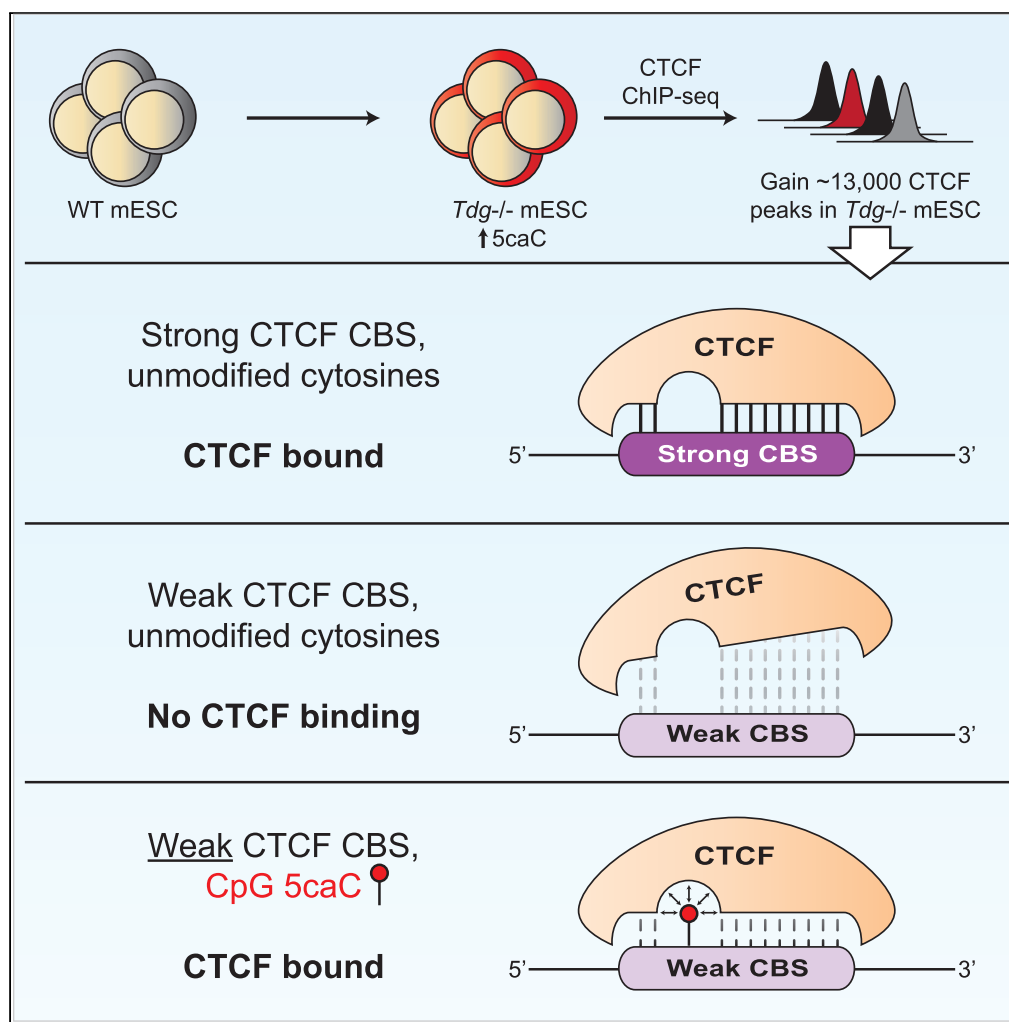**Article**

TET-Catalyzed 5-Carboxylcytosine Promotes CTCF Binding to Suboptimal Sequences Genome-wide

Kyster K. Nanan,
David M. Sturgill,
Maria F. Prigge,
Morgan Thenoz,
Allissa A. Dillman,
Mariana D.
Mandler, Shalini
Oberdoerffer

shalini.oberdoerffer@nih.gov

HIGHLIGHTS

TET-catalyzed 5-carboxylcytosine (5caC) strengthens CTCF association with DNA

5caC accumulation in genomic DNA through *Tdg* deletion produced ~13,000 new CTCF sites

Gains in CTCF binding were most pronounced at locations with suboptimal sequence motifs

5caC promoted CTCF binding to otherwise refractory sequences *in vitro*

DATA AND CODE AVAILABILITY

GSE123101

Article

TET-Catalyzed 5-Carboxylcytosine Promotes CTCF Binding to Suboptimal Sequences Genome-wide

Kyster K. Nanan,[1,2,5] David M. Sturgill,[1,5] Maria F. Prigge,[1] Morgan Thenoz,[1,3] Allissa A. Dillman,[1,4] Mariana D. Mandler,[1] and Shalini Oberdoerffer[1,6,*]

SUMMARY

The mechanisms supporting dynamic regulation of CTCF-binding sites remain poorly understood. Here we describe the TET-catalyzed 5-methylcytosine derivative, 5-carboxylcytosine (5caC), as a factor driving new CTCF binding within genomic DNA. Through a combination of *in vivo* and *in vitro* approaches, we reveal that 5caC generally strengthens CTCF association with DNA and facilitates binding to suboptimal sequences. Dramatically, profiling of CTCF binding in a cellular model that accumulates genomic 5caC identified ~13,000 new CTCF sites. The new sites were enriched for overlapping 5caC and were marked by an overall reduction in CTCF motif strength. As CTCF has multiple roles in gene expression, these findings have wide-reaching implications and point to induced 5caC as a potential mechanism to achieve differential CTCF binding in cells.

INTRODUCTION

CTCF is an 11 zinc-finger DNA-binding protein that regulates multiple critical genomic functions, including promoting long-range interactions between distal regions of the genome and insulating areas of active transcription from inactive regions (Phillips and Corces, 2009). Profiling of CTCF binding in human cells suggests tens of thousands of binding sites, more than half of which display tissue specificity (Chen et al., 2012; The ENCODE Project Consortium, 2012; Wang et al., 2012). Regulation of CTCF binding at variable locations is largely achieved through dynamic DNA methylation, wherein overlapping 5-methylcytosine (5mC) inhibits CTCF association with DNA (Bell and Felsenfeld, 2000; Hark et al., 2000; Maurano et al., 2015; Shukla et al., 2011; The ENCODE Project Consortium, 2012; Wang et al., 2012). Among the described functions for variable CTCF sites, we identified a role in alternative pre-mRNA splicing (Shukla et al., 2011). CTCF binding within actively transcribed genes transiently obstructs RNA polymerase II (pol II) elongation, thereby kinetically favoring spliceosome assembly at weak upstream splice sites (Shukla et al., 2011). In contrast, inhibition of CTCF binding through overlapping 5mC shifts splicing to competing downstream sites through loss of pol II pausing (Shukla et al., 2011). However, the mechanisms that dynamically regulate CTCF exchange in alternative splicing and other tissue-specific activities remained unknown.

We recently determined that the alpha-ketoglutarate-dependent dioxygenases, TET1 and TET2, support CTCF function in splicing regulation by antagonizing overlapping 5mC at CTCF-binding sites (Marina et al., 2016). The TET proteins catalyze active DNA demethylation through successive oxidation of 5mC to 5-hydroxymethycytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxycytosine (5caC) (Ito et al., 2011; Tahiliani et al., 2009). In the final step in the demethylation pathway, 5caC is converted to cytosine through the base-excision repair factor thymine DNA glycosylase (TDG) (He et al., 2011; Maiti and Drohat, 2011). In contrast, reduced TET activity results in increased 5mC at CTCF-binding sites and associated exclusion of dependent upstream exons from spliced mRNA due to CTCF eviction (Marina et al., 2016). Curiously, the underlying DNA at these splicing-associated CTCF sites was not fully unmethylated but was rather marked by a steady level of 5caC (Marina et al., 2016). Although 5caC levels in genomic DNA are generally low, we readily detected the oxidized derivative within CTCF sites in actively dividing primary peripheral lymphocytes (Marina et al., 2016). Biochemical characterization confirmed CTCF interaction with 5caC-containing DNA *in vitro* that was, unexpectedly, enhanced when compared with unmethylated DNA (Marina et al., 2016). However, the significance of CTCF association with 5caC *in vivo* remained unclear.

Here we directly examine whether and how 5caC influences CTCF binding in cells. By utilizing a cellular system that accumulates 5caC within genomic DNA (Cortazar et al., 2011), we observe a dramatic expansion in locations of CTCF binding, genome-wide. Characterization of CTCF binding at these *de novo* sites revealed unique features, including an enrichment for overlapping 5caC and loosening of the consensus CTCF-binding motif. Likewise, 5caC was found enriched at low-motif CTCF sites in primary T cells. CTCF and 5caC profiling in primary T cells and biochemical analysis support the notion that 5caC reinforces CTCF binding in suboptimal contexts. Together, these results provide a rationale for a perplexing aspect of CTCF biology: CTCF physically interacts with DNA methyltransferases (DNMTs) that establish 5mC in genomic DNA and the TET proteins that site-specifically oxidize 5mC in a pathway to demethylation (Dubois-Chevalier et al., 2014; Guastafierro et al., 2008; Zampieri et al., 2012). Our results raise the intriguing possibility that CTCF association with these factors acts to reinforce its own binding, and potentially create a platform for the dynamic regulation of CTCF binding as DNMT and TET levels vary during development.

RESULTS

De Novo CTCF Sites Overlap with 5caC-Rich Regions in *Tdg*−/− Cells

Given the multiple critical roles played by CTCF in the nucleus, a deeper understanding of the molecular determinants that drive CTCF binding is warranted. We recently identified 5caC as one such factor, wherein we observed that purified CTCF showed increased interaction with 5caC-containing when compared with unmethylated DNA in electrophoretic mobility shift assays (EMSA) (Marina et al., 2016). However, whether 5caC enhances CTCF binding within the complex environment of chromosomal DNA was unclear. To begin to address CTCF/5caC binding *in vivo*, we turned to knockout mouse embryonic stem cells (mESCs) lacking the base-excision factor Tdg (*Tdg*−/−) (Figure 1A) (Cortazar et al., 2011). Others previously established that Tdg depletion can be leveraged to boost the otherwise low level of 5caC within genomic DNA, without compromising other aspects of base-excision repair (Cortazar et al., 2011; He et al., 2011). Loss of Tdg signal in knockout mESCs was confirmed by immunoblotting (Figure 1B). Of note, mESCs were maintained on wild-type (*Tdg*+/+) mouse embryonic fibroblast (MEF) feeders to avoid unintended and uneven differentiation. Accordingly, residual Tdg detection in the knockout population can be attributed to minor MEF contamination during cell harvest. However, given that 5caC levels are exceedingly low in differentiated tissues (Ito et al., 2011; Wu and Zhang, 2017; Zhu et al., 2018), it is unlikely that residual fibroblasts substantially contributed to our analysis at the genome-wide level. In direct support of the integrity of the genomic results, MEF-specific mRNAs were severely depleted in RNA sequencing from wild-type and *Tdg*−/− mESCs (Figure S1A) and strain-specific single-nucleotide polymorphisms were appropriately assigned in sequencing of DNA and RNA inputs from wild-type and *Tdg*−/− mESCs (Figure S1B). Importantly, 5caC levels were substantially increased in *Tdg*−/− mESCs and compared to wild-type mESC genomic DNA, as evidenced by dot blot with 5caC-specific antibody (Figure 1C).

Having confirmed the elevated presence of 5caC in *Tdg*−/− cells, we next explored the consequence to CTCF binding genome-wide. CTCF chromatin immunoprecipitation sequencing (ChIP-seq) was performed in duplicate in wild-type and *Tdg*−/− mESCs (Figures 1D and S1C, Table S1). Consistent with previous reports, peak-calling indicated ~40,000 CTCF peaks that were commonly detected in both mESC populations (Chen et al., 2012; The ENCODE Project Consortium, 2012; Wang et al., 2012). However, CTCF sites that were uniquely detected in one population or the other showed a dramatic imbalance: whereas ~1,900 sites were present in wild-type but not in *Tdg*−/− mESCs, nearly ~13,000 sites were found in *Tdg*−/− but not wild-type mESCs (Figure 1D, Table S2). In other words, Tdg loss was associated with a substantial increase in CTCF binding genome-wide. This increase in CTCF binding was not related to a change in expression, as immunoblotting revealed comparable CTCF protein levels in wild-type and *Tdg*−/− mESCs (Figure 1E). With this in mind, it is notable that comparison of CTCF ChIP-seq read density within the *Tdg*−/− induced sites showed that CTCF binding was not entirely absent in wild-type cells, but rather failed to reach the threshold for peak detection (Figure 1F). Importantly, CTCF signal was not observed in randomly shuffled sites, thus establishing the validity of the called peaks (Figure 1F). Likewise, CTCF-ChIP-seq peaks showed high replicate concordance (Figure S1D). These findings suggest that *de novo* CTCF binding in *Tdg*−/− cells results from cellular change unrelated to CTCF abundance that reinforces binding at otherwise weak sites.

