5084–5094 *Nucleic Acids Research, 2021, Vol. 49, No. 9* *Published online 20 April 2021*
doi: 10.1093/nar/gkab276

Preferential CEBP binding to T:G mismatches and increased C-to-T human somatic mutations

Jie Yang[1,†], John R. Horton[1,†], Kadir C. Akdemir[2], Jia Li[3], Yun Huang[3], Janani Kumar[1], Robert M. Blumenthal[4,*], Xing Zhang[1,*] and Xiaodong Cheng[1,*]

[1]Department of Epigenetics and Molecular Carcinogenesis, University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA, [2]Departments of Genomic Medicine and Neurosurgery, University of Texas MD Anderson Cancer Center, Houston, TX, USA, [3]Center for Epigenetics & Disease Prevention, Institute of Biosciences and Technology, College of Medicine, Texas A&M University, Houston, TX 77030, USA and [4]Department of Medical Microbiology and Immunology, and Program in Bioinformatics, The University of Toledo College of Medicine and Life Sciences, Toledo, OH 43614, USA

Received March 04, 2021; Editorial Decision April 02, 2021; Accepted April 07, 2021

ABSTRACT

DNA cytosine methylation in mammals modulates gene expression and chromatin accessibility. It also impacts mutation rates, via spontaneous oxidative deamination of 5-methylcytosine (5mC) to thymine. In most cases the resulting T:G mismatches are repaired, following T excision by one of the thymine DNA glycosylases, TDG or MBD4. We found that C-to-T mutations are enriched in the binding sites of CCAAT/enhancer binding proteins (CEBP). Within a CEBP site, the presence of a T:G mismatch increased CEBPβ binding affinity by a factor of >60 relative to the normal C:G base pair. This enhanced binding to a mismatch inhibits its repair by both TDG and MBD4 in vitro. Furthermore, repair of the deamination product of unmethylated cytosine, which yields a U:G DNA mismatch that is normally repaired via uracil DNA glycosylase, is also inhibited by CEBPβ binding. Passage of a replication fork over either a T:G or U:G mismatch, before repair can occur, results in a C-to-T mutation in one of the daughter duplexes. Our study thus provides a plausible mechanism for accumulation of C-to-T human somatic mutations.

INTRODUCTION

Genomic DNA is constantly being damaged by insults that range from UV irradiation or (aging-associated) oxidative stress, to interactions with environmental mutagens and cancer chemotherapeutic drugs (1). Somatic mutations accumulate if not repaired prior to DNA replication. As a result, accrual of somatic mutations is a likely consequence of anything that reduces DNA accessibility by the DNA repair machinery. This impaired access can result from three-dimensional chromatin organization, nucleosome occupancy, binding of transcription factors (TFs) and other stable protein-DNA interactions (2–5). Characterization of somatic mutations from cancer genomes has identified mutational signatures, including 49 single-base-substitution (SBS) patterns with probable biological origins ((6) and references therein). Some SBS signatures include all six types of substitutions (C-to-A/G/T or T-to-A/C/G; in which the mutated basepair is represented by the pyrimidine) (Supplementary Figure S1A). Other SBS signatures are dominated by one type; for instance, C→T, C→A, or T→G (Supplementary Figure S1B–D). In contrast, some signatures are missing one or two types of SBSs, particularly T→G (Supplementary Figure S1E). There are substantial differences in the numbers and types of SBSs across tumors examined by the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium (7) (Supplementary Figure S1F).

Here, we focus on evidence for a plausible mechanism to explain methylation effects on substitutions, involving methylation of C (yielding 5-methylcytosine, 5mC). The proposed mechanism involves limiting repair enzyme access to the deamination product of 5mC (T), boosting C→T substitutions, which are the most abundant substitution. The suggested aetiology of C→T mutation begins with spontaneous or APOBEC-mediated deamination of C→U (generating a G:U mismatch) or of 5mC→T (generating a G:T mismatch) (8–11). These two mismatches can be repaired, beginning with base excision by repair enzymes that remove T or U opposite to G (thymine DNA glycosylases – TDG and MBD4 – for T; and uracil DNA glycosylase –

*To whom correspondence should be addressed. Email: XCheng5@mdanderson.org
Correspondence may also be addressed to Xing Zhang. Email: XZhang21@mdanderson.org
Correspondence may also be addressed to Robert M. Blumenthal. Email: Robert.Blumenthal@utoledo.edu
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

UDG – for U) (12,13). However, if the mismatch is inaccessible to repair enzymes, or the latter are reduced in activity or amount, then DNA polymerases can pair A with the inappropriate T (or U) in the template, resulting in C→T mutation (Figure 1A). Here, we examine for the first time the effects of binding by a regulatory protein (CEBPβ) on C→T mutation rates, after we noticed enhanced frequencies of such mutations within CEBP binding sites.

MATERIALS AND METHODS

Somatic mutation mapping with the CEBP footprints

In order to evaluate the somatic mutation occurrence rate in human TF binding sites, we utilized our (KCA) recently-published cancer somatic mutation data, generated from 3000 human tumor and matched-normal whole-genome sequences (2). To identify the CEBP binding sites in the genome, we downloaded 'consensus_footprints_and_collapsed_motifs_hg38.bed' from the ENCODE study (14). We converted the obtained CEBP footprints from this file to hg19 human genome assembly by using the UCSC genome browser LiftOver tool, which resulted in 85 204 footprints. We overlapped the somatic mutation data with these footprints to estimate the CEBP binding site mutation rates in human cancers. Next, we generated an estimate of the expected number of mutations by shuffling randomly the CEBP footprints on the same chromosome and calculating the total number of mutations at the random locations. We performed this shuffling exercise 1000 times and an empirical *P*-value was calculated.

Somatic mutations in MCF-7

The CEBPβ enriched regions in MCF-7 breast cancer cell line was generated from dataset GSM1010889 (total peaks = 92 896; within exons = 12 259; with C→T mutation = 2143). The mutations in coding regions (exons) in breast cancer were identified using the dataset available in Cosmic (total SPNs of exons = 1 160 476; total C→T of exons = 197 281; C→T within CEBPβ peaks = 2311). We performed integrative analysis between these two datasets, to calculate the C→T mutations within CEBPβ binding regions that are located within exons. Motif analysis was performed using Multiple Expression motifs for Motif Elicitation (MEME) (15). A summary of assessments of CEBPβ binding specificity is available (16).

Protein expression and purification

The highly purified recombinant proteins used in this study, except for UDG, were all characterized previously in our laboratories: CEBPβ (pXC1599, residues 269–344) (17), TCF4 (pXC2002, residues 569–628) (18), MBD4 (pXC1063, residues 411–554) (19,20), TDG (pXC1168, residues 1–306) (21–23). Uracil–DNA glycosylase (UDG) was purchased form NEB (catalog # M0280S).

Fluorescence-based DNA binding assay

Fluorescence polarization assays were performed using a Synergy 4 microplate reader (BioTek) to measure DNA binding affinity. The 6-carboxy-fluorescein (FAM)-labeled double stranded DNA probe (5 nM) was incubated with increasing amounts of proteins (monomer concentration 0.6 nM to 10 μM) for 15 min in 20 mM Tris (pH 7.5), 5% (v/v) glycerol, 500 mM NaCl. GraphPad Prism software (version 7.0) was used to do curve fitting. K_D values were calculated as [mP] = [maximum mP] × [C]/(K_D + [C]) + [baseline mP], where mP is millipolarization, [C] is protein concentration, and ΔmP = ([mP] – [baseline mP]). Error bars represent the standard deviation from two independent experiments, each done in quadruplicate. For those binding curves that did not reach saturation, the lower limit of the binding affinity was estimated.

Isothermal titration calorimetry

ITC experiments were performed at 25°C using a MicroCal PEAQ-ITC automated system (Malvern instrument Ltd). Double stranded oligonucleotides and protein were diluted in buffer (20 mM Tris, pH 7.5 and 500 mM NaCl, 5% glycerol). DNA was maintained in the sample cell and the proteins were injected into the cell by syringe. The amount of each injection was 2 μl with continuous stirring (750 rpm) and the reference power was set to 8 μcal/s. The duration of each injection was fixed at 4 s, and the spacing time between the injections was 200 s in order to achieve equilibrium. For each oligo, a reference titration of buffer to DNA, protein to buffer and buffer to buffer were subtracted from experimental data to control for heat of dilution and non-specific binding. Binding constants were calculated by fitting the data using the ITC data analysis module supplied by the manufacturer.

DNA glycosylase activity assays

Indicated amounts of CEBPβ or TCF4, with 40 nM FAM-labeled 32 nt-DNA (either a CEBP-binding site or a random control sequence), were incubated in reaction buffer (20 mM Tris, pH 8.0, 1 mM EDTA, 1 mM TCEP, 0.1 mg/ml BSA) at room temperature for 10 min. Addition of 200 nM DNA glycosylase (TDG, MBD4 or UDG) started the reaction. The reactions were incubated at room temperature for 60 min and quenched by addition of 0.1 M NaOH with heating to 95°C for 10 min. Samples were mixed with 2× loading buffer (98% formamide, 1 mM EDTA and trace amount of bromophenol blue and xylene cyanole) and heated at 95°C for 10 min and cooled on ice. A 5-μl sample was loaded onto a 10 cm × 10 cm denaturing PAGE gel containing 15% acrylamide, 7 M urea and 24% formamide in 1 × Tris-borate-EDTA (TBE). The gel was run at 1× TBE buffer at 200 V for 35 min. A Bio-Rad ChemiDoc MP Imaging system was used to scan the gel.

Crystallography

An Art Robbins Gryphon Crystallization Robot was used to set up screens of the sitting drop at ∼19°C via vapor diffusion method. For crystallization of CEBPβ in the presence of 16-bp G:T mismatch oligonucleotide (5′- AGG ATT GTG CAA TAT A-3′ and 3′-T TCC TAA CGC GTT ATA-5′ where T:G mismatch is in bold and underlined), equal

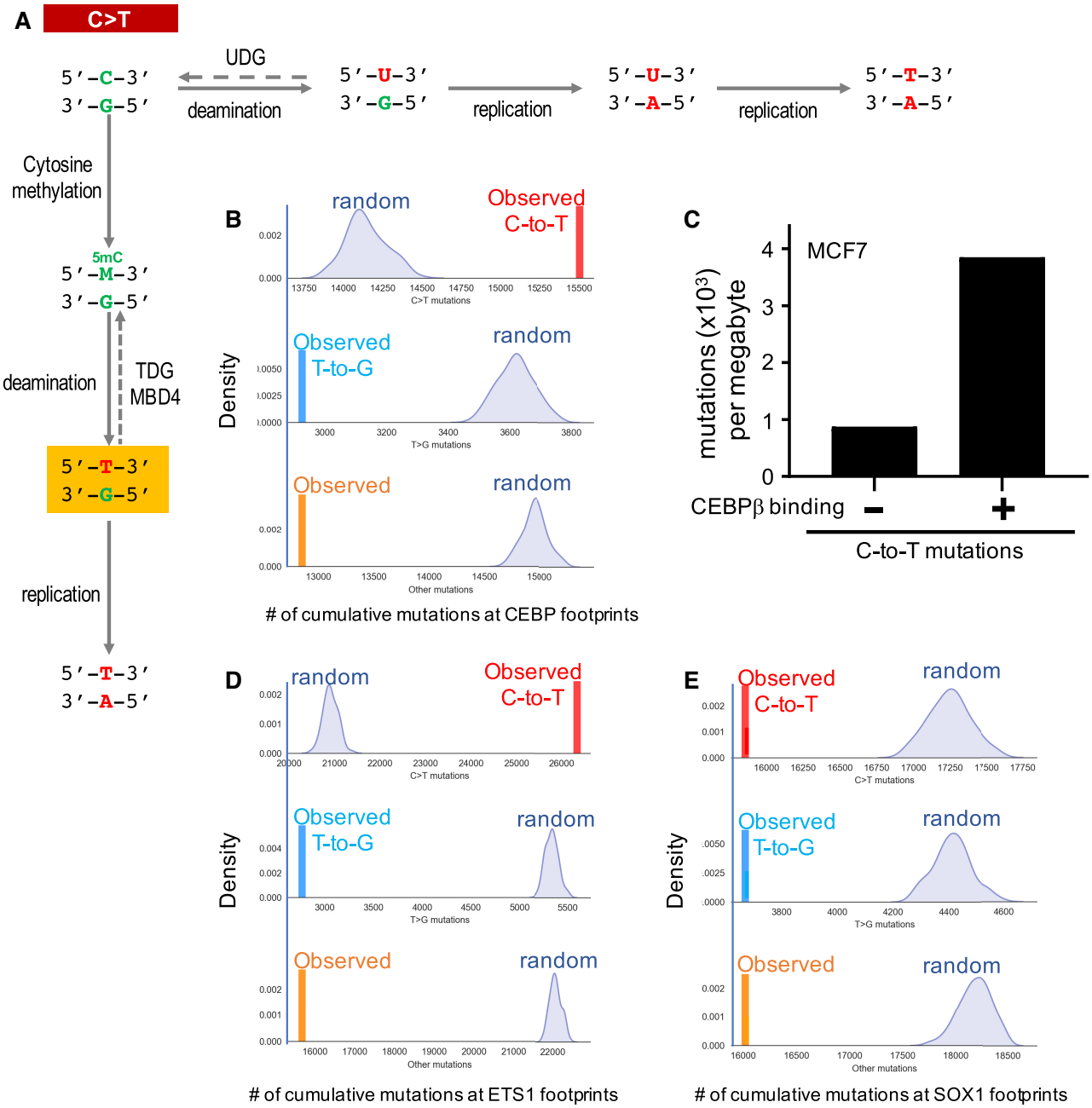


Figure 1. C→T mutations are enriched in CEBP binding sites. (**A**) Examples of single-base-substitutions (SBS) associated with deamination in DNA. The deamination products of cytosine and 5-methylcytosine (5mC or M) create G:U or G:T mismatches, respectively. Binding of the mismatch by TFs, limiting access to the damaged site. (**B**) Overlap of somatic mutations with CEBP footprints (85 204 sites with 20-bp length). Observed mutation rates within CEBP footprints are shown with a vertical bar (red for C→T; blue for T→G and orange for other four mutation types; note that the axis scales differ). The empirical *P*-value (0.001) was derived from 1000 randomly generated (curves) overlaps. The Y-axis is the density of the randomness in arbitrary units. (**C**) C→T mutation rates in exon with and without CEBPβ enrichment in MCF7 breast cancer cells (using dataset GSM1010889). (**D**) Overlap of somatic mutations with footprints of ETS1 (119 962 binding sites) with enriched C→T mutations. (**E**) No enrichment of mutations within SOX1 binding sites (116 875). X-axis is the number of cumulative mutations at the TF footprints. The Y-axis is the density of the randomness at the curve.

amounts of purified protein (2 mM) and double-stranded oligonucleotide (2 mM) were incubated at 4°C for 30 min in 20 mM Tris (pH 7.5), 300 mM NaCl, 5% (v/v) glycerol and 0.5 mM TCEP before crystallization. Crystals were observed under many conditions and X-ray diffraction data were collected from the crystals that formed in solutions 50 mM ammonium sulfate, 50 mM Bis–Tris pH 6.5, 30% (v/v) pentaerythritol ethoxylate (15/4 EO/OH).

Single crystals were flash frozen in liquid nitrogen by equilibrating in a cryoprotectant buffer containing the crystallization solution and 20% (v/v) ethylene glycol. X-ray diffraction data were collected at the SER-CAT beamline 22ID of the Advanced Photon Source at Argonne National Laboratory at wavelength of 1.0 Å. Crystallographic datasets were processed with HKL2000 (24). Molecular replacement was performed with the PHENIX PHASER module (25) by using respectively the known structures of the human CEBPβ (PDB 1GU4) as a search model. Structure refinement was performed with PHENIX Refine (26) with 5% randomly chosen reflections for the validation by the R_{free} value. COOT (27) was used for the manual building of the structure model and corrections between refinement rounds. DNA models were built into difference density during the first several rounds of refinement for the two complex structures. Structure quality was analyzed during PHENIX refinements and finally validated by the PDB validation server (28). Molecular graphics were generated by using PyMol (Schrödinger, LLC).

RESULTS

Human C→T mutations are enriched in CEBP binding sites

Previous studies have suggested that nucleotide excision repair can be compromised by the binding of TFs to DNA (3,4,29–31). We analyzed and mapped mutations, obtained from whole-genome sequencing data of 42 different cancer types (2), onto the ENCODE TF footprints. These ENCODE footprints represent TF occupancy, at nucleotide resolution, from hundreds of human cell and tissue types (14). We noted a strong correlation between C→T mutations and CEBP footprint locations. CCAAT/enhancer binding proteins (CEBPs) constitute a six-member family of TFs, that regulate gene expression in a variety of cells/tissues/organs at different developmental stages, under both physiological and pathological conditions (32–35). CEBP proteins have been described as being both tumor promoters and tumor suppressors (36).

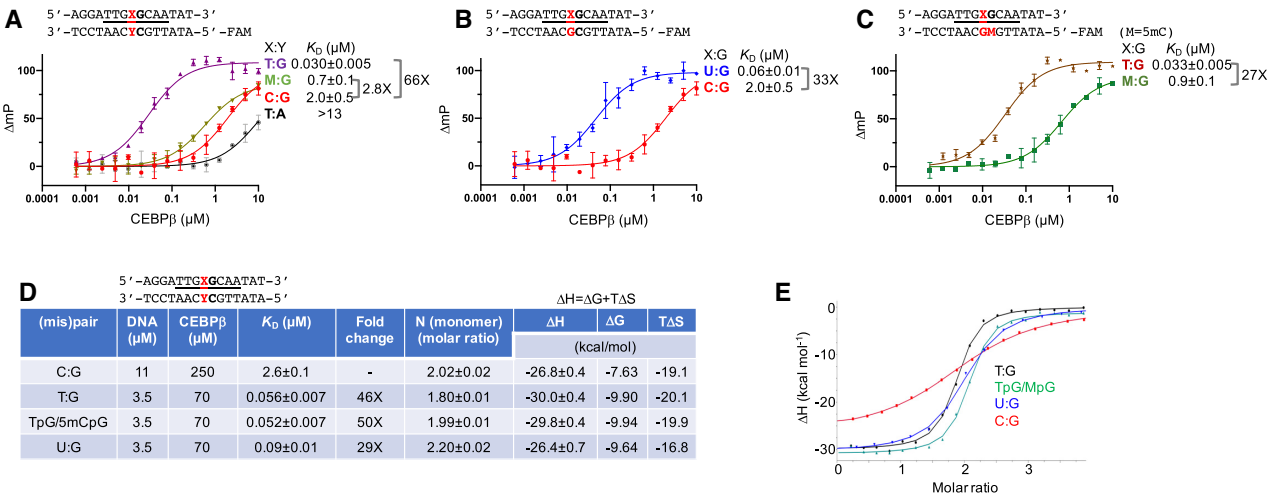
The human genome includes ∼85 200 CEBP 8-bp binding sites, the palindromic consensus for which (TTGC|GCAA) includes two of each of the four normal base pairs, and has at its center a methylatable CpG dinucleotide. As a control, we randomly selected the same number (85 200) of 8-bp sequence segments, and compared the two sets of sites for the presence of mutations. After repeatedly performing this analysis with 1000 independently-chosen sets of 85 200 random 8-bp sequence segments, we noted that C→T mutations are significantly enriched among CEBP binding sites compared to the randomly selected regions (Figure 1A; $P < 0.001$). Interestingly, all the other five potential mutations are significantly underrepresented in CEBP sites, relative to the random sequences (Figure 1B; $P < 0.001$). As noted above, the CEBP family consists of six isoforms (α-to-ζ) (32). Focusing on one CEBP family member, in MCF-7 breast cancer cells, and using the Catalogue Of Somatic Mutations In Cancer (Cosmic) database, we found that C→T mutations are preferentially enriched within identified CEBPβ binding regions, compared to randomly-selected regions (Figure 1C).

The enrichment of C→T mutations is not unique to CEBP. As with CEBP binding sites, ETS1 sites can contain a methylatable CpG dinucleotide and, indeed, are also enriched for C→T mutations (Figure 1D). Previous studies noted a significant increase in the mutation rate within ETS1 binding sites (3), and reported that ETS-related mutation hotspots exhibit strong increases in UV-induced cyclobutane pyrimidine dimers (37,38), which are more prone to undergo spontaneous deamination (39), resulting in C-to-T mutations. For comparison, no mutation enrichment was observed within SOX1 binding sites which do not include a CpG motif (Figure 1E).

CEBPβ DNA binding domain has a significant increased affinity for T:G mismatch

In order to understand the connection of DNA mutation to the binding of CEBP proteins, we used the isolated human CEBPβ basic leucine-zipper DNA binding domain and an oligonucleotide containing the consensus sequence (TTGC|GCAA). We used two biophysical assays (fluorescence polarization and isothermal titration calorimetry) to measure the dissociation constants (K_D) of CEBPβ to DNA (Figure 2 and Supplementary Figure S2). In agreement with previous observations, CEBPβ accommodates 5mC at the central CpG site, with a modestly-increased (<3×) binding affinity relative to the unmodified cognate sequence (Figure 2A) (17,40,41). This relatively small increase in binding of methylated DNA corresponds well with the genome-wide occupancy of CEBPβ irrespective of 5mC levels in H1 human embryonic stem cells (17).

As the deamination of 5mC yields T, we next measured the binding of CEBPβ to oligos containing a G:T mismatch. Under the same conditions, the G:T mismatch oligos exhibited greatly increased binding affinity, relative to the normal Watson-Crick G:C base pair, by a factor of ∼60 (from 2 μM to 30 nM) (Figure 2A). This large increase in binding affinity (∼30×) is still evident for a G:U mismatch (Figure 2B), and remains if the G following the mismatch is paired to 5mC on the opposite strand (Figure 2C). This result suggests that the G:T or G:U mismatch has a very substantial positive effect on CEBPβ binding to DNA. If the mismatch is not repaired in time, a round of replication would generate C:G to T:A substitution (to TTGTGCAA), which reduces CEBPβ binding affinity (K_D > 13 μM or >7× lower than for C:G; Figure 2A). DNA cytosine methylation generates 5mC which, like thymine, has a methyl group at the pyrimidine 5-carbon position, and increases the binding affinity modestly (<3×) but it does not change the relative order of binding affinities. Thus, CEBPβ binds DNA in the decreasing order of affinity with mismatch (T:G) >> normal base pair (5mC:G or C:G) > normal but



(mis)pair	DNA (μM)	CEBPβ (μM)	K_D (μM)	Fold change	N (monomer) (molar ratio)	ΔH	ΔG	TΔS
						(kcal/mol)		
C:G	11	250	2.6±0.1	-	2.02±0.02	-26.8±0.4	-7.63	-19.1
T:G	3.5	70	0.056±0.007	46X	1.80±0.01	-30.0±0.4	-9.90	-20.1
TpG/5mCpG	3.5	70	0.052±0.007	50X	1.99±0.01	-29.8±0.4	-9.94	-19.9
U:G	3.5	70	0.09±0.01	29X	2.20±0.02	-26.4±0.7	-9.64	-16.8

Figure 2. CEBPβ binding DNA. (**A–C**) Binding affinities of CEBPβ with oligos containing normal base pairs (C:G, 5mC:G and T:A) or mismatches (T:G and U:G), as measured by fluorescence polarization assays. (**D**, **E**) Tabular and graphical summary of concentrations of DNA and CEBPβ (monomer) used in the titration and derived binding affinity (K_D), stoichiometry (N) and entropy and enthalpy of the binding reaction by isothermal titration calorimetry (Supplementary Figure S2).

substituted base pair (T:A). This observation has probable significance beyond CEBPβ alone, as other members of the basic leucine-zipper (bZIP) family of TFs would accommodate DNA cytosine methylation within their recognition sequences (42), and may thus have similar binding preferences for the T:G product of 5mC:G deamination.

Structure of CEBPβ DNA binding domain in complex with T:G mismatch

To understand how CEBPβ binds G:T mismatches preferentially, we next co-crystallized the CEBPβ DNA-binding domain with a 15-bp duplex oligo (with a 3′ overhang nucleotide A or T) containing the G:T mismatch within the central CpG dinucleotide. The sequence chosen for co-crystallization is the same (except for the mismatch) as was used for characterizing CEBPβ bound to normal DNA (PDB 1GU4), and the resulting structure has the same resolution of 1.8 Å in the same space group (Supplementary Table S1). Except for the side chain of residue Arg289 and the mismatched thymine, there are no global changes in the overall structure of the dimeric CEBPβ and the associated DNA conformation between the two complexes, with root-mean-square deviations of <0.3 Å across 112 pairs of residues.

Each monomer recognizes half of the CEBP element by occupying the major groove of the DNA (Figure 3A). For the unaltered half site, Arg289 takes two alternative conformations, shifting between two neighboring guanines at base pair positions 3 and 4 (Figure 3B), as observed previously (17). For the altered half site, Arg289 is fixed onto the G:T mismatch at base pair position 2 (Figure 3C). For the DNA component, the major difference between the two structures lies in the shift from a three-H-bond G:C base pair at position 2 to a two-H-bond G:T mismatch, resulting in movement by ∼1.2 Å of the T away from the opposite G (Figure 3D). This movement might locally destabilize the DNA duplex were it to occur in naked DNA. However, in addition to the conventional bidentate contacts between Arg289 and G, an additional H-bond forms between the O4 keto oxygen atom of T and the Arg289 guanidinium group (Figure 3E), forming a tighter G-Arg-T triad interaction upon protein binding. In addition, a water-mediated interaction in the DNA minor groove bridges between the N2 atom of G and O2 atom of T (Figure 3E).

We note that Arg289 is highly conserved, among vertebrates from Mammalia to Chondrichthyes, among both α and β CEBP orthologs (see Supplementary Figure S5 of reference (17)). For comparison, a positively charged residue (Arg61 in pol η and Lys679 in pol ν), located in the active sites of human DNA polymerases, stabilizes a T:G mismatch in the major groove (43,44) (Figure 3F). In pol ν, mutants of Lys679 to alanine or to threonine have full activity during normal Watson-Crick base pairing, but poorly incorporate T opposite template G (45), which seems consistent with our structural observation on CEBP Arg269 stabilizing a T:G mismatch.

CEBPβ binding prevents G:T and G:U mismatch repairs *in vitro*

Two mammalian DNA glycosylases, TDG and MBD4, excise the mismatched U or T (the deamination products of C and 5mC, respectively) opposite to G (12,13). On the basis of the strong binding to DNA mismatches we observed, we asked whether CEBPβ interferes with G:T mismatch repair *in vitro*. We incubated the recombinant glycosylase domains of human TDG or MBD4 with a fluorescently-labelled oligonucleotide duplex, containing a single G:T mismatch within a CEBP consensus element. We observed that an abasic site was generated by the removal of the mispaired pyrimidine, by glycosidic bond cleavage (lane 2 in Figure 4A-B). However, in the presence of CEBPβ, the glycosylase activity was inhibited when CEBPβ concentration was at or above that of the DNA probe (lanes 7 and 8 of Figure 4A, B). Similarly, the glