S. P.; Faller, D.V. Histone Deacetylase Inhibitors Reverse CpG Methylation by Regulating DNMT1 Through ERK Signaling. *Anticancer Res*. **2011**, *31,* 2723–2732.

(19) Poplawski, A.; Hu, K.; Lee, W.; Natesan, S.; Peng, D.; Carlson, S.; Molecular Insights into the Recognition of N-terminal Histone Modifications by the BRPF1 Bromodomain. J. Mol. Biol. **2014**, *426,* 1661–76.

(20) Glozak, M. A.; Seto, E.; Histone Deacetylases and Cancer. *Oncogene* **2007**, 26, 5420–5432.

(21) Gao, H.; Yu, Z.; Bi, D.; Jiang, L.; Cui, Y.; Sun, J.; Ma, R. Akt/PKB Interacts with the Histone H3 Methyltransferase SETDB1 and Coordinates to Silence Gene Expression. *Mol. Cell. Biochem.* **2007**, *305*, 35-44.

(22) David M. Knipe.; Anna Cliffe. Chromatin Control of Herpes Simplex virus Lytic and Latent Infection. *Nat. Rev. Microbiol.* **2008**, *6,* 211-221.

(23) Li, H.; Rauch, T.; Chen, Z. X.; Szabo, P. E.; Riggs, A. D.; Pfeifer, G. P. The Histone Methyltransferase SETDB1 and the DNA Methyltransferase DNMT3A Interact Directly and Localize to Promoters Silenced in Cancer Cells. *J. Biol. Chem.* **2006**, *281,* 19489-19500.

(24) Meehan, R. R.; Lewis, J. D.; McKay, S.; Kleiner, E. L.; Bird, A. P. Identification of a Mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell* **1989**, *58*, 499-507.

(25) Lewis, J. D.; Meehan, R. R.; Henzel, W. J.; Maurer-Fogy, I.; Jeppessen, P.; Kleain, F.; Bird, A. Purification, Sequence, and Cellular Localization of a Novel Chromosomal Protein that Binds to Methylated DNA. *Cell* **1992**, *69,* 905-914.

(26) Nan, X.; Campoy, F. J.; Bird, A. MeCP2 is a Transcriptional Repressor with Abundant Binding Sites in Genomic Chromatin. *Cell* **1997**, *88*, 471-481.

(27) Nan, X.; Meehan, R. R.; Bird, A. Dissection of the Methyl-CpG Binding Domain from the Chromosomal Protein MeCP2. *Nucleic Acids Res.* **1993**, 21, 4886-4892.

(28) Hendrich, B.; Bird, A. Identification and Characterization of a Family of Mammalian Methyl-CpG Binding Proteins. *Mol. Cell. Biol.* **1998**, *18*, 6538-6547.

(29) Fujita, N.; Shimotake, N.; Ohki, I.; Chiba, T.; Saya, H.; Shirakawa, M.; Nakao, M. Mechanism of Transcriptional Regulation by Methyl-CpG Binding Protein MBD1. *Mol. Cell Biol*. **2000**, *20,* 5107-5118

(30) Ng, H. H.; Zhang, Y.; Hendrich, B.; Johnson, C. A.; Turner, B. M. Erdjument-Bromage H.. MBD2 is a Transcriptional Repressor Belonging to the MeCP1 Histone Deacetylase Complex. *Nat. Genet.* **1999**, *23,* 58–61.

(31) Perez-Torrado, R.; Yamada, D., Defossez, P. A. Born to Bind: the BTB Protein-Protein Interaction Domain. *BioEssays* **2006**, *28,* 1194-1202

(32) Daniel, J. M.; Reynolds, A. B. The Catenin p120(ctn) Interacts with Kaiso, a Novel BTB/POZ Domain Zinc Finger Transcription Factor. *Mol. Cell. Biol*. **1999**, *19,* 3614-3623.

(33) Filion, G. J. P.; Zhenilo, S.; Salozhin, S.; Yamada, D.; Prokhortchouk, E.; Defossez, P-A. A Family of Human Zinc Finger Proteins that Bind Methylated DNA and Repress Transcription. *Mol. Cell. Biol.* **2006**, *26*, 169-181.

(34) Daniel, J. M.; Spring, C. M.; Crawford, H. C.; Reynolds, A. B.; Baig, A. The p120(ctn)-Binding Partner Kaiso is a Bi-Modal DNA-Binding Protein that Recognizes Both a Sequence-Specific Consensus and Methylated CpG Dinucleotides. *Nucleic Acids Res*. **2002**, *30,* 2911-2919.

(35) Prokhortchouk, A. V.; Sansom, O.; Selfridge, J.; Caballero, I. M.; Salozhin, S.; Aithozhina, D.; Cerchietti, L.; Meng, F. G.; Augenlicht, L. H.; Mariadason, J. M.; Hendrich, B.; Melnick, A.; Prokhortchouk, E.; Clarke, A.; Bird, A. Kaiso-Deficient Mice Show Resistance to Intestinal Cancer. *Mol. Cell. Biol.* **2006**, *26,* 199-208.

(36) Lopes, EC.; Valls, E.; Figueroa, M. E.; Mazur, A.; Meng, F-G.; Chiosis, G.; Laird, P. W.; Schreiber-Agus, N.; Greally, J. M.; Prokhortchouk, E.; Melnick, A.; Kaiso Contributes to DNA Methylation-Dependent Silencing of Tumor Suppressor Genes in Colon Cancer Cell Lines. *Cancer Res*. **2008**, *68,* 7258-7263.

(37) Vermeulen, J. F.; van de Ven, R. A.; Ercan, C.; van der Groep, P.; van der Wall, E.; Bult, P.; Christgen, M.; Lehmann, U.; Daniel, J.; van Diest, P. J.; Derksen, P. W. Nuclear Kaiso Expression is Associated with High Grade and Triple-Negative Invasive Breast Cancer. *PLoS One* **2012**, *7*, e37864.

(38) Jones, J.; Wang, H.; Zhou, J.; Hardy, S.; Turner, T.; Austin, D.; He, Q.; Wells, A.; Grizzle, W. E.; Yates, C. (2012) Nuclear Kaiso Indicates Aggressive Prostate Cancers and Promotes Migration and Invasiveness of Prostate Cancer Cells. *Am. J. Pathol.* **2012**, *181*, 1836-1846

(39) Bassey-Archibong, B. I.; Kwiecien, J. M.; Milosavljevic, S. B.; Hallett, R. M.; Rayner, L. G.; Erb, M. J.; Crawford-Brown, C. J.; Stephenson, K. B.; Bédard, P. A.; Hassell, J. A.; Daniel, J. M. (2016) Kaiso Depletion Attenuates Transforming Growth Factor-β Signaling and Metastatic Activity of Triple-Negative Breast Cancer Cells. *Oncogenesis* **2016**, *5,* e208

(40) Rhee, H. S.; Pugh, B. J. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell* **2011**,147, 1408-1419.

(41) Terooatea, W. T.; Pozner, A.; Buck-Koehntop, B. A. PAtCh-Cap: Input Strategy for Improving Analysis of ChIP-exo Data Sets and Beyond. *Nucleic Acids Res.* **2016**, *44*, 21, e159.

(42) Yoon, H. G.; Chan, D. W.; Reynolds, A. B.; Qin, J. Wong. N-CoR mediates DNA methylation-dependent repression through a methyl CpG binding protein Kaiso. *J. Mol. Cell* **2003**, *12,* 723-734

(43) Dai S-D.; Wang, Y.; Zhang, J.-Y.; Zhang, D.; Zhang, P.-X.; Jiang, G.-Y., Han, Y.; Zhang, S.; Ciu, Q.-Z.; Wang, E.-H.; Upregulation of d-Catenin is Associated with Poor Prognosis and Enhances Transcriptional Activity Through Kaiso in Non-Small-Cell Lung Cancer. *Cancer Sci.* **2011,** *102,* 95-103.

(44) Donaldson, N. S.; Pierre, C. C.; Antsey, M. I.; Robinson, S. C.; Weerawardane, S. M.; Daniel, J. M. Kaiso Represses the Cell Cycle Gene Cyclin D1 Via Sequence-Specific and Methyl-CpG-Mependent Mechanisms. PLoS One 2012, 7, e50398.

(45) Zhang B-Z.; Gu, L.-K.; Deng, D.-J. Methylation Specific Binding Activity of Zinc Finger Protein Kaiso.*Chin. J. Prev. Med.* **2007**, *41,* 43-46.

(46) Blattler, A.; Yao, L.;Wang, Y.; Ye, Z.; Jin, V. X.;Farnham, P. J. ZBTB33 Binds

Unmethylated Regions of the Genome Associated with Actively Expressed Genes. *Epigenet. Chromat.* **2013**, *6,* 18-25.

(47) Pozner, A.; Terooatea, W. T.; Buck-Koehntop, A. B. Cell Specific Kaiso (ZBTB33) Regulation of Cell Cycle Through Cyclin D1 and Cyclin E1. *J. Biol. Chem*. **2016**, *291*, 24538-24550.

CHAPTER 2

PAtCh-Cap: INPUT STRATEGY FOR IMPROVING ANALYSIS OF ChIP-exo DATA SETS AND BEYOND

Introduction

ChIP-seq (chromatin immunoprecipitation coupled with next-generation sequencing) has emerged as a powerful and widely used methodology for defining the site-specific localization of transcription factors and histone marks in the context of the cellular genome.[1-3] Since inception of this core technique nearly a decade ago, several improvements have been implemented to expand the capabilities,[4-11] reduce costs[12,13] or maximize resolution.[14-16] In many of the recently enhanced ChIP-based methods including ChIP-exo, ChIP-nexus, lobChIP and ChIPmentation,[12-15] several steps of sample and/or library preparation are performed on bead-bound immunoprecipitated chromatin, posing a challenge in generating a similarly treated input control required for downstream bioinformatic analysis and data quality assessments.

Numerous reports have focused on the sources and methods for removal of artifacts in ChIP-based data sets, many of which highlight the necessity and critical importance of having a proper input control.[2,17-22] For ChIP-seq, input controls are normally derived from isolated cellular DNA that has been cross-linked, sheared, and

Reprinted (adapted) with permission from Terooatea, T. W.; Pozner, A.; Buck-Koehntop, B. A. Nucleic Acids Res. **2016**, *44*, 21, e159.

ideally chemically treated in an analogous manner to DNA immunoprecipitated for the protein of interest. When subjected to high-throughput sequencing in parallel with the ChIP sample, the input control informs on the genomic locations of technique specific artifact peaks that exist in the ChIP-seq data set. As such, the use of an input control has become a core component of the communally agreed upon standards and guidelines for ChIP-seq experiments.[20] In lieu of having a comparable input, less ideal methods for artifact removal must be implemented such as: 1) utilizing an IgG control, which typically pulls down comparably less DNA resulting in lower library complexity and significant sequencing biasing relative to the ChIP sample;[19,20] 2) relying on only filtering ChIP-seq derived blacklisted peaks, which is unable to eliminate technique specific false positives; and/or 3) applying alternative peak caller strategies in which p-values may provide less reliable false discovery rates (FDRs), reducing confidence in the statistical significance of identified peaks.[22] Thus, a general procedure for producing a matched input control that can facilitate technique specific artifact removal would greatly increase the quality and confidence of information gained from the above listed modified ChIP-based methodologies that perform additional bead-bound processing steps.[12-15]

Indeed, bioinformatics treatments commonly utilized in ChIP-seq data analysis, such as blacklist filtering and duplicate read removal, have proven to be inappropriate for eliminating artifacts in ChIP-exo data.[17,23] It is understood that due to the narrow peak distributions observed in high-resolution methods such as ChIP-exo, removal of duplicate reads would discard essential peak information,[17,23] whereas the failure of blacklisting to improve these data sets is likely a direct result of technique specific variances in artifact generation. Similarly, it has been well established that ChIP-chIP and ChIP-seq data sets

have variable artifact signatures.[24] While methods like ChIP-exo[15] and ChIP-nexus[14] have undoubtedly increased resolution over standard ChIP-seq, the high number of washing and digestion steps in conjunction with decreased library complexity[23] result in significant false positive peaks that can considerably impact downstream data analysis. This may be of minimal concern for histone proteins, high occupancy transcription factors, or transcription factors for which the consensus binding motif is well characterized. However, for analysis of transcription factors where the consensus sequence has yet to be identified, or for proteins that have a globally low genomic occupancy, the persistence of false positives in these data present a significant barrier in reliably identifying peaks with high confidence and discerning de novo binding motifs.

Here, we report a method for nonspecifically capturing cross-linked chromatin complexes via protein carboxylate groups that allows for the DNA to be subjected to all downstream chemical treatments in parallel with bead-bound chromatin separately immunoprecipitated for the target of interest. This input control method, termed protein attached chromatin capture (PAtCh-Cap), is designed to be facile and universally applicable to any of the current[12-15] and future ChIP-based techniques that perform additional chemical and library preparation steps on bead-bound chromatin. Applying our input control method to the analysis of CTCF ChIP-exo data demonstrated that we were able to selectively remove artifacts in both pericentromeric and gene proximal regions, significantly increasing confidence in peak identification, revealing previously unidentifiable peaks and affording the capability of performing a de novo motif search analysis. This improved analysis capability within a high-resolution ChIP-exo data set was essential for the identification of a novel CTCF motif that appears to have an

independent cellular function.

Materials and methods

Cell culturing

HeLa cells (from the laboratory of Prof. C. J. Burrows; University of Utah) were cultured in Dulbecco's modified Eagle medium supplemented with 4.5 g/L glucose, 10% fetal bovine serum, and 2 mM glutamine and maintained in a humidified incubator at 37 °C and 5% CO_2. Cell counting and viability analysis was performed on a Countess Automated Cell Counter (Thermo Fisher Scientific). Cell line authentication to confirm lack of cross-contamination was routinely verified by short tandem repeat (STR) DNA profiling.

ChIP-exo

For each of the two ChIP-exo replicates, 20 × 106 HeLa cells were fixed with 1% formaldehyde for 15 min to cross-link protein:DNA complexes, followed by a quench with 125 mM glycine. The IP, exonuclease digestions and library generation procedures were all performed using a commercially available ChIP-exo Kit (Active Motif) following the manufacturer's instructions with the few noted modifications. A Diagenode Bioruptor Standard sonication device (run at max amplitude for 5 × 15 min in ice water) was used to shear the cross-linked DNA to 100–400 bp fragments. Cell debris was removed by centrifugation and the supernatant containing the solubilized chromatin DNA:protein complexes were isolated. Prior to further treatment, 10% of the sheared chromatin sample volume was removed from each replicate for input sample preparation

(see PAtCh-Cap section below). For the IP step, protein G coated magnetic beads were prefunctionalized with CTCF antibody (Millipore) prior to incubation with the sheared chromatin sample. DNA purification after reverse cross-linking was performed with the MinElute PCR Purification Kit (Qiagen). It should be noted that the library preparations performed with the Active Motif ChIP-exo Kit are designed to be compatible with the Illumina sequencing platform.[25] Final purified DNA libraries were sequenced by the High-Throughput Genomics Core within the University of Utah Huntsman Cancer Institute using the Illumina HiSeq 2000 platform.

Protein attached chromatin capture (PAtCh-Cap) for ChIP-exo

Following two series of washes with 0.01 M PBS (pH 7.4), 50 μg of M-280 streptavidin coated Dynabeads (Thermo Fisher Scientific; equivalent to the number of beads utilized per reaction in the Active Motif ChIP-exo Kit) were conjugated with 10 μl of 50 nM EZ-link amine-PEG3-Biotin (Thermo Fisher Scientific; PEG = polyethylene glycol) in 0.01 M PBS (pH 7.4) at room temperature for 20 min. These beads were selected as they have the same size and core material composition as the protein G coated magnetic beads utilized in the Active Motif ChIP-exo Kit. The biotinylated beads were then washed twice with 0.01 M PBS (pH 7.4) and once with 0.1 mM MES (pH 5.0) to remove any nonconjugated material. For each of the two replicates, the input sample (obtained as discussed above) was combined with 300 μl of 0.1 M EDC (1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride) along with the prefunctionalized biotinylated magnetics beads and incubated in 10 ml 0.1 mM MES buffer (pH 5.0) for 3 h at room temperature on a mechanical rotator. Once covalently bound to the magnetic

beads, the input samples were treated identically as described above for the CTCF ChIP-exo samples utilizing reagents and materials from the Active Motif ChIP-exo Kit.

Preparation of input DNA for ChIP-seq

For each of the three replicates, 20 x 10^6 HeLa cells were fixed with 1% formaldehyde for 15 min to cross-link protein: DNA complexes, followed by a quench with 125 mM glycine. Cells were washed with cold 0.01 M PBS (pH 7.4) and lysed for 10 min at 40 °C in cell lysis buffer (50 mM HEPES (pH 8.0), 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100). Cellular nuclei were then washed (10 mM Tris-HCl (pH 8.0), 200mM EDTA), centrifuged and resuspended in nuclear lysis buffer (50 mM Tris-HCl (pH 8.0), 100 mM NaCl, 10 mM EDTA, 1% SDS). A Diagenode Bioruptor Standard sonication device (run at max amplitude for 5 x 15 min. in ice water) was used to shear the cross-linked DNA to 100 - 400 bp fragments. Proteinase K digestion was performed overnight at 55 °C. DNA was purified with the MinElute PCR Purification Kit (Qiagen). DNA quantification, library construction, and sequencing were all performed by the High-Throughput Genomics Core within the University of Utah Huntsman Cancer Institute. The input DNA libraries were sequenced using the Illumina HiSeq 2000 platform.

RNA interference and RNA-seq

HeLa cells were transfected with either a scrambled siRNA or one of two CTCF siRNAs (Thermo Fisher Scientific) in triplicate for each siRNA using Lipofectamine RNAiMAX (Thermo Fisher Scientific) for 24 h. Cells were washed with 0.01 M PBS

(pH 7.4) and resuspended in TRIzol (Thermo Fisher Scientific) prior to RNA extraction with the Direct-zol RNA Kit (Zymo Research). Prior to submission for high-throughput sequencing analysis, an aliquot of the RNA from each sample was reversed transcribed using the High Capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific). The amount of CTCF was then determined in each sample by quantitative real-time PCR (qRT-PCR), utilizing HPRT1 as a normalization control, to ensure that sufficient CTCF knock-down was achieved. In parallel, protein was extracted from HeLa cells after siRNA transfection utilizing NP-40 buffer (50 mM Tris (pH 8.0), 150 mM NaCl, 1.0% NP-40) supplemented with protease inhibitors (Roche) and separated by gel electrophoresis. Proteins were then transferred to nitrocellulose membranes and immunoblotted using a CTCF antibody (Millipore) following standard procedures, to ensure sufficient knockdown was also achieved at the protein level. For RNAseq, RNA quality control measurements, purification, library construction, and sequencing were all performed by the High-Throughput Genomics Core within the University of Utah Huntsman Cancer Institute. In short, RNA quality was measured on a Bioanalyzer RNA 6000 Nano Chip. Total RNA was then further purified with the RiboMinus Eukaryote Kit for RNA-seq (Thermo Fisher Scientific). Small and long directional RNA-seq libraries were then constructed using Illumina TruSeq Stranded mRNA Sample Prep with poly(A) selection and sequenced with a 50 bp single-end run on the Illumina HiSeq 2000 platform.

Sequencing data analyses

ChIP-exo fastq files were aligned to the human genome (hg19) using Novoalign (Novocraft, Inc.) and the following parameters: -0 SAM –r. Sam files were then sorted and indexed with SAMtools.[26] Peak calling was performed with and without the input control by MACS2[27] using the following standard parameters: model fold settings of 5 to 50 q-values with a cutoff of 0.05. Bedgraph outputs from MACS2 were viewed on the Integrated Genome Browser (IGB).[28] The MACS2 bdgcmp command was further performed to obtain the read coverage tracks after ChIP-exo normalization to the PAtCh-Cap derived input control. RNA-seq fastq files were similarly aligned to the human genome (hg19) using Novoalign (Novocraft, Inc.) and peak called with the USeq suite.[21] RNA-seq reads were aligned with all known and theoretical splice junctions using the following parameters: -r All 50 -t 40 -o SAM 90 –k. The USeq NovoalignParser application was then used to parse the alignment files into binary point data by setting the posterior probability to 0 and alignment score threshold to 60. The MultipleReplicaDefinedRegionScanseqs USeq application, which utilizes the DESeq R package,[29] identified statistically significant differentially expressed genes between cells treated with the scrambled siRNA and CTCF depleted cells.

Bioinformatics analyses

To identify the high confidence ChIP-exo peaks, the peak confidence level was plotted against the ranked peak number. Upon doing this, a clearly observable inflection point separated a subset of peaks with the highest intensity from the overall peaks analyzed. Therefore, in this analysis high-confidence CTCF peaks were selected from

enriched regions characterized by a –log(q-value) higher than the inflection point and a line with a slope of –tan 1 to the curve. Further, ChIP-seq input identified blacklisted genomic regions from the DAC, DER and UHS lists[30] were intersected with the above determined high-confidence pools for both the CTCF ChIP-exo data with and without input treatment. The DAC and DER blacklisted regions were downloaded from the UCSC table browser (http: //hgwdev .cse. ucsc. edu/ cgi-bin /hg FileUi?db =hg19 &g= wgEncode Mapability)[31] whereas the UHS regions were extracted from the following site: https://sites.google.com/site/anshulkundaje/projects/blacklists.

High confidence peaks were then subjected to de novo motif discovery analysis using RSAT[32-34] with the following parameters: peak -motifs -v 1 -title bedfile -i $RSAT -max_seq_len 200 -markov auto -disco oligos,positions,local_words -nmotifs 8 -minol 6 -maxol 6 -no_merge_lengths -2str -origin center -motif_db jaspar_core_vertebrates. In addition to the expected CTCF consensus sequence, two additional motifs were identified from the de novo motif analysis in the input treated ChIP-exo data. FIMO[35] was then used to identify matches for all of these motifs in all peak regions (‘fimo -bgfile flanking.bg–motif 1 motif.meme.txt) with a default p-value threshold of 10^{-4}. To determine the preferential spacing and co-localization of the additionally identified motifs relative to the CTCF consensus sequence, the CTCF core motif was extended by ±30 nucleotides and analyzed by SpaMo (spamo -png -bgfile -dumpseqs -inc 1 meme.motif.txt).[36] Average logos representing all the extended motifs around the centralized CTCF consensus were created using MEME.[37] The read tag and nucleotide base heatmaps as well as aggregate plots were generated using in-house python scripts. Each pool of sequences containing the motifs of interest were trimmed to the same size

and centered at their CTCF core motifs using the USeq scoreSequences application. All instances in which the identified CTCF motifs were localized within gene promoters (defined as +/-1000 bps around the transcription start site (TSS)) were intersected with the genes identified to be differentially expressed from the above RNA-seq analysis (0.05 adjusted Benjamini–Hochberg p-value). For each set of filtered differentially expressed genes, Ingenuity Pathway Analysis (IPA, www.ingenuity.com) was performed to identify uniquely significant biological pathways correlated with each motif.

Results and discussion

Development of a method for nonspecific protein mediated chromatin capture

Development of a general method to generate matched input controls for ChIP-based techniques that perform additional preparation steps on bead-bound immunoprecipitated chromatin complexes required that two major challenges be addressed. First, it was necessary to identify a way to nonspecifically pull-down a random sampling of cross-linked chromatin complexes that could be affixed to magnetic beads with a sufficient affinity to remain associated throughout the following treatment steps. Second, it was ideal to determine a strategy for affixing these chromatin complexes to the magnetic beads via the proteins, leaving the chromatin DNA accessible for further processing. These criteria eliminated the possibility of utilizing an antibody based approach as no single protein-specific antibody would be capable of immunoprecipitating a completely unbiased background representation of the chromatin complexes in a given sample with adequate affinity.

Thus, to facilitate protein mediated bead-bound capture of cross-linked chromatin

complexes, we took advantage of the readily available amine functionalized pegylated biotin reagents commonly utilized for the conjugation of EDC (1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride) activated protein carboxylates. In short, after standard cross-linking, chromatin isolation and DNA shearing procedures, an aliquot of the sample (typically 10%) is removed to generate the input control. In parallel, streptavidin coated magnetic beads are conjugated with an amine functionalized pegylated biotin (amine-PEG3-biotin). These preconjugated beads are then incubated with the isolated cross-linked chromatin complexes in the presence of EDC producing bead-bound chromatin complexes covalently linked via protein carboxylate groups (Figure 2.1). These pulled-down complexes can then be subjected to additional chemical processing steps in parallel with bead-bound chromatin separately immunoprecipitated for the target of interest. Compared to other chromatin capture methods that pull-down on modified DNA,[38,39] which may result in disruption of protein:DNA interactions prior to cross-linking and/or mask accessibility of the chromatin fragments, our PAtCh-Cap method achieves the goal of nonspecifically pulling-down chromatin bound proteins, leaving the cross-linked DNA freely accessible for additional treatments.

Application of PAtCh-Cap to generate an input control for technique specific artifact removal in ChIP-exo data

As a proof-of-principle, we applied our PAtCh-Cap strategy to produce a ChIP-exo input control that could be utilized in downstream bioinformatics analysis of CTCF genomic occupations in HeLa cells. After cross-linking and shearing, 10% of the chromatin complex sample volume was removed for the input control and treated with

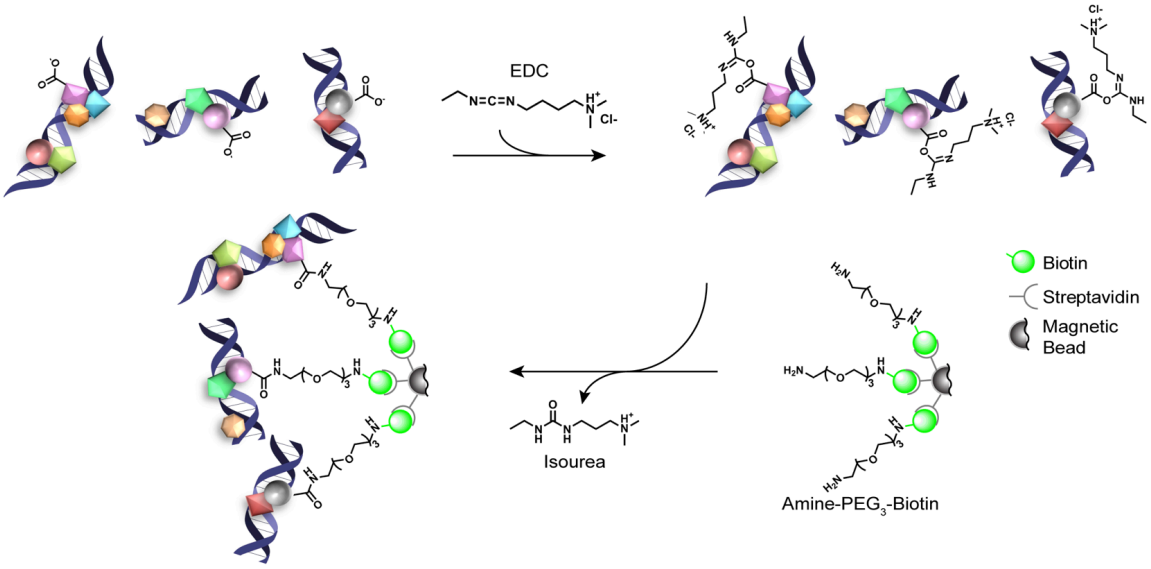


Figure 2.1. Schematic overview for the protein attached chromatin capture (PAtCh-Cap). Streptavidin coated magnetic beads are first conjugated with amine-PEG3-biotin (where PEG is polyethylene glycol). After standard cross-linking, chromatin isolation and DNA shearing procedures, 10% of the sample volume is removed and incubated with the preconjugated beads in the presence of EDC (1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride). EDC reacts with protein carboxylate groups, forming unstable o-acylisourea activated esters that can undergo nucleophilic attack by the primary amines on the amine-PEG3-biotin prosthetics. This forms a covalent amide linkage with proteins in the chromatin complexes and releases an isourea by-product. Once bead-bound, these chromatin complexes can be subjected to additional chemical processing steps in parallel with bead-bound chromatin that was separately immunoprecipitated for the target of interest.

the PAtCh-Cap method, generating the bead-bound input control as described above. In parallel, protein G coated magnetic beads were conjugated with CTCF antibody and utilized to selectively immunoprecipitate CTCF-bound chromatin fragments from theremaining cross-linked chromatin complex pool. At this stage, both the bead-bound input control and isolated CTCF-containing chromatin complexes were simultaneously and identically subjected to all subsequent exonuclease digestions, library preparation steps, reverse cross-linking, purification procedures, quality assessments, and high-throughput sequencing.

To assess whether the use of sequencing data from a similarly treated input control was able to sufficiently remove artifacts and improve peak identification within our CTCF ChIP-exo data set, we first defined pools of high-confidence peaks from this data with and without input treatment. These high-confidence pools were defined from enriched regions by plotting the -log(q-value) versus the ranked peak number and only considering peaks above the inflection point and a line with the slope of –tan 1 to the curve (Figure 2.2A). Once identified, the high-confidence peaks from the pools with and without input treatment were intersected and found to have a high degree of overlap, though in each case there were also a significant number of peaks that were individual to each pool (Figure 2.2B). To discern whether artifact peaks were eliminated from the ChIP-exo data set, we first performed a weight matrix search[21] for the core CTCF motif within each set of high-confidence peaks. This search revealed an approximate 20% increase in peaks containing the CTCF motif out of total high-confidence peaks called for the input treated data relative to the nontreated data set (Figure 2.2B). The majority of

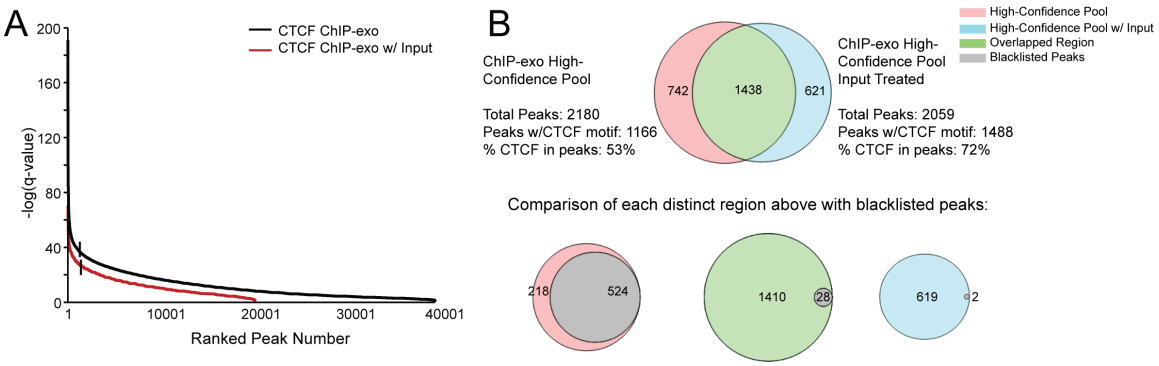


Figure 2.2. Application of PAtCh-Cap to CTCF ChIP-exo data allowed for significant artifact removal and improved confidence in peak identification. (A) To identify high-confidence CTCF peaks, all peaks called with a 0.05 q-value threshold from the ChIP-exo data with (red) and without (black) input treatment were plotted as the –log(q-value) versus ranked peak number. High confidence peaks were determined to be those characterized by a –log(q-value) higher than the inflection point and a line with a slope of –tan 1 to the curve (denoted by vertical lines). (B) Venn diagram demonstrating the overlap of high-confidence peaks identified from data sets with and without input treatment (top). The number of CTCF motifs found within each pool is denoted and clearly shows that the percentage of CTCF containing peaks relative to the total increases substantially after input treatment. Venn diagrams for the overlap of blacklisted peaks with each of the above regions (bottom).

remaining peaks that do not harbor the core CTCF motif in the cleaned ChIP-exo data likely represent genomic occupations for the numerous CTCF protein interacting partners.[40] Together, these findings indicate that bioinformatic treatment of the CTCF ChIP-exo data with the input control not only removes a significant number of artifact peaks but also allowed for identification of additional CTCF-containing peaks that were otherwise masked.

As further confirmation, we separately intersected known ChIP-seq blacklisted peaks with the above determined high-confidence pools for the CTCF ChIP-exo data with and without input treatment. Blacklisted peaks constitute known genomic artifact regions systematically observed in input controls from standard ChIP-seq data and it has become an acceptable practice to exclude these reads out of ChIP-seq data sets.[30,41–43] As can be seen in Figure 2.2B, nearly all of the blacklisted peaks are present in the pool of ChIP-exo peaks excluded by the input control treatment. This clearly demonstrates that the majority of these blacklisted peaks are captured by the ChIP-exo input control and that as previously observed, the removal of blacklisted peaks alone is not sufficient to remove all ChIP-exo specific artifacts.[17]

Furthermore, we prepared a ChIP-seq input control from our HeLa cells and compared this with our ChIP-exo input. Comparisons of enrichment profiles and genome-wide correlation analysis between these input controls demonstrates that there are a number of artifact peaks that are common between the two methods, though there are many more ChIP-exo specific false positives (Figure 2.3A (green loci) and Figure 2.3B). Indeed, intersection of the identified peaks in the ChIP-seq and ChIP-exo input controls results in nearly all of the ChIP-seq input peaks being captured by the ChIP-exo input

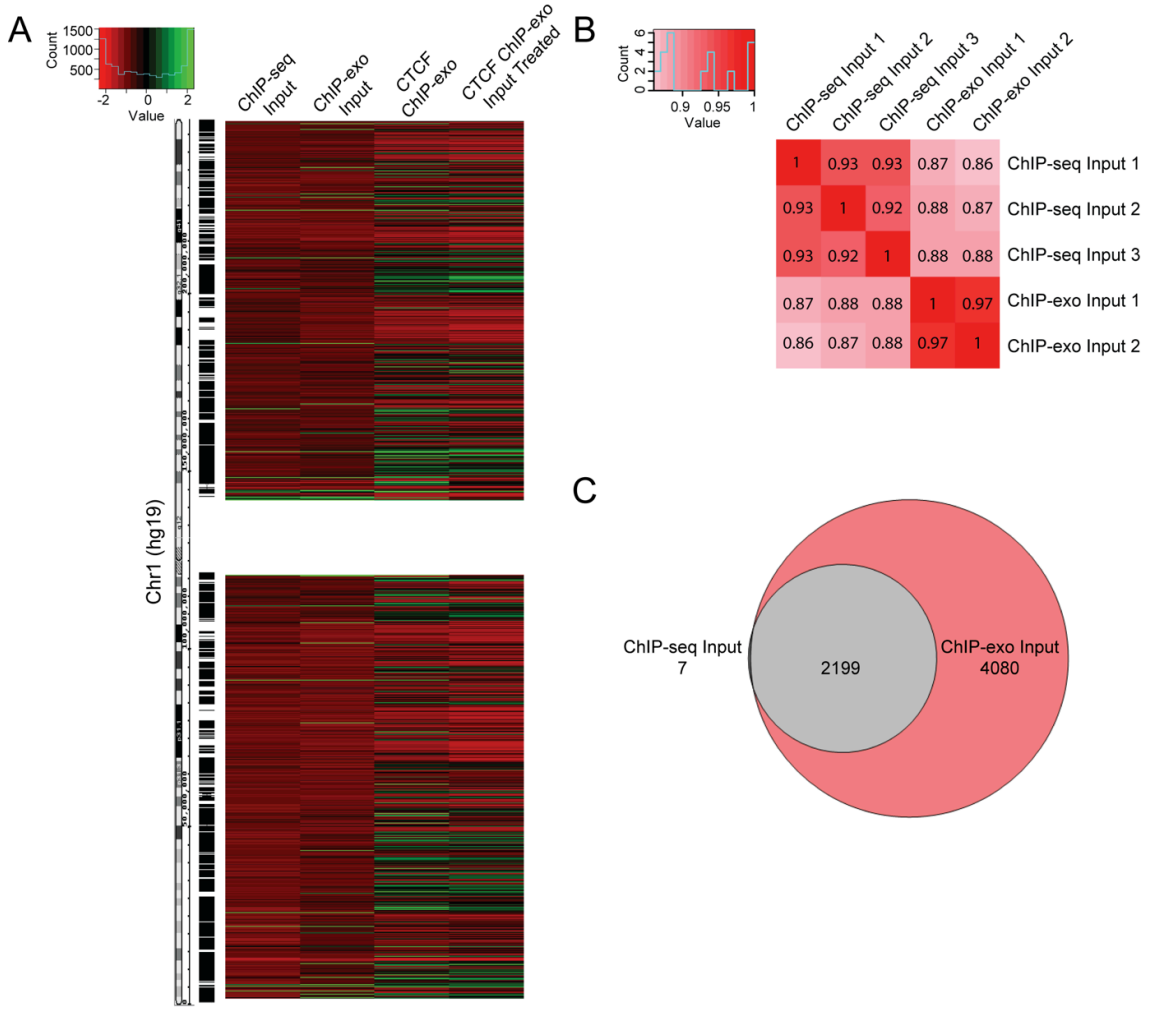


Figure 2.3. Technique specific removal of false positive peaks in ChIP-exo data. (A) Color-coded heat map enrichment profiles visualized within the Integrated Genome Browser (IGB) for the standard ChIP-seq input, the bead-bound ChIP-exo input, and the ChIP-exo profiles without and with input treatment mapped onto chromosome 1. (B) Genome-wide correlations (1,000 bp windows) for standard ChIP-seq inputs compared to the bead-bound ChIP-exo inputs in HeLa cells. (C) Venn diagram showing the intersect of peaks (gray) within the standard ChIP-seq input (not observable) and the ChIP-exo input (red) generated by the PAtCh-Cap method.

control (Figure 2.3C). Additionally, example read coverage tracks of CTCF ChIP-exo with and without input treatment indicate that use of the input control dramatically cleans up artifacts not only within pericentromeric, but also gene proximal regions (Figure 2.4). Analysis of genomic sequences underneath the peaks removed by the ChIP-exo input determined that these sites do not contain the CTCF motif, whereas remaining peaks do.

Finally, we sought to determine whether the increased incidence of artifact peaks in the ChIP-exo data was a consequence of biasing from the exonuclease digestion or GC content during sequencing. Comparative analysis of the nucleotide frequency and GC content plots from the ChIP-exo input controls and CTCF ChIP-exo sequencing data relative to standard ChIP-seq input controls indicates that there is a minor biasing common between all of the ChIP-exo data sets that likely results from the exonucleases used in the processing of these samples (Figure 2.5). Nonetheless, this observed exonuclease biasing in the ChIP-exo samples is not sufficient to account for the significant increase in generated artifacts relative to ChIP-seq data sets. Thus, it is likely that the remainder of these artifacts result from nonspecific interactions that occur during the additional processing steps performed on the bead-bound chromatin samples. This further suggests that the other bead-bound ChIP-based methods12-14 may also generate an increase in technique specific artifacts relative to standard ChIP-seq and that a strategy such as PAtCh-Cap to produce a matched input is needed. Together all of the above findings clearly demonstrate that ChIP-exo data sets harbor significant technique specific false positive peaks and that our PAtCh-Cap method was not only able to remove these artifacts, but also improved confidence in peak identification.

It should be noted that while we clearly demonstrated that use of an appropriate

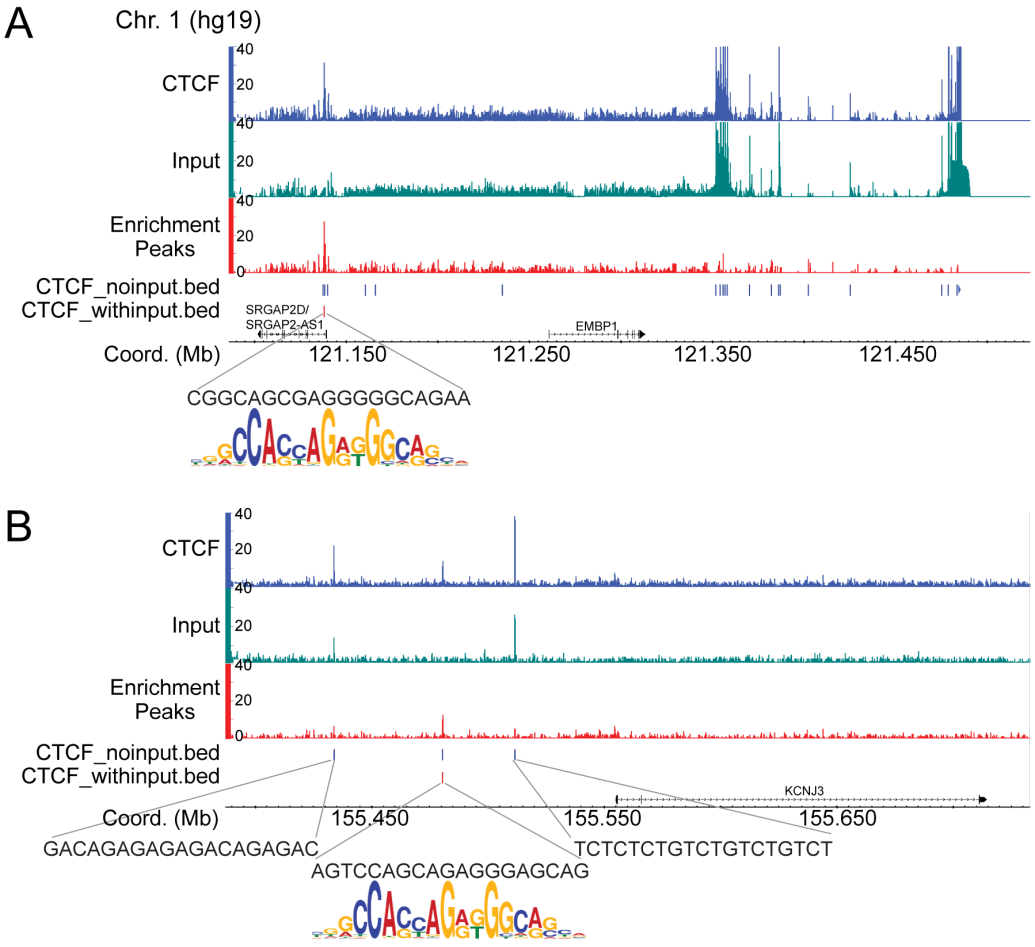


Figure 2.4. Representative CTCF ChIP-exo read coverage tracks for the pericentromeric region of chromosome 1. (A) and the promoter of the KCNJ3 gene (B). The CTCF reads (blue) were normalized to the reads from the input control (green) using MACS2[27] to generate the enrichment read coverage tracks (red). Peaks identified by the MACS2 peak caller (represented in the bed tracks) are denoted as red or blue vertical lines for the CTCF ChIP-exo data sets with and without input treatment, respectively. Analysis of the genomic sequences underneath the remaining peaks after input treatment (vertical red lines) definitively showed that these sites contain the core CTCF motif as evidenced by alignment of the CTCF sequence logo beneath.

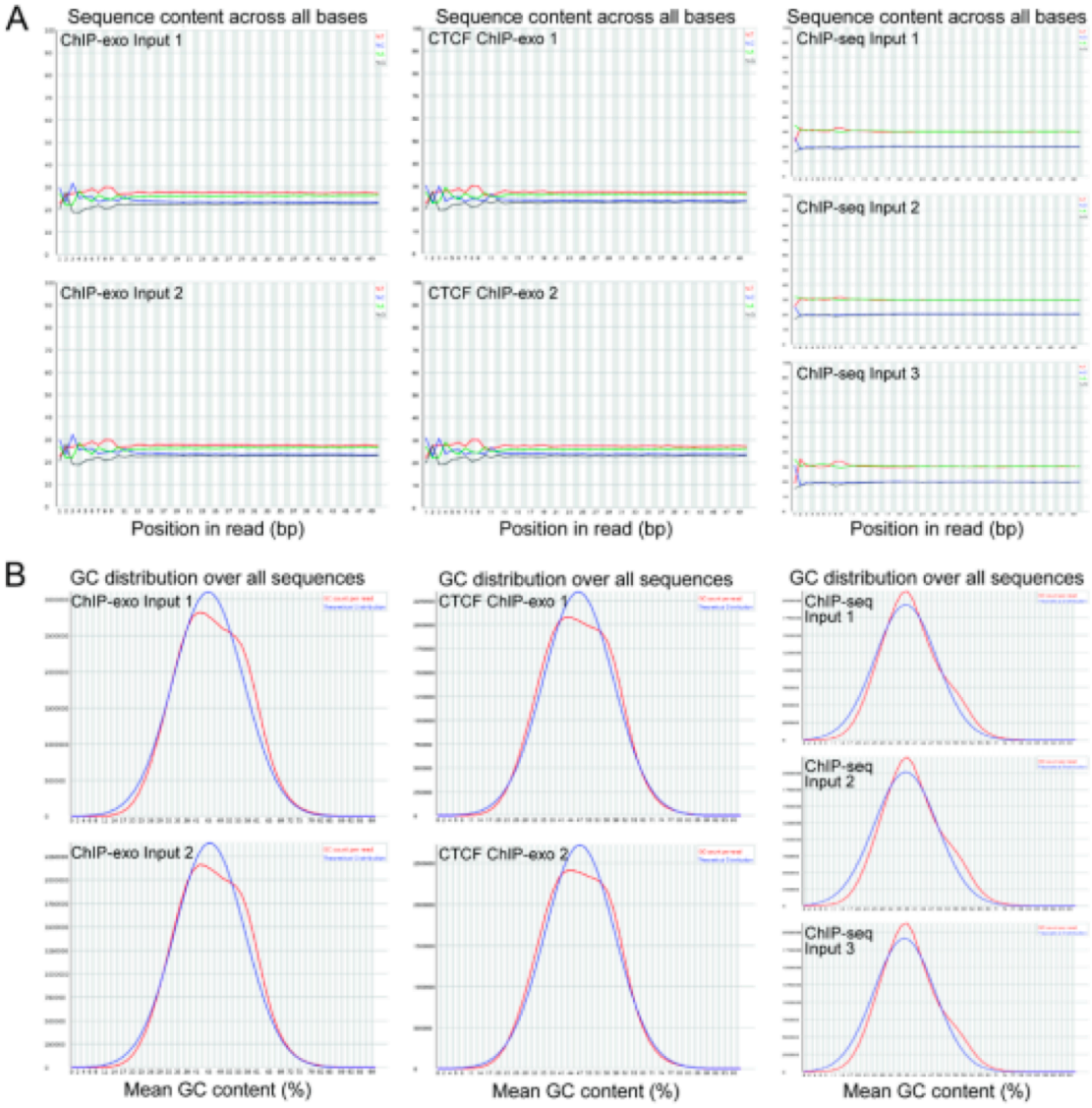


Figure 2.5. FastQC analysis for the ChIP-exo input, CTCF ChIP-exo, and ChIP-seq input replicates. (A) FastQC plots of nucleotide frequency as a function of base position for the ChIP-exo input, CTCF ChIP-exo, and ChIP-seq input replicates. The four colors represent the bases A (green), T (red), G (black) or C (blue). (B) FastQC plots showing that the GC count per read (red) for each ChIP-exo input, CTCF ChIP-exo, and ChIP-seq input data set relative to the theoretical distribution (blue).

input control can dramatically improve artifact removal in ChIP-exo data, new bioinformatics tools will need to be developed that are capable of implementing input treatments while maintaining full resolution of ChIP-exo data sets. For the analysis presented here, we utilized the MACS2 peak caller[26] as it has the built-in capability of determining peak enrichment over input controls and using local statistics to reduce biasing and calculate empirical FDRs. However, peak calling is administered by performing read-shifting to account for the offset in forward and reverse strand reads. In contrast, peak callers specifically designed for analysis of ChIP-exo data sets, such as GeneTrack[41] or MACE,[42] rely on retention of strand information by using 5' cross-link borders on each strand to define "peak-pairs" that result in identification of high-resolution protein footprints.[23] Currently, neither of these ChIP-exo specific peak callers are capable of readily normalizing these data sets to input controls. Thus, it would be ideal to have a peak caller suite that can normalize the data relative to an input control without compromising resolution by eliminating strand information.

Improved ChIP-exo data analysis identified a novel CTCF binding site

Once peaks within the high-confidence pools for the ChIP-exo data with and without input treatment were identified as described above (Figure 2.2A), we subjected each pool to a de novo motif search analysis.[32-34] This search returned not only the core CTCF motif, but also its previously characterized flanking 5′-site[15,43] for both the untreated and input treated ChIP-exo CTCF data sets (Figure 2.6A). It has been shown previously that depending on the genomic occupation context, various combinations of the 11 CTCF zinc finger subsets recognize the CTCF core alone or in conjunction with

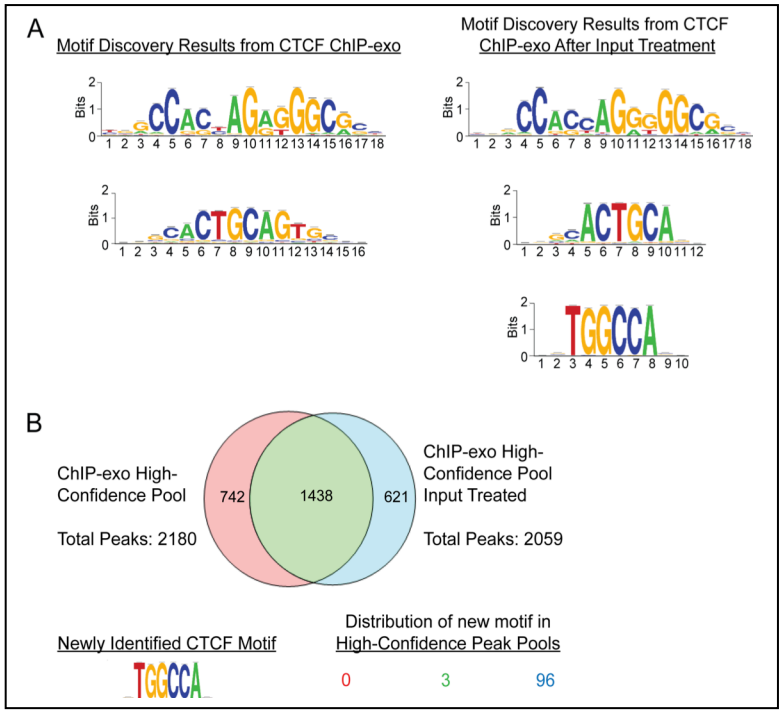


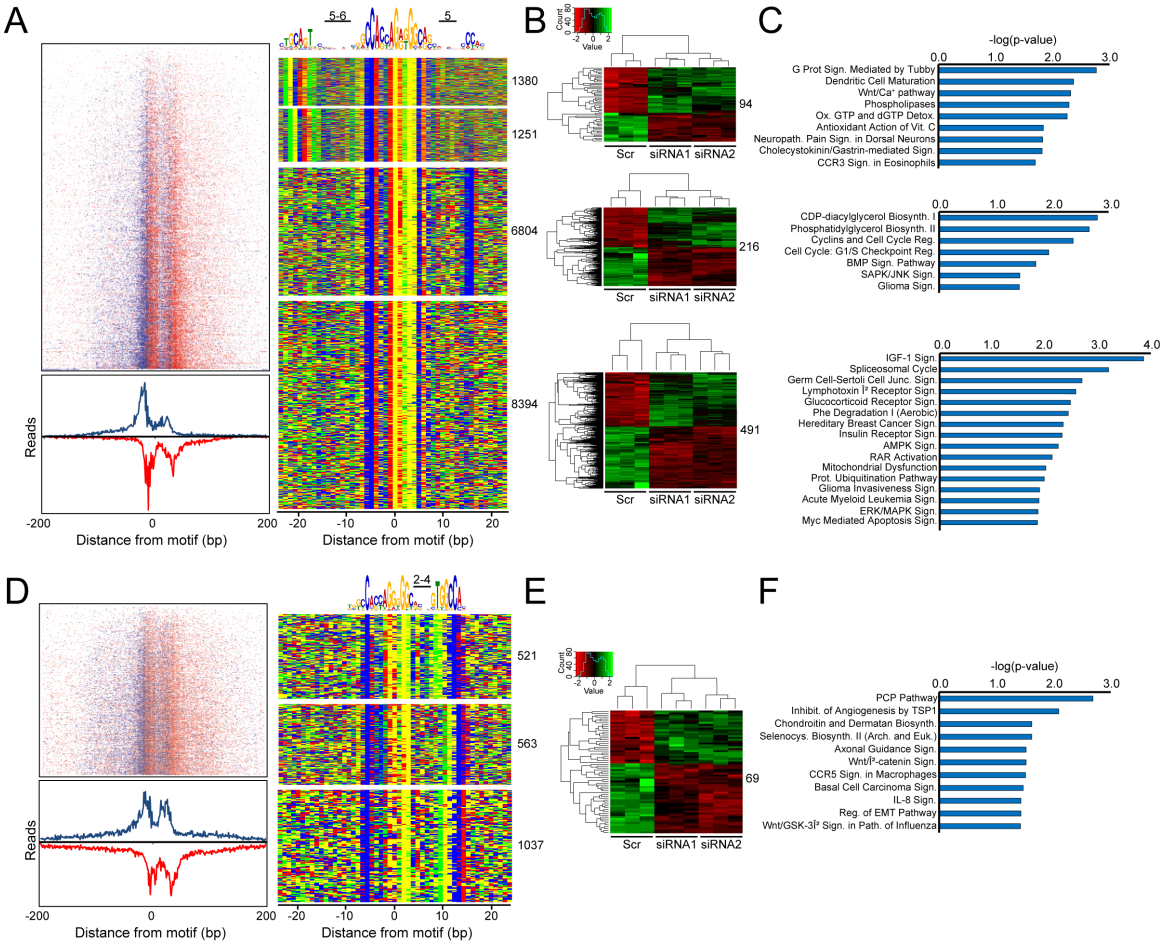
Figure 2.6. De novo motif search analysis. (A) Results for the de novo motif search analysis performed by RSAT on the highconfidence CTCF peaks identified from the ChIP-exo data without and with input treatment (refer to Figure 2A). The newly identified motif is only observed from the de novo search conducted on the input treated highconfidence pool. (B) Venn diagram from Figure 2B showing the overlap of high-confidence peaks identified from data sets with and without input treatment (top). Distribution analysis of the newly identified CTCF motif within each of these pools (color coded numbers; bottom) clearly shows that this sequence is found nearly exclusively within the population of high-confidence peaks that were only identified after input treatment (blue). Enrichment profiles depicting CTCF genomic localizations represented by both aggregate plots and heat maps (left panels in Figure 2.7A and 2.7D) also confirmed CTCF occupancy at all sites identified within the input treated ChIP-exo data set identified from the de novo search, including the new consensus site which was found to be localized 2-4 base pairs 3' of the CTCF core (Figure 2.7A and 2.7D). In addition, a previously characterized flanking 3′-site to the CTCF core sequence was also found.[15,43]

conserved 5′- and 3′-flanking regions.[15,43] This modularity in CTCF motif recognition is believed to allow for tunability in strength of the DNA binding interaction and subsequently chromatin residence time.[43] Surprisingly, we also identified a novel consensus site in the de novo motif search, but only for the CTCF data set in which the input control was utilized (Figure 2.6A). This observation was confirmed by the fact that the peaks containing this new motif were nearly all localized within the high-confidence peak population that was only identified after input treatment of the ChIP-exo data (Figure 2.6B). Thus, identification of this novel motif was only possible when the ChIP-exo specific artifact peaks were removed prior to de novo motif analysis.

Next, a comprehensive weight matrix analysis was performed to identify all peaks containing the core CTCF sequence within the complete input treated CTCF ChIP-exo data set, which resulted in 19,950 hits. To validate that all sequences found in the de novo motif search (Figure 2.7A) were localized proximal to the core CTCF site, a spaced motif analysis (SpaMo) was used.[36] This search demonstrated that there was a significantly enriched spacing between the CTCF motif and the other two conserved sequences.

To determine the biological relevance of this newly identified 3′-flanking sequence, we analyzed RNA-seq data from CTCF depleted HeLa cells. CTCF depletion was confirmed by both immunoblot and qRT-PCR (Figure 2.8). In the absence of CTCF, a substantial number of genes showed a significant transcriptional alteration relative to the control (Figure 2.9). Independent intersection of these altered genes with the CTCF ChIP-exo data with and without input treatment showed that many more promoter proximal peaks overlapped within the untreated ChIP-exo data set (Table 2.1). However, the number of peaks actually containing the core CTCF motif was essentially the same as

Figure 2.7. From the input treated CTCF ChIP-exo data set, (A) read tag distributions around all genomic CTCF-bound sites shown in the four binned motif combinations (right panel) were centered on the midpoint of the CTCF consensus to generate a heat map (top left),which is summed below as an aggregate plot. Denoted in blue and red are the sense and antisense strand read enrichments around the core CTCF motif, respectively. The centralized CTCF core sequence and adjacent motifs are depicted above a color map representation of 50 bp DNA stretches containing the various motif combinations (right panel). (B) Heat maps from RNA-seq data depicting gene transcripts exhibiting a two-fold up- (green) or down-regulation (red) after CTCF depletion relative to the scrambled siRNA control (Scr). For each motif group, CTCF promoter occupation sites (defined as ±1000 bps around the transcription start site (TSS)) were intersected with the RNA-seq data and resulting altered gene sets were binned as individual heat maps. (C) Each gene set from (B) was subjected to Ingenuity Pathway Analysis (IPA, www.ingenuity.com) to identify biological pathways uniquely modulated by each of the CTCF motif combinations. (D–F) The same analyses in (A–C) were performed separately on the core CTCF consensus with the newly identified 3′-CTCF motif.


A
5-6
5
Reads
-200
0
200
Distance from motif (bp)
-20
-10
0
10
20
Distance from motif (bp)
1380
1251
6804
8394
B
Count
Value
Scr
siRNA1
siRNA2
94
216
491
C
-log(p-value)
0.0
1.0
2.0
3.0
4.0
G Prot Sign. Mediated by Tubby
Dendritic Cell Maturation
Wnt/Ca+ pathway
Phospholipases
Ox. GTP and dGTP Detox.
Antioxidant Action of Vit. C
Neuropath. Pain Sign. in Dorsal Neurons
Cholecystokinin/Gastrin-mediated Sign.
CCR3 Sign. in Eosinophils
CDP-diacylglycerol Biosynth. I
Phosphatidylglycerol Biosynth. II
Cyclins and Cell Cycle Reg.
Cell Cycle: G1/S Checkpoint Reg.
BMP Sign. Pathway
SAPK/JNK Sign.
Glioma Sign.
IGF-1 Sign.
Spliceosomal Cycle
Germ Cell-Sertoli Cell Junc. Sign.
Lymphotoxin I² Receptor Sign.
Glucocorticoid Receptor Sign.
Phe Degradation I (Aerobic)
Hereditary Breast Cancer Sign.
Insulin Receptor Sign.
AMPK Sign.
RAR Activation
Mitochondrial Dysfunction
Prot. Ubiquitination Pathway
Glioma Invasiveness Sign.
Acute Myeloid Leukemia Sign.
ERK/MAPK Sign.
Myc Mediated Apoptosis Sign.
D
2-4
Reads
-200
0
200
Distance from motif (bp)
-20
-10
0
10
20
Distance from motif (bp)
521
563
1037
E
Count
Value
Scr
siRNA1
siRNA2
69
F
-log(p-value)
0.0
1.0
2.0
3.0
PCP Pathway
Inhibit. of Angiogenesis by TSP1
Chondroitin and Dermatan Biosynth.
Selenocys. Biosynth. II (Arch. and Euk.)
Axonal Guidance Sign.
Wnt/I²-catenin Sign.
CCR5 Sign. in Macrophages
Basal Cell Carcinoma Sign.
IL-8 Sign.
Reg. of EMT Pathway
Wnt/GSK-3I² Sign. in Path. of Influenza

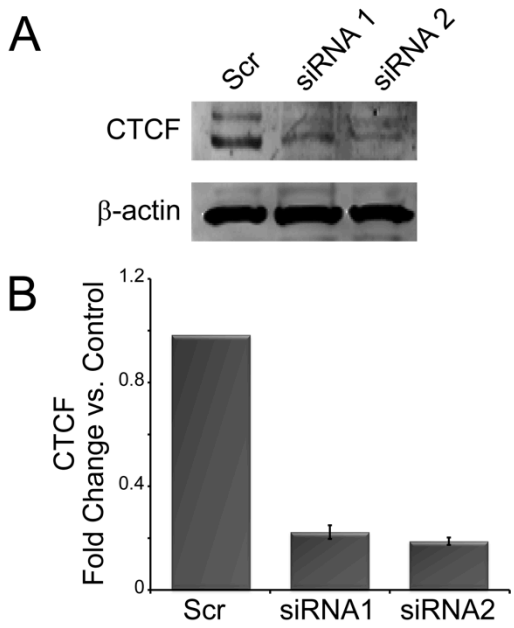


Figure 2.8. Assessment of CTCF depletion via RNA interference in HeLa cells. (A) Representative immunoblot indicating CTCF protein depletion after HeLa cells were treaed with two different siRNAs relative to a scrambled siRNA (Scr) control. (B) Sufficient CTCF depletion at the mRNA level in HeLa cells was further confirmed by qRT-PCR. mRNA levels are represented as the fold C_T change relative to *HPRT1* and a scrambled siRNA(normalized to 1).

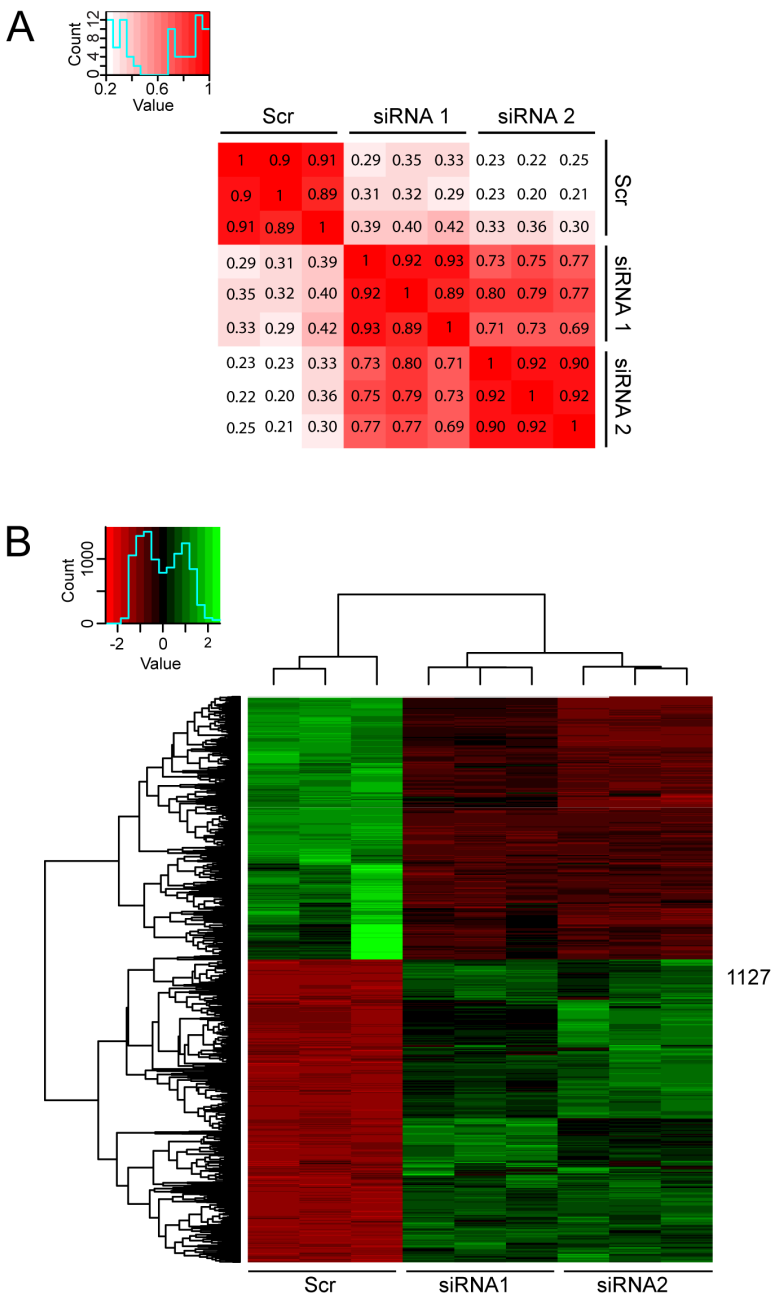


Figure 2.9. RNA-seq analysis after CTCF depletion in HeLa cells. (A) Genome-wide correlations (1,000 bp window) of the RNA-seq data acquired for each of the scrambled and CTCF siRNA replicates. (B) Comprehensive heat map from the RNA-seq data depicting gene transcripts exhibiting a two-fold up- (green) or down-regulation (red) after CTCF depletion relative to the scrambled siRNA control (p-value cut-off of 0.05).

Table 2.1. Incidence of CTCF motif within ±1000 bps of regulated gene promoter TSS

	CTCF ChIP-exo	CTCF ChIP-exo Input Treated
Total genes altered after CTCF depletion (from RNA-seq)	1127	
Number of ChIP-exo peaks intersected with above altered genes	771	448
Number of intersected peaks containing core CTCF motif within promoter	398	396
% CTCF motif	52%	88%

the input treated data set. Thus, the percentage of real peaks containing the CTCF core localized within promoters of regulated genes was improved >30% after input treatment. This analysis highlights the importance and necessity for being able to remove technique specific artifacts from ChIP-exo data sets and further demonstrates that our PAtCh-Cap derived ChIP-exo input control is able to perform this task remarkably well.

In the context of the input treated CTCF ChIP-exo data, analysis of the altered genes binned by the various motif combinations denoted in Figure 2.7A and 2.7D showed that there was a clear set of transcriptional alterations that were specific to each of these bins, including the core CTCF consensus in conjunction with the newly identified 3′-site (Figure 2.7B and 2.7E). As further evidence, analysis of the biological pathways associated with each of the binned gene sets demonstrated that several pathways were uniquely regulated by this newly identified extended CTCF motif (Figure 2.7C and 2.7F). While this analysis is not able to delineate direct versus indirect transcriptional regulation by CTCF, it does establish that CTCF occupation at sites containing the CTCF core in conjunction with the newly identified 3′ -flanking sequences independently modulates the transcriptional outcome of certain genes.

Here we have provided a simple and convenient method for generating technique specific input controls that affords increased confidence in peak identification and the ability to perform de novo motif searches for ChIP-based methods that utilize additional bead-bound processing steps. We anticipate that the presented PAtCh-Cap input control strategy will remove a barrier preventing advanced bead-bound ChIP-based techniques such as ChIP-exo from becoming mainstream, allowing them to be accessible to a broader range of proteins and increasing the level at which we can interrogate interesting

biological questions. Indeed, using our approach afforded identification of a novel-binding motif for the very well characterized protein CTCF that appears to have an independent physiological function; the significance of which will require further investigation.

References

(1) Johnson, D. S.; Mortazavi, A.; Myers, R. M.; Wold, B. Genome-Wide Mapping of In Vivo Protein-DNA Interactions. *Science* **2007**, *316*, 1497-1502.

(2) Park, P. J. ChIP-Seq: Advantages and Challenges of a Maturing Technology. *Nat. Rev. Genet.* **2009**, *10*, 669-680.

(3) Roberstson, G.; Hirst, M.; Bainbridge, M.; Bilenky, M.; Zhao, Y.; Zeng, T.; Euskirchen, G.; Bernier, B.; Varhol, R.; Delaney, A.; Thiessen, N.; Griffith, O. L.; He, A.; Marra, M.; Snyder, M.; Jones, S. Genome-Wide Profiles of STAT1 DNA Association Using Chromatin Immunoprecipitation and Massively Parallel Sequencing. *Nat. Methods* **2007**, *4*, 651-657.

(4) Adli, M.; Bernstein, B. E. Whole-Genome Chromatin Profiling from Limited Numbers of Cells Using Nano-ChIP-Seq. *Nat. Protoc.* **2011**, *6*, 1656-1668.

(5) Balakrishnan, L.; Milavetz, B. Dual Agarose Magnetic (DAM) ChIP. *BMC Res. Notes* **2009**, *2*, 250-257.

(6) Jakobsen, J. S.; Bagger, F. O.; Hasemann, M. S.; Schuster, M. B.; Frank, A.-K.; Waage, J.; Vitting-Seerup, K.; Porse, B. T. Amplification of Pico-Scale DNA Mediated by Bacerial Carrier DNA for Small-Cell-Number Transcription Factor ChIP-Seq. *BMC Genetics* **2014**, *16*, 46-56.

(7) Lara-Astiaso, D.; Weiner, A.; Lorenzo-Vivas, E.; Zaretsky, I.; Jaitin, D. A.; David, E.; Keren-Shaul, H.; Mildner, A.; Winter, D.; Jung, S.; Friedman, N.; Amit, I. Chromatin State Dynamics During Blood Formation. *Science* **2014**, *345*, 943-949.

(8) Sachs, M.; Onodera, C.; Blaschke, K.; Ebata, K. T.; Song, J. S.; Ramalho-Santos, M. Bivalent Chromatin Marks Devlopment Regulatory Genes in the Mouse Embryonic Germline In Vivo. *Cell Reports* **2013**, *3*, 1777-1784.

(9) Shankaranarayanan, P.; Mendoza-Parra, M.-A.; Walia, M.; Wang, L.; Li, N.; Trindade, L. M.; Gronemeyer, H. Single-Tube Linear DNA Amplification (LinDa) for Robust ChIP-Seq. *Nat. Methods* **2011**, *8*, 565-567.

(10) Shen, J.; Jiang, D.; Fu, Y.; Wu, X.; Guo, H.; Feng, B.; Pang, Y.; Streets, A. M.; Tang, F.; Huang, Y. H3K4me3 Epigenomic Landscape Derived from ChIP-Seq of 1000 Mouse Early Embryonic Cells. *Cell Res.* **2015**, *25*, 143-147.

(11) Zwart, W.; Koornstra, R.; Wesseling, J.; Rutgers, E.; Linn, S.; Carroll, J. S. A Carrier-Assisted ChIP-Seq Method for Estrogen Receptor-Chromatin Interactions from Breast Cancer Core Needle Biopsy Samples. *BMC Genomics* **2013**, *14*, 232.

(12) Schmidl, C.; Renderio, A. F.; Sheffield, N. C.; Bock, C. ChIPmentation: Fast, Robust, Low-Input ChIP-Seq for Histones and Transcription Factors. *Nat. Methods* **2015**, *12*, 963-965.

(13) Wallerman, O.; Nord, H.; Bysani, M.; Borghini, L.; Wadelius, C. LobChIP: from Cells to Sequencing Ready ChIP Libraries in a Single Day. *Epigenet. Chromat.* **2015**, *8*, 25-33.

(14) He, Q.; Johnston, J.; Zeitlinger, J. ChIP-Nexus Enables Improved Detection of In Vivo Transcription Factor Binding Footprints. *Nat. Biotechnol.* **2015**, *33*, 395-401.

(15) Rhee, H. S.; Pugh, B. J. Comprehensive Genome-Wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell* **2011**, *147*, 1408-1419.

(16) Skene, P. J.; Henikoff, S. A Simple Method for Generating High-Resolution Maps of Genome-Wide Protein Binding. *Elife* **2015**, *4*, e09225.

(17) Carroll, T. S.; Liang, Z.; Salama, R.; Stark, R.; de Santiago, I. Impact of Artifact Removal on ChIP Quality Metrics in ChIP-Seq and ChIP-Exo Data. *Front. Genet.* **2014**, *5*, 1-11.

(18) Chen, Y.; Negre, N.; Li, Q.; Mieczkowska, J. O.; Slattery, M.; Liu, T.; Zhang, Y.; Kim, T.-K.; He, H. H.; Zieba, J.; Ruan, Y.; Bickel, P. J.; Myers, R. M.; Wold, B. J.; White, K. P.; Lieb, J. D.; Liu, X. S. Systematic Evaluation of Factors Influencing ChIP-Seq Fidelity. *Nat. Methods* **2012**, *9*, 609-614.

(19) Kidder, B. L.; Hu, G.; Zhao, K. ChIP-Seq: Technical Considerations for Obtaining High Quality Data. *Nat. Immunol.* **2013**, *12*, 918-922.

(20) Landt, S. G.; Marinov, G. K.; Kundaje, A.; Kheradpour, P.; Pauli, F.; Batzoglou, S.; Bernstein, B. E.; Bickel, P.; Brown, J. B.; Cayting, P.; Chen, Y.; DeSalvo, G.; Epstein, C.; Fisher-Aylor, K. I.; Euskirchen, G.; Gerstein, M.; Gertz, J.; Hartemink, A. J.; Hoffman, M. M.; Iyer, V. R.; Jung, Y. L.; Karmaker, S.; Kellis, M.; Kharchenko, P. V.; Li, Q.; Liu, T.; Liu, X. S.; Ma, L.; Milosavljevic, A.; Myers, R. M.; Park, P. J.; Pazin, M. J.; Perry, M. D.; Raha, D.; Reddy, T. E.; Rozowsky, J.; Shoresh, N.; Sidow, A.; Slattery, M.; Stamatoyannopoulos, J. A.; Tolstorukov, M. Y.; White, K. P.; Xi, S.; Farnham, P. J.; Lieb, J. D.; Wold, B. J.; Snyder, M. ChIP-Seq Guidelines and Practices of the ENCODE and ModENCODE Consortia. *Gen. Res.* **2012**, *22*, 1813-1831.

(21) Nix, D. A.; Courdy, S. J.; Boucher, K. M. Empirical Methods for Controlling False Positives and Estimating Confidence in ChIP-Seq Peaks. *BMC Bioinformatics* **2008**, *9*, 523-531.

(22) Pepke, S.; Wold, B.; Mortazavi, A. Computation for ChIP-Seq and RNA-Seq Studies. *Nat. Methods* **2009**, *6*, S22-S32.

(23) Mahony, S.; Pugh, B. F. Protein-DNA Binding in High-Resolution. *Crit. Rev. Biochem. Mol. Biol.* **2015**, *50*, 269-283.

(24) Ho, J. W. K.; Bishop, E.; Karchenko, P. V.; Negre, N.; White, K. P.; Park, P. J. ChIP-ChIP Versus ChIP-Seq: Lessons for Experimental Design and Data Analysis. *BMC Genomics* **2011**, *12*, 134.

(25) Serandour, A. A.; Brown, G. D.; Cohen, J. D.; Carroll, J. S. Development of an Illumina-Based ChIP-Exonuclease Method Provides Insight Into FoxA1-DNA Binding Properties. *Genome Biol.* **2013**, *14*, R147.

(26) Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **2009**, *25*, 2078-2079.

(27) Feng, J.; Liu, T.; Qin, B.; Zhang, Y.; Liu, X. S. Identifying ChIP-Seq Enrichment Using MACS. *Nat. Protocols* **2012**, *7*, 1728-1740.

(28) Nicol, J. W.; Helt, G. A.; Blanchard, S. G. Jr.; Raja, A.; Loraine, A. E. The Integrated Genome Browser: Free Software for Distribution and Exploration of Genome-Scale Datasets. *Bioinformatics* **2009**, *25*, 2730-2731.

(29) Love, M. I.; Huber, W.; Anders, S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* **2014**, *15*, 550.

(30) The ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **2012**, *489*, 57-74.

(31) Dreszer, T. R.; Karolchik, D.; Zweig, A. S.; Hinrichs, A. S.; Raney, B. J.; Kuhn, R. M.; Meyer, L. R.; Wong, M.; Sloan, C. A.; Rosenbloom, K. R.; Roe, G.; Rhead, B.; Pohl, A.; Malladi, V. S.; Li, C. H.; Learned, K.; Kirkup, V.; Hsu, F.; Harte, R. A.; Guruvadoo, L.; Goldman, M.; Giardine, B. M.; Fujita, P. A.; Diekhans, M.; Cline, M. S.; Clawson, H.; Barber, G. P.; Haussler, D.; James, K. W. The UCSC Genome Browser Database: Extensions and Updates. 2011. *Nucleic Acids Res.* **2012**, *40*, D918-D923.

(32) Medina-Rivera, A.; Defrance, M.; Sand, O.; Herrmann, C.; Castro-Mondragon, J.; Delerce, J.; Jaeger, S.; Blanchet, C.; Vincens, P.; Caron, C.; Staines, D. M.; Contreras-Moreira, B.; Artufel, M.; Charbonnier-Khamvongsa, L.; Hernandez, C.; Thieffry, D.; Thomas-Chollier, M.; van Helden, J. RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res.* **2015**, *43*, W50-W56.

(33) Thomas-Chollier, M.; Defrance, M.; Medina-Rivera, A.; Sand, O.; Herrmann, C.; Thieffry, D.; van Helden, J. RSAT 2011: Regulatory Sequence Analysis Tools. *Nucleic Acids Res.* **2011**, *39*, 86-91.

(34) Thomas-Chollier, M.; Sand, O.; Turatsinze, J. V.; Janky, R.; Defrance, M.; Vervisch, E.; Brohee, S.; van Helden, J. RSAT: Regulatory Sequence Analysis Tools. *Nucleic Acids Res.* **2008**, *36*, W119-W127.

(35) Grant, C. E.; Bailey, T. L.; Noble, W. S. FIMO: Scanning for Occurrences of a Given Motif. *Bioinformatics* **2011**, *27*, 1017-1018.

(36) Whitington, T.; Frith, M. C.; Johnson, J.; Bailey. T. L. Inferring Transcription Factor Complexes from ChIP-Seq Data. *Nucleic Acids Res.* **2011**, *39*, e98.

(37) Machanick, P.; Bailey, T. L. MEME-ChIP: Motif Analysis of Large DNA Datasets. *Bioinformatics* **2011**, *27*, 1696-1697.

(38) Kliszczak, A. E.; Rainey, M. D.; Harhen, B.; Boisvert, F. M.; Santocanale, C. DNA Mediated Chromatin Pull-Down for the Study of Chromatin Replication. *Sci. Rep.* **2011**, *1*, 95-114.

(39) Sirbu, B.; Couch, F. B.; Feigerle, J. T.; Bhaskara, S.; Hiebert, S. W.; Cortez, D. Analysis of Protein Dynanmics at Active, Stalled, and Collapsed Replication Forks. *Genes Dev.* **2011**, *25*, 1320-1327.

(40) Zlatanova, J.; Caiafa, P. CTCF and Its Protein Partners: Divide and Rule? *J. Cell Sci.* **2009**, *122*, 1275-1284.

(41) Albert, I.; Wachi, S.; Jiang, C.; Pugh, F. B. GeneTrack - A Genomic Data Processing and Visualization Framework. *Bioinformatics* **2008**, *24*, 1305-1306.

(42) Wang, L.; Chen, J.; Wang, C.; Uuskula-Reimand, L.; Chen, K.; Medina-Rivera, A.; Young, E. J.; Zimmermann, M. T.; Yan, H.; Sun, Z.; ZHang, Y.; Wu, S. T.; Huang, H.; Wilson, M. D.; Kocher, J.-P. A.; Li, W. MACE: Model Based Analysis of ChIP-Exo. *Nucleic Acids Res.* **2014**, *42*, e156.

(43) Nakahashi, H.; Kwon, K.-R. K.; Resch, W.; Vian, L.; Dose, M.; Stavreva, D.; Hakim, O.; Pruett, N.; Nelson, S.; Yamane, A.; Qian, J.; Dubois, W.; Welsh, S.; Phair, R. D.; Pugh, B. F.; Lobanenkov, V.; Hager, G. L.; Casellas, R. A Genome-Wide Map of CTCF Multivalency Redefines the CTCF Code. *Cell Reports* **2013**, *3*, 1678-1689.

CHAPTER 3

LIMITATIONS OF CURRENT ChIP AND HIGH-THROUGHPUT SEQUENCING TECHNIQUES TO INVESTIGATE ZBTB33 GLOBAL GENOMIC OCCUPATIONS

Introduction

Numerous studies have shown that *in cell*, ZBTB33 (also known as Kaiso) binds both methylated CpG DNA sites and a sequence specific unmethylated DNA target, termed the Kaiso binding site (KBS).[1-6] However, a recent high-throughput ChIP-seq study of ZBTB33 genomic occupancy contradicted these findings.[7] Namely, ZBTB33 ChIP-seq analyses in two hematopoietic cell lines (GM12878 and K562) showed preferential binding to an unmethylated palindromic DNA sequence TCTCGCGAGA and to several known transcription factor binding motifs.[7] The inconsistency between the various studies demonstrating physiological ZBTB33 functions upon binding to methylated CpG and KBS and the observed ZBTB33 genomic occupations determined by the above mentioned ChIP-seq study highlight potential limitations in this method. It was thus thought that the low signal to noise ratio and low base pair resolution limitation of ChIP-seq may have reduced accurate detection of the minimal methylated ZBTB33 recognition sites within CpG islands (CGIs). Therefore, we hypothesized that ChIP-exo,[8]

a higher resolution high-throughput sequencing technique, would be valuable in mapping the global *in cell* ZBTB33 DNA occupancy sites and more specifically define the methylated DNA recognition motif.

Unlike ChIP-seq, ChIP-exo is able to locate transcription factor DNA binding motifs with single base pair precision. However, due to the lack of appropriate input control for removing persisting background signals specific to ChIP-exo, the method produces disproportionately far too many false positives making it disadvantageous for studying low occupancy transcription factors and transcription factors for which the recognition motif is not yet well defined, such as ZBTB33. For that reason, we developed a novel method termed PAtCh-Cap (Protein Attached Chromatin Capture), that generates a specific input control which significantly reduces false positives, increases confidence in peak identification and affords de novo motif searching analysis.[9]

Nonetheless, application of our new PAtCh-Cap method to ChIP-exo analysis of ZBTB33 in HeLa cells failed to identify ZBTB33-specific genomic occupations. Further investigations revealed fundamental issues regarding the ChIP step specific for proteins such as ZBTB33. First, it has been shown that the strength of the DNA:protein interaction after formaldehyde cross-linking varies dramatically from one transcription factor to another. Indeed, recent studies have demonstrated that the strongest covalent cross-linkage occurs between lysine primary amines and guanine bases.[10] Analysis of the structure of ZBTB33 in complex with its cognate methylated and KBS DNA targets showed that it does not have a binding interface lysine capable of forming such ideal cross-linking interactions.[11] Furthermore, it has been shown that the reversal of formaldehyde protein:DNA cross-links increases rapidly with increasing temperature.[12]

This is significant as the ChIP-exo protocol includes several high temperature incubation steps capable of reversing these cross-links. Therefore, we suggest that the unsuccessful ZBTB33 ChIP-exo attempts were due to premature reverse cross-linking as a result of insufficient protein:DNA cross-linking strength and prolonged high temperature exposure during the ChIP-exo procedure.

Unfortunately, applying the ChIP-seq technique to identify ZBTB33 global genomic occupation did not produce better results. Although this method does not include high temperature steps that would compromise the cross-link, as discussed above, it is unable to identify ZBTB33 specific genomic occupation sites. While the KBS and mCpG ZBTB33 targets can be detected through ChIP-qPCR,[3-4] ChIP-seq analysis appears to mainly enrich DNA motifs for other well-known transcription factors such as CTCF.[7] Thus, we hypothesize that ZBTB33 ChIP-seq may only report on indirect protein:protein:DNA interactions. Here we detail our efforts to identify the reasons for why ChIP-based methods do not work for transcription factors such as ZBTB33.

Materials and methods

Cell culturing

HeLa cells were cultured in DMEM supplemented with 4.5 g/L glucose, 2 mM glutamine, and 10% FBS and maintained at 37 °C under 5% CO_2. Cell counting and viability analysis was performed on a Countess Automated Cell Counter (Thermo Fisher Scientific). Cell line authentication was routinely verified by short tandem repeat (STR) DNA profiling.

Construction of Ad-ZBTB33-ZBTB33-HA vector and adenovirus infection of HeLa cells

The *ZBTB33* open reading frame (ORF), obtained from GenScript, was fused to a C-terminal HA-tag and transferred into the pcDNA3.1 mammalian expression vector (Thermo Fisher Scientific) utilizing standard cloning strategies. To put *ZBTB33* under control of its endogenous promoter, the original CMV promoter was replaced with a DNA fragment PCR-amplified from genomic human DNA encompassing 2 kb of DNA upstream of the *ZBTB33* transcription start site (TSS). Construction of the Ad-ZBTB33-HA vector and the viral lysate was prepared and titered by SignaGen Laboratories. HeLa cell monolayers were infected with the Ad-ZBTB33-HA construct at a multiplicity of infection (MOI) of 10 infectious units per cell for 16 h prior to preparation for additional experiments.

ChIP-seq of ZBTB33 in HeLa cells

For each duplicate, 20 x 10^6 HeLa cells with or without infection with the Ad-ZBTB33-HA vector, were fixed with 1% formaldehyde for 15 min to cross-link the protein:DNA complexes, followed by a quench with 125 mM glycine. Cells were washed with cold PBS (pH 7.4) and lysed for 10 min at 40 °C in cell lysis buffer (50 mM HEPES (pH 8.0), 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100). Cellular nuclei were then washed (10 mM Tris-HCl (pH 8.0), 200mM EDTA), centrifuged and resuspended in nuclear lysis buffer (50 mM Tris-HCl (pH 8.0), 100 mM NaCl, 10 mM EDTA, 1% SDS). A Diagenode Bioruptor Standard sonication device (run at max amplitude for 5 x 15 min in ice water) was used to shear the cross-linked DNA to

100-400 bp fragments. Cell debris was pelleted by centrifugation and supernatant containing the solubilized chromatin protein:DNA complexes were isolated. Prior to further treatment, 1% of the sheared chromatin sample volume was removed from each replicate for input sample preparation. For the IP step, Protein G coated magnetic beads were prefunctionalized with either ZBTB33 (6F8, Santa Cruz) or HA-tag antibodies prior to incubation with the sheared chromatin sample. Proteinase K digestion was performed overnight at 55 °C. DNA was purified with the MinElute PCR Purification Kit (Qiagen). DNA quantification, library construction, and sequencing using the Illumina HiSeq 2000 platform were all performed by the High-Throughput Genomics Core within the University of Utah Huntsman Cancer Institute.

Bioinformatics analyses

ChIP-seq fastq files were aligned to the human genome (hg19) using Novoalign (Novocraft, Inc.) and the following parameters: -0 SAM –r. Sam files were then sorted and indexed with SAMtools.[13] Peak calling was performed by MACS2[14] using the following standard parameters: model fold settings of 5 to 50 q-values with a cutoff of 0.05. Bedgraph outputs from MACS2 were viewed in the Integrated Genome Browser (IGB).[15] The MACS2 bdgcmp command was further performed to obtain the read coverage tracks.

ChIP-qPCR

HeLa cells were cross-link treated and chromatin was sheared as described above for the ChIP-seq samples. The cell lysis, chromatin isolation and immunoprecipitation

(IP) reaction procedures were performed using materials from the Zymo-Spin ChIP Kit (Zymo Research) following the manufacturer's instructions. The IP was carried out using an antibody against ZBTB33 (6F8, Santa Cruz). After quantitating the amount of genomic DNA (PicoGreen assay), equal amounts of ChIP amplicons for each sample were prepared with the SYBR Green Real-Time PCR Master Mix (Thermo Fisher Scientific) and analyzed by qPCR on a QuantStudio 6 Flex Real-Time PCR System (Thermo Fisher Scientific). The fold enrichment was determined based on the cycle differences after normalization to input DNA. A region from the ELL3 gene was amplified as a negative control.

Co-immunoprecipitation

HeLa cells were collected in ice-cold PBS and lysed in nondenaturing RIPA buffer (10 mM Tris-HCl (pH 7.2), 150 mM NaCl, 1% Triton X-100, 0.1% SDS, 1% sodium deoxycholic acid, and 5 mM EDTA) supplemented with an EDTA-free protease inhibitor cocktail (Roche). Protein lysates were cross-linked with dimethyl pimelimidate and incubated overnight at 4 °C with Protein G Dynabeads (Thermo Fisher Scientific) preconjugated with an antiZBTB33 antibody (6F8, Santa Cruz). Adequately washed bound material was subsequently eluted with 50 mM glycine (pH 2.8) and boiled for 10 min. Eluates were separated by gel electrophoresis, transferred to nitrocellulose membrane, and immunoblotted with ZBTB33 (Santa Cruz), CTCF (Millipore), or PARP1 (Rockland) antibodies to analyze the immunoprecipitants and input samples.

Results and discussion

Failure of ChIP-exo for analysis of ZBTB33 genomic occupations

The need for a high-resolution technique to identify minimal global ZBTB33 methylated CpG recognition sites within CpG islands suggested that ChIP-exo would be the ideal method. To achieve this level of resolution, the ChIP-exo method utilizes a lambda exonuclease to digest unprotected DNA around cross-linked protein:DNA complexes to map the protein binding DNA footprint.[8] Consequently, the nuclease digestion of background DNA yields high signal to noise ratio peaks. Thus, we utilized ChIP-exo in conjunction with our newly developed PAtCh-Cap method[9] to determine the genome-wide binding occupancies of ZBTB33. However, after ChIP-exo, high-throughput sequencing and extensive bioinformatics analysis, we observed only artifact peaks and no incidence of ZBTB33-specific genomic occupations. This outcome led us to further investigate why this method failed for ZBTB33.

One major way in which the ChIP-exo procedure differs from standard ChIP-seq[16,17] is that there are several chemical processing steps that occur on bead-bound chromatin samples, each of which requires incubation of the samples at higher temperatures ranging from 25 – 37 °C for up to 5 h. Recent reports have demonstrated that reversal of formaldehyde cross-links between protein and DNA is temperature dependent (Figure 3.1).[12] Further, there are only a couple of specific amino acid:nucleotide interactions that result in strong formaldehyde generated cross-linking interactions, with linkages between lysine and guanosine bases being the strongest (Figure 3.2).[10]

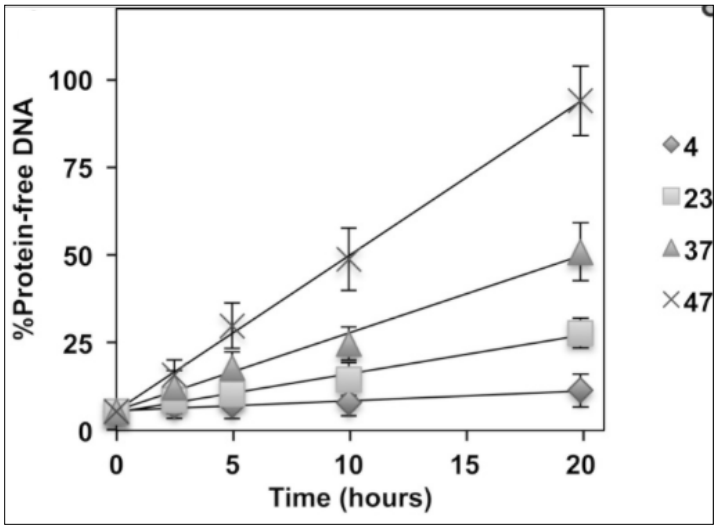


Figure 3.1. Temperature dependence of the formaldehyde mediated protein:DNA cross-link reversal rates. The percentage of protein-free DNA measured by qPCR is plotted versus time for samples at four different temperatures: 4, 23, 37, and 47 °C. The error bars are the standard deviation of qPCR signals. A best fit linear regression is shown for each temperature trace. Adapted from (12).

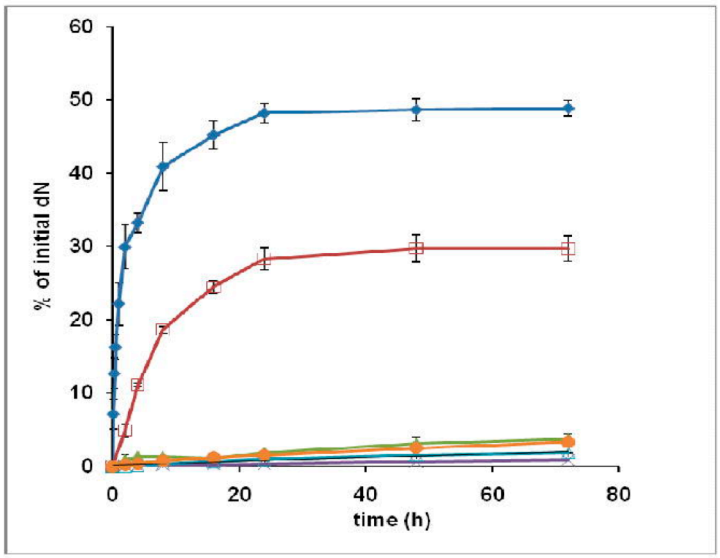


Figure 3.2. Time dependent formation of cross-links between amino acids and nucleosides after formaldehyde treatment. Formation of cross-links between 4 mM amino acids and 4 mM nucleosides treated with 50 mM formaldehyde at various times. ♦, sum of Lys products; □, Cys-CH_2-dG; ▲, Cys-CH_2-dA; ×, Cys-CH_2-dC; △, His-CH_2-dA; and ●, Trp-CH_2-dG. Relative yields are based on integration of HPLC peak areas in UV traces recorded at 254 nm. Adapted from (10).

Thus, we utilized the high-resolution structure of ZBTB33 bound to both its methylated and sequence-specific unmethylated DNA targets[11] to determine whether ZBTB33 has the ability to form the strongest lysine:guanine cross-links (Figure 3.3). As suspected, ZBTB33 is not capable of forming these strong formaldehyde cross-links. Similar analyses on well-known protein structures with successful ChIP-exo and ChIP-seq data sets reveals that they all are able to make these idealized lysine:guanosine formaldehyde cross-links.[8] Thus, it is likely that the inability of ZBTB33 to form these strong cross-links with its DNA targets makes it susceptible to premature cross-link reversal caused by the numerous high temperature incubation steps in the ChIP-exo procedure, resulting in a loss of the ZBTB33-specific genomic occupation sites.

Additional issues with analysis of ZBTB33 genomic occupations by ChIP-seq

The discrepancies between various reports on the physiological functions of ZBTB33 mediated by both *in cell* KBS and methyl-CpG site recognition and the observed lack of genomic occupations at these sites by ChIP-seq led us to investigate this further. More specifically, while ChIP-qPCR studies report that ZBTB33 binds to methylated DNA and KBS sequences[3-4] ChIP-seq data from two cell lines (GM12878 and K562) showed only preferential bindings to an unmethylated palindromic DNA sequence TCTCGCGAGA and several sequences belonging to other transcription factors (Figure 3.4).[7] Thus, we sought to evaluate whether these inconsistencies originated from differences in the cell lines used or the quality of the ChIP-grade ZBTB33 antibody.

To address these questions, we performed ChIP-seq in another cell line (HeLa) and compared results in these cells using the commercially available ChIP-grade

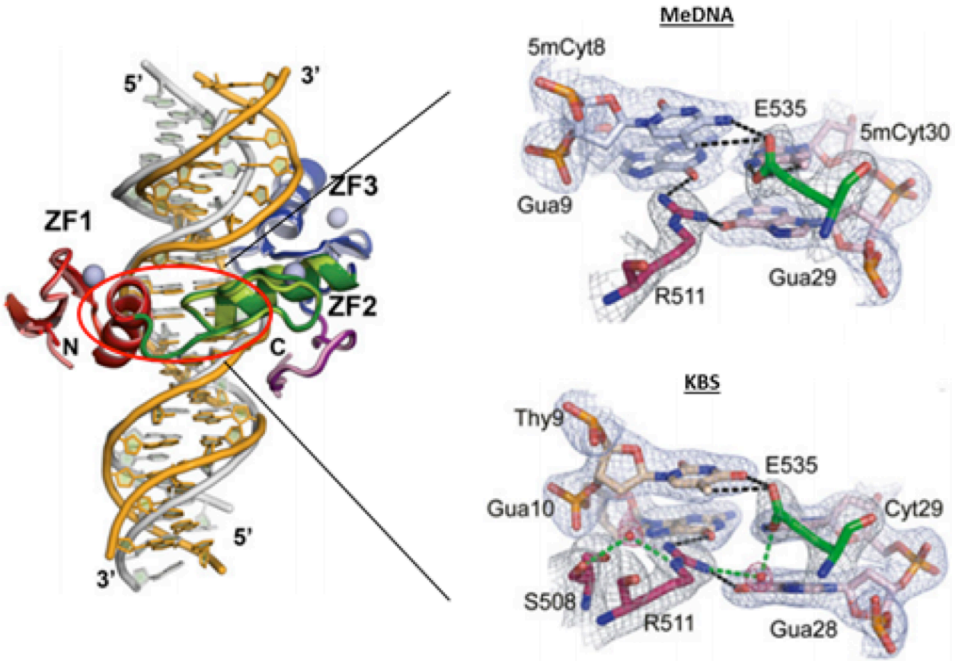


Figure 3.3. Crystal structures of ZBTB33 in complex with both its methylated and sequence specifc nonmethylated DNA targets. Crystal structures of ZBTB33 in complex with both its cognate methylated CpG (dark protein colors, light blue DNA) and the KBS sequence (light protein colors, yellow DNA). These structures highlight the lack of lysine and cysteine residues at the DNA binding interface capable of cross-linking with a guanosine bases. Adapted from (11).

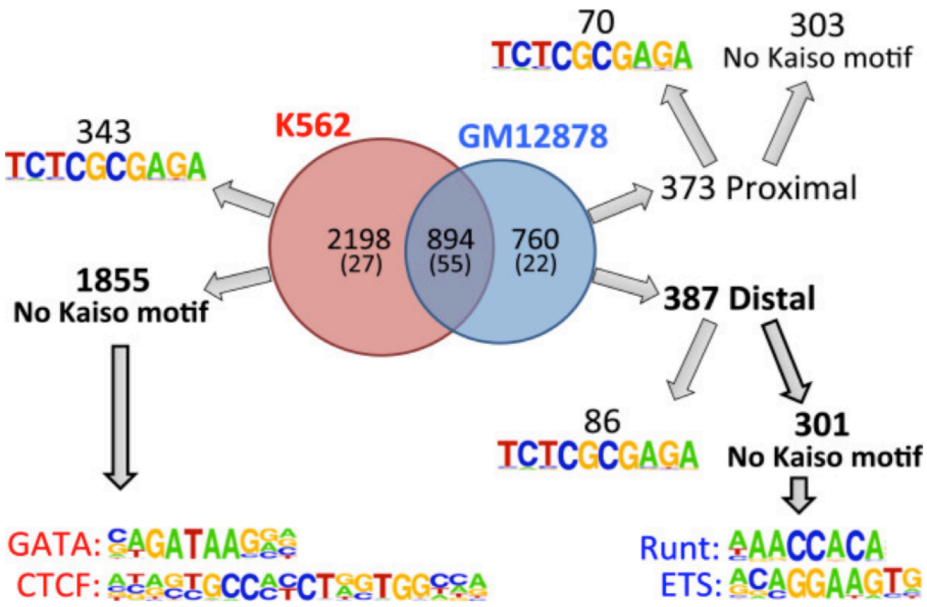


Figure 3.4. Motif analyses of cell type-specific ZBTB33 (Kaiso) binding sites. ZBTB33 peaks from GM12878 and K562 were compared to identify common and cell type-specific binding sites (the median tag height is shown in parentheses). GM12878- and K562-unique peaks were divided into two sets of peaks, those that contain and those that lack a ZBTB33 motif; GM12878-unique peaks were further separated into promoter proximal and promoter distal sets. A motif analysis was then performed for each set of peaks, identifying motifs for the GATA and CTCF families of transcription factors in the K562-unique peaks and motifs for the ETS and Runt families of transcription factors in distal GM12878 peaks. Adapted from (7).

antibody to IP endogenous ZBTB33 or an HA ChIP-grade antibody to IP C-terminally tagged ZBTB33 (ZBTB33-HA). Once prepared, we first utilized ChIP-qPCR to analyze whether ZBTB33 occupied known target sites prior to submitting these samples for library preparation and high-throughput sequencing. Similar to previous reports,[1,3] ZBTB33 clearly demonstrated genomic occupations at both KBS and methylated-CpG containing DNA sites (Figure 3.5A).

Nonetheless, ChIP-seq analysis on these same samples demonstrated that the ChIP-seq results with both endogenous and HA-tagged ZBTB33 recapitulated previously reported findings (Figure 3.6),[7] indicating that the ZBTB33 ChIP-grade antibody does not participate in a high level of nonspecific interactions and that the lack of KBS and methyl-CpG occupation is not cell type specific. Surprisingly, ChIP-qPCR analysis of the ChIP-seq samples after library preparation showed that there was no longer genomic occupation of the sites observed prior to preparation of the library (Figure 3.6B). This finding is significant and suggests that the inability to identify ZBTB33 genomic occupation sites may result from a loss of this information during one or more of the library preparation steps.

It is worth noting that unlike ChIP-exo, all of the ChIP-seq data sets for ZBTB33 clearly show highly enriched occupation sites within numerous genes. Yet, advanced de novo motif searches within these data sets do not identify known ZBTB33 DNA targets.[7] Intriguingly, these motif-searching analyses do identify consensus sequences for the known ZBTB33 binding partner CTCF.[18,19] Further, while the unmethylated palindromic sequence TCTCGCGAGA has been attributed to the genomic target of ZBTB33, recent reports have suggested that this recognition sequence may actually belong to PARP1.[20]

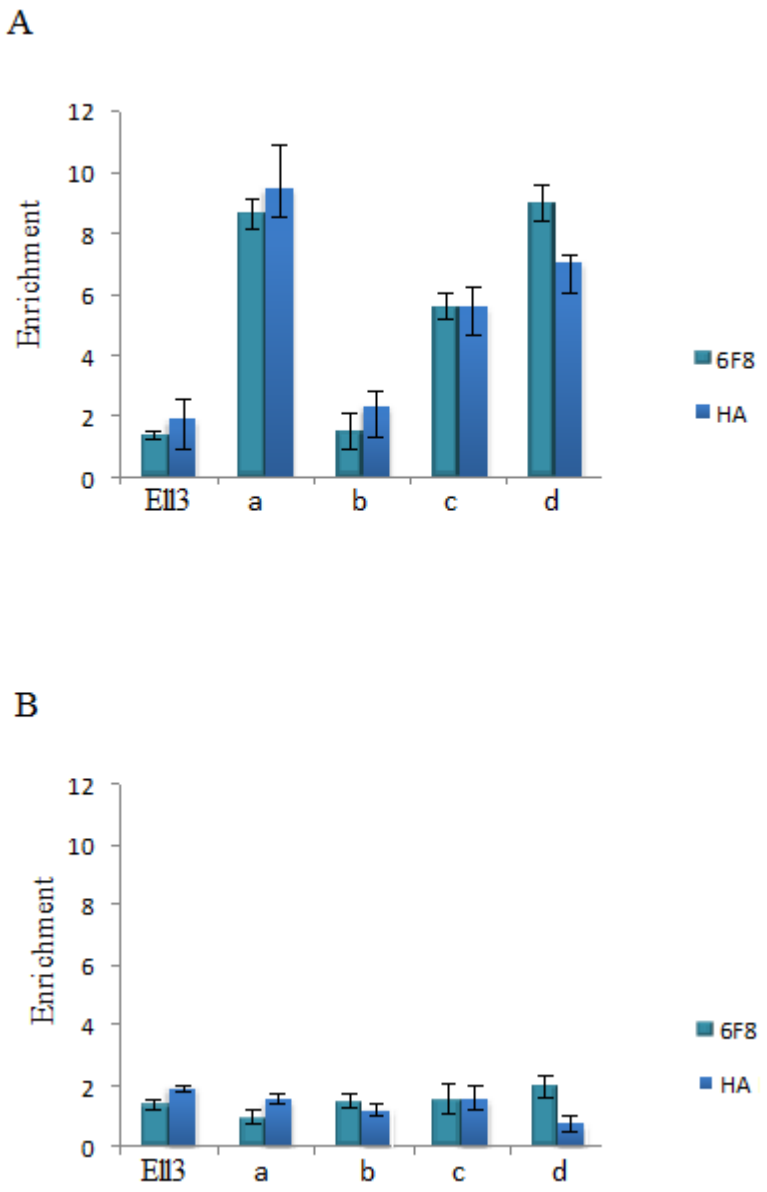


Figure 3.5. ZBTB33 ChIP-qPCR at multiple promoter sites within the CCND1 and CCNE1 gene promoters. (A) qPCR amplifications before the ChIP-seq DNA library preparation. (B) qPCR amplifications after the ChIP-seq DNA library preparation. Letters (a, b, c, and d) after CCND1 and CCNE1 refer to promoter locations as described in Figure 4.5.

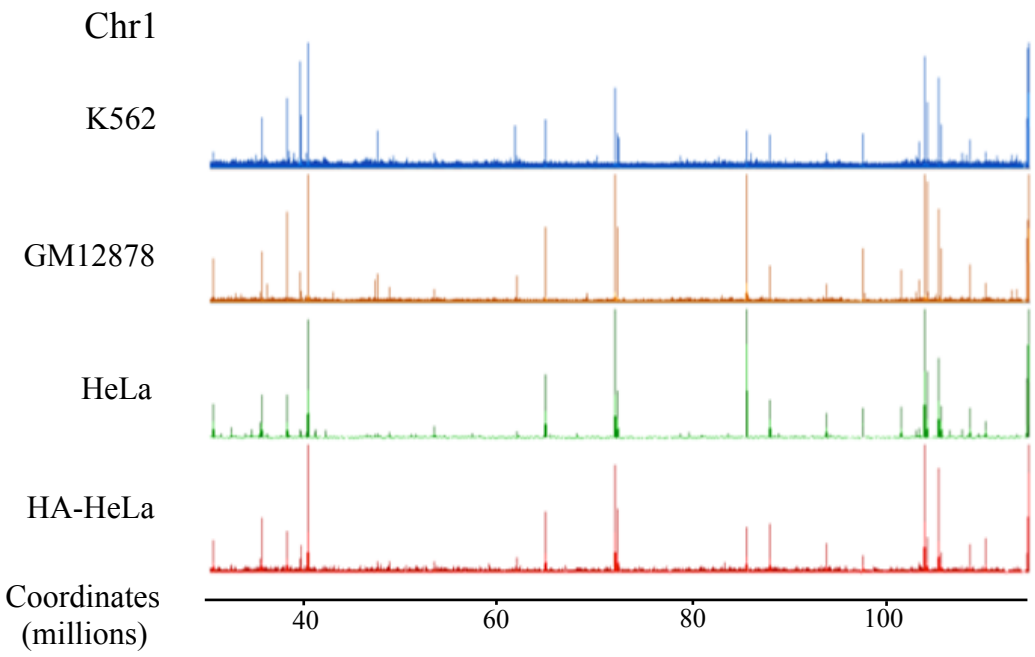


Figure 3.6. Representative ZBTB33 and ZBTB33-HA ChIP-seq profiles at chromosome 1. ZBTB33 ChIP-seq profiles at chromosome 1 (Chr 1) in three cell lines: K562, GM12878, and HeLa. Additionally, ChIP-seq for ZBTB33-HA was performed using an HA-specific antibody in HeLa cells (HA-HeLa).

Indeed, there are a number of reports that indicate that PARP1 exhibits sequence-specific DNA binding capability.[20,21] Together, these observations suggest that perhaps the ZBTB33 ChIP-seq data sets are reporting on indirect genomic occupations through ZBTB33-specific protein interaction partners and not on direct ZBTB33 genomic occupation sites.

To investigate this, we first confirmed through co-immunoprecipitation (co-IP) analysis that in HeLa cells, ZBTB33 does form a complex with CTCF (Figure 3.7A). This is consistent with previous reports that show a direct interaction between ZBTB33 and CTCF.[22] Further, many reports indicate that CTCF and PARP1 work together to mediate DNA methylation status and chromatin organization.[23,24] Thus, given that CTCF is a binding partner for both ZBTB33 and PARP1, we performed a co-IP to see if ZBTB33 and PARP1 associate. Indeed, we show for the first time that ZBTB33 and PARP1 do associate within the same protein complex (Figure 3.7B).

Conclusion

In our attempts to determine the global genomic occupations of ZBTB33 in various cell lines through ChIP-based high-throughput sequencing methods, we have identified two major issues. First, the lack of strong formaldehyde generated protein:DNA cross-links in ZBTB33:DNA complexes limits the utility of ChIP-exo due to the numerous high temperature processing steps that result in premature cross-link reversal. Second, it also appears that there may be steps during the ChIP-seq library preparation that reduce down-stream sequencing amplification of ZBTB33 genomic occupation sites.

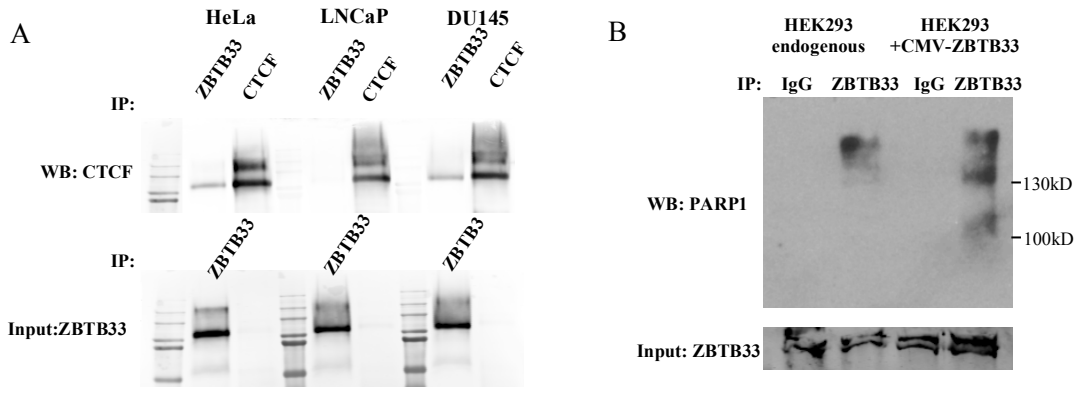


Figure 3.7. Co-immunoprecipitations of ZBTB33 with PARP1 and CTCF. (A) Co-immunoprecipitation of ZBTB33 and CTCF in HeLa, LNCaP, and DU145 cells. (B) Co-immunoprecipitation of ZBTB33 and PARP1 in HEK93 cells using either endogenous ZBTB33 levels or overexpression of ZBTB33.

Unfortunately, the protein:DNA cross-linking issue for proteins such as ZBTB33 will be difficult to resolve. It appears that this protein represents a group of transcription factor that does not harbor the capability of forming the high-affinity formaldehyde mediated protein:DNA cross-links. Numerous cross-linking agents exist for protein:DNA interactions, however, only formaldehyde has been shown to be amenable to ChIP procedures.[25] Thus, there is an urgent need to identify alternative protein:DNA cross-linking agents that can be utilized in ChIP-based experiments. Furthermore, in order to investigate the technical problems caused during library preparation, a step-by-step analysis of this process will be required.

References

(1) Daniel, J. M.; Spring, C. M.; Crawford, H. C.; Reynolds, A. B.; Baig, A. The p120ctn-Binding Partner Kaiso is a Bi-Modal DNA-Binding Protein that Recognizes Both a Sequence-Specific Consensus and Methylated CpG Dinucleotides. *Nucleic Acids Res.* **2002**, *30*, 2911-2919.

(2) Filion, G. J. P.; Zhenilo, S.; Salozhin, S.; Yamada, D.; Prokhortchouk, E.; Defossez, P-A. A Family of Human Zinc Finger Proteins that Bind Methylated DNA and Repress Transcription. *Mol. Cell. Biol.* **2006**, *26*, 169-18.

(3) Donaldson, N. S.; Pierre, C. C.; Antsey, M. I.; Robinson, S. C.; Weerawardane, S. M.; Daniel, J. M. Kaiso Represses the Cell Cycle Gene Cyclin D1 Via Sequence-Specific and Methyl-CpG-Mependent Mechanisms. *PLoS One* **2012**, *7*, e50398.

(4) Kim, S. W.; Park, J. I.; Spring, C. M.; Sater, A. K.; Ji, H.; Otchere, A. A.; Daniel, J. M.; McCrea, P. D. Non-Canonical Wnt Signals are Modulated by the Kaiso Transcriptional Repressor and p120-Catenin. *Nat. Cell Biol.* **2004**, *6*, 1212-1220.

(5) Kelly, K. F.; Otchere, A. A.; Graham, M.; Daniel, J. M. Nuclear Import of the BTB/POZ Transcriptional Regulator Kaiso. *J. Cell Sci.* **2004**, *117*, 6143-6152.

(6) Kelly, K. F.; Spring, C. M.; Otchere, A. A.; Daniel, J. M. NLS-Dependent Nuclear Localization of p120ctn is Necessary to Relieve Kaiso-Mediated Transcriptional Repression. *J. Cell Sci.* **2004**, *117*, 2675-2686.

(7) Blattler, A.; Yao, L.; Wang, Y.; Ye, Z.; Jin, V. X.; Farnham, P. J. ZBTB33 Binds Unmethylated Regions of the Genome Associated with Actively Expressed Genes. *Epigenet. Chromat.* **2013**, *6,* 18-25.

(8) Rhee, H. S.; Pugh, B. J. Comprehensive Genome-Wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell* **2011**, *147,* 1408-1419.

(9) Terooatea, W. T.; Pozner, A.; Buck-Koehntop, B. A. PAtCh-Cap: Input Strategy for Improving Analysis of ChIP-Exo Data Sets and Beyond. *Nucleic Acids Res.* **2016**, *44*, 21, e159.

(10) Lu, K.; Ye, W.; Zhou, L.; Collins, L. B.; Chen, X.; Gold, A.; Ball, L. M.; Swenberg, J. A. Structural Characterization of Formaldehyde-Induced Cross-Links Between Amino Acids and Deoxynucleosides and Their Oligomers. *J. Am. Chem. Soc.* **2010**, 1*32*, 3388-3399.

(11) Buck-Koehntop, B. A.; Stanfield, R. L.; Ekiert, D. C.; Martinez-Yamout, M. A.; Dyson, H. J.; Wilson, I. A.; Wright, P. E. Molecular Basis for Recognition of Methylated and Specific DNA Sequences by the Zinc Finger Protein Kaiso. *Proc. Natl. Acad. Sci.*

2012, 1*09,* 15229-15234.

(12) Kennedy-Darling, J.; Smith, L. M. Measuring the Formaldehyde Protein-DNA Cross-Link Reversal Rate. *Anal. Chem* **2014**, *86*, 5678-5681.

(13) Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **2009**, *25*, 2078-2079.

(14) Feng, J.; Liu, T.; Qin, B.; Zhang, Y.; Liu, X. S. Identifying ChIP-Seq Enrichment Using MACS. *Nat. Protocols* **2012**, *7,* 1728-1740.

(15) Nicol, J. W.; Helt, G. A.; Blanchard, S. G. Jr.; Raja, A.; Loraine, A. E. The Integrated Genome Browser: Free Software for Distribution and Exploration of Genome-Scale Datasets. *Bioinformatics* **2009**, *25*, 2730-2731.

(16) Chen, Y.; Negre, N.; Li, Q.; Mieczkowska, J. O.; Slattery, M.; Liu, T.; Zhang, Y.; Kim, T.-K.; He, H. H.; Zieba, J.; Ruan, Y.; Bickel, P. J.; Myers, R. M.; Wold, B. J.; White, K. P.; Lieb, J. D.; Liu, X. S. Systematic Evaluation of Factors Influencing ChIP-Seq Fidelity. *Nat. Methods* **2012**, *9,* 609-614.

(17) Kidder, B. L.; Hu, G.; Zhao, K. ChIP-Seq: Technical Considerations for Obtaining High Quality Data. *Nat. Immunol.* **2013**, *12,* 918-922.

(18) Zlatanova, J.; Caiafa, P. CTCF and Its Protein Partners: Divide and Rule? *J. Cell Sci.* **2009**, *122,* 1275-1284.

(19) Nakahashi, H.; Kwon, K.-R. K.; Resch, W.; Vian, L.; Dose, M.; Stavreva, D.; Hakim, O.; Pruett, N.; Nelson, S.; Yamane, A.; Qian, J.; Dubois, W.; Welsh, S.; Phair, R. D.; Pugh, B. F.; Lobanenkov, V.; Hager, G. L.; Casellas, R. A Genome-Wide Map of CTCF Multivalency Redefines the CTCF Code. *Cell Reports* **2013**, *3,* 1678-1689.

(20) Mikula, M.; Gaj, P.; Dzwonek, K.; Rubel, T.; Karczmarski, J.; Paziewska, A.; Dzwonek, A.; Bragoszewski, P.; Dadlez, M.; Ostrowski, J. Comprehensive Analysis of the Palindromic Motif TCTCGCGAGA: A Regulatory Element of the HNRNPK Promoter. *DNA Res.* **2010**, *17,* 245–260.

(21) Guo, G.; Rödelsperger, C.; Digweed, M.; Robinson, P. N. Regulation of Fibrillin-1 Gene Expression by Sp1. *Gene* **2013**, *527*. 448-455.

(22) Defossez, P. A.; Kelly, K. F.; Filion, G. J.; Pérez-Torrado, R.; Magdinier, F.; Menoni, H.; Nordgaard, C. L.; Daniel, J. M, Gilson, E. The Human Enhancer Blocker CTC-Binding Factor Interacts with the Transcription Factor Kaiso. *J. Biol. Chem.* **2005**, *52*, 43017-43023.

(23) Guastafierro, T.; Cecchinelli, B.; Zampieri, M.; Reale, A.; Riggio, G.; Sthandier, O.; Zupi, G.; Calabrese, L.; Caiafa P. CCCTC-Binding Factor Activates PARP-1 Affecting DNA Methylation Machinery. *J. Biol. Chem.* **2008**, *283,* 21873–21880.

(24) Wacker, D. A.; Ruhl, D. D.; Balagamwala, E. H.; Hope, K. M.; Zhang, T.; Kraus, W. L. The DNA Binding and Catalytic Domains of Poly(ADP-Ribose) Polymerase 1 Cooperate in the Regulation of Chromatin Structure and Transcription. *Mol. Cell Biol.* **2007**, *27,* 7475-7485.

(25) Hoffman, E. A.; Frey, B. L.; Smith, L. M.; Auble, D. T. Formaldehyde Crosslinking: A Tool for the Study of Chromatin Complexes. *J. Biol. Chem.* **2015**, *290*, 26404-26411.

CHAPTER 4

CELL SPECIFIC KAISO (ZBTB33) REGULATION OF CELL CYCLE THROUGH *CYCLIN D1* AND *CYCLIN E1*

Introduction

Cytosine methylation in the CpG context (mCpG) is a prevalent and essential epigenetic modification required for the maintenance of genomic stability, control of gene expression, and the regulation of chromatin structure. Misappropriation of genomic DNA methylation patterns has consequently been associated with cancer promotion and progression.[1,2] Indeed, it has become increasingly evident that nearly all cancers exhibit aberrant alterations in DNA methylation preceding tumorigenesis. Further, specific patterns of gene hyper-methylation leading to inappropriate transcriptional regulation appear to vary between tumor type and grade.[3] In gene promoters, DNA methylation elicits transcriptional regulation by either preventing or promoting transcription factor (TF) binding. Specifically, mCpG sites can be high-affinity targets for specialized TFs, termed methyl-CpG binding proteins (MBPs), which mediate chromatin remodeling and transcriptional control.[4]

ZBTB33 (also known as Kaiso), is the founding member of the zinc finger (ZF) family of MBPs and has been shown to regulate gene expression that is causally linked to

---

tumorigenesis.[5-11] Importantly, ZBTB33 and its two orthologs, ZBTB4 and ZBTB38, are unique amongst other MBPs in that they exhibit bimodal DNA recognition, specifically targeting both methylated and sequence-specific nonmethylated DNA sites.[12-14] Intriguingly, ZBTB33 has been shown to recognize methylated DNA and its sequence-specific site, termed the Kaiso binding site (KBS; TCCTGCNA), utilizing the same set of three Cys_2His_2 ZFs.[12,15] Further, this bimodal DNA recognition has afforded ZBTB33 the capability of functioning as both a transcriptional repressor and activator depending on the sequence context and cellular phenotype.[16-23]

There is mounting evidence that increased ZBTB33 expression and corresponding transcriptional activities are associated with cancer progression. In colon cancer cells, methyl-dependent repression of *CDKN2A* by ZBTB33 protects these cells from cell cycle arrest.[19] Similarly, *ZBTB33*-null mice crossed with $APC^{MIN/+}$ mice demonstrated resistance to intestinal cancer, implying a direct correlation between the presence of ZBTB33 and the disease state.[9] ZBTB33 also displays a dynamic subcellular localization mediated by its binding partner p120 catenin ($p120^{ctn}$)[24,25] that appears to play a role in its disease inducing potential. Specifically, a cytoplasmic-to-nuclear shift has been directly linked to the more aggressive phenotypes of breast and prostate cancer (PCa) patient tumors.[5-7,10] Indeed, a high nuclear ZBTB33 presence correlated with the pathologies, histology, and grading in invasive cohorts of high-grade and basal/triple-negative breast cancers.[5,7,10] Similarly, in PCa an increase in ZBTB33 protein expression and nuclear localization correlated with higher tumor grades and Gleason scores, with the highest levels of ZBTB33 being found in metastatic PCas.[6] However, the subcellular localization of ZBTB33 appears to be variable depending on the tumor type, as the presence of

cytoplasmic ZBTB33 has been linked to poor prognosis in nonsmall cell lung cancer[26,27] and pancreatic cancer.[28] Combined, these findings suggest that various cancer types are able to modulate the expression levels, cellular localizations, and transcriptional responses of ZBTB33 to differentially provide a context specific survival advantage. Nonetheless, there is little mechanistic insight into how various cellular phenotypes harness the transcriptional capabilities of ZBTB33 to differentially promote and progress the disease state.

Interestingly, a significant number of findings have implicated ZBTB33 in the regulation of cellular proliferation. *ZBTB33*-knockout mice exhibited increased body weight and size, due to splenomegaly resulting from increased splenocyte proliferation,[18] and a dramatic reduction of lateral ventricles indicative of increased embryonic neuronal stem cell proliferation.[28] Conversely, the small intestinal crypt of transgenic mice over-expressing ZBTB33 exhibited decreased cell proliferation.[29] Further, various ZBTB33 depletion studies have shown a consequential enhancement of cellular proliferation in lung carcinomas (BE1, LTEP-A-2, SPC-A-1),[26] HCT 116 cell colon carcinomas,[16] SK-LMS-1 vulva leiomyosarcoma cells,[30] HEK293 embryonic kidney fibroblasts,[31] and K562 blast crisis chronic myeloid leukemia cells when additionally depleted of p120^{ctn}.[32] In contrast, various lines of evidence have also demonstrated a proproliferative function for ZBTB33. Indeed, ZBTB33 depletion sensitizes Colo320 and HCT 116 colon cancer cell lines to cell cycle arrest after release from serum starvation[19] and induces decreased cellular proliferation in PC3 PCa cells.[11]

Given the evident role for ZBTB33 in regulating cellular proliferation in cancer, we initiated studies to mechanistically interrogate the differential cell cycle responses

mediated by the transcriptional activities of ZBTB33 in two different cell lines, HeLa and HEK293, both of which have been used extensively for studies of the cell cycle. Collectively, our data demonstrate that ZBTB33 transcriptionally regulates the G1-phase transition, though ZBTB33 acts as a proproliferative factor in HeLa cells, and an antiproliferative in HEK293 cells. Specifically, we have determined that ZBTB33 directly occupies the promoter regions of cyclin D1 and cyclin E1 in a KBS and methyl-specific manner, respectively, to enhance cyclin expression in HeLa cells. This ensures appropriate Retinoblastoma (RB1) phosphorylation and E2F transcriptional activity, facilitating an accelerated G1- to S-phase transition. In contrast, in HEK293 cells ZBTB33 indirectly regulates cyclin E abundance resulting in reduced RB1 hyper-phosphorylation leading to decreased E2F activity and a decelerated transition through G1-phase.

Experimental procedures

Cell culturing

HeLa and HEK293 cells were cultured in DMEM supplemented with 4.5 g/L glucose, 2 mM glutamine, and 10% FBS and maintained at 37°C and 5% CO2. Cell counting and viability analysis was performed on a Countess Automated Cell Counter (Thermo Fisher Scientific). Cell line authentication to confirm lack of cross-contamination was routinely verified by short tandem repeat (STR) DNA profiling.

Plasmids and luciferase reporter assay

The *ZBTB33* open reading frame (ORF), obtained from GenScript, was transferred into the pcDNA3.1 mammalian expression vector (Thermo Fisher Scientific); pCMV-ZBTB33. To put *ZBTB33* under control of its endogenous promoter, the original CMV promoter was replaced with a DNA fragment PCR amplified from genomic human DNA encompassing 2 Kb of DNA upstream of the *ZBTB33* transcription start site (TSS); pZBTB33^{2K}-ZBTB33. To generate the *cyclin D1*-luciferase reporter plasmids, an 1150 bp DNA fragment upstream of the *cyclin D1* TSS was PCR amplified from genomic human DNA and cloned into the pGL3-Basic Vector (Promega). Mutation of the -1067 KBS was completed utilizing standard cloning strategies. The E2F transcription reporter, 6xE2F-luciferase[33] was co-transfected with *ZBTB33* siRNAs or overexpressing vectors. The Dual-Luciferase Reporter Assays (Promega) were performed 48 h after transfection according to manufacturer's directions and measured on a luminometer. A CMV-Renilla construct was used for normalizing the efficiency of the transfection.

RNA interference and RNA/plasmid transfection

For all ZBTB33 depletion experiments, HeLa or HEK293 cells were transfected with either 30 pmol per well (6-well plate) of a scrambled siRNA or one of two siRNAs designed against the ZBTB33 transcript (Thermo Fisher Scientific) using Lipofectamine RNAiMAX (Thermo Fisher Scientific). Depending on the experiment, RNA or protein was extracted from the cells at either 24 or 48 h after transfection as detailed below. For mRNA transfection, 30 pmol of GFP mRNA (TRILINK Biotech.) was transfected into a 6-well plate using Lipofectamine RNAiMAX. Cells were collected 48 h after transfection

and GFP expression was followed by FACS analysis. Plasmid transfections in HeLa cells were performed by either using the Neon Transfection System (Thermo Fisher Scientific) following the manufacturer's protocol (2x pulse voltage 1005 V, pulse width 35 ms) or Lipofectamine 3000 (Thermo Fisher Scientific). For HEK293 cells, plasmid transfections were completed either utilizing a calcium phosphate procedure, as described previously,[33] or Lipofectamine 3000 (Thermo Fisher Scientific).

Immunoblotting

Cells were collected in ice-cold PBS and then lysed and sonicated in 1x RIPA buffer (10 mM Tris HCl (pH 7.20), 150 mM NaCl, 1% Triton X-100, 0.1% SDS, 1% sodium deoxycholic acid, and 5 mM EDTA) supplemented with an EDTA-free protease inhibitor cocktail (Roche). Samples were separated by gel electrophoresis, transferred to nitrocellulose or PVDF membranes and immunoblotted.

Antibodies (Abs)

For primary antibodies, the following were used: β-actin (Santa Cruz, sc-47778), Cyclin D1 (Cell signaling, 2926), Cyclin E (Thermo Fisher, MA5-14336), E2F1 (ABGENT, AP7593b), E2F2 (Assay Biotechnology, C11025), E2F3 (ABGENT, AP14598c), E2F4 (BETHYL Lab, A302-134A), E2F5 (Elabscience, ENT1445), PHH3 (Millipore, 06-570), RB1 (BD Biosciences, 554136), RB1 pT821 (Abcam, ab32015), RB1 pS780 (Abcam, ab173289), ZBTB33 (Santa Cruz, sc-23871). For secondary antibodies, the following were used: goat anti-rabbit DyLight 800 (Rockland), goat antimouse IRDye 680LT (LI-COR), goat antimouse IgG (H+L) Alexa Fluor 488

(Thermo Fisher, A-11001), goat anti-rabbit IgG (H+L) Alexa Fluor 546 (Thermo Fisher, A-11035).

Semiquantitative RT-PCR (sq-RT-PCR)

Transfected cells were washed with PBS. Total RNA was isolated by TRIzol (Thermo Fisher Scientific) and reverse transcribed using the high capacity cDNA reverse transcription kit (Thermo Fisher Scientific) following the manufacturer's protocols. The PCR products were analyzed on a 2% agarose gel. Relevant bands were excised, purified, and sequenced for validation. A primer list is available in Table 4.1.

Immunofluorescence, microscopy, and quantitative fluorescence intensity analysis

Cells were fed with 10 mM ethynyl deoxyuridine (EdU) under regular growth conditions for 30 min and then fixed in 3.7% ice cold formaldehyde in PBS for 15 min followed by a 5-min incubation with 100% MeOH at -30°C. Cells were permeabilized with PBST (0.5% Triton X-100 in PBS) for 20 min at room temperature (RT) before S-phase detection using the Click-iT EdU Alexa Fluor 488 Imaging kit (Molecular Probes), following the manufacturer's protocol. Subsequent to adequate washings in 3% BSA in PBS, cells were blocked with PBST containing 10% NGS for 1 h at 4°C. Primary Abs were incubated overnight at 4°C in PBST with 5% NGS and secondary Abs were incubated in the same buffer for 1 h at RT (Table 4.2 for antibody list). Images were acquired using a Zeiss Axiovert 200M microscope and captured with SlideBook software (Intelligent Imaging Innovations). Quantifications of ZBTB33 mean fluorescence intensity was performed on two randomly collected fields per cell line using ImageJ

Table 4.1. Primers used for semiquantitive RT-PCR

Gene	Sense	Antisense
β-actin	agctacgagctgcctgacgg	gatccacacggagtacttgcg
cyclin D1	gtgcagaaggaggtcctgcc	gcttgttcaccaggagcagc
cyclin E1	gaggaaggcaaacgtgaccg	gctcaagaaagtgctgatccc
ZBTB33	tggagcgacgtttaaagaaggg	ctgaaagaatattcttgtgagcc

Table 4.2. Primers used for ChIP-qPCR

Gene	Sense	Antisense
CCND1 -1067* (Figure 4.5, A and B; a)	tttacatctgcttaagtttgcg	ttagaatttgccctgggact
CCND1 +69* (Figure 4.5, A and B; b)	cacacggactacaggggagtt	ctcggctctcgcttctgctg
CCNE1 (Figure 4.5, D and E; c)	cttggtccaggcactatgcc	ggctccacaggacctgaccc
CCNE1 (Figure 4.5, D and E; d)	gggtcaggtcctgtggagcc	cccgcaagcccttcctggc
CCNE1 (Figure 4.5, D and E; e)	cagaaaggtcttcagagagcc	cgccaggcacgccctccc
CCNE1 (Figure 4.5, D and E; f)	cgccgtgtttacattccac	gacgcgggagaagtctg
CCNE1 (Figure 4.5, D and E; g)	caaacgtgaccgttgtgagta	gatcaggttatgagctccgttc
ELL3 (Figure 4.5, B and E)	tcactcagggagcgcctcattatt	atagcctcctgtgttctgccact
PSMD5 (Figure 4.5E)		

*Sequences are according to Ref. 16.

software as described previously.[33]

Cell proliferation and apoptosis

For both fluorometric proliferation and apoptotic assays, cells were harvested 48 h post-siRNA transfection. Fluorometric cell proliferation and apoptosis analyses were performed following the manufacturer's protocols for the CyQUANT NF Cell Proliferation Assay and CellEvent Caspas-3/7 Green Flow Cytometry Assay kits, respectively (Thermo Fisher Scientific). Cell growth curves were performed on siRNA-transfected cells that were grown in 6-cm plates for 48 h prior to transferring to 24-well culture dishes (HeLa; $5x10^4$ cells/well, HEK293; $1.5x10^4$ cells/well), and then cultured for an additional 0-4 days. Cell numbers were counted daily using a Countess Automated Cell Counter (Thermo Fisher Scientific).

Cell cycle analysis by FACS

Cells were harvested 48 h post-siRNA transfection and fixed/permeabilized in ice-cold 1:1 PBS:MeOH-acetone [4:1 (v:v)] while vortexing. Subsequent to adequate washings in FACS buffer (0.5% BSA, 0.05% NaN3 in PBS), cells were incubated over night with 0.1% Triton X-100, 50 mg/ml RNase A, and 50 mg/ml propidium iodide (Sigma) in FACS buffer at 4 °C. Cells were analyzed with a FACScan flow cytometer (BD Biosciences), and cell cycle gating was examined using FLOWJO software.

Statistical analysis

All above mentioned experiments were repeated three to four times independently of each other. Data were analyzed with Excel software using Student's *t*-test (Microsoft). *P*-values (two-tailed) of 0.05 were considered statistically significant. For multiple group analyses, $\chi 2$-values were applied.

ChIP-qPCR

HeLa cells were fixed with 1% formaldehyde for 10 min to cross-link protein:DNA complexes, followed by a quench with 125 mM glycine. The cell lysis, chromatin isolation, and immunoprecipitation (IP) reaction procedures were performed using materials from the commercially available Zymo-Spin ChIP Kit (Zymo Research) following the manufacturer's instructions. In short, after lysis, washing and chromatin isolation, a Diagenode Bioruptor Standard sonication device (run at max amplitude for 5 x 15 min in ice water) was used to shear the cross-linked DNA down to 100 - 400 bp fragments. The IP was carried out using an Abs against ZBTB33 (6F8, Santa Cruz). After quantitating the amount of genomic DNA by PicoGreen assay, equal amounts of ChIP amplicons for each sample were prepared with the SYBR Green Real-Time PCR Master Mix (Thermo Fisher Scientific) and analyzed by qPCR on a QuantStudio 6 Flex Real-Time PCR System (Thermo Fisher Scientific). The fold enrichment was determined based on the cycle differences after normalization to input DNA. A region from the ELL3 gene was also amplified as a negative control. The qRT-PCR primers are listed in Table 4.2. Sequencing validated the authenticity of the cyclin E1 promoter PCR products.

RNA-seq

To achieve ZBTB33 depletion, $3x10^6$ HeLa cells per replicate were transfected with either *ZBTB33* specific or control siRNAs as discussed above. Twenty four hours after transfection, cells were washed with 0.01 M PBS (pH 7.4) and resuspended in TRIzol (Thermo Fisher Scientific) prior to RNA extraction with the Direct-zol RNA Kit (Zymo Research). Prior to sequencing analysis, RNA aliquots were reverse transcribed and the amount of *ZBTB33* determined by qPCR as described above, utilizing *HPRT1* as a normalization control, to ensure that sufficient *ZBTB33* knock-down was achieved. For RNA-seq, RNA quality control measurements, purification, library construction and sequencing were all performed by the High-Throughput Genomics Core within the University of Utah Huntsman Cancer Institute (HCI). In short, RNA quality was measured on a Bioanalyzer RNA 6000 Nano Chip. Small and long directional RNA-seq libraries were constructed using the Illumina TruSeq Standard mRNA Sample Preparation Kit v2 with poly(A) selection. Libraries were sequenced with a 50 bp single-end run on the Illumina HiSeq 2000 platform.

Whole genome shotgun bisulfite sequencing (WGSBS)

Genomic DNA was isolated from $20x10^6$ HeLa cells using the DNeasy Blood and Tissue Kit (QIAGEN). DNA quality control measurements, bisulfite (BS) conversion, purification, library construction, and sequencing were all performed by the High-Throughput Genomics Core within the University of Utah HCI. The BS conversion and library construction was performed using the EpiGnome/TruSeq DNA Methylation Kit (Illumina). Briefly, genomic DNA was denatured and BS converted in a reaction

containing unmethylated lambda DNA (Promega) as a control. Following purification, adapter-ligated DNA was generated using EpiGnome polymerase. Adapter-ligated DNA molecules were enriched by 10 cycles of PCR, concentration normalized and sequenced with a 125-cycle paired-end run on the Illumina HiSeq 2000 platform.

Sequencing data analyses

RNA-seq and WGSBS fastq files were aligned to the human genome (hg19) using Novoalign (Novocraft, Inc.) and peak called with the USeq suite.[34] RNA-seq reads were aligned with all known and theoretical splice junctions using the following parameters: -r All 50 -t 40 -o SAM 90 –k. The USeq NovoalignParser application was then used to parse the alignment files into binary point data by setting the posterior probability to 0 and alignment score threshold to 60. The MultipleReplicaDefinedRegionScanseqs USeq application, which utilizes the DESeq R package,[35] identified statistically significant differentially expressed genes between cells treated with the scrambled and ZBTB33 siRNAs. The WGSBS fastq data were aligned in Novoalign bisulfite mode using the following parameters: -oSAM -rRandom -t240 -h120 -b2. The sam alignments were sorted with the Picard SortSam.jar script and all duplicates removed using the Picard MarkDuplicates.jar script for WGSBS datasets. The results were converted into .bam files and indexed using Samtools.[36] Peak calling was performed with the USeq NovoalignBisulfiteParser to parse the alignments into four binary PointData sets containing the number of observed converted Cyt and nonconverted Cyt at each reference Cyt site sequenced in the genome for both the plus and minus strands.

Bioinformatics analysis

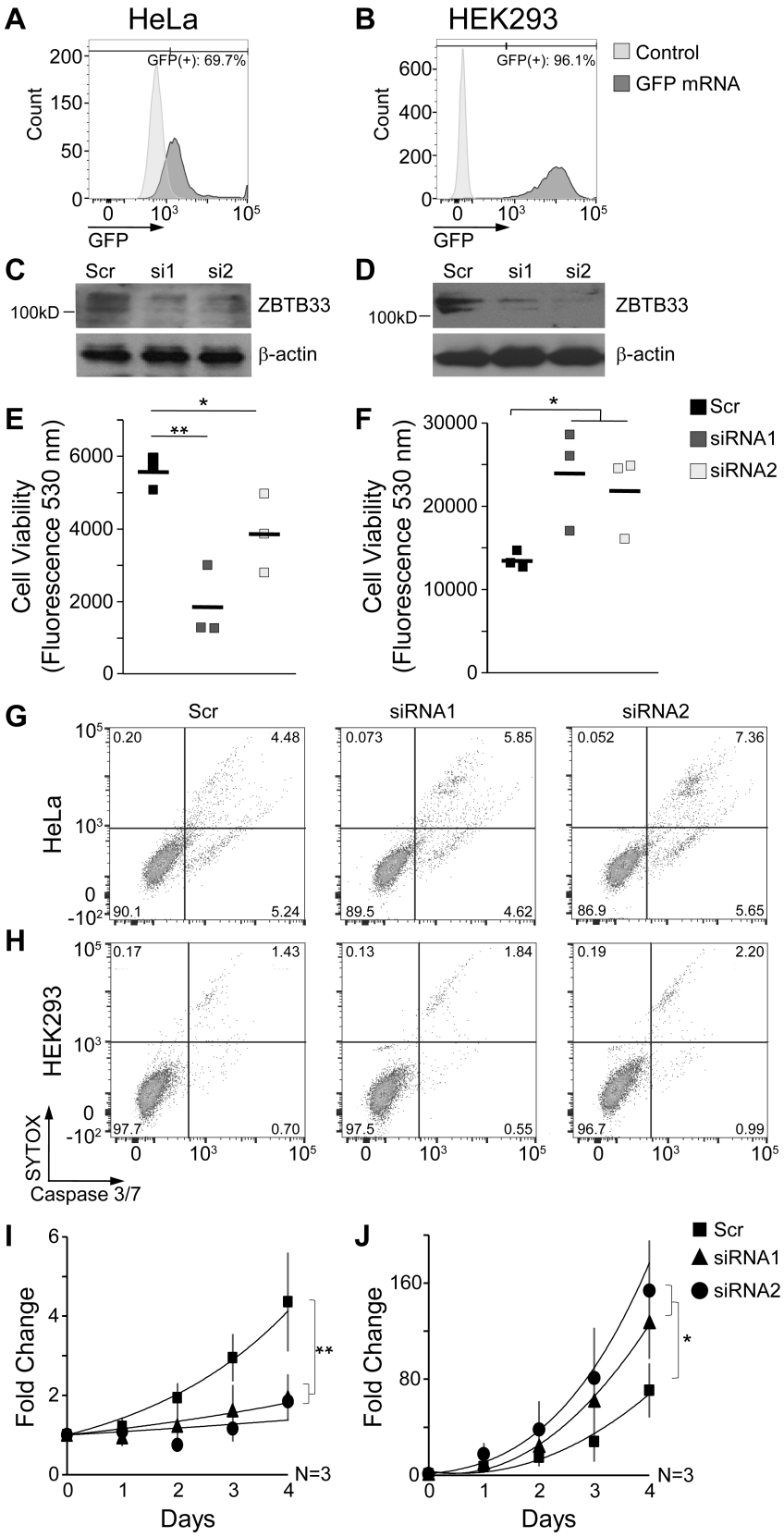
IPA (www.ingenuity.com) and GSEA[37] analyses tools were used to compute significant biological pathways enriched in the depleted ZBTB33 RNA-seq dataset relative to the control. For IPA analysis, only genes with ≥2-fold expression differences and *P*-values lower than 0.01 calculated by USeq (see above) were used. The obtained canonical pathways with false discovery rate (FDR) values lower than 0.01 were considered significant. GSEA was run to identify significant up- and down-regulated expression patterns using the c2.all.v3.0.symbols.gmt gene set collection from the Broad Institute Molecular Signatures Database. Only gene sets with a minimum size of 10 in the MsigDB v3.0 c2 curated database were selected. The Kolmogorov–Smirnov statistics with a cumulative null distribution of 1000 permutations was used to calculate the enrichment scores of each gene set. Furthermore, all gene sets with a Benjamini–Hochberg and GSEA FDR values lower than 0.01 were deemed significant.

Results

ZBTB33 is required for proper HeLa cell proliferation but has an inhibitory effect on HEK293 cell growth

ZBTB33 depletion studies were performed by using two different *ZBTB33* targeting siRNA sequences or a scrambled (Scr) siRNA control in both HeLa and HEK293 cells. The efficiency of RNA transfection was measured and determined to be ~70% and ~96% in HeLa and HEK293 cells, respectively (Figure 4.1A and 4.1B). Despite the fact that HeLa cells have been transformed by the human papilloma virus (HPV) E7 oncogene,[38] they preserve an evident cell cycle progression regulation[39] for

Figure 4.1. ZBTB33 depletion induces differential proliferation trends in HeLa and HEK293 cells. (A and B) FACS analysis of GFP expression in HeLa and HEK293 cells 48 h after GFP mRNA transfection. (C and D) Immunoblot analyses of ZBTB33 protein expression in HeLa and HEK293 cells after transfection with either a scrambled (Scr) or ZBTB33 targeting siRNAs.(E and F) Fluorometric quantitation of cell viability after ZBTB33 depletion. (G and H) FACS analysis of apoptosis in ZBTB33 depleted HeLa and HEK293 cells after 48 h. (I and J) growth curves of HeLa and HEK293 cells after ZBTB33 depletion. *P<0.05, **P<0.005 by student's *t*-test

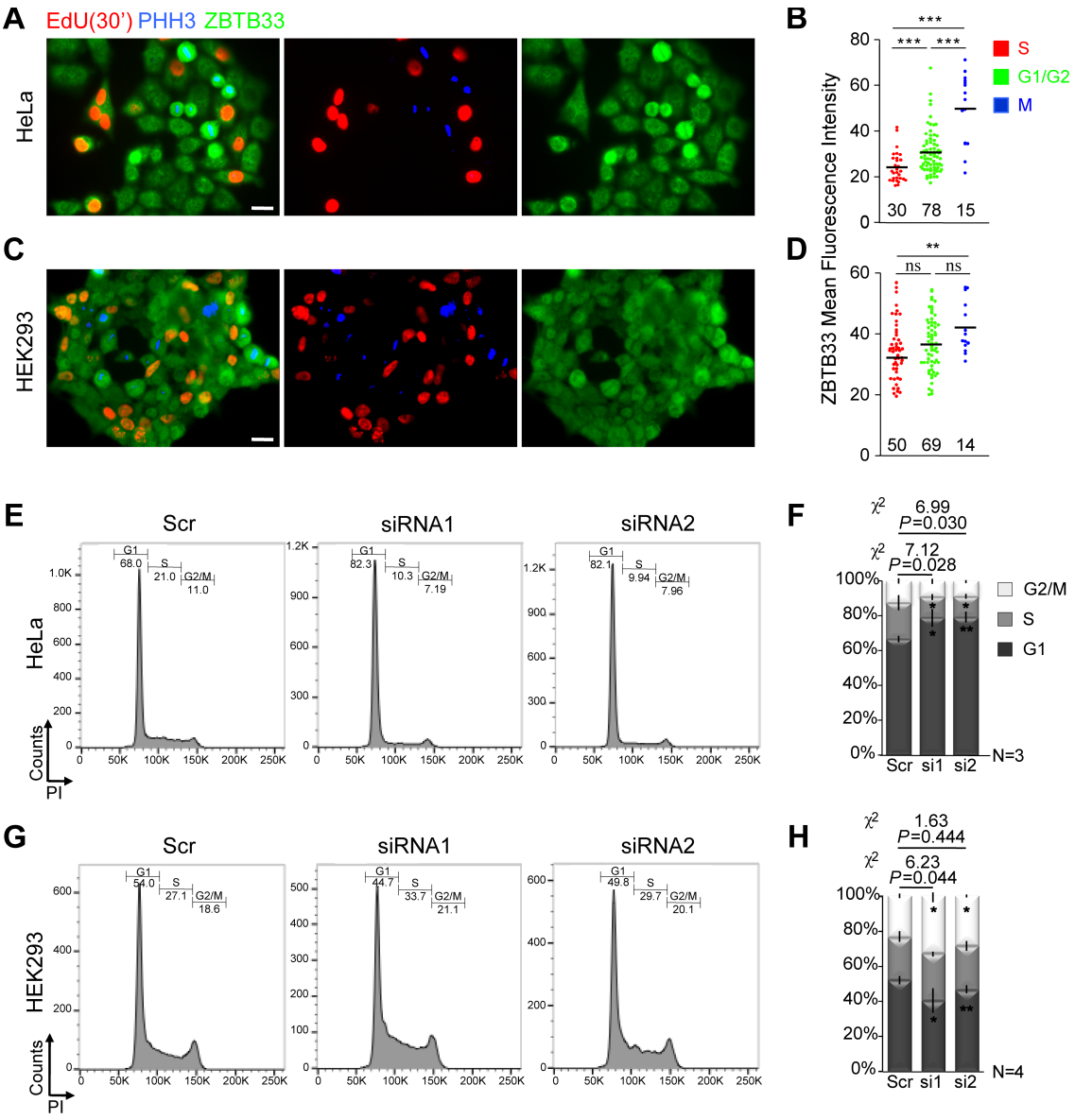

A
HeLa
GFP(+): 69.7%
Count
GFP
B
HEK293
GFP(+): 96.1%
Control
GFP mRNA
C
Scr
si1
si2
100kD
ZBTB33
β-actin
D
E
Cell Viability (Fluorescence 530 nm)
F
Scr
siRNA1
siRNA2
G
HeLa
H
HEK293
SYTOX
Caspase 3/7
I
Fold Change
Days
N=3
J

which their response to ZBTB33 depletion can be evaluated. Immunoblot analyses confirmed that appreciable ZBTB33 depletion was achieved at the protein level 48 h after siRNA transfection (Figure 4.1C and 4.1D). Subsequent fluorescence quantification of cell viability in ZBTB33 depleted HeLa cells showed a marked reduction (Figure 4.1E). In contrast, but consistent with previous observations,[31] ZBTB33 depleted HEK293 cells exhibited a significant increase in cell viability (Figure 4.1F). Notably, for both cell lines the observed alterations in cell viability appeared to occur without deviations in the extent of cell death (Figure 4.1 G and H), indicating that ZBTB33 depletion in HeLa cells leads to decelerated cell proliferation while stimulating HEK293 proliferation. As further confirmation, direct cell counting similarly showed that ZBTB33 depletion induced substantial deceleration of HeLa proliferation (Figure 4.1I), and a significant acceleration of HEK293 proliferation (Figure 4.1J).

ZBTB33 regulates the G1- to S-phase transition

Next, we sought to identify which cell cycle checkpoint(s) is/are regulated by ZBTB33 through monitoring changes in ZBTB33 protein levels during each cell cycle phase. Double immunostaining for ZBTB33 and the M-phase marker PHH3 was conducted on HeLa and HEK293 cells in combination with a 30 min exposure to EdU for S-phase labeling (Figure 4.2A and C). Quantification of the mean fluorescence intensities in HeLa cells showed significant differences in ZBTB33 abundance between the various cell cycle phases (Figure 4.2B), whereas in HEK293 cells the amount of ZBTB33 was nearly identical across all cell cycle phases (Figure 4.2D). These findings indicate that in HeLa cells ZBTB33 protein expression levels are modulated throughout the cell cycle

Figure 4.2. ZBTB33 abundance correlates with the regulation of cell cycle progression in HeLa cells. (A and C) HeLa and HEK293 cells, respectively, 30 min after EdU pulse under regular growth conditions to visualize cells in S-phase, were additionally stained for phospho-histone 3 (PHH3-blue, M-phase) and ZBTB33 (green). From these images, the mean fluorescence intensity (normalized with cell area) of ZBTB33 staining per cell in S-, G1-, and M-phase was plotted (B and D). Because the extent of G1 is by far longer than G2 duration, we approximated that PHH3 and EdU negative cells were overwhelmingly in G1. Two random fields were analyzed per cell line. Numbers in the plots reflect the number of cells measured within each cell cycle phase. Scale bars, 20 μm. (E and G) FACS analysis of cell cycle phase distributions after ZBTB33 depletion in HeLa, and HEK293 cells, respectively. From triplicates of the FACS data, the percent populations of cells in each cell cycle phase were plotted (F and H). Error bars reflect the mean ± S.D. and * $p < 0.05$, **$p < 0.005$, ***$p < 0.0005$ by student's *t*-test; ns not significant.


A
EdU(30') PHH3 ZBTB33
HeLa
C
HEK293
B
ZBTB33 Mean Fluorescence Intensity
S
G1/G2
M
30 78 15
D
50 69 14
E
Scr
siRNA1
siRNA2
HeLa
G1 68.0 S 21.0 G2/M 11.0
G1 82.3 S 10.3 G2/M 7.19
G1 82.1 S 9.94 G2/M 7.96
Counts
PI
F
χ2 6.99 P=0.030
χ2 7.12 P=0.028
G2/M
S
G1
Scr si1 si2
N=3
G
HEK293
G1 54.0 S 27.1 G2/M 18.6
G1 44.7 S 33.7 G2/M 21.1
G1 49.8 S 29.7 G2/M 20.1
H
χ2 1.63 P=0.444
χ2 6.23 P=0.044
N=4

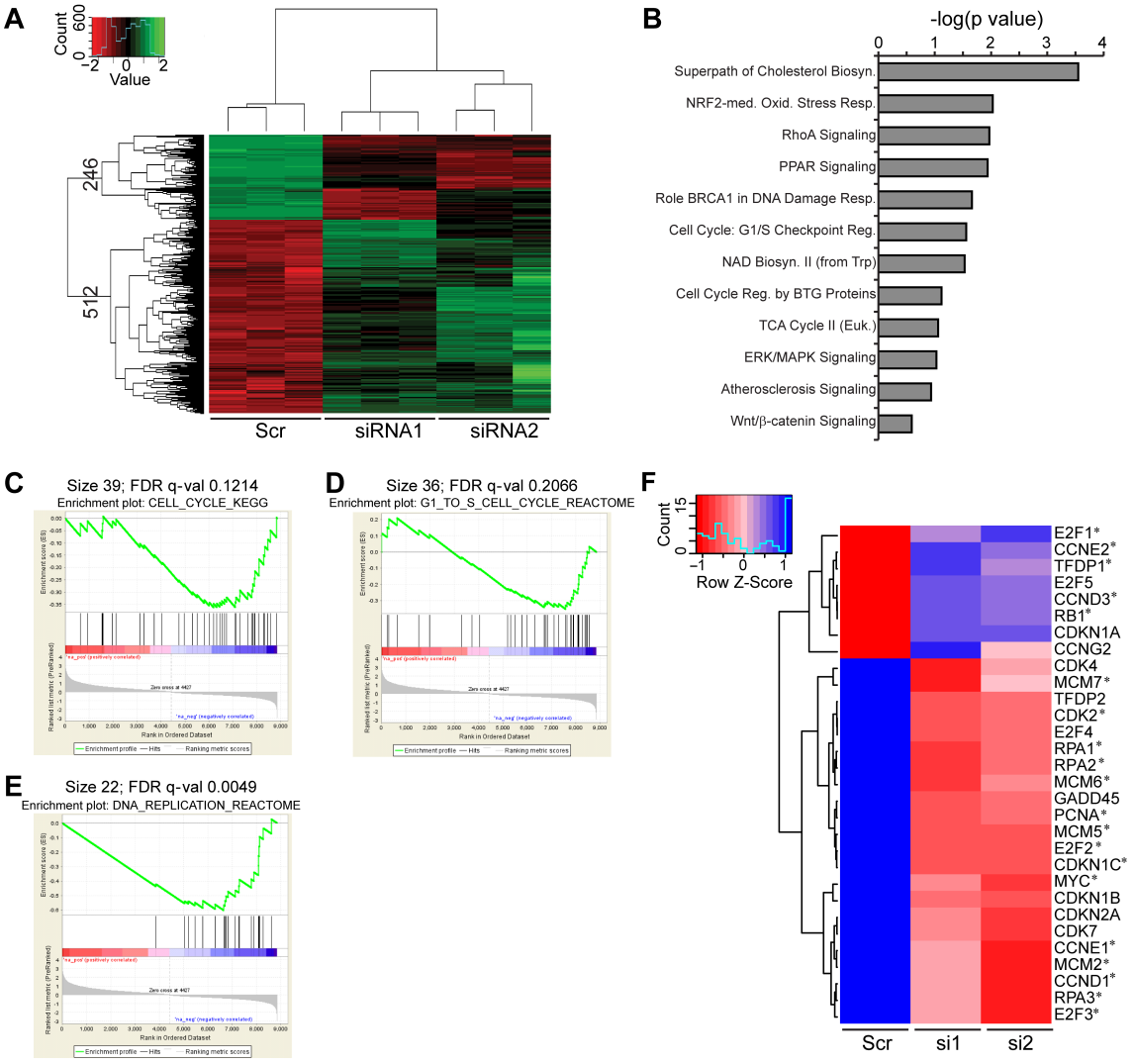
course and that ZBTB33 mainly plays a role in G1- and M-phase progression.

To interrogate this further, cell distribution across the different cell cycle phases were determined by FACS analysis after ZBTB33 depletion using propidium iodide staining. Significantly, HeLa cells exhibited a substantially increased G1-phase population at 48 h post-ZBTB33 depletion concomitant with a decreased S-phase relative to the control (Figure 4.2E and F), indicating that ZBTB33 depletion was sufficient to induce a G1 arrest in HeLa cells. ZBTB33 depleted HEK293 cells on the other hand, showed a moderately significant decrease in G1-phase cell population relative to the control (Figure 4.2G and H). Thus, these observations suggested that ZBTB33 is necessary for proper G1-phase progression in HeLa cells while it has an inhibitory effect on the G1-phase in HEK293 cells.

Gene expression profiling in ZBTB33 depleted HeLa cells

To begin delineating the mechanism by which ZBTB33 modulates HeLa cell growth, RNA-seq was utilized to compare global transcriptome alterations between normal and ZBTB33 depleted cells. In the absence of ZBTB33, a substantial number of genes (758) demonstrated a significant transcriptional alteration relative to the control. Interestingly, it can be seen that while ZBTB33 transcriptional activities are largely repressive in HeLa cells (512 genes), there are an appreciable number of genes in which ZBTB33 also appears to mediate transcriptional activation (246 genes) (Figure 4.3A). Thus, in the global context these RNA-seq results reaffirm that ZBTB33 indeed has the capability of being a bimodal regulator of transcriptional activity in HeLa cells. To identify the most significant biological pathways associated with the transcriptionally

Figure 4.3. Identification of ZBTB33-regulated transcriptional signature in HeLa cells. (A) Unsupervised hierarchical clustering heatmap from RNA-seq data depicting gene transcripts exhibiting a two-fold up- (green) or downregulation (red) after ZBTB33 depletion relative to the control. (B) Altered genes depicted in (A) were subjected to Ingenuity Pathway Analysis (IPA) to identify biological pathways uniquely modulated by ZBTB33. Only the 12 most statistically significant (all $P < 0.05$) downregulated pathways that were also identified in our GSEA analysis are depicted. (C-E) GSEA correlation plots[37] depicting ZBTB33 regulation of cell cycle genes (C), genes associated with G1- to S-phase transition (D), and (E) genes associated with DNA replication. The barcode plot indicates the position of the genes in each gene set; red and blue represent positive and negative Pearson correlations with ZBTB33 depletion, respectively. (F) Hierarchical clustering heatmap illustrating differential expression of genes affecting cell cycle in ZBTB33 depleted HeLa cells. The heatmap color represents Z-score normalized isotig number of for each gene. Blue indicates high gene expression while red indicates low. Genes labeled with an asterisk are established E2F targets


A
Count
600
0
−2
0
2
Value
246
512
Scr
siRNA1
siRNA2
B
-log(p value)
0
1
2
3
4
Superpath of Cholesterol Biosyn.
NRF2-med. Oxid. Stress Resp.
RhoA Signaling
PPAR Signaling
Role BRCA1 in DNA Damage Resp.
Cell Cycle: G1/S Checkpoint Reg.
NAD Biosyn. II (from Trp)
Cell Cycle Reg. by BTG Proteins
TCA Cycle II (Euk.)
ERK/MAPK Signaling
Atherosclerosis Signaling
Wnt/β-catenin Signaling
C
Size 39; FDR q-val 0.1214
Enrichment plot: CELL_CYCLE_KEGG
D
Size 36; FDR q-val 0.2066
Enrichment plot: G1_TO_S_CELL_CYCLE_REACTOME
E
Size 22; FDR q-val 0.0049
Enrichment plot: DNA_REPLICATION_REACTOME
F
Count
15
0
−1
0
1
Row Z-Score
E2F1*
CCNE2*
TFDP1*
E2F5
CCND3*
RB1*
CDKN1A
CCNG2
CDK4
MCM7*
TFDP2
CDK2*
E2F4
RPA1*
RPA2*
MCM6*
GADD45
PCNA*
MCM5*
E2F2*
CDKN1C*
MYC*
CDKN1B
CDKN2A
CDK7
CCNE1*
MCM2*
CCND1*
RPA3*
E2F3*
Scr
si1
si2

altered genes in our RNA-seq data, we performed both IPA and GSEA analyses (Figure 4.3B-4.3E and Table 4.3). Both of these analyses indicated that as the immunofluorescence and FACS data predicted, ZBTB33 clearly modulates genes associated with regulation of the G1- to S-phase transition in HeLa cells.

ZBTB33 regulates cyclin D1 and cyclin E1 expression

Having established a definitive causal link between ZBTB33 transcriptional activities and the G1- to S-phase transition in HeLa cells, we sought to identify the cell cycle regulators that are most likely direct ZBTB33 gene targets. Thus, a list of candidate cell cycle regulator genes was mined from the RNA-seq data using the above discussed GSEA results. From this analysis, 30 candidate genes were found to have a statistically significant differential expression between the control and ZBTB33 depleted cells (Figure 4.3F). Of particular note, two master regulators of G1- to S-phase transition, cyclin D1 and cyclin E1,[40] were amongst the most significantly downregulated genes upon ZBTB33 depletion in HeLa cells. Direct repression of cyclin D1 by ZBTB33 has been reported previously in a number of cancerous cell lines,[16,17] however, our findings here suggest that ZBTB33 is transcriptionally activating cyclin D1 in HeLa cells. Similarly, in A549 and SPC cells, ZBTB33 transcriptional activities have been associated with repression of cyclin E, though it has not been definitively discerned as to whether this is a direct or indirect consequence.[17]

To compare our HeLa cell findings with HEK293 cells and to investigate whether the observed changes in mRNA transcript levels translate at the protein level, we performed comparative semiquantitative RT-PCR and immunoblot analyses from lysates

Table 4.3. Gene lists from the GSEA correlation plots depicted in Figure 4.3C-E

Cell Cycle Kegg	G1 to S Cell Cycle Reactome	DNA Replication Reactome
ATM CCND3 CCNE1 CCNE2 CDK4 CDKN2A E2F1 E2F2 E2F3 E2F4 E2F5 GADD45A RB1 RBL1 TFDP1		
	NACA POLE2 PRIM1 RPA1 RPA2 RPA3	
CDC6		CDC6
CDK2 MCM2 MCM5 MCM6 MCM7 PCNA		
CDAN1	CCND1	GMNN
CDC14A	CCNG2	MCM10
CHEK1	CDK7	POLD4
CHEK2	CDKN1B	RFC2
DTX4	CREB3L1	RFC3
HDAC2	CREB3L3	RFC4
HDAC3	MNAT1	RPS27A
HDAC4	MYC	UBB
HDAC5	TFDP2	UBC
HDAC6		
MAD1L1		
MAD2L1		
MAD2L2		
PTPRA		
SMAD4		
TBC1D8		
TGFB1		

of control and ZBTB33 depleted cells. In HeLa cells, the quantities of cyclin D1 and cyclin E1 at both the mRNA and protein levels were significantly lower in ZBTB33 depleted cells (Figure 4.4A-4.4D). In contrast, transcript levels of cyclin D1 and cyclin E1 were seemingly unaffected by ZBTB33 depletion in HEK293 cells (Figure 4.4A). Interestingly, at the protein level cyclin D1 also exhibited no change after ZBTB33 depletion (Figure 4.4B and 4.4F), however, cyclin E displayed a significant two-fold increase in HEK293 cells (Figure 4.4B and 4.4E). cyclin E cellular abundance is regulated predominantly at the levels of gene transcription and ubiquitin-dependent proteolysis.[41] Accordingly, the discrepancy between increased cyclin E protein levels and unaffected mRNA expression suggests that posttranscriptional ubiquitin-dependent regulation of this protein is activated in ZBTB33 depleted HEK293 cells. Together, these findings indicate that in HeLa cells ZBTB33 plays a direct role in transcriptionally mediating cellular levels of cyclin D1 and cyclin E1, whereas in HEK293 cells transcription of these genes appears to be ZBTB33 independent.

ZBTB33 directly occupies both cyclin D1 and cyclin E1 promoter regions in HeLa Cells

It was shown previously that repression of the cyclin D1 gene in MCF-7 and HCT 116 cells required dual occupation by ZBTB33 at both a -1067 location that contains the KBS as well as a mCpG containing site at the +69 position.[16] It was further confirmed that occupation at the +69 position is methyl-dependent. To resolve the perceived discrepancy by which ZBTB33 mediates transcriptional activation of cyclin D1 in HeLa cells and repression in MCF-7 and HCT 116 cells, we performed ChIP-qPCR at the -

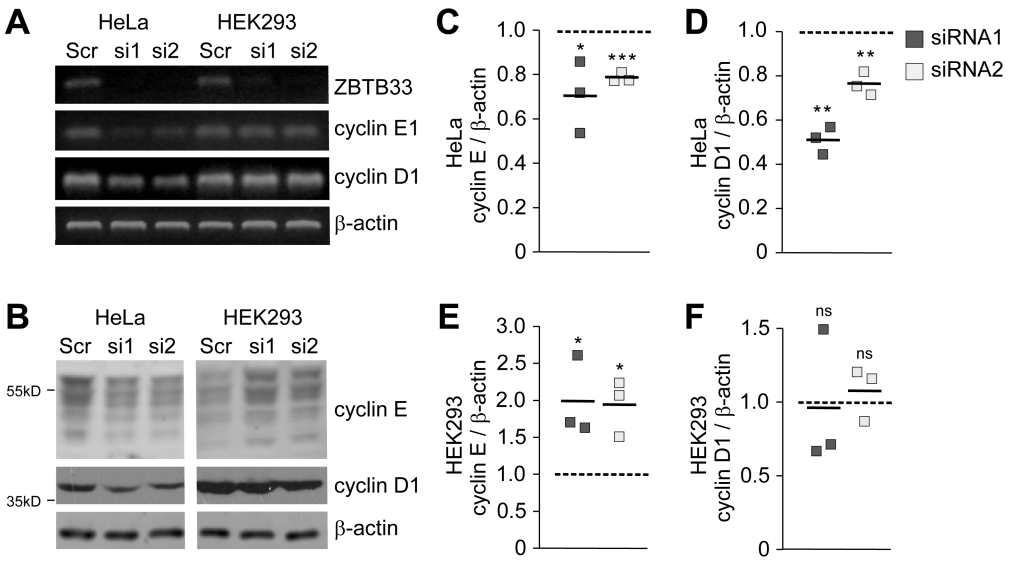


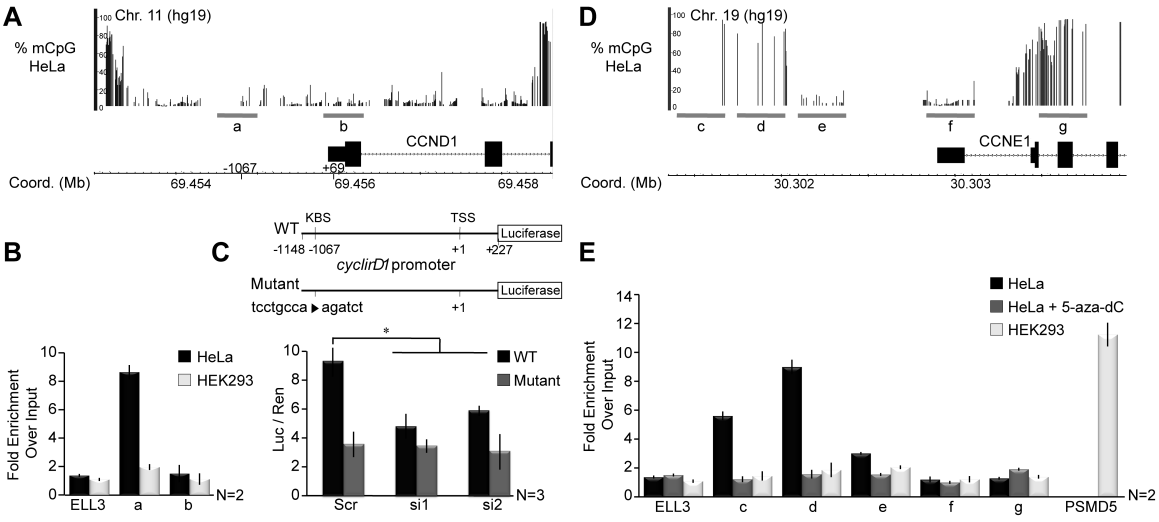
Figure 4.4. ZBTB33 regulates cyclin E1 and cyclin D1 expression. (A) Relative mRNA expression levels of *ZBTB33*, cyclin E1, and cyclin D1 in HeLa and HEK293 cells determined after ZBTB33 depletion by semiquantitative RT-PCR. (B) Immunoblot analysis for ZBTB33, cyclin E and cyclin D1 in HeLa and HEK293 cells after ZBTB33 depletion. The low molecular weight cyclin E bands are due to protease or alternative splicing processing.[33,43] (C-F) Band intensities from three independent immunoblots were quantitated and normalized to β-actin to determine the protein expression levels of cyclin E and cyclin D1 after ZBTB33 depletion relative to the Scr control in both HeLa and HEK293 cells. *$p<0.05$, **$p<0.01$, ***$p<0.005$ by Student's *t*-test; not significant (ns).

1067 and +69 loci and analyzed these findings in the context of global methylome mapping in HeLa cells. From ChIP-qPCR we observed a significant occupation of ZBTB33 at the -1067 KBS containing site, but no occupation at the CpG-containing +69 site (Figure 4.5A and B). Interestingly, global methylome analysis demonstrated that in HeLa cells the cyclin D1 promoter harbors a low methylation incidence (Figure 4.5A). This finding is consistent with the lack of ZBTB33 occupation at the +69 CpG site and reaffirms that ZBTB33 indeed binds this region in a methyl-dependent manner.

To corroborate these findings, we assessed the ability of ZBTB33 to regulate luciferase expression under control of either a minimal wild-type (WT) or mutated cyclin D1 promoter that considerably altered the -1069 KBS core (Figure 4.5C). As anticipated, luciferase activity of the mutated cyclin D1 promoter was significantly reduced relative to the WT (Figure 4.5C). Moreover, co-transfection of either the WT or mutant cyclin D1 promoter-reporter plasmid with ZBTB33 siRNAs resulted in a similar decrease in luciferase activity (Figure 4.5C). Combined, these data substantiate that the ability of ZBTB33 to transcriptionally activate cyclin D1 in HeLa cells is solely dependent on the integrity of the -1067 KBS loci.

Finally, we sought to determine if ZBTB33 also directly occupies the cyclin E1 promoter. Unlike the cyclin D1 promoter, cyclin E1 does not have a KBS site within approximately -1000 base pairs of the TSS, but according to our methylome analysis does have regions of high CpG methylation (Figure 4.5D). ChIP-qPCR analysis for ZBTB33 binding in several of these regions demonstrated that it indeed appears to preferentially occupy mCpG sites in the promoter but not within intergenic regions (Figure 4.5E). Moreover, occupation at this site is methyl-dependent as treatment of HeLa cells with the

Figure 4.5. ZBTB33 directly occupies and regulates cyclin D1 and cyclin E1 gene promoters. (A and D) WGSBS read coverage tracks depicting percent DNA methylation (mCpG) levels at cyclin D1 and cyclin E1 promoters. Regions within each promoter amplified by ChIP-qPCR are indicated (gray bars). (B) ChIP-qPCR analysis at the cyclin D1 promoter in HeLa and HEK293 cells at known ZBTB33 occupation sites, including a KBS-containing location at -1067 and a CpG containing site at the +69 position. (C) Scheme depicting design of the wild type (WT) and KBS-mutated minimal cyclin D1 promoter-luciferase reporter plasmids (top). The minimal WT or KBS-mutated cyclin D1 promoter-luciferase reporter plasmids were co-transfected with *ZBTBT33* siRNAs, and luciferase activity was quantitated (bottom). Luciferase reads were normalized to *Renilla* activity. (E) ChIP-qPCR analysis for ZBTB33 occupation of the cyclin E1 promoter in HeLa and HEK293 cells at various CpG-containing regions. The *PSMD5* gene promoter was selected as a positive control in HEK293 cells based on publicly available ZBTB33 ChIP-seq data (GSM1334009 and GSM803504). Error bars reflect mean_S.D.; *, $p < 0.02$ by Student's *t* test


A
% mCpG
HeLa
Chr. 11 (hg19)
a
b
CCND1
-1067
+69
Coord. (Mb)
69.454
69.456
69.458
D
% mCpG
HeLa
Chr. 19 (hg19)
c
d
e
f
g
CCNE1
Coord. (Mb)
30.302
30.303
B
Fold Enrichment Over Input
HeLa
HEK293
ELL3
a
b
N=2
C
WT
KBS
TSS
-1148 -1067
+1
+227
Luciferase
cyclirD1 promoter
Mutant
tcctgcca ▸ agatct
+1
Luc / Ren
WT
Mutant
Scr
si1
si2
N=3
E
HeLa
HeLa + 5-aza-dC
HEK293
Fold Enrichment Over Input
ELL3
c
d
e
f
g
PSMD5
N=2

DNA methyltransferase inhibitor 5-aza-2-deoxycytidine abrogates ZBTB33 binding (Figure 4.5E). Although promoter methylation is typically associated with gene silencing, recent large-scale DNA methylation profiling studies have shown that depending on the cellular context, hyper-methylated promoters are also found near highly transcribed genes.[44-46] Thus, our findings suggest that ZBTB33 has the ability to function as a methyl-dependent activator of gene activity. To the best of our knowledge, this is the first evidence that ZBTB33 directly occupies and transcriptionally modulates cyclin E1 expression to affect cell cycle progression. Together, our findings demonstrate that in HeLa cells ZBTB33 is a direct transcriptional regulator of two key cell cycle genes that function in the G1- to S-phase transition. Finally, ChIP-qPCR analysis of cyclin D1 and cyclin E1 promoter occupancy by ZBTB33 in HEK293 cells reveals a lack of ZBTB33 association with these promoters (Figure 4.5B and 4.5E), further confirming that ZBTB33 does not regulate cyclin D1 transcription and indirectly regulates cyclin E abundance in HEK293 cells.

Aberrant RB1-E2F activity in ZBTB33 depleted cells

To establish a causal relationship between ZBTB33 regulation of cyclin D1 and cyclin E1 transcription, HeLa cell proliferation and G1- to S-phase transition, we assessed phosphorylation of RB1 and subsequent E2F activity in ZBTB33 depleted cells. The cyclin D1-cyclin E-RB1-E2F pathway functions as a bi-stable switch regulating G1-phase transition upon cellular commitment to proliferate.[47] At the beginning of G1, E2F is bound and repressed by RB1. With sufficient stimulation, activities of cyclin D/Cdk4,6 and cyclin E/Cdk2 complexes are successively induced to phosphorylate RB1 and relieve

its repression of E2F.[40,48,49] Of note, it has been previously determined that the HPV E7 oncoprotein binding site on RB1 is separable.

As a first analysis, we utilized a pan-RB1 antibody to discern whether global levels of RB1 protein were affected in either HeLa or HEK293 cells after ZBTB33 depletion. HEK293 cells harbor a mixture of both hyper- and hypo-phosphorylated RB1 isoforms (Figure 4.6A).[51] By comparison, HeLa cells typically show a predominant single RB1 band,[52,53] consistent with the hyper-phosphorylated form as observed in these studies (Figure 4.6A). Notably, Figure 4.6A clearly shows that in either cell line, there are no detectably significant changes in total pan-RB1 protein levels indicating that ZBTB33 depletion did not sequester RB1 for further HPV E7 proteolysis, or reduce RB1 degradation in HEK293 cells.

Next, an antibody specific to RB1 phosphorylated at S780 ($pRB1^{S780}$) was selected as this site is exclusively phosphorylated by the cyclin D/Cdk(4/6) complexes. Similarly, an antibody specific to phosphorylated T821 ($pRB1^{T821}$) was used to evaluate cyclin E/Cdk2 activity. Consistent with the observed decrease in cyclin D1 levels (Figure 4.4B and 4.4D), ZBTB33 depleted HeLa cells exhibited reduced $pRB1^{S780}$ (Figure 4.6A). Although ZBTB33-depleted HeLa cells harbored reduced cyclin E levels (Figure 4.6B) and 4.6C), they did not show a corresponding significant decrease in $pRB1^{T821}$ (Figure 4.6A). These unexpected results may be attributed to compensatory activity of the cyclin A/Cdk2 complex, which can also act to phosphorylate T821.[54]

Proper activation of the E2F family of transcription factors is essential for adequate G1- to S-phase transition. At early G1-phase, E2F proteins are bound to and repressed by hypo-phosphorylated RB1 (Figure 4.7). With sufficient stimulation, the

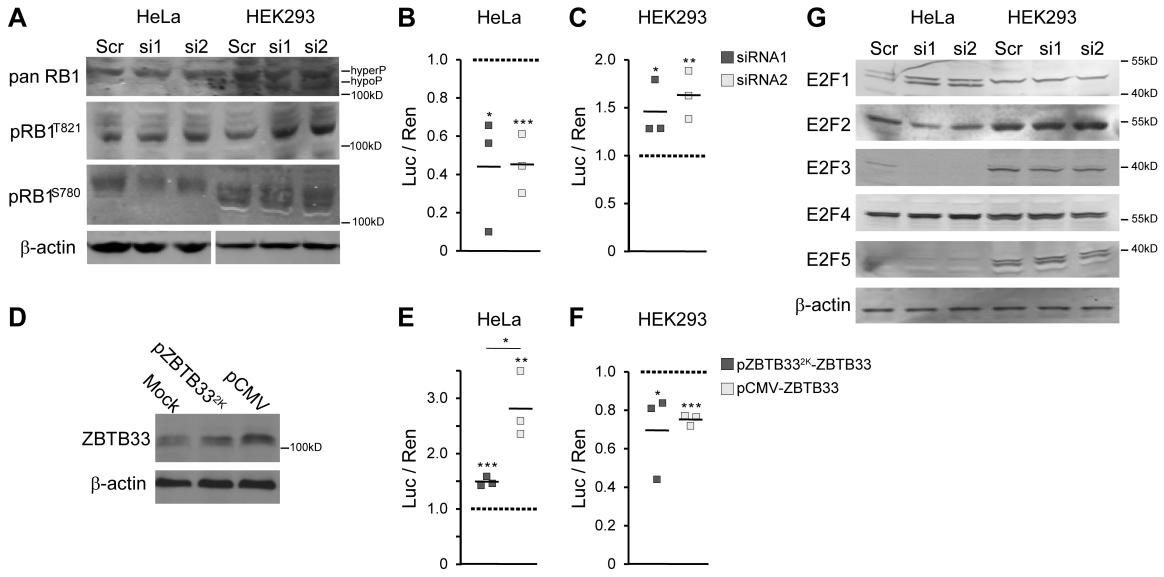


Figure 4.6. ZBTB33 regulates the RB1-E2F pathway. (A) Representative immunoblots for whole (pan) and phosphorylated RB1 in HeLa and HEK293 cells (100 μg and 12 μg total protein per lane, respectively) after ZBTB33 depletion. Note, HeLa and HEK293 β-actin immunoblots were analyzed using different exposure intensities. (B, C) A 6xE2F-luciferase reporter ≠was transfected into control and ZBTB33-depleted HeLa and HEK293 cells and luciferase activity quantitated. Luciferase reads, normalized to Renilla activity, represent endogenous E2F activity. (D) Immunoblot analysis for ZBTB33 in HeLa cells transfected with ZBTB33 overexpression plasmids under control of either a 2-Kb minimal endogenous ZBTB33 promoter or a highly active CMV promoter. (E, F) A 6xE2F-luciferase reporter was transfected into mock-control or ZBTB33 overexpressing cells demonstrating dose-dependent E2F activity in HeLa cells, and a decreased E2F activity in HEK293 cells. (G) Representative immunoblots for E2F1-5 proteins in HeLa and HEK293 cells after ZBTB33 depletion*p < 0.05, **p < 0.01, ***p < 0.005 by Student's *t*-test.

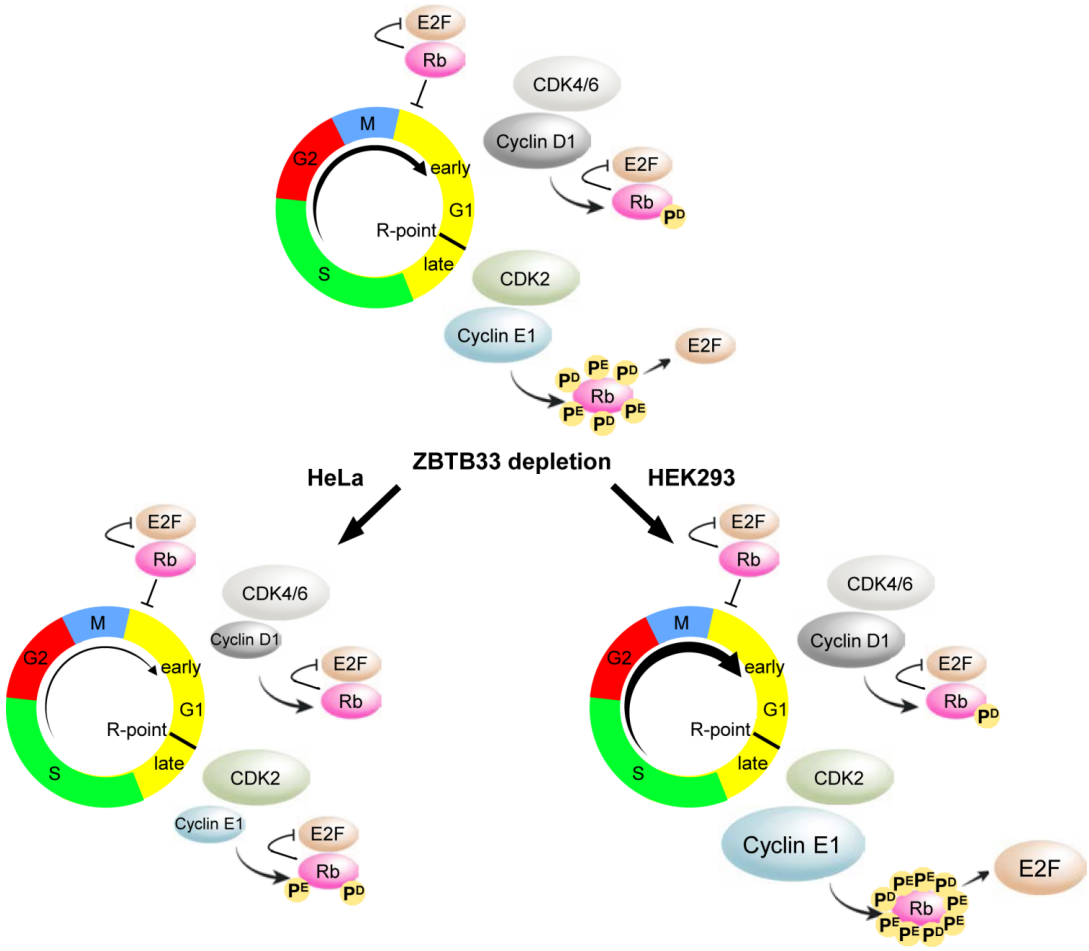


Figure 4.7. Diagram illustrating the canonical and proposed ZBTB33 mediated cyclin D1/cyclin E1/RB1/E2F pathways. The canonical cyclin D1/cyclin E1/RB1/E2F pathway regulates G1- to S-phase progression (top) in comparison to proposed models for the consequences of ZBTB33 depletion on these pathways in HeLa (left) and HEK293 (right) cells.

activity of cyclin D/Cdk(4,6) and cyclin E/Cdk2 complexes are induced to initiate RB1 hyper-phosphorylation, relieving E2F repression.[55] Given that ZBTB33 depletion was from the regions of RB1 required to impose cell cycle arrest. As evidence, mutant RB1 proteins deficient in E7 binding retain their ability to induce cell cycle arrest and repress transcription of E2F target genes.[50]

As a first analysis, we utilized a pan-RB1 antibody to discern whether global levels of RB1 protein were affected in either HeLa or HEK293 cells after ZBTB33 depletion. HEK293 cells harbor a mixture of both hyper- and hypo-phosphorylated RB1 isoforms (Figure 4.6A).[51] By comparison, HeLa cells typically show a predominant single RB1 band,[52,53] consistent with the hyper-phosphorylated form as observed in these studies (Figure 4.6A). Notably, Figure 4.6A clearly shows that in either cell line, there are no detectably significant changes in total pan-RB1 protein levels indicating that ZBTB33 depletion did not sequester RB1 for further HPV E7 proteolysis, or reduce RB1 degradation in HEK293 cells.

Next, an antibody specific to RB1 phosphorylated at S780 ($pRB1^{S780}$) was selected as this site is exclusively phosphorylated by the cyclin D/Cdk(4/6) complexes. Similarly, an antibody specific to phosphorylated T821 ($pRB1^{T821}$) was used to evaluate cyclin E/Cdk2 activity. Consistent with the observed decrease in cyclin D1 levels (Figure 4.4B and 4.4D), ZBTB33 depleted HeLa cells exhibited reduced $pRB1^{S780}$ (Figure 4.6A). Although ZBTB33-depleted HeLa cells harbored reduced cyclin E levels (Figure 4.6B) and 4.6C), they did not show a corresponding significant decrease in $pRB1^{T821}$ (Figure 4.6A). These unexpected results may be attributed to compensatory activity of the cyclin A/Cdk2 complex, which can also act to phosphorylate T821.[54]

Proper activation of the E2F family of transcription factors is essential for adequate G1- to S-phase transition. At early G1-phase, E2F proteins are bound to and repressed by hypo-phosphorylated RB1 (Figure 4.7). With sufficient stimulation, the activity of cyclin D/Cdk(4,6) and cyclin E/Cdk2 complexes are induced to initiate RB1 hyper-phosphorylation, relieving E2F repression.[55] Given that ZBTB33 depletion was demonstrated to alter RB1 phosphorylation, we sought to determine the effect of ZBTB33 depletion and overexpression on E2F activity utilizing a 6xE2F luciferase reporter assay.[33] As expected, ZBTB33 depleted HeLa cells displayed significantly reduced endogenous E2F activity (Figure 4.6B), and concomitantly a concentration-dependent increase in E2F activity with ZBTB33 overexpression (Figure 4.6D and 4.6E). Notably, the observed increase in Cyclin E abundance in ZBTB33 depleted HEK293 cells (Figure 4.4B and 4.4E) resulted in an increase of $pRB1^{T821}$ (Figure 4.6A). Accordingly, ZBTB33 depletion in HEK293 cells resulted in increased E2F activity (Figure 4.6C), while ZBTB33 overexpression inhibited E2F function (Figure 4.3F). Interestingly, many of the genes identified by the RNA-seq analysis to be dysregulated by ZBTB33 depletion in HeLa (Figure 4.3F) are also established E2F targets.[42] Furthermore, RNA-seq analysis also showed differential expression of several E2F transcription factors upon ZBTB33 depletion in HeLa cells (Figure 4.3F). In general, immunoblot analyses confirmed these observations at the protein level, particularly showing increased E2F1 and decreased E2F2 and E2F3 proteins after reduction of ZBTB33 (Figure 4.6G). Nonetheless, a direct ZBTB33 regulation of E2F gene transcription is difficult to discern given that the E2F proteins also regulate their own expression.[51] While this provides some complexity in interpreting the E2F luciferase reporter assay, it is clear that the net

E2F activity is differentially modulated by ZBTB33 between HeLa and HEK293 cells.

Discussion

Dysregulated expression of G1-phase cyclins has been correlated with the genesis of a large proportion of human malignancies.[56] In particular, tumor upregulation of cyclin D1 and cyclin E is generally correlated with poor prognosis. Indeed, the activation of cdk4 and cdk6 by cyclin D1 along with the association of cyclin E with cdk2, function as gatekeepers at the G1 restriction checkpoint (R-point) to regulate G1- to S-phase transition and initiation of DNA replication. Consequently, overexpression of these genes has been implicated in the acceleration of cell cycle transition through the R-point that promotes genetic instability and uncontrolled cell proliferation by eliminating p53/p21 regulatory constraints.[56]

Interestingly, the observed increase in ZBTB33 expression and nuclear localization associated with high-grade and poor prognosis in human breast and prostate tumors appears to correlate with tumor phenotypes harboring increased cyclin D1 and cyclin E levels. ZBTB33 has been reported to play a critical role in TGFβ signaling regulation of metastasis in breast cancer, such that increasing levels of ZBTB33 are associated with decreased breast cancer metastasis-free survival times.[5,7,10] Likewise, cyclin D1 is one of the most commonly amplified genes and frequently overexpressed proteins in breast cancer,[57] and the level of cyclin E expression is used as a powerful clinical predictor of survival status in breast cancer patients.[58] Elevated levels of either ZBTB33 or cyclin D1 in primary prostate tumors have also been shown to coincide with increased tumor aggressivity and Gleason score.[59-61] Further, cyclin D1 and cyclin E have

been identified as critical co-activators of androgen-dependent transcription and cell cycle progression in PCa cells, suggesting that their aberrant expression in tumors may contribute to persistent activation of androgen receptor function, even during androgen ablation therapy.[60,62]

While it has been determined that ZBTB33 can transcriptionally modulate cyclin D1 and cyclin E in several cell lines,[16,17] it has thus far been shown to directly or indirectly transcriptionally repress these cyclins and subsequently reduce their proliferation capabilities. However, these findings are contradictory to the established proproliferative characteristic of G1 cyclins observed in high-grade tumors. Interestingly, similar to our observations in HeLa cells, ZBTB33 depletion induced a decreased cellular proliferation in PC3 cells, a cell line model for high grade PCa.[11] Given the increasing evidence that various cancer phenotypes are able to differentially modulate the expression levels, cellular localizations and transcriptional responses of ZBTB33, and the correlation between ZBTB33 activities in the more aggressive forms of breast and prostate tumors, it stands to reason that in addition to its established role in transcriptionally inducing epithelial-to-mesenchymal transition (EMT) programming in high grade breast and prostate tumors,[59,63-65] ZBTB33 may also mediate cellular proliferation through upregulation of G1 cyclins. As evidence for a possible mechanism by which ZBTB33 may mediate proproliferative activities in cancer, we show here for the first time a positive causal correlation between cell proliferation and ZBTB33, cyclin D1 and cyclin E expression levels that appears to be cell-type specific (Figure 4.7).

Repression of cyclin D1 was previously shown to require ZBTB33 cooccupation at a -1067 KBS-containing loci as well as a mCpG site located at +69.[16] In HeLa cells we

observed that activation of cyclin D1 was facilitated solely through occupation at the -1067 KBS region and that occupation of the +69 was abrogated by the lack of DNA methylation at this site. Thus, these findings suggest that altering the methylation status at the +69 site may provide a novel mechanism by which different cell types can direct the transcriptional activities of ZBTB33 to be repressive or activating at the cyclin D1 promoter. In contrast, activation of the cyclin E1 promoter appears to be solely reliant on methyl-dependent CpG recognition. Further, we demonstrated in HeLa cells that ZBTB33 is absolutely necessary for proper cyclin D1 and cyclin E expression and that it subsequently plays a proproliferative role by positively regulating the RB1-E2F pathway (Figure 4.7). Conversely, ZBTB33 depletion in HEK293 cells resulted in an accelerated cell proliferation due to an indirect increase in cyclin E abundance (Figure 4.7), suggesting that in HEK293 ZBTB33 normally functions to restrain cyclin E protein levels. Thus, together we have provided novel mechanistic insight into how various cell phenotypes may be able to differentially tune the transcriptional activities of ZBTB33 at cyclin D1 and cyclin E1 promoters to alter the rates of cellular proliferation in cancer.

References

(1) Esteller, M. Aberrent DNA Methylation as a Cancer-Inducing Mechanism. *Annu. Rev. Pharmacol. Toxicol.* **2005**, *45*, 629-656.

(2) Jones, P. A.; Baylin, S. B. The Epigenomics of Cancer. *Cell* **2007**, *128*, 683-692.

(3) Costello, J. F.; Fruhwald, M. C.; Smiraglia, D. J.; Rush, L. J.; Robertson, G. P.; Gao, X.; Wright, F. A.; Feramisco, J. D.; Peltomaki, P.; Lang, J. C.; Schuller, D. E.; Yu, L.; Bloomfield, C. D.; Caligiuri, M. A.; Yates, A.; Nishikawa, R.; Huang, H.-J. S.; Petrelli, N. J.; Zhang, X.; O'Dorisio, M. S.; Held, W. A.; Cavenee, W. K.; Plass, C. Aberrant CpG-Island Methylation has Non-Random and Tumour-Type−Specific Patterns. *Nat. Genet.* **2000**, *24*, 132-138.

(4) Klose, R. J.; Bird, A. P. Genomic DNA Methylation: the Mark and Its Mediators. *Trends Biochem. Sci.* **2006**, *31*, 89-97.

(5) Bassey-Archibong, B. I.; Kwiecien, J. M.; Milosavljevic, S. B.; Hallett, R. M.; Rayner, L. G.; Erb, M. J.; Crawford-Brown, C. J.; Stephenson, K. B.; Bédard, P. A.; Hassell, J. A.; Daniel, J. M. Kaiso Depletion Attenuates Transforming Growth Factor-β Signaling and Metastatic Activity of Triple-Negative Breast Cancer Cells. *Oncogenesis* **2016**, *5*, e208.

(6) Jones, J.; Wang, H.; Zhou, J.; Hardy, S.; Turner, T.; Austin, D.; He, Q.; Wells, A.; Grizzle, W. E.; Yates, C. Nuclear Kaiso Indicates Aggressive Prostate Cancers and Promotes Migration and Invasiveness of Prostate Cancer Cells. *Am. J. Pathol.* **2012**, *181*, 1836-1846.

(7) Jones, J.; Wang, H.; Karanam, B. Theodore, S.; Dean-Colomb, W.; Welch, D. R.; Grizzle, W.; Yates, C. Nuclear Localization of Kaiso Promotes the Poorly Differentiated Phenotype and EMT in Infiltrating Ductal Carcinomas. *Clin. Exp. Metastasis.* **2014**, *31*, 497-510.

(8) Pierre, C. C.; Longo, J.; Mavor, M.; Milosavljevic, S. B.; Chaudhary, R.; Gilberth, E.; Yates, C.; Daniel, J. M. Kaiso Overexpresison Promotes Intestinal Inflammation and Potentiates Intestinal Tumorigenesis in $Apc^{Min/+}$ Mice. *Biochim. Biophys. Acta.* **2015**, *1852*, 1846-1855.

(9) Prokhortchouk, A. V.; Sansom, O.; Selfridge, J.; Caballero, I. M.; Salozhin, S.; Aithozhina, D.; Cerchietti, L.; Meng, F. G.; Augenlicht, L. H.; Mariadason, J. M.; Hendrich, B.; Melnick, A.; Prokhortchouk, E.; Clarke, A.; Bird, A. Kaiso-Deficient Mice Show Resistance to Intestinal Cancer. *Mol. Cell. Biol.* **2006**, *26*, 199-208.

(10) Vermeulen, J. F.; van de Ven, R. A. H.; Ercan, C.; van der Groep, P.; van der Wall, E.; Bult, P.; Christgen, M.; Lehmann, U.; Daniel, J.; van Diest, P. J.; Derksen, P.

W. B. Nuclear Kaiso Expression is Associated with High Grade and Triple-Negative Invasive Breast Cancer. *PLoS One* **2012**, *7*, e37864.

(11) Wang, H.; Liu, W.; Black, S.; Turner, O.; Daniel, J. M.; Dean-Colomb, W.; He, Q. P.; Davis, M.; Yates, C. Kaiso, a Transcriptional Repressor, Promotes Cell Migration and Invasion of Prostate Cancer Cells through Regulation of miR-31 Expression. *Oncotarget* **2016**, *7*, 5677-5689.

(12) Buck-Koehntop, B. A.; Stanfield, R. L.; Ekiert, D. C.; Martinez-Yamout, M. A.; Dyson, H. J.; Wilson, I. A.; Wright, P. E. Molecular Basis for Recognition of Methylated and Specific DNA Sequences by the Zinc Finger Protein Kaiso. *Proc. Natl. Acad. Sci.* **2012**, *109*, 15229-15234.

(13) Daniel, J. M.; Spring, C. M.; Crawford, H. C.; Reynolds, A. B.; Baig, A. The $p120^{ctn}$-Binding Partner Kaiso is a Bi-Modal DNA-Binding Protein that Recognizes both a Sequence-Specific Consensus and Methylated CpG Dinucleotides. *Nucleic Acids Res.* **2002**, *30*, 2911-2919.

(14) Filion, G. J. P.; Zhenilo, S.; Salozhin, S.; Yamada, D.; Prokhortchouk, E.; Defossez, P-A. A Family of Human Zinc Finger Proteins that Bind Methylated DNA and Repress Transcription. *Mol. Cell. Biol.* **2006**, *26*, 169-181.

(15) Buck-Koehntop, B. A.; Martinez-Yamout, M. A.; Dyson, H. J.; Wright, P. E. Kaiso Uses all Three Zinc Fingers and Adjacent Sequence Motifs for High Affinity Binding to Sequence-Specific and Methyl-CpG DNA Targets. *FEBS Lett.* **2012**, *586*, 734-739.

(16) Donaldson, N. S.; Pierre, C. C.; Antsey, M. I.; Robinson, S. C.; Weerawardane, S. M.; Daniel, J. M. Kaiso Represses the Cell Cycle Gene *Cyclin D1* via Sequence-Specific and Methyl-CpG-Dependent Mechanisms. *PLoS One* **2012**, *7*, e50398.

(17) Jiang, G.; Wang, Y.; Dai, S.; Liu, Y.; Stoecker, M.; Wang, E. P120-Catenin Isoforms 1 and 3 Regulate Proliferation and Cell Cycle of Lung Cancer Cells via Beta-Catenin and Kaiso Respectively. *PLoS One* **2012**, *7*, e30303.

(18) Koh, D. I.; Yoon, J. H.; Kim, M. K.; An, H.; Kim, M. Y.; Hur, M. W. Kaiso is a Key Regulator of Spleen Germinal Center Formation by Repressing Bcl6 Expression in Splenocytes. *Biochem. Biophys. Res. Commun.* **2013**, *442*, 177-182.

(19) Lopes, E. C.; Valls, E.; Figueroa, M. E.; Mazur, A.; Meng, F-G.; Chiosis, G.; Laird, P. W.; Schreiber-Agus, N.; Greally, J. M.; Prokhortchouk, E.; Melnick, A. Kaiso Contributes to DNA Methylation-Dependent Silencing of Tumor Suppressor Genes in Colon Cancer Cell Lines. *Cancer Res.* **2008**, *68*, 7258-7263.

(20) Pierre, C. C.; Longo, J.; Bassey-Archibong, B. I.; Hallett, R. M.; Milosavljevic,

S.; Beatty, L.; Hassell, J. A.; Daniel, J. M. Methylation-Dependent Regulation of Hypoxia Inducible Factor-1 Alpha Gene Expression by the Transcription Factor Kaiso. *Biochim. Biophys. Acta* **2015**, *1849*, 1432-1441.

(21) Prokhortchouk, A.; Hendrich, B.; Jorgensen, H.; Ruzov, A.; Wilm, M.; Georgiev, G.; Bird, A.; Prokhortchouk, E. The p120 Catenin Partner Kaiso is a DNA Methylation-Dependent Transricptional Repressor. *Genes Dev.* **2001**, *15*, 1613-1618.

(22) Rodova, M.; Kelly, K. F.; VanSuan, M.; Daniel, J. M.; Werle, M. J. Regulation of the Rapsyn Promoter by Kaiso and d-Catenin. *Mol. Cell. Biol.* **2004**, *24*, 7188-7196.

(23) Spring, C. M.; Kelly, K. F.; O'Kelly, I.; Graham, M.; Crawford, H. C.; Daniel, J. M. The catenin p120^{ctn} Inhibits Kaiso-Mediated Transcriptional Repression of the b-Catenin/TCF Target Gene. *Exp. Cell. Res.* **2005**, *305*, 253-265.

(24) Daniel, J. M.; Reynolds, A. B. The Catenin p120^{ctn} Interacts with Kaiso, a Novel BTB/POZ Domain Zinc Finger Transcription Factor. *Mol. Cell. Biol.* **1999**, *19*, 3614-3623.

(25) Kelly, K. F.; Spring, C. M.; Otchere, A. A.; Daniel, J. M. NLS-Dependent Nuclear Localization of p120^{ctn} is Necessary to Relieve Kaiso-Mediated Transcriptional Repression. *J. Cell Sci.* **2004**, *117*, 2675-2686.

(26) Dai, S.-D.; Wang, Y.; Miao, Y.; Zhao, Y.; Zhang, Y.; Jiang, G-Y.; Zhang, P-X.; Yang, Z-Q.; Wang, E-H. Cytoplasmic Kaiso is Associated with Poor Prognosis in Nonsmall Cell Lung Cancer. *BMC Cancer* **2009**, *9*, 2471-2497.

(27) Dai, S.-D.; Wang, Y.; Jiang, G-Y.; Zhang, P-X.; Dong, X-J.; Wei, Q.; Xu, H-T.; Li, Q-C.; Zhao, C.; Wang, E-H. Kaiso is Expressed in Lung Cancer: Its Expression and Localization is Affected by p120ctn. *Lung Cancer* **2010**, *67*, 205-215.

(28) Kulikov, A. V.; Korostina, V. S.; Kulikova, E. A.; Fursenko, D. V.; Akulov, A. E.; Moshkin, M. P.; Prokhortchouk, E. B. Knockout Zbtb33 Gene Results in an Increased Locomotion, Exploration and Pre-Pulse Inhibition in Mice. *Behav. Brain Res.* **2016**, *297*, 76-83.

(29) Chaudhary, R.; Pierre, C. C.; Nanan, K.; Wojtal, D.; Morone, S.; Pinelle, C.; Wood, G. A.; Robine, S.; Daniel, J. M. The POZ-ZF Transcription Factor Kaiso (ZBTB33) Induces Inflamation and Progenitor Cell Differentiation in the Murine Intestine. *PLoS One* **2013**, *8*, e74160.

(30) Soubry, A.; Staes, K.; Parthoens, E.; Noppen, S.; Stove, C.; Bogaert, P.; van Hengel, J.; van Roy, F. The Transcriptional Repressor Kaiso Localizes at the Mitotic Spindle and is a Constituent of the Pericentriolar Material. *PLoS One* **2010**, *5*, e9203.

(31) Koh, D.-I.; Han, D.; Ryu, H.; Choi, W.-I.; Jeon, B.-N.; Kim, M.-K.; Kim, Y.; Kim, J. Y.; Parry, L.; Clarke, A. R.; Reynolds, A. B.; Hur, M.-W. KAISO, a Critical Regulator of p53-Mediated Transcription of CDKN1A and Apoptotic Genes. *Proc. Natl. Acad. Sci.* **2014**, *111*, 15078-15083.

(32) Cofre, J.; Menezes, J. R.; Pizzatti, L.; Abdelhay, E. Knock-down of Kaiso Induces Proliferation and Blocks Granulocytic Differentiation in Blast Crisis of Chronic Myeloid Leukemia. *Cancer Cell Int.* **2012**, *12*, 28.

(33) Capecchi, M. R.; Pozner, A. ASPM Regulates Symmetric Stem Cell Division by Tuning Cyclin E Ubiquitination. *Nat. Commun.* **2015**, *6*, 8763.

(34) Nix, D. A.; Courdy, S. J.; Boucher, K. M. Empirical Methods for Controlling False Positives and Estimating Confidence in ChIP-Seq Peaks. *Bioinformatics* **2008**, *9*, 523-531.

(35) Love, M. I.; Huber, W.; Anders, S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* **2014**, *15*, 550.

(36) Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **2009**, *25*, 2078-2079.

(37) 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.; Mesirov, J. P. Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proc. Natl. Acad. Sci.* **2005**, *102*, 15545-15550.

(38) Helt, A. M.; Galloway, D. A. Mechanisms by Which DNA Tumor Virus Oncoproteins Target the Rb Family of Pocket Proteins. *Carcinogenesis* **2003**, *24*, 159-169.

(39) Dorner, D.; Vlcek, S.; Foeger, N.; Gajewski, A.; Makolm, C.; Gotzmann, J.; Hutchison, C. J.; Foisner, R. Lamina-Associated Polypeptide 2alpha Regulates Cell Cycle Progression and Differentiation Via the Retinoblastoma-E2F Pathway. *J. Cell Biol.* **2006**, *173*, 83-93.

(40) Sherr, C. J. G1 Phase Progression: Cycling on Cue. *Cell* **1994**, *79*, 551-555.

(41) Clurman, B. E.; Sheaff, R. J.; Thress, K.; Groudine, M.; Roberts, J. M. Turnover of Cyclin E by the Ubiquitin-Proteasome Pathway is Regulated by cdk2 Binding and Cyclin Phosphorylation. *Genes Dev.* **1996**, *10*, 1979-1990.

(42) Bracken, A. P.; Ciro, M.; Cocito, A.; Helin, K. E2F Target Genes: Unraveling the Biology. *Trends Biochem. Sci.* **2004**, *29*, 409-417.

(43) Wang, X. D.; Rosales, J. L.; Magliocco, A.; Gnanakumar, R.; Lee, K. Y. Cyclin E in Breast Tumors is Cleaved into Its Low Molecular Weight Forms by Calpain. *Oncogene* **2003**, *22*, 769-774.

(44) Spruijt, C. G.; Vermeulen, M. DNA Methylation: Old Dog, New Tricks? *Nat. Struct. Mol. Biol.* **2014**, *21*, 949-954.

(45) Iurlaro, M.; Ficz, G.; Oxley, D.; Raiber, E. A.; Bachman, M.; Booth, M. J.; Andrews, S.; Balasubramanian, S.; Reik, W. A Screen for Hydroxymethylcytosine and Formylcytosine Binding Proteins Suggests Functions in Transcription and Chromatin Regulation. *Genome Biol.* **2013**, *14*, R119.

(46) Hu, S.; Wan, J.; Su, Y.; Song, Q.; Zeng, Y.; Nguyen, H. N.; Shin, J.; Cox, E.; Rho, H. S.; Woodard, C.; Xia, S.; Liu, S.; Lyu, H.; Ming, G. L.; Wade, H.; Song, H.; Qian, J.; Zhu, H. DNA Methylation Presents Distinct Binding Sites for Human Transcription Factors. *Elife* **2013**, *2*, e00726.

(47) Yao, G.; Lee, T. J.; Mori, S.; Nevins, J. R.; You, L. A Bistable Rb-E2F Switch Underlies the Restriction Point. *Nat. Cell Biol.* **2008**, *10*, 476-482.

(48) Sherr, C. J.; Roberts, J. M. Living With or Without Cyclins and Cyclin-Dependent Kinases. *Genes Dev.* **2004**, *18*, 2699-2711.

(49) Bartek, J.; Bartkova, J.; Lukas, J. The Retinoblastoma Protein Pathway and the Restriction Point. *Curr. Opin. Cell Biol.* **1996**, *8*, 805-814.

(50) Dick, F. A.; Sailhamer, E.; Dyson, N. J. Mutagenesis of the pRB Pocket Reveals that Cell Cycle Arrest Functions are Separable from Binding to Viral Oncoproteins. *Mol. Cell Biol.* **2000**, *20*, 3715-3727.

(51) Chano, T.; Saji, M.; Inoue, H.; Minami, K.; Kobayashi, T.; Hino, O.; Okabe, H. Neuromuscular Abundance of RB1CC1 Contributes to the Nonproliferating Enlarged cell Phenotype Through both RB1 Maintenance and TSC1 Degradation. *Int. J. Mol. Med.* **2006**, *18*, 425-432.

(52) Vance, K. W.; Shaw, H. M.; Rodriguez, M.; Ott, S.; Goding, C. R. The Retinoblastoma Protein Modulates Tbx2 Functional Specificity. *Mol. Biol. Cell.* **2010**, *21*, 2770-2779.

(53) Goodwin, E. C.; DiMaio, D. Repression of Human Papillomavirus Oncogenes in HeLa Cervical Carcinoma Cells Causes the Orderly Reactivation of Dormant Tumor Suppressor Pathways. *Proc. Natl. Acad. Sci.* **2000**, *97*, 12513-12518.

(54) Zarkowska, T.; Mittnacht, S. Differential Phosphorylation of the Retinoblastoma Protein by G1/S Cyclin-Dependent Kinases. *J. Biol. Chem.* **1997**, *272*, 12738-12746.

(55) Classon, M.; Harlow, E. The Retinoblastoma Tumour Suppressor in Development and Cancer. *Nat. Rev. Cancer* **2002**, *2*, 910-917.

(56) Malumbres, M.; Barbacid, M. Cell Cycle, CDKs and Cancer: a Changing Paradigm. *Nat. Rev. Cancer* **2009**, *9*, 153-166.

(57) Roy, P. G.; Thompson, A. M. Cyclin D1 and Breast Cancer. *Breast* **2006**, *15*, 718-727.

(58) Lents, N. H.; Baldassare, J. J. CDK2 and Cyclin E Knockout Mice: Lessons From Breast Cancer. *Trends Endocrinol. Metab.* **2004**, *15*, 1-3.

(59) Jones, J.; Wang, H.; Zhou, J.; Hardy, S.; Turner, T.; Austin, D.; He, Q.; Wells, A.; Grizzle, W. E.; Yates, C. Nuclear Kaiso Indicates Aggressive Prostate Cancers and Promotes Migration and Invasiveness of Prostate Cancer Cells. *Am. J. Pathol.* **2012**, *181*, 1836-1846.

(60) Comstock, C. E.; Revelo, M. P.; Buncher, C. R.; Knudsen, K. E. Impact of Differential Cyclin D1 Expression and Localisation in Prostate Cancer. *Br. J. Cancer* **2007**, *96*, 970-979.

(61) Ozbek, E.; Mizrak, B.; Ozbek, M.; Buyukberber, S.; Davarci, M. Cyclin-D1 Protooncogen Expression in Prostate Cancer. *Turk. J. Cancer* **2000**, *30*, 15-23.

(62) Yamamoto, A.; Hashimoto, Y.; Kohri, K.; Ogata, E.; Kato, S.; Ikeda, K.; Nakanishi, M. Cyclin E as a Coactivator of the Androgen Receptor. *J. Cell Biol.* **2000**, *150*, 873-880.

(63) Bassey-Archibong, B. I.; Kwiecien, J. M.; Milosavljevic, S. B.; Hallett, R. M.; Rayner, L. G.; Erb, M. J.; Crawford-Brown, C. J.; Stephenson, K. B.; Bedard, P. A.; Hassell, J. A.; Daniel, J. M. Kaiso Depletion Attenuates Transforming Growth Factor-Beta Signaling and Metastatic Activity of Triple-Negative Breast Cancer Cells. *Oncogenesis* **2016**, *5*, e208.

(64) Jones, J.; Wang, H.; Karanam, B.; Theodore, S.; Dean-Colomb, W.; Welch, D. R.; Grizzle, W.; Yates, C. Nuclear Localization of Kaiso Promotes the Poorly Differentiated Phenotype and EMT in Infiltrating Ductal Carcinomas. *Clin. Exp. Metastasis* **2014**, *31*, 497-510.

(65) Vermeulen, J. F.; van de Ven, R. A.; Ercan, C.; van der Groep, P.; van der Wall, E.; Bult, P.; Christgen, M.; Lehmann, U.; Daniel, J.; van Diest, P. J.; Derksen, P. W. Nuclear Kaiso Expression is Associated with High Grade and Triple-Negative Invasive Breast Cancer. *PLoS One* **2012**, *7*, e37864.