Based on our previous EMSA showing enhanced CTCF binding in the presence of 5caC, we specifically queried CTCF sites in *Tdg*−/− cells for emergent 5caC. To this end, we examined 5caC-methylated
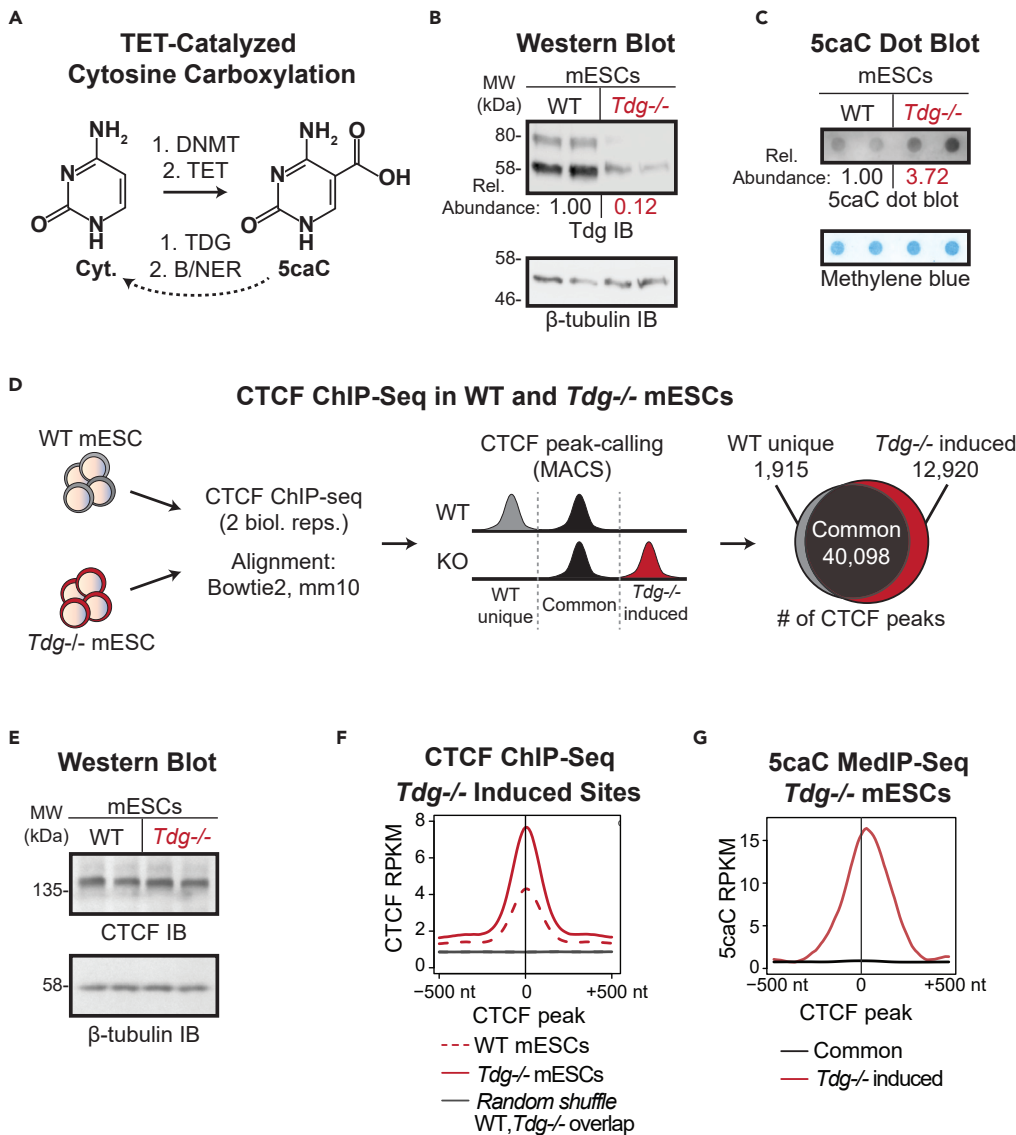
Figure 1. 5caC Enhances CTCF Binding to Genomic DNA

(A) Schematic of cytosine and 5-carboxylcytosine (5caC) molecular structures.

(B) Tdg immunoblot in wild-type (WT) and *Tdg−/−* mESCs. Replicates are derived from independently maintained cultures. Relative Tdg abundance when compared with WT control is determined through β-tubulin normalization.

(C) 5caC dot blot in WT and *Tdg−/−* mESC genomic DNA. Replicates as in (B) with methylene blue staining as loading control. Relative 5caC abundance when compared with WT control is determined through methylene blue normalization.

(D) CTCF peaks determined through ChIP-seq in WT and *Tdg−/−* mESCs were parsed by occurrence in only WT or *Tdg−/−* cells (unique) or in both populations (common). The Venn diagram depicts the number of CTCF peaks per category.

(E) CTCF immunoblot in WT and *Tdg−/−* mESCs.

(F) CTCF ChIP-seq signal in WT and *Tdg−/−* mESCs centered on locations uniquely detected in *Tdg−/−* cells. Reduced CTCF signal in WT mESCs failed to reach the threshold for peak calling. Overlapping gray lines indicate CTCF signal in randomly shuffled sites from WT and *Tdg−/−* cells.

(G) Averaged 5caC meDIP-seq in *Tdg−/−* mESCs centered on CTCF peaks unique to *Tdg−/−* cells or commonly detected in *Tdg−/−* and WT mESCs.

DNA-immunoprecipitation sequencing (meDIP-seq) data from *Tdg*-depleted mESCs (Shen et al., 2013). In total, we identified >172,000 discrete genomic regions with 5caC enrichment in *Tdg−/−* cells. Intuitively, the bulk of these 5caC-enriched locations were not punctuated by overlapping CTCF, as would be

expected within the genomic environment wherein modified cytosines exist in numerous sequence contexts and diverse transcription factors are in competition for limited binding sites. Nevertheless, co-occurrence was statistically enriched when compared with random genomic bins (Figure S1E). Critically, when reciprocally focused within CTCF binding sites (CBS), a remarkable association with 5caC emerges: a robust 5caC signal was observed directly within Tdg−/− induced CTCF peaks (Figure 1G). Of note, enrichment for overlapping 5caC was not exclusive to Tdg−/− induced sites when compared with random genomic bins, but was rather also observed in CBS that were commonly detected in wild-type and Tdg−/− cells (Figure S1F). It is relevant in this regard that overall CTCF signal was also increased at common CBS, although to a lesser extent (Figures S1F and S1G). Overall, these data raise the intriguing possibility that 5caC globally reinforces CTCF binding in chromosomal DNA, allowing otherwise low-penetrance CBS to reach the threshold for detection and potentially biological relevance.

Tdg−/− Induced CTCF Sites Display Unique Molecular Features In Vivo

Having established the presence of 5caC-rich induced CTCF sites in Tdg−/− cells, we next explored their molecular basis and biological relevance. In particular, we examined for unique features in the induced subset as influenced by motif strength. Our rationale derived from the following two observations: (1) CTCF association with DNA does not adhere to a strict consensus sequence (Nakahashi et al., 2013; Rhee and Pugh, 2011) and (2) weak CTCF binding was observed in wild-type mESCs at locations that were robustly detected in Tdg−/− cells (Figure 1F). These findings reflect a nuanced aspect of CTCF biology: the 11 zinc fingers associate with DNA to varying extents, resulting in degrees of binding strength and a relatively degenerate motif (Hashimoto et al., 2017; Nakahashi et al., 2013; Rhee and Pugh, 2011). We thus reasoned that overlapping 5caC at Tdg−/− induced CTCF sites may reflect a positive role for the cytosine modification in suboptimal contexts. Empirically determined CTCF-binding sites in wild-type and Tdg−/− mESCs were accordingly assigned motif scores through the CTCFBSDB 2.0 database (Figure 2A, schematic) (Ziebarth et al., 2013). Considering that 5caC was globally enriched in Tdg−/− induced CTCF sites (Figures 1G, S1D, and S1E), it is notable that the de novo sites were generally characterized by reduced motif strength when compared with commonly detected peaks (Figure 2B). To specifically interrogate the relationship between 5caC and motif strength, CTCF peaks were segregated into groups representing low, mid, and high motif scores (bottom quartile, middle two quartiles, and top quartile, respectively) (Figure 2A). This grouping placed 13,183 sites in the low-motif group. Examination of 5caC content within CTCF sites not only confirmed a general enrichment for 5caC across the spectrum of Tdg−/− induced sites but also revealed an inverse relationship to CTCF motif strength, wherein 5caC levels were highest in the low-motif subset (Figure 2C). Browser shot examples of genomic data depicting increased CTCF binding at low-motif locations associated with overlapping 5caC in Tdg−/− cells are shown (Figure S2A). As 5caC is most frequently observed within CpG dinucleotide contexts, we further examined CpG content within CTCF sites parsed by motif strength. In agreement with increased substrate density, low-motif sites were characterized by higher overall CpG density when compared with mid- or high-motif locations, thus providing a rationale for the increased prevalence of 5caC (Figure 2D). These findings support a role for 5caC in promoting new CTCF binding in Tdg−/− cells, particularly to suboptimal motifs.

The above-mentioned data establish the presence of 5caC-rich CTCF sites in Tdg−/− cells, but to confirm functionality, we examined for co-occurrence of RNA polymerase II (pol II). We and others have shown that CTCF binding within genic DNA transiently obstructs pol II elongation, resulting in local accumulation of pol II at CTCF sites (Lu and Tang, 2012; Paredes et al., 2013; Shukla et al., 2011). Consistent with the general distribution of CTCF binding, alignment to genomic features showed that ~45% of the Tdg−/− induced sites occur within gene bodies (Figure S2B, Table S3). To assess pol II levels within the induced sites, we performed pol II ChIP-seq in Tdg−/− cells. General examination of pol II read density within both common and induced genic CTCF sites revealed a clear accumulation that is consistent with bona fide CTCF binding, which was absent in randomized regions (Figure 2E). To further explore the relationship to 5caC, pol II occurrence was examined within genic Tdg−/− induced CTCF sites segregated on the basis of overlapping 5caC (5caC-rich (+) or 5caC-poor (−)). Accumulating pol II was detected in both classes of CTCF sites, whereas levels were markedly elevated for the 5caC(+) sites (Figure 2F). Although somewhat unexpected, this finding is consistent with our previous demonstration of increased CTCF binding to a 5caC-containing probe in EMSA assay when compared with unmodified probe (Marina et al., 2016). Thus, the observed increase in pol II density at 5caC-rich CTCF sites in Tdg−/− cells raises the possibility that 5caC both promotes and strengthens CTCF binding in vivo. Importantly, examination of pol II in 5caC-rich versus poor
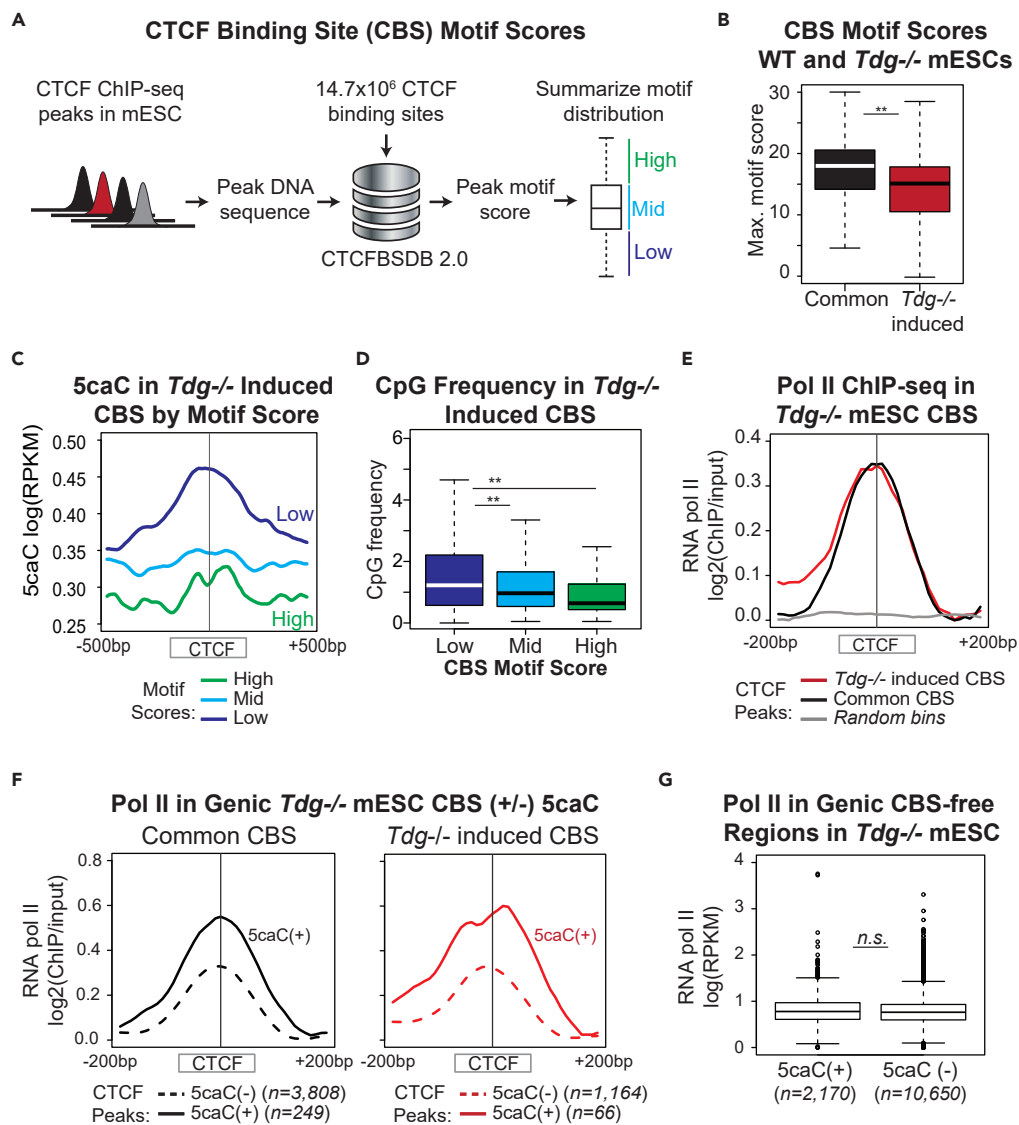
Figure 2. Low-Motif CTCF Peaks Exhibit Greatest Gains in Binding at 5caC-rich DNA

(A) CTCF ChIP-seq peaks in WT and Tdg−/− mESCs were assigned motif scores through the CTCFBSDB 2.0 algorithm and segregated into high-, mid-, and low-motif bins.

(B) Box-and-whisker diagram of maximum motif score distribution in common versus Tdg−/− induced CTCF peaks. **p < 2.2 × 10^{-16}, Mann-Whitney U test.

(C) Tdg−/− mESC 5caC meDIP-seq within empirically determined CTCF peaks, segregated by motif score. meDIP-seq abundance is given in natural log scale.

(D) CpG dinucleotide frequency within Tdg−/− mESC CTCF peaks, segregated by motif score. **p < 2.2 × 10^{-16}, Mann-Whitney U test.

(E) Tdg−/− mESC RNA pol II ChIP-seq density within common and Tdg−/− induced CTCF peaks relative to randomly permutated (shuffled) control peaks. Within each window, bins are normalized so that aggregate signal reflects local polymerase stalling.

(F) Tdg−/− mESC Pol II ChIP-seq in common and Tdg−/− induced CTCF sites as in (E), segregated on the basis of overlapping 5caC. Analysis is performed on the subset of CTCF sites located within expressed genes.

(G) Tdg−/− mESC pol II ChIP-seq at 5caC-rich (+) and 5caC-poor (−) regions not associated with proximal CTCF binding. n.s. = not significant (p > 0.05), Mann-Whitney U test.

genic regions that are not marked by proximal CTCF showed no distinction, demonstrating that accumulating pol II is not a general feature of 5caC-rich DNA (Figures 2G and S2C). Together, these results establish both the presence and functionality of 5caC-associated CTCF sites in Tdg−/− cells.
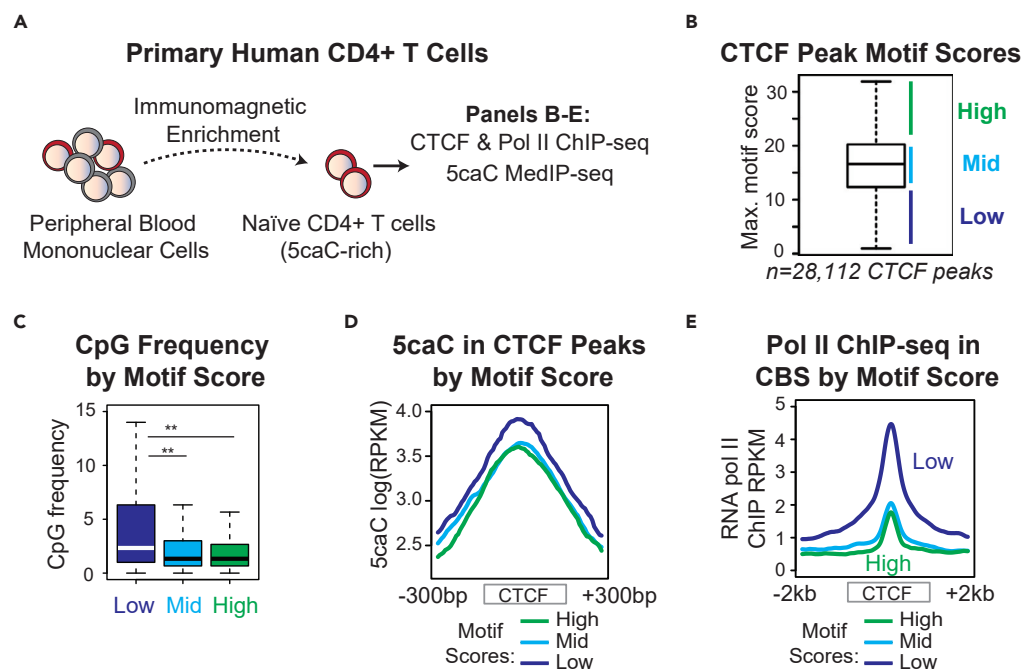
Figure 3. 5caC Associates with Low-Motif CTCF-Binding Sites in Primary Human CD4+ T Cells

(A) Naive CD4+ T cells were isolated from human peripheral blood through immunomagnetic enrichment and subjected to genome-wide analyses.

(B) Distribution of maximum CTCFBSDB 2.0-determined motif scores at CTCF peaks detected in naive T cells. Peaks were segregated into high, mid, and low bins corresponding to scores within the top quartile, middle two quartiles, and bottom quartile, respectively.

(C) CpG dinucleotide frequency within CD4+ T cell CTCF peaks, segregated by motif score. **$p < 2.2 \times 10^{-16}$, Mann-Whitney U test.

(D and E) 5caC meDIP-seq (D) and RNA pol II ChIP-seq (E) within determined CD4+ T cell CTCF peaks, segregated by motif score; (D) includes sites with 5caC abundance greater than genomic baseline (determined by shuffled control).

Of note, mESCs are qualitatively distinct from differentiated tissues on several accounts. Most relevant to the current analysis, embryonic stem cells are uniquely characterized by non-CpG methylation and overall higher levels of 5mC-oxidized derivatives (Guo et al., 2014; Huang et al., 2014; Ito et al., 2010, 2011; Koh et al., 2011; Lister et al., 2009; Ramsahoye et al., 2000). Thus, to examine the generality of our findings outside of mESCs, we turned to primary human lymphocytes. We previously co-detected 5caC and CTCF at pre-mRNA splicing-associated regions in naive CD4+ T lymphocytes (Marina et al., 2016). TET1 and TET2 expression was high in naive CD4+ T cells, whereas levels decreased upon activation (Marina et al., 2016). Accordingly, we isolated naive CD4+ T cells from peripheral blood for genome-wide analysis of CTCF and 5caC through ChIP-seq and meDIP-seq, respectively (Figure 3A). As in the mESC analysis, experimentally determined CTCF sites were parsed based on motif strength into low-, mid-, and high-scoring groups (Figure 3B) with the bottom 25% of scores (7,016 CBS) comprising the low-motif group. Concordant with the mESC results, low-motif CTCF sites in CD4+ T cells were marked by increased CpG density when compared with the mid- and high-scoring groups (Figure 3C). Intriguingly, the low-motif cohort showed the greatest change in CTCF occupancy during the naive to activated transition, wherein a net decrease in CTCF binding was observed (Figure S3A). In support of developmental regulation through overlapping 5caC, low-motif CTCF sites in naive CD4+ T cells were further marked by elevated overlapping 5caC when compared with the mid- and high-scoring groups (Figure 3D). A browser shot example depicting reduced 5caC coinciding with decreased CTCF binding at a low-motif CBS in activated versus naive T cells is shown (Figure S3B). Given that it is commonly accepted that 5caC levels within genomic DNA are too low for basal detection, 5caC occurrence within CTCF sites in an unperturbed cellular setting is noteworthy and attests to the relevance of association. Indeed, analysis of pol II ChIP-seq performed in naive CD4+ T cells revealed a clear enrichment for pol II within low-motif CTCF sites, which also showed the highest levels of overlapping 5caC (Figure 3E). Of note, pol II levels were highest in the low-motif

set firmly establishing these sites as locations of CTCF binding with a bona fide role in the regulation of a biological function.

CTCF Binds Motif-Free DNA in the Presence of 5caC *In Vitro*

Our genome-wide data strongly support a role for 5caC in driving CTCF binding in a chromosomal setting. However, to move beyond correlations, we turned to *in vitro* systems for biochemical characterization of CTCF association with 5caC-containing DNA. In particular, based on the observed association of CTCF with 5caC-rich DNA at low-motif sites, we examined whether 5caC promotes CTCF binding in suboptimal contexts through EMSA assays. Although EMSA varies from chromosomal DNA in the use of linear DNA probes, concerns related to artificiality are mitigated by the fact that CTCF binds to nucleosome-free DNA *in vivo* (Chen et al., 2012; Fu et al., 2008; Magbanua et al., 2015; Teif et al., 2014). Importantly, we previously revealed a surprising preference for 5caC-containing when compared with unmethylated DNA in CTCF EMSA: overall complex formation was increased in the presence of overlapping 5caC, and unlabeled 5caC was a generally more effective competitor when compared with unmethylated competitor (Marina et al., 2016). These results were observed with multiple probes and distinct CTCF protein sources including recombinant CTCF and FLAG-tagged CTCF purified from cell culture (Marina et al., 2016). These findings are consistent with the genome-wide analysis of CTCF binding in *Tdg−/−* cells and highlight EMSA as an appropriate method for examining CTCF interaction with carboxylated DNA *in vitro*.

To validate the relationship between 5caC and CTCF motif strength uncovered in the genome-wide analysis, we generated EMSA probes embodying distinct CTCF-binding modalities in the presence and absence of 5caC. To represent strong ("CBS-high") and weak ("CBS-low") motifs, we utilized a CTCF-binding site located within exon 5 of the *CD45* gene in either a wild-type or mutated context. We previously performed extensive characterization of the *CD45* probes and showed that the incorporation of point mutations within the CTCF core to generate the CBS-low probe abolished CTCF binding in EMSA with unmodified DNA (Marina et al., 2016; Shukla et al., 2011). Importantly, the introduced mutations do not alter CpG or general cytosine density in double-stranded DNA. For reference, the CTCFBSDB 2.0 algorithm revealed a score of 18.16 associated with the CBS-high CTCF-binding site, whereas the CBS-low probe produced a score of 7.26 (Figure 4A). EMSA was thus performed with FLAG-tagged CTCF purified from HEK293T lysates and radiolabeled *CD45* probes. 5caC incorporation was accomplished through (1) PCR amplification in the presence of 2′-deoxy-5-carboxycytidine 5′-triphosphate (5-carboxy-dCTP) or dCTP and (2) commercial synthesis. The PCR approach results in a 72-bp probe centered on the CTCF-binding site with uniformly modified cytosines, whereas commercial synthesis yields a 41mer with a total of three modified cytosines per DNA stand, occurring exclusively in a CpG context (Figure 4A). Although CTCF binding is generally compromised on shorter probes such as the 41mer (Marina et al., 2016), both approaches yielded consistent and clear results. In the presence of unmethylated DNA, CTCF bound to the CBS-high, but not the CBS-low, site, as evidenced by complex formation in phosphorimager analysis (Figure 4B). Percent shift was calculated as the amount of label in complex with CTCF relative to free label (Figure 4B). In contrast, CTCF formed a robust and specific complex with both the CBS-high and CBS-low probes in the presence of overlapping 5caC, as demonstrated by cold competition and supershift with anti-CTCF antibody (Figure 4B). Importantly, cold competition revealed a reproducible degree of distinctness between the unmodified and 5caC-containing complexes. In the PCR-generated probes with uniform incorporation of cytosine or 5caC, unmodified competitor more effectively competed unmodified complexes, whereas carboxylated competitor more effectively competed carboxylated complexes (Figure 4B, left). This distinction may reflect a cumulative change in DNA structure or charge in heavily modified substrates, which in this case includes 5caC outside of a CpG context. In support of this notion, in examining the 41mers with CpG-restricted modification, CBS-high carboxylated and unmodified probes were comparably efficient at competing CTCF association with unmodified CBS-high DNA, whereas 5caC-containing competitor was more effective at disrupting complexes involving carboxylated DNA. Furthermore, whereas the unmodified CBS-low probe was an ineffective competitor in any circumstance, the low-motif 5caC-containing probe was on par with the carboxylated CBS-high competitor (Figure 4B, right). Overall, these EMSA results highlight the capacity of 5caC to overcome the sequence penalty associated with weak-motif CTCF sites.

Of note, the genome-wide analysis of *Tdg−/−* induced CTCF sites revealed sequences for which the likelihood of CTCF binding was less than expected in random sequence (negative motif scores). Likewise, an unbiased mass spectrometry study that identified CTCF as a 5caC-specific reader utilized a probe entirely
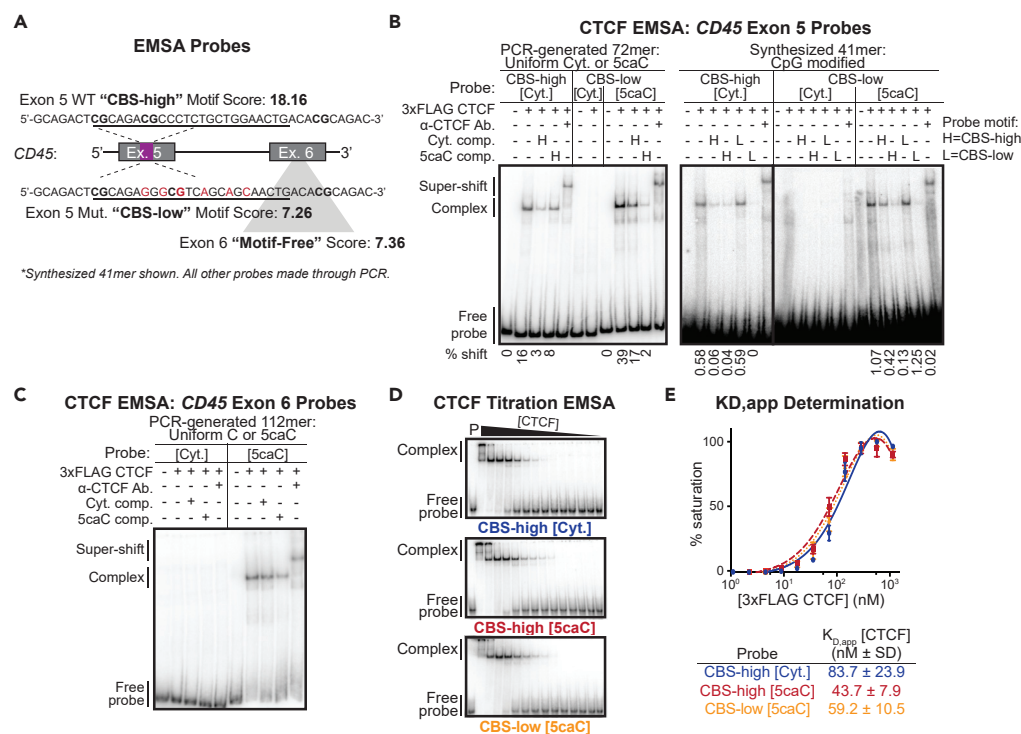
Figure 4. 5caC Enhances CTCF Association with Weak Binding Sites *In Vitro*

(A) Schematic of EMSA probes derived from the human *CD45* gene. Point mutations were incorporated into a strong-motif CTCF site (CBS-high) within exon 5 to generate an analogous weak-motif probe (CBS-low). A region devoid of CTCF binding *in vivo* within exon 6 was selected as a "motif-free" control. CTCFBSDB 2.0 motif scores are shown. Underlined sequences indicate the exon 5 CTCF-binding site location, red text highlights mutated residues, and sequences in bold denote CpG locations.

(B) EMSA with affinity-purified CTCF and radiolabeled PCR-generated 72mer or commercially synthesized 41mer DNA probes with uniform or CpG-restricted cytosine [Cyt.] and 5-carboxylcytosine [5caC], respectively. Unlabeled CBS-high (H) or CBS-low (L) motif cold-competitor DNA added to indicated lanes; 10X and 100X molar excess used for 72mer and 41mer, respectively. Supershift performed with α-CTCF antibody. Quantification of bound probe indicated as % shift below figure; representative of n ≥ 3.

(C) EMSA using affinity-purified CTCF and motif-free PCR-generated *CD45* exon 6 probes of varying carboxylation status; 10X molar excess of unlabeled *CD45* exon 6 cold competitor where indicated; representative of n = 2.

(D) Titration EMSA performed using 41mer *CD45* exon 5 DNA probes (3.76 nM); P indicates lane containing probe alone, CTCF concentration: 342.3–0.0418 nM.

(E) Saturation binding curves, derived from EMSAs in (D), and summary chart of apparent K_D values for each 41mer probe; markers on curves indicate mean of two (CBS-high [Cyt.]) or three (CBS-high [5caC], CBS-low [5caC]) independent replicates, and error bars indicate standard deviation.

lacking elements consistent with known CTCF-interacting sequences (−8.04 motif score) (Spruijt et al., 2013; Ziebarth et al., 2013). To examine whether we could recapitulate CTCF binding to seemingly motif-free DNA in the presence of 5caC, we performed additional EMSA with a sequence lacking any characteristics of CTCF binding. Specifically, *CD45* exon 6 does not contain any computationally predicted CTCF-binding sites (7.36 motif score, Figure 4A) and shows no evidence of CTCF binding in ChIP-qPCR and ChIP-seq (Marina et al., 2016; Shukla et al., 2011; The ENCODE Project Consortium, 2012). Exon 6 EMSA probes were generated by PCR in the presence of dCTP or 5-carboxy-dCTP. As predicted, purified CTCF did not interact with the unmethylated exon 6 probe under established binding conditions. In contrast, the 5caC-containing exon 6 probe formed a robust and specific complex with CTCF (Figure 4C). This unexpected finding clearly demonstrates the positive impact of 5caC on CTCF binding *in vitro*.

The above EMSAs establish that CTCF binding is qualitatively enhanced in the presence of 5caC. To further quantify the strength of association, we pursued relative binding affinity determination ($K_{D,apparent(app)}$) through saturation binding experiments involving purified CTCF and radiolabeled *CD45* exon 5 41mers.
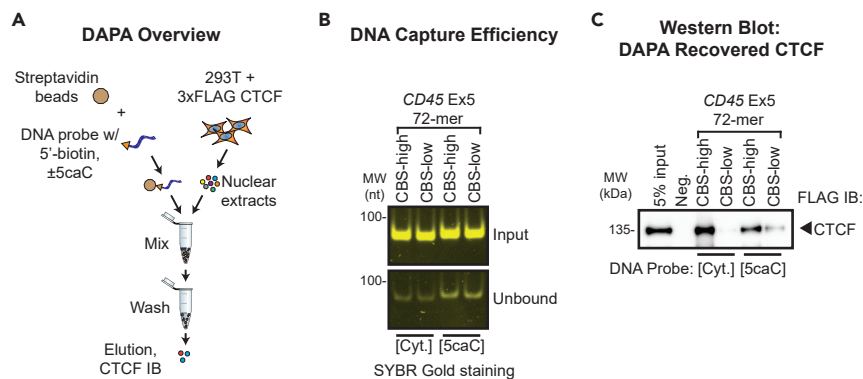
Figure 5. Affinity Purification of CTCF from Nuclear Lysates with 5caC-Containing DNA

(A) Overview of CTCF DNA affinity precipitation assay (DAPA). Bead-bound DNA probes generated in the presence or absence of 5caC were used to enrich CTCF from nuclear lysates.

(B) SYBR Gold-staining to assess DNA probe capture efficiency on streptavidin magnetic beads. *CD45* exon 5 probes representing strong (CBS-high) and weak (CBS-low) CTCF motifs generated through PCR in the presence of dCTP (Cyt.) or 5-carboxy-dCTP (5caC) were used.

(C) Immunoblotting of DAPA-recovered CTCF through the various probes.

EMSA was performed with a fixed amount of *CD45* probe representing CBS-high (+/− CpG 5caC) and CBS-low (+CpG 5caC) CTCF sites and decreasing levels of purified CTCF (Figure 4D) (Heffler et al., 2012). The unmodified CBS-low probe was excluded from this analysis as CTCF binding was not detected in standard EMSA. Saturation binding curves were generated through percent shift to determine relative CTCF binding affinity as it relates to motif strength and 5caC (Heffler et al., 2012). Consistent with the enhanced binding visualized in standard EMSA, incorporation of 5caC into the three CpGs in the CBS-high probe strengthened CTCF binding and resulted in a near 2-fold increase in affinity when compared with the unmodified CBS-high probe. Remarkably, the presence of 5caC within the CBS-low probe yielded an intermediate $K_{D,app}$ that was strengthened when compared with the unmodified CBS-high probe but moderately reduced relative to the 5caC-containing CBS-high probe (Figure 4E). These $K_{D,app}$ values are within the established range for CTCF association with DNA (Hashimoto et al., 2017; Li et al., 2017; Martinez et al., 2014; Plasschaert et al., 2014). Taken together, these data corroborate the *in vivo* observations that overlapping 5caC, even in a minimal CpG setting, is sufficient to promote CTCF binding to weak consensus motifs.

An inherent limitation of EMSA relates to the fact that protein:DNA interactions are analyzed in the absence of additional variables at the cellular level. For example, while CTCF may interact with weak binding sites in the presence of 5caC in isolation, other factors may occupy such sites *in vivo*. To address this possibility, the CBS-high and CBS-low exon 5 probes were used as bait in DNA affinity purification assay (DAPA) (Figure 5A). PCR-generated 72mer probes were biotinylated and immobilized on streptavidin-coated magnetic beads. Of note, capture of the 5caC-containing probes was slightly reduced relative to the unmodified probes, as assessed through SYBR Gold staining of the unbound portion (Figure 5B). Nevertheless, recovery of CBS-low versus CBS-high probes was comparable per modification state, allowing for direct comparisons as related to motif strength (Figure 5B). Incubation with nuclear extracts from HEK293T cells expressing FLAG-CTCF allowed for the capture and subsequent elution of associated proteins. Immunoblotting of DAPA eluates from the unmodified DNA probes demonstrated a robust interaction between CTCF and the probe containing a CBS-high motif, whereas binding was not observed for the CBS-low probe (Figure 5C). In contrast, CTCF was recovered through incubation with both the CBS-high and CBS-low 5caC-containing probes. Consistent with the determined $K_{D,app}$ CTCF association was reduced for the CBS-low 5caC-containing probe, but was clearly visible (Figure 5C). Relatedly, while the uneven streptavidin immobilization precludes direct comparison between unmodified and 5caC-containing probes, the overall reduction in CTCF recovery through the 5caC-containing probes may reflect competition for binding with other factors that shape the binding landscape *in vivo*. Indeed, numerous nuclear factors exhibited 5caC-specific binding in an unbiased mass spectrometry screen (Spruijt et al., 2013). Overall, these DAPA results confirm that CTCF interacts with suboptimal DNA motifs in the presence of 5caC within the complex cellular milieu.
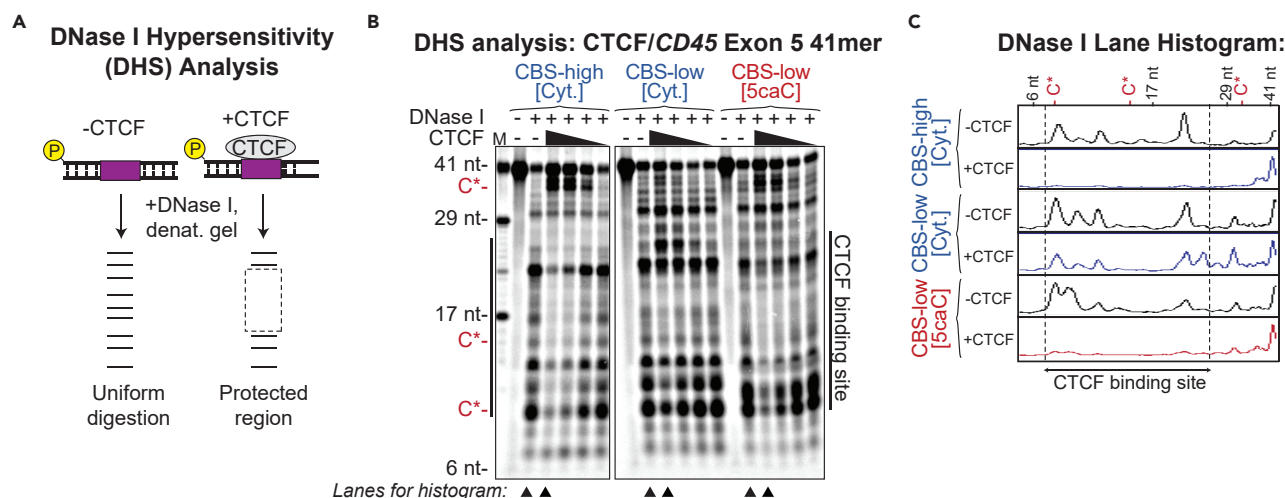
A DNase I Hypersensitivity (DHS) Analysis

B DHS analysis: CTCF/*CD45* Exon 5 41mer

C DNase I Lane Histogram:

Figure 6. DNase I Hypersensitivity (DHS) Analysis of CTCF Associated with 5caC-Containing DNA

(A) Overview of CTCF DNase I footprinting with variably carboxylated radiolabeled DNA probes representing strong (CBS-high) and weak (CBS-low) CTCF motifs.

(B) Gel analysis of CTCF DNase I footprinting performed with the *CD45* exon 5 probe, ±CpG 5caC. The DNase I reaction contained 342.3–42.8 nM CTCF and 7.52 nM DNA probe. The location of the CTCF-binding site is highlighted through the black bars. Carboxylated cytosine residues are indicated by C*. M signifies oligonucleotide marker, and arrowheads show lanes used to generate histograms.

(C) Lane histogram densitometry analysis of overall probe digestion patterns in the presence and absence of CTCF as indicated in (B). Significant protection from cleavage is seen in the region overlapping the CTCF-binding site for the CBS-high[Cyt.] and CBS-low[5caC] probes.

CTCF Generates a Strong Footprint at Weak Binding Motifs in the Presence of 5caC

The sum of the genome-wide analysis of *Tdg−/−* cells and general EMSA results clearly demonstrate that 5caC enhances CTCF binding in suboptimal contexts. However, although these approaches quantitatively assess the extent of CTCF association, they do not indicate whether binding is qualitatively distinct. To establish whether CTCF occupies a similar or unique expanse of DNA in the presence of 5caC, we thus pursued *in vitro* DNase I footprinting analysis. DNase I footprinting relies on time- and concentration-dependent DNase I cleavage according to availability (Brenowitz et al., 2001; Carey et al., 2013; Hampshire et al., 2007; Leblanc and Moss, 2015). This capacity hinges on the fact that DNA molecules adopt an inherent structure that renders certain regions more or less exposed. Incubation of end-labeled probe with a protein of interest in the presence of DNase I can thus inform on the protein-binding site through the region that is protected from cleavage (Brenowitz et al., 2001; Carey et al., 2013; Hampshire et al., 2007; Leblanc and Moss, 2015) (Figure 6A).

It is well established that CTCF binds to nucleosome-free DNA, and previous studies have demonstrated CTCF footprints of >20 bp (Chen et al., 2012; Filippova et al., 2001; Fu et al., 2008; Magbanua et al., 2015; Teif et al., 2014). We thus pursued DNase I footprinting with the synthesized *CD45* exon 5 41mer probe in which 5caC is restricted to three CpGs located within and adjacent to the CTCF-binding core (Figure 4A). Of note, DNase I cleavage displays some sequence preference and 5caC protrusion into the major grove of the DNA double helix is known to minorly alter base-pairing thermodynamic stability (Dai et al., 2016; Herrera and Chaires, 1994; Szulik et al., 2015). Accordingly, as the interrogated 41mers differ in both absolute sequence context and modified nucleotide composition, it is unsurprising that gel analysis of DNase I hypersensitivity (DHS) patterns in the absence of CTCF showed some distinctions (Figure 6B). In this regard, DNase I footprinting effectively informs whether select nucleotides display resistance to cleavage when compared with other locations within a particular probe. Importantly, CTCF addition to the DNase I reaction involving the unmodified strong motif probe resulted in a protected region encompassing the known location of CTCF binding (Figure 6B). In contrast, cleavage persisted within the weak CTCF motif-binding core in the presence of increasing CTCF (Figure 6B). Of note, consistent with basal CTCF detection at *Tdg−/−* induced locations in wild-type mESC cells (Figure 1F), minor protection from cleavage was observed upon CTCF addition to the weak motif probe. However, overall digestion patterns were virtually indistinguishable in lane histogram densitometry analysis in the presence or absence of CTCF (Figure 6B). With this in mind, it is remarkable that the 5caC-containing weak motif generated a CTCF footprint that effectively mirrored the strong unmodified probe (Figures 6B and 6C). In addition to validating that overlapping 5caC does indeed enable CTCF binding in suboptimal contexts, these DHS results indicate

that binding is qualitatively similar to unmodified strong binding sites, suggesting a related mode of interaction. Taken into consideration along with the EMSA results, these *in vitro* data complement the genome-wide *in vivo* results and solidify the role of 5caC in reconstituting CTCF association at weak sequence motifs.

DISCUSSION

Once considered a stable hereditary mark, it is now appreciated that DNA methylation is dynamically regulated to shape and define gene expression in a cell-specific manner (Luo et al., 2018). Tissue-specific changes in methylation often occur within gene bodies (Deaton et al., 2011; Maunakea et al., 2010), wherein we previously determined a role in pre-mRNA splicing that is achieved through modulation of CTCF binding (Marina et al., 2016; Shukla et al., 2011). We showed that genic CTCF promotes inclusion of weak exons in spliced mRNA through local RNA polymerase II pausing, whereas DNA methylation has the opposite effect (Shukla et al., 2011). Our efforts to understand how dynamic methylation at CTCF sites is achieved further uncovered a role for TET-catalyzed oxidized 5mC derivatives (5oxiC) (Marina et al., 2016). In particular, we asked how methylation is modulated at specific CTCF sites while leaving others unaffected. Given that CTCF is critical to numerous cellular processes, including general nuclear architecture, precise control of variable binding would be of tantamount relevance (Dixon et al., 2012; Handoko et al., 2011; Rao et al., 2014; Vietri Rudan et al., 2015; Zuin et al., 2014). We previously determined that splicing-associated "dynamic" CTCF sites are marked by overlapping oxidized derivatives, whereas static sites are unmethylated (Marina et al., 2016). Considering that CTCF has been biochemically associated with both the DNMT enzymes (that would evict CTCF) (Guastafierro et al., 2008; Zampieri et al., 2012) and the TET enzymes (that would facilitate CTCF binding) (Dubois-Chevalier et al., 2014) these results raised the possibility that 5oxiC facilitates CTCF binding and bookmarks locations of future CTCF eviction as TET activity declines. In other words, CTCF association with DNMT1 would ensure methylation of proximal CpGs post-replication, whereas the TET enzymes would subsequently oxidize 5mC and enable CTCF binding. However, a problematic aspect of this model related to the fact that CTCF is incapable of binding the abundant 5mC oxidized derivative 5hmC (Marina et al., 2016), and the downstream derivatives 5fC and 5caC are lowly detected in genomic DNA (Ito et al., 2011; Lu et al., 2015).

Our demonstrations associating CTCF and 5caC reconcile these observations: CTCF robustly interacts with 5caC, *in vitro* and *in vivo*, and 5caC is readily detected at CTCF sites. This latter point further suggests that CTCF may protect 5caC from removal through the base-excision repair enzymes (He et al., 2011). However, the truly unexpected aspect of this work relates to the observation that CTCF binding is seemingly enhanced through overlapping 5caC, suggesting a novel mode of binding. These results are consistent with a previous unbiased mass spectrometry study employing a short CpG-carboxylated DNA probe, wherein CTCF was identified as a 5caC-specific reader (Spruijt et al., 2013). As in our EMSA experiments, this DNA fragment lacked a computationally identifiable CTCF-binding site. Combined with the genome-wide observation that CTCF-binding sites with overlapping 5caC are generally characterized by lower motif scores, these findings suggest unique binding determinants. Although the precise biophysical bases guiding CTCF binding at unmethylated versus 5caC-rich DNA are unclear, our DHS data comparing unmodified CBS-high to carboxylated CBS-high and CBS-low probes indicate that binding in these variably modified contexts is conformationally similar and motif-centric. These findings suggest that rather than creating a true *de novo* environment involving novel CTCF contacts with unique sequences, 5caC rather stabilizes CTCF association with suboptimal motifs that are otherwise insufficient for protein retention. Indeed, it is well established that transcription factors dynamically engage substrate DNA (Voss and Hager, 2014). 5caC in this regard may subtly alter K_{on} versus K_{off} rates to reach a favorable equilibrium for bona fide CTCF binding. This prospect is in line with the EMSA results demonstrating increased K_D in the presence of 5caC as well as the genome-wide data showing globally enhanced CTCF binding in the presence of overlapping 5caC.

As for how 5caC generates favorable conditions for CTCF binding, charge-based stabilization involving the negatively charged carboxylate group of 5caC and cation-loaded zinc fingers of CTCF may play a role. Alternatively, the presence of 5caC within genomic DNA may facilitate a double-helical structure that enables CTCF binding. Detailed biophysical studies will be required to resolve the precise mechanistic underpinnings of sequence- and modification-specific CTCF DNA binding. Additional investigation will also be required to determine why only a subset of 5caC-rich locations promote enhanced CTCF association. In this regard, the canonical CTCF motif logo may be principally driven by binding to high-affinity

unmodified sites, thus obscuring the influence of lower prevalence dynamic sites. In support of this premise, 1,490 determined locations of CTCF binding in T cells yielded motif scores that were lower than the "motif-free" exon 6 EMSA probe (score of 7.36). In addition to highlighting the relevance of 5caC localization *in vivo* (exon 6 is not marked by 5caC), these findings suggest nuances within CTCF motifs that are driven by elements outside of strict sequence. We believe 5caC to be one such factor. Relatedly, CTCF possesses a variety of binding partners such that proximal sequence motifs that engage associated factors may influence CTCF recruitment to otherwise weak sites (Holwerda and de Laat, 2013; Parelho et al., 2008). Finally, it is likely that other factors occupy the 5caC-rich sequences *in vivo*. Indeed, CTCF is only one of many factors that showed preferential association with 5caC-containing DNA in an unbiased mass spectrometry screen (Spruijt et al., 2013). All in all, additional investigation will be required to determine the basis of 5caC-associated CTCF binding *in vivo*.

Although comprehensive analysis examining the physiological impact of widespread *Tdg*−/− induced CTCF sites is pending, one can infer that the observed increase in binding will impact diverse CTCF functions. CTCF has a demonstrated role in numerous aspects of nuclear biology including chromatin insulation, long-range chromosomal interactions, and gene expression regulation (Ong and Corces, 2014). With respect to the latter point, we provide evidence herein that the low-motif, 5caC-rich CTCF sites show a strong impact on RNA polymerase II pausing, intuiting a role in transcription elongation. Curiously, a recent global run-on sequencing study described a function for 5caC in reduced RNA polymerase II elongation in *Tdg*−/− cells (Wang et al., 2015). Our description of ~13,000 novel CTCF sites upon *Tdg* deletion raises the possibility that emergent CTCF contributed to the overall reduction in pol II processivity in *Tdg*−/− cells. Indeed, in our hands, examination of pol II occupancy at 5caC-rich regions that were not marked by CTCF binding showed no elevation when compared with 5caC-poor regions. Further analyses will be required to conclusively determine the source of altered elongation in *Tdg*−/− cells.

In sum, we describe a role for 5caC in modulating CTCF binding in cells. Given that 5caC levels vary during development (Wheldon et al., 2014), these results have significant implications to dynamic CTCF binding during tissue differentiation. A detailed analysis of CTCF and 5caC co-occurrence during organismal development will inform on the extent to which 5caC shapes CTCF tissue specificity. Importantly, our findings reported herein raise the possibility that CTCF sites may be engineered in the genome through targeted 5caC. As CRISPR/Cas9 technology continues to advance, one can envision applications both in studying specific CTCF sites and the overall impact of induced CTCF on nuclear architecture.

Limitations of the Study

In this study, we provide evidence that 5caC stabilizes CTCF binding to suboptimal DNA sequence contexts *in vitro* and *in vivo*. However, the net impact on cellular function was not assessed. Whether variations in 5caC levels that occur during development or in response to specific stimuli influence gene expression or genomic architecture through enhanced CTCF binding is unclear. In addition, subtle distinctions between our *in vitro* and *in vivo* results suggest that CTCF is in competition with other 5caC-sensitive transcription factors in a chromosomal setting. The identity of such factors and how they shape the transcriptional landscape in concert with CTCF remains to be determined.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

DATA AND CODE AVAILABILITY

The accession number for the sequence data reported in this paper is GEO: GSE123101.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2019.07.041.

ACKNOWLEDGMENTS

iScience

CellPress

AUTHOR CONTRIBUTIONS

Conceptualization, S.O.; Methodology, K.K.N., M.T., and S.O.; Data analysis and curation, D.M.S. and A.A.D.; Investigation and validation, K.K.N., M.F.P., A.A.D., and M.D.M.; Writing, K.K.N, D.M.S., and S.O.; Supervision and funding acquisition, S.O.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Bell, A.C., and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature *405*, 482–485.

Brenowitz, M., Senear, D.F., and Kingston, R.E. (2001). DNase I footprint analysis of protein-DNA binding. Curr. Protoc. Mol. Biol. 12.4.1–12.4.16. Chapter 12, Unit 12.4. https://doi.org/10.1002/0471142727.mb1204s07.

Carey, M.F., Peterson, C.L., and Smale, S.T. (2013). DNase I footprinting. Cold Spring Harb. Protoc. *2013*, 469–478.

Chen, H., Tian, Y., Shu, W., Bo, X., and Wang, S. (2012). Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. PLoS One *7*, e41374.

Cortazar, D., Kunz, C., Selfridge, J., Lettieri, T., Saito, Y., MacDougall, E., Wirz, A., Schuermann, D., Jacobs, A.L., Siegrist, F., et al. (2011). Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. Nature *470*, 419–423.

Dai, Q., Sanstead, P.J., Peng, C.S., Han, D., He, C., and Tokmakoff, A. (2016). Weakened N3 hydrogen bonding by 5-formylcytosine and 5-carboxylcytosine reduces their base-pairing stability. ACS Chem. Biol. *11*, 470–477.

Deaton, A.M., Webb, S., Kerr, A.R., Illingworth, R.S., Guy, J., Andrews, R., and Bird, A. (2011). Cell type-specific DNA methylation at intragenic CpG islands in the immune system. Genome Res. *21*, 1074–1086.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dubois-Chevalier, J., Oger, F., Dehondt, H., Firmin, F.F., Gheeraert, C., Staels, B., Lefebvre, P., and Eeckhoute, J. (2014). A dynamic CTCF chromatin binding landscape promotes DNA hydroxymethylation and transcriptional induction

of adipocyte differentiation. Nucleic Acids Res. *42*, 10943–10959.

Filippova, G.N., Thienes, C.P., Penn, B.H., Cho, D.H., Hu, Y.J., Moore, J.M., Klesert, T.R., Lobanenkov, V.V., and Tapscott, S.J. (2001). CTCF-binding sites flank CTG/CAG repeats and form a methylation-sensitive insulator at the DM1 locus. Nat. Genet. *28*, 335–343.

Fu, Y., Sinha, M., Peterson, C.L., and Weng, Z. (2008). The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet. *4*, e1000138.

Guastafierro, T., Cecchinelli, B., Zampieri, M., Reale, A., Riggio, G., Sthandier, O., Zupi, G., Calabrese, L., and Caiafa, P. (2008). CCCTC-binding factor activates PARP-1 affecting DNA methylation machinery. J. Biol. Chem. *283*, 21873–21880.

Guo, J.U., Su, Y., Shin, J.H., Shin, J., Li, H., Xie, B., Zhong, C., Hu, S., Le, T., Fan, G., et al. (2014). Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. Nat. Neurosci. *17*, 215–222.

Hampshire, A.J., Rusling, D.A., Broughton-Head, V.J., and Fox, K.R. (2007). Footprinting: a method for determining the sequence selectivity, affinity and kinetics of DNA-binding ligands. Methods *42*, 128–140.

Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W., Ye, C., Ping, J.L., Mulawadi, F., et al. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. Nat. Genet. *43*, 630–638.

Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M., and Tilghman, S.M. (2000). CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature *405*, 486–489.

Hashimoto, H., Wang, D., Horton, J.R., Zhang, X., Corces, V.G., and Cheng, X. (2017). Structural basis for the versatile and methylation-dependent binding of CTCF to DNA. Mol. Cell *66*, 711–720.e3.

He, Y.F., Li, B.Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., et al. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science *333*, 1303–1307.

Heffler, M.A., Walters, R.D., and Kugel, J.F. (2012). Using electrophoretic mobility shift assays to measure equilibrium dissociation constants: GAL4-p53 binding DNA as a model system. Biochem. Mol. Biol. Educ. *40*, 383–387.

Herrera, J.E., and Chaires, J.B. (1994). Characterization of preferred deoxyribonuclease I cleavage sites. J. Mol. Biol. *236*, 405–411.

Holwerda, S.J., and de Laat, W. (2013). CTCF: the protein, the binding partners, the binding sites and their chromatin loops. Philos. Trans. R. Soc. Lond. B Biol. Sci. *368*, 20120369.

Huang, Y., Chavez, L., Chang, X., Wang, X., Pastor, W.A., Kang, J., Zepeda-Martinez, J.A., Pape, U.J., Jacobsen, S.E., Peters, B., et al. (2014). Distinct roles of the methylcytosine oxidases Tet1 and Tet2 in mouse embryonic stem cells. Proc. Natl. Acad. Sci. U S A *111*, 1361–1366.

Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C., and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature *466*, 1129–1133.

Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C., and Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science *333*, 1300–1303.

Koh, K.P., Yabuuchi, A., Rao, S., Huang, Y., Cunniff, K., Nardone, J., Laiho, A., Tahiliani, M., Sommer, C.A., Mostoslavsky, G., et al. (2011). Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. Cell Stem Cell *8*, 200–213.

Leblanc, B.P., and Moss, T. (2015). In vitro DNase I footprinting. Methods Mol. Biol. *1334*, 17–27.

Li, W., Shang, L., Huang, K., Li, J., Wang, Z., and Yao, H. (2017). Identification of critical base pairs required for CTCF binding in motif M1 and M2. Protein Cell 8, 544–549.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315–322.

Lu, J., and Tang, M. (2012). CTCF-dependent chromatin insulator as a built-in attenuator of angiogenesis. Transcription 3, 73–77.

Lu, X., Han, D., Zhao, B.S., Song, C.X., Zhang, L.S., Dore, L.C., and He, C. (2015). Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. Cell Res. 25, 386–389.

Luo, C., Hajkova, P., and Ecker, J.R. (2018). Dynamic DNA methylation: in the right place at the right time. Science 361, 1336–1340.

Magbanua, J.P., Runneburger, E., Russell, S., and White, R. (2015). A variably occupied CTCF binding site in the ultrabithorax gene in the Drosophila bithorax complex. Mol. Cell. Biol. 35, 318–330.

Maiti, A., and Drohat, A.C. (2011). Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. J. Biol. Chem. 286, 35334–35338.

Marina, R.J., Sturgill, D., Bailly, M.A., Thenoz, M., Varma, G., Prigge, M.F., Nanan, K.K., Shukla, S., Haque, N., and Oberdoerffer, S. (2016). TET-catalyzed oxidation of intragenic 5-methylcytosine regulates CTCF-dependent alternative splicing. EMBO J. 35, 335–355.

Martinez, F.P., Cruz, R., Lu, F., Plasschaert, R., Deng, Z., Rivera-Molina, Y.A., Bartolomei, M.S., Lieberman, P.M., and Tang, Q. (2014). CTCF binding to the first intron of the major immediate early (MIE) gene of human cytomegalovirus (HCMV) negatively regulates MIE gene expression and HCMV replication. J. Virol. 88, 7389–7401.

Maunakea, A.K., Nagarajan, R.P., Bilenky, M., Ballinger, T.J., D'Souza, C., Fouse, S.D., Johnson, B.E., Hong, C., Nielsen, C., Zhao, Y., et al. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466, 253–257.

Maurano, M.T., Wang, H., John, S., Shafer, A., Canfield, T., Lee, K., and Stamatoyannopoulos, J.A. (2015). Role of DNA methylation in modulating transcription factor occupancy. Cell Rep. 12, 1184–1195.

Nakahashi, H., Kieffer Kwon, K.R., Resch, W., Vian, L., Dose, M., Stavreva, D., Hakim, O., Pruett, N., Nelson, S., Yamane, A., et al. (2013). A genome-wide map of CTCF multivalency redefines the CTCF code. Cell Rep. 3, 1678–1689.

Ong, C.T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. 15, 234–246.

Paredes, S.H., Melgar, M.F., and Sethupathy, P. (2013). Promoter-proximal CCCTC-factor binding is associated with an increase in the transcriptional pausing index. Bioinformatics 29, 1485–1487.

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., et al. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell 132, 422–433.

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

Plasschaert, R.N., Vigneau, S., Tempera, I., Gupta, R., Maksimoska, J., Everett, L., Davuluri, R., Mamorstein, R., Lieberman, P.M., Schultz, D., et al. (2014). CTCF binding site sequence differences are associated with unique regulatory and functional trends during embryonic stem cell differentiation. Nucleic Acids Res. 42, 774–789.

Ramsahoye, B.H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A.P., and Jaenisch, R. (2000). Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc. Natl. Acad. Sci. U S A 97, 5237–5242.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680.

Rhee, H.S., and Pugh, B.F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell 147, 1408–1419.

Shen, L., Wu, H., Diep, D., Yamaguchi, S., D'Alessio, A.C., Fung, H.L., Zhang, K., and Zhang, Y. (2013). Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. Cell 153, 692–706.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature 479, 74–79.

Spruijt, C.G., Gnerlich, F., Smits, A.H., Pfaffeneder, T., Jansen, P.W., Bauer, C., Munzel, M., Wagner, M., Muller, M., Khan, F., et al. (2013). Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. Cell 152, 1146–1159.

Szulik, M.W., Pallan, P.S., Nocek, B., Voehler, M., Banerjee, S., Brooks, S., Joachimiak, A., Egli, M., Eichman, B.F., and Stone, M.P. (2015). Differential stabilities and sequence-dependent base pair opening dynamics of Watson-Crick base pairs with 5-hydroxymethylcytosine, 5-formylcytosine, or 5-carboxylcytosine. Biochemistry 54, 1294–1305.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., et al. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 324, 930–935.

Teif, V.B., Beshnova, D.A., Vainshtein, Y., Marth, C., Mallm, J.P., Hofer, T., and Rippe, K. (2014). Nucleosome repositioning links DNA (de) methylation and differential CTCF binding during stem cell development. Genome Res. 24, 1285–1295.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D.T., Tanay, A., and Hadjur, S. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. 10, 1297–1309.

Voss, T.C., and Hager, G.L. (2014). Dynamic regulation of transcriptional states by chromatin and transcription factors. Nat. Rev. Genet. 15, 69–81.

Wang, H., Maurano, M.T., Qu, H., Varley, K.E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., et al. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res. 22, 1680–1688.

Wang, L., Zhou, Y., Xu, L., Xiao, R., Lu, X., Chen, L., Chong, J., Li, H., He, C., Fu, X.D., et al. (2015). Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex. Nature 523, 621–625.

Wheldon, L.M., Abakir, A., Ferjentsik, Z., Dudnakova, T., Strohbuecker, S., Christie, D., Dai, N., Guan, S., Foster, J.M., Correa, I.R., Jr., et al. (2014). Transient accumulation of 5-carboxylcytosine indicates involvement of active demethylation in lineage specification of neural stem cells. Cell Rep. 7, 1353–1361.

Wu, X., and Zhang, Y. (2017). TET-mediated active DNA demethylation: mechanism, function and beyond. Nat. Rev. Genet. 18, 517–534.

Zampieri, M., Guastafierro, T., Calabrese, R., Ciccarone, F., Bacalini, M.G., Reale, A., Perilli, M., Passananti, C., and Caiafa, P. (2012). ADP-ribose polymers localized on Ctcf-Parp1-Dnmt1 complex prevent methylation of Ctcf target sites. Biochem. J. 441, 645–652.

Zhu, Q., Stoger, R., and Alberio, R. (2018). A lexicon of DNA modifications: their roles in embryo development and the germline. Front. Cell Dev. Biol. 6, 24.

Ziebarth, J.D., Bhattacharya, A., and Cui, Y. (2013). CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. Nucleic Acids Res. 41, D188–D194.

Zuin, J., Dixon, J.R., van der Reijden, M.I., Ye, Z., Kolovos, P., Brouwer, R.W., van de Corput, M.P., van de Werken, H.J., Knoch, T.A., van IJcken, W.F., et al. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc. Natl. Acad. Sci. U S A 111, 996–1001.

Supplemental Information

TET-Catalyzed 5-Carboxylcytosine

Promotes CTCF Binding to Suboptimal

Sequences Genome-wide

Kyster K. Nanan, David M. Sturgill, Maria F. Prigge, Morgan Thenoz, Allissa A. Dillman, Mariana D. Mandler, and Shalini Oberdoerffer
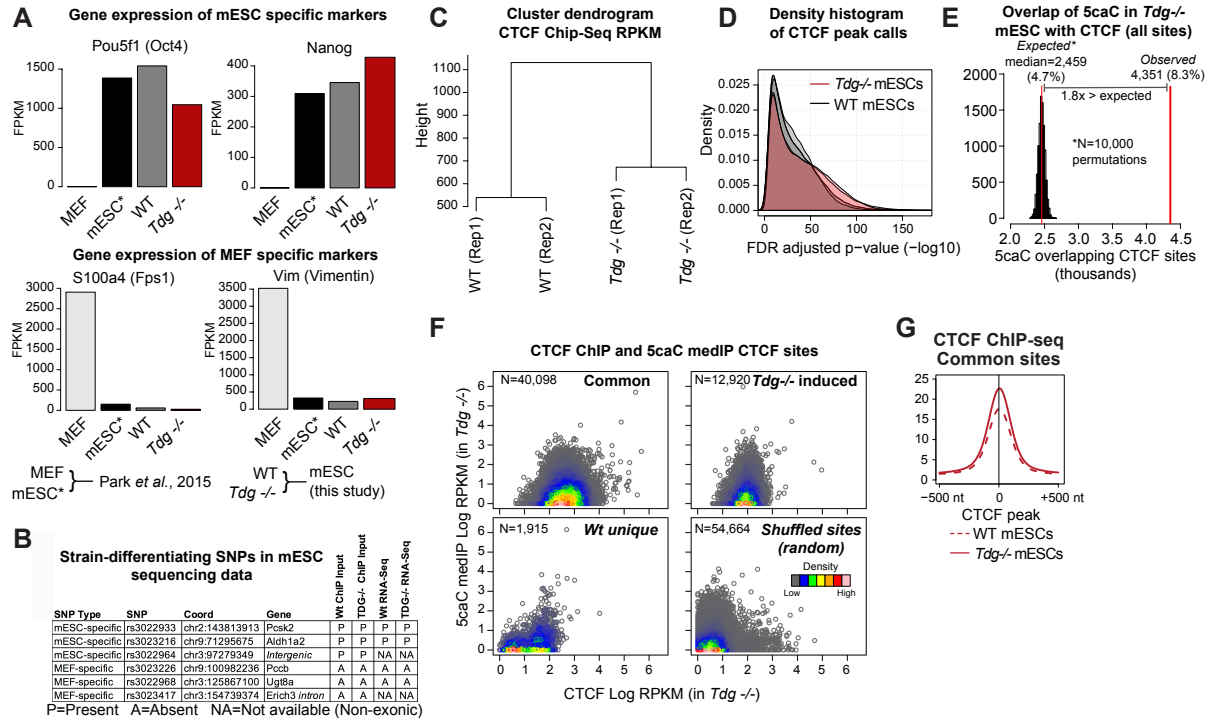
Figure S1, related to Figure 1. Genome-wide analysis of CTCF and 5caC in wildtype and *Tdg-/-* mESCs. A) Expression of mESC (top) and MEF (bottom) specific genes in RNA-seq from wildtype and *Tdg-/-* mESCs as compared to published mESC and MEF results (Park et al., 2015). B) Table of identifying SNP presence or absence in mESC sequencing data. C) Hierarchical clustering (Ward's minimum variance method) of mESC ChIP-Seq samples. Euclidean distances between samples were calculated based on RPKM values within CTCF sites. D) Frequency histogram of CTCF ChIP-seq peak significance values. Both replicates of each sample are shown. E) Co-occurrence of CTCF and 5caC in *Tdg-/-* genome-wide data as compared to size-matched random intervals. Histogram of the number of CTCF sites overlapping 5caC in 10,000 trials of randomly shuffling 5caC peaks and CTCF sites. Assembly gaps, centromeres, and chrY are excluded from shuffled locations. Thin red line indicates the median of these trials, and thick red line indicates the value of the observed overlap in all called CTCF peaks in *Tdg-/-* mESCs. F) Density scatter plot of CTCF vs 5caC abundance within CTCF sites in *Tdg-/-* cells. CTCF sites are compared to size-matched random intervals. G) CTCF ChIP-seq signal in WT and *Tdg-/-* mESCs centered on CTCF peaks in common between WT and *Tdg-/-* cells.
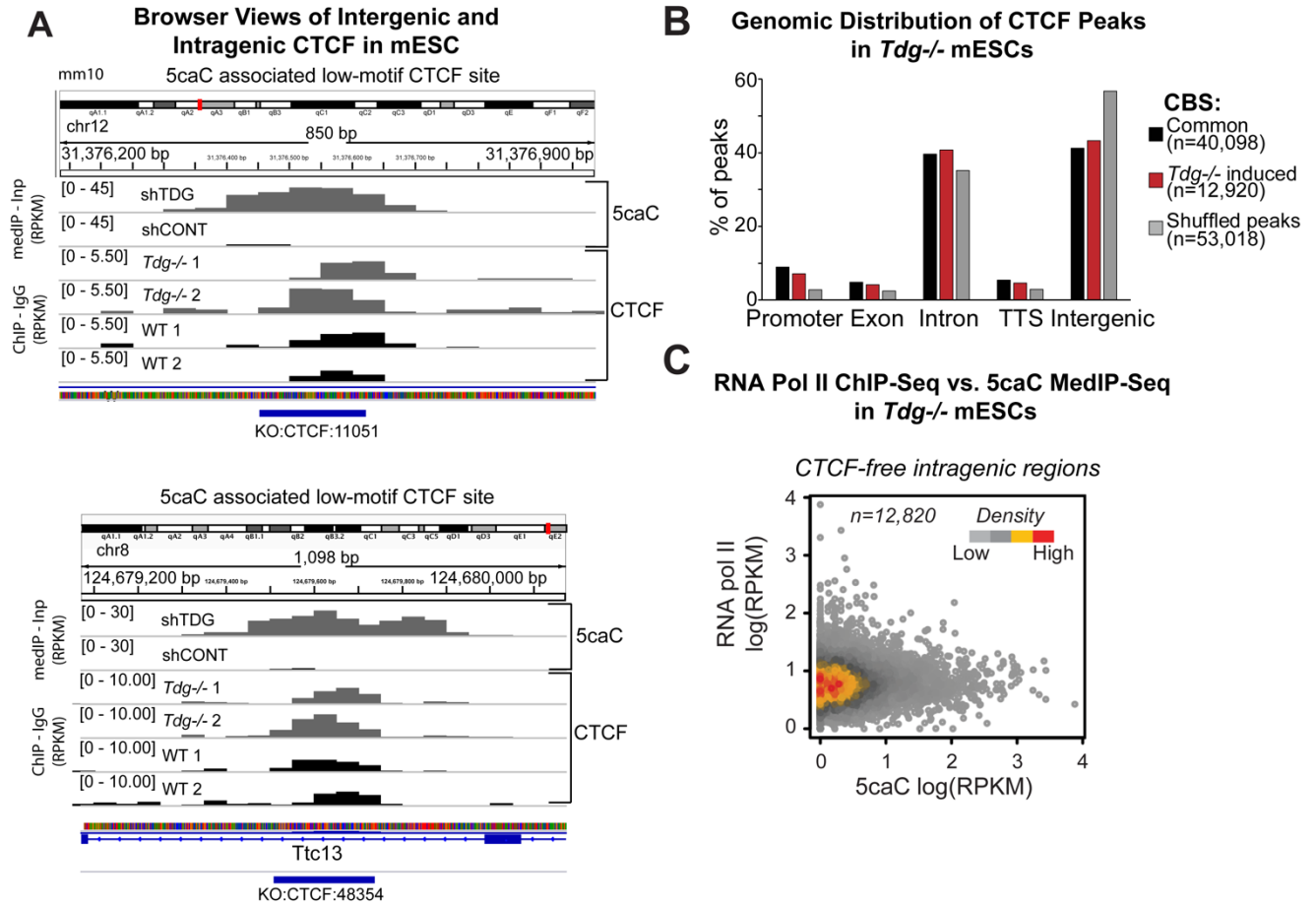
Figure S2, related to Figure 2. Genome-wide analysis of CTCF, 5caC and RNA pol II in *Tdg-/-* mESCs. A) Browser views of 5caC and CTCF in intragenic and intergenic CTCF sites in mESCS. B) Genomic distribution of common and *Tdg-/-* induced CTCF ChIP-seq peaks throughout the murine genome, compared to randomly permutated (shuffled) peaks. C) Scatter plot comparison of RNA pol II ChIP-seq and 5caC medIP-seq densities at CTCF-free intragenic regions in *Tdg-/-* mESC. Increasing 5caC levels do not correlate with increased pol II.
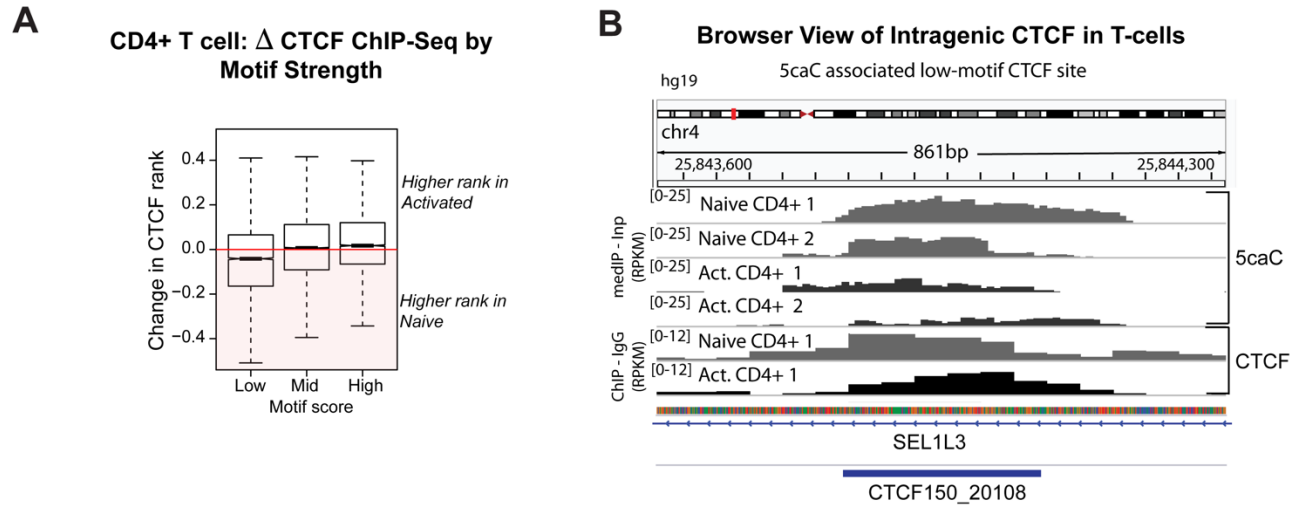
Figure S3, related to Figure 3. 5caC and CTCF dynamics in T-cells. A) Comparison of change in CTCF abundance, measured by change in rank, for CTCF sites grouped by motif score. B) Browser view of 5caC and CTCF in T-cells.

TRANSPARENT METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING
Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Shalini Oberdoerffer (shalini.oberdoerffer@nih.gov). There are no restrictions on any data or materials presented in this paper.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell culture and transfection
293T acquired from the ATCC were cultured in DMEM (Gibco) supplemented with 2 mM L-glutamine (Gibco) and 10% bovine calf serum (HyClone, #SH30073.03). *Tdg*-/- mouse embryonic stem cells (mESCs) and wildtype control mESCs were a gift from Dr. Primo Schär. Complete mESC media was composed of KnockOut DMEM (Gibco), supplemented with 2 mM GlutaMAX (Gibco), non-essential amino acids (Gibco), 0.1 mM 2-mercaptoethanol (SIGMA), 1,000 U/ml LIF (Millipore, #ESG1106), and 15% ES-tested FBS (Gibco, #10439-024). mESC were co-cultured with mitotically-arrested mouse embryonic fibroblast (MEF; produced in-house or purchased from Millipore, #PMEF-CF) on gelatin-coated (Millipore, #ES-006-B) dishes. MEF depletion was achieved by three 10-minute rounds of serial-plating onto uncoated tissue culture dishes. Human CD4+ human T lymphocytes were acquired and cultured as described in Marina *et al.*, 2016. Plasmids were transfected into 293T using Lipofectamine 2000 (Invitrogen, #11668019) according to the manufacturer's instructions.

METHOD DETAILS

Antibodies
The following antibodies were used for immunoblotting: anti-β-tubulin (Cell Signaling Technology; CST, #2146), anti-CTCF (CST, #3418), anti-FLAG (Sigma, #F3165), anti-TDG (abcam, #ab154192), HRP-linked anti-mouse IgG (Jackson ImmunoResearch, #115-035-003), HRP-linked anti-rabbit IgG (CST, #7074). The following antibodies were used for ChIP: anti-CTCF (CST, #3418), anti-RNA pol II (Millipore, #05-623), monoclonal rabbit isotype control IgG (DA1E) (CST, #3900), normal rabbit polyclonal control IgG (CST, #2729). Anti-5caC (abcam, #ab185492) was used for DNA dot blot and 5caC meDIP. Densitometric analyses of immunoblots were performed using Bio-Rad ImageLab software; all signals were background-corrected prior to normalization to β-tubulin loading control.

5caC meDIP and dot blot
MeDIP in primary human CD4+ T cells was performed as described in Shukla et al., 2011 using anti-5caC antibody. Genomic DNA was purified from MEF-depleted *Tdg*-/- and control mESC using PureLink Genomic DNA Mini Kit (Invitrogen, #K182001) according to the manufacturer's instructions. Dot blot was performed as described by the Cell Signaling Technology protocol for product #36836. Briefly, genomic DNA was sheared between 200 and 500 bp by sonication (Bioruptor Twin, Diagenode). 10 µg fragmented DNA was heated at 95 °C for 10 minutes in 200 µl DNA denaturing buffer (100 mM NaOH, 10 mM EDTA), neutralized with 200 µl 20X SSC (3 M NaCl, 300 mM sodium citrate, pH 7.0), and chilled on ice for 5 minutes. DNA was applied to

nylon membrane (GE Healthcare, #RPN303B) using a Bio-Dot apparatus (Bio-Rad), air-dried then fix to membrane by UV crosslinking (Stratalinker 2400, Stratagene). Membrane was blocked at room temperature for 1 hour in PBS containing 0.05% Tween-20 (PBS-T), 2.5% BSA, and 2.5% non-fat dry milk and probed overnight at 4 °C with 5caC antibody (abcam, #ab185492) diluted 1:1000 in blocking buffer. After washing and incubation with HRP-linked anti-rabbit 2° antibody diluted 1:1000 in blocking buffer, signal detection was performed using SuperSignal West Femto ECL substrate (Thermo Scientific, #34095), imaged using ChemiDoc MP (Bio-Rad). Genomic DNA samples were applied to a separate nylon membrane as described above and methylene blue was used as a total DNA stain that served as loading control. Densitometric quantification was performed using ImageLab software (Bio-Rad), wherein 5caC dotblot signals were background-corrected prior to normalization to methylene blue.

Chromatin immunoprecipitation-sequencing (ChIP-seq)
ChIP-seq was performed using two biological replicates that were cultured independently. Prior to chromatin immunoprecipitation, antibodies (5 µl polyclonal or 5 µg monoclonal) were pre-bound to 200 µl Protein G magnetic beads (Invitrogen, #10004D) by overnight incubation in PBS with 5% BSA; beads were washed and resuspended in PBS with 5% BSA. MEF-depleted mESCs were fixed with 1% formaldehyde (Sigma-Aldrich, # 252549) at room temperature for 5 min and quenched with 125 mM glycine (ICN Biomedical, #ICN808822). Cell membranes were lysed with cold NP-40 buffer (1% NP40, 150 mM NaCl, 50 mM Tris–HCl; pH 8.0) and nuclei collected by centrifugation at $12,000 \times g$ for 1 min at 4 °C. Nuclear pellets were resuspended at a concentration of 2E8 cells/ml in ChIP sonication buffer (1% SDS, 10 mM EDTA, 50 mM Tris–HCl; pH 8.0) supplemented with protease (Thermo Scientific, #78430) and phosphatase (Calbiochem, #524627) inhibitors. Chromatin was sheared into 150-400 bp fragments by sonication (Bioruptor Twin, Diagenode). Debris was pelleted by centrifugation and cleared chromatin was diluted 10-fold in ChIP dilution buffer (1.1% Triton X-100, 0.01% SDS, 167 mM NaCl, 1.2 mM EDTA, 16.7 mM Tris–HCl; pH 8.1). The prepared antibody-bound beads were added to 1 ml diluted chromatin containing 2E7 million cell equivalents and incubated overnight with rotation at 4 °C. Immune complexes were washed 5 times with LiCl wash buffer (250 mM LiCl, 1% NP-40, 1% sodium deoxycholate, 100 mM Tris-HCl; pH 7.5) and once with TE (0.1 mM EDTA, 10 mM Tris-HCl; 7.5). Beads were resuspended in IP Elution Buffer (1% SDS, 0.1 M NaHCO$_3$) and crosslinks reversed by overnight incubation at 65 °C. DNA was purified by column purification (QIAGEN, # 28106) and subjected to Illumina sequencing at the Advanced Technology Research Facility at NCI Frederick. Libraries were constructed with the Illumina TruSeq ChIP Sample Prep Kit (#IP-202-1012-1024) and sequenced on an Illumina NextSeq instrument for seventy-six cycles in single-end mode, using NextSeq High Output v.2 chemistry.

Nuclear extracts
Nuclear extracts were prepared from 293T cells, either untransfected or transfected with plasmid encoding 3xFLAG-hCTCF as described in Jacobs *et al.*, 1993 (Jacobs et al., 1993). Hypotonic lysis buffer A (20 mM HEPES; pH 7.9, 10 mM KCl, 1 mM EDTA, 1 mM DTT) and nuclear extraction buffer B (20 mM HEPES; pH 7.9, 400 mM NaCl, 1 mM EDTA, 1mM DTT) were supplemented with 4 µM ZnCl$_2$, protease inhibitor (Thermo Scientific, #78429) and Calbiochem phosphatase inhibitors (EMD-Millipore, #524627). Protein quantification was performed using Bradford assay (Bio-Rad #5000201).

DNA affinity precipitation assay (DAPA)

Double-stranded *CD45* exon 5 DNA probes were generated by amplification of plasmid DNA template with Phusion polymerase (NEB, #M0530L) using 5' biotinylated forward (5'-CCT CAC CTT CCC ACG CAC GCA GAC TC-3') and unlabeled reverse primers (5'-GGA GCC GCT GAA TGT CTG CGT GTC AGT TC-3') (Integrated DNA Technologies). PCR performed in the presence of dCTP and 5-carboxy-dCTP (Trilink, #N-2063) produced unmodified or uniformly carboxylated DNA probes, respectively. Biotinylated probes were immobilized on streptavidin magnetic beads (Invitrogen, #11205D) according to the manufacturer's instructions; DNA capture efficiency was evaluated by monitoring unbound DNA present in the supernatant by SYBR Gold (Invitrogen, # S11494) staining after gel electrophoresis. DAPA was performed by combining 20 µg nuclear extract with 50 mg DNA-bound streptavidin beads in 100 µl DAPA buffer (10 mM Tris-HCl; pH 8.5, 50 mM KCl, 0.1 mM ZnCl2, 0.1% NP-40, 0.05 mg/ml BSA, 2 mM DTT, 1 ug/ml poly-dI:dC) supplemented with HALT protease inhibitor (Thermo Scientific, #78429) and Calbiochem phosphatase inhibitor (EMD-Millipore, #524627) cocktails. After overnight incubation with rotation at 4 °C, beads were washed 4 times with 100 µl DAPA buffer, heated at 95 °C for 10 minutes in Laemmli sample buffer, and subjected to immunoblotting using standard techniques.

CTCF purification, EMSA, and relative KD determination

CTCF was affinity-purified from 293T cells transfected with a 3xFLAG-tagged CTCF expression construct using FLAG M2 agarose beads as described in Marina *et al.*, 2016 with the exception that the NaCl concentration in lysis and wash buffers was increased to 500 mM. Purified CTCF was quantified relative to Precision Plus Protein standards (Bio-Rad, #1610363) in a Coomassie-stained gel imaged on Bio-Rad ChemiDoc MP and analyzed with Bio-Rad ImageLab software. *CD45* exon 5 72-mer was produced using PCR amplification as indicated above; *CD45* exon 6 was produced by PCR amplification of template plasmid DNA using unlabeled forward (5'-AGC ACC TTT CCT ACA GAC CCA GTT-3') and reverse (5'TGT TCG CTG TGA TGG TGG TGT T-3') primers. DNA probes were labeled using γ-^{32}P-ATP (Perkin Elmer #NEG502A250UC) and PNK (NEB #M0201L) and EMSA was performed as described in Marina *et al.*, 2016; cold competition assays utilized 10x and 100x molar excess of unlabeled 72-mer and 41-mer DNA probes, respectively. Relative K_D was determined using EMSA as described in Heffler et al., 2012 (Heffler et al., 2012). Briefly, CTCF EMSA was first performed by incubating two-fold serial dilutions of purified 3xFLAG-CTCF with a fixed amount of radiolabeled DNA probe. The fraction of CTCF-bound DNA was determined using background subtracted densitometric signals as follows; fraction bound = complex/(complex + free probe). Saturation binding curves were generated by plotting fraction bound relative to CTCF concentration and non-linear regression was used to obtain relative K_D values (Prism 7, GraphPad). Data are represented as mean ± SD for at least 2 replicates.

DNase I hypersensitivity (DHS) assay and quantitative DHS analysis

The sense strand of each commercially-synthesized 41-nt *CD45* exon 5 probe was individually radiolabeled using γ-32P-ATP and PNK as described above and annealed to equimolar quantities of its unlabeled complementary strand in annealing buffer (10 mM Tris, pH 8.0, 50 mM NaCl, 1 mM EDTA) by heating to 95 °C and allowing to cool slowly to room temperature. Free nucleotides were separated from the annealed probe using the QIAquick Nucleotide Removal Kit (QIAGEN, #28304). Probes were quantified using Qubit dsDNA HS Assay Kit (Invitrogen,

#Q32851). DHS assay was performed by combining CTCF and radiolabeled DNA fragment in CTCF DHS buffer (10 mM Tris-HCl; pH 8.5, 0.5 mM CaCl2, 2.5 MgCl2, 50 mM KCl, 0.1 mM ZnCl2, 0.1% NP-40, 0.05 mg/ml BSA, 2 mM DTT, 1 ug/ml poly-dI:dC) and incubating on ice for 30 minutes. DNase I (1.25E-3U/ul; NEB, #M0303L) was added and samples were incubated at room temperature for the indicated time intervals. DNase I digestion was stopped by addition of equal volume of 2X formamide loading buffer (95% formamide, 20 mM EDTA, 0.025% each of xylene cyanol, bromophenol blue, orange G) and heating at 95 °C for 3 minutes before loading on a pre-run 15% 7M urea gel. Gel was exposed to phosphor screen (Molecular Dynamics) for 24 hours; signals were detected using PhosphorImager (Storm 840, Molecular Dynamics).

Bioinformatics
Genomic alignment, peak-calling, annotation, and accession numbers
ChIP-seq reads from our experiments and published data were aligned to the relevant reference genome (mm10 or hg19) using Bowtie version 2.2.9 (Langmead and Salzberg, 2012) with default parameters; alignment statistics are presented in Table S1. Peak calling was performed with MACS (v 2.2.1) (Zhang et al., 2008), using sonicated chromatin (input) as the control. Peaks were called separately for each replicate. Peaks called on chromosome Y were removed. Merged peak calls from replicates constituted the reference set, from which a set of consensus peaks was defined based on presence in individual replicates; all subsequent analyses were performed using these peak coordinates (Table S2). Replicate concordance was measured based on genomic read abundance, and ranged from r=0.89 (meDIP) to r=0.94 (CTCF ChIP-seq). Peak annotation was determined for the mm10 genome using HOMER version 4.10 with default settings (Table S3); genomic annotation of an equal number of shuffled peaks was used to derive observed/expected values (Heinz et al., 2010). Human CD4+ T cell CTCF ChIP-seq data were previously reported in Marina *et al.*, 2016 (GEO accession #GSE74850) (Marina et al., 2016). 5caC meDIP from control and *Tdg*-depleted mESC were reported in Shen et al., 2013 (GEO accession #GSE46111) (Shen et al., 2013). CD4+ T pol II ChIP-Seq data were obtained from Barski et al., 2007 (Barski et al., 2007). CTCF and RNA pol II ChIP-seq data from *Tdg-/-* and wildtype control mESCs are accessible via GSE123101; 5caC meDIP from primary human CD4+ T cells accessible via GSE46111.

Data from two independent biological replicates were used for all computational analyses relating to Tdg WT and _Tdg-/-_ mESC. To evaluate replicate concordance, we compared genomic read abundance between samples, and each pair of replicates had a minimum Pearson's correlation coefficient > 0.94 for CTCF ChIP-Seq, and 0.89 for 5caC medIP. In unsupervised hierarchical clustering, replicates cluster together while WT and TDG-/- samples segregated (Fig. S1C). To confirm the absence of MEF contamination, we compared our results to expression data in DBTMEE database (Park et al., 2015)

To generate aggregate plots from sequencing data, deepTools was used to compute coverage summaries from bigwig files (Ramirez et al., 2014). Bedtools was used for coordinate intersections (Quinlan and Hall, 2010). When intersecting features with transcript models to define intergenic/intragenic, we used Ensembl rel. 83 for mm10. To define an expressed gene in mESC we aligned reads from poly(A) RNA-seq performed in wildtype and *Tdg-/-* mESCs (GSE123101) with Tophat v2.1.1 (Trapnell et al., 2009) and quantified gene expression with Cufflinks 2.2.1 (Trapnell et al., 2010). Genes with fragments per kilobase per million mapped reads (FPKM) greater than 1 were considered expressed genes. Genic definitions for T-cell data

were taken from Marina *et al.*, 2016 (Marina et al., 2016). In aggregate plots over discrete features with pol II ChIP data, each bin was normalized to the minimum value within its respective window, to give an indication of local polymerase stalling.

Evaluation of CTCF consensus binding sites
CTCF binding sites were detected and scored using the "Scan" feature in CTCFBSDB 2.0 (Table S2) from peak sequence (Ziebarth et al., 2013). This tool takes user-input sequence and queries a set of curated Position Weight Matrices (PWMs) to generate match scores. Raw database results were imported into R for downstream filtering and analysis. Unless otherwise indicated, only the maximum motif score for each input sequence is reported. To calculate CpG frequencies, in house perl scripts were used to count CG frequency relative to all other dinucleotide combinations as described in Marina *et al.*, 2016 (Marina et al., 2016).

SUPPLEMENTAL REFERENCES

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Heffler, M.A., Walters, R.D., and Kugel, J.F. (2012). Using electrophoretic mobility shift assays to measure equilibrium dissociation constants: GAL4-p53 binding DNA as a model system. Biochem Mol Biol Educ *40*, 383-387.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576-589.

Jacobs, Y., Vierra, C., and Nelson, C. (1993). E2A expression, nuclear localization, and in vivo formation of DNA- and non-DNA-binding species during B-cell development. Mol Cell Biol *13*, 7321-7333.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods *9*, 357-359.

Marina, R.J., Sturgill, D., Bailly, M.A., Thenoz, M., Varma, G., Prigge, M.F., Nanan, K.K., Shukla, S., Haque, N., and Oberdoerffer, S. (2016). TET-catalyzed oxidation of intragenic 5-methylcytosine regulates CTCF-dependent alternative splicing. EMBO J *35*, 335-355.

Park, S.J., Shirahige, K., Ohsugi, M., and Nakai, K. (2015). DBTMEE: a database of transcriptome in mouse early embryos. Nucleic Acids Res *43*, D771-776.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res *42*, W187-191.

Shen, L., Wu, H., Diep, D., Yamaguchi, S., D'Alessio, A.C., Fung, H.L., Zhang, K., and Zhang, Y. (2013). Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. Cell *153*, 692-706.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology *28*, 511-515.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., *et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology *9*, R137.

Ziebarth, J.D., Bhattacharya, A., and Cui, Y. (2013). CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. Nucleic Acids Res *41*, D188-194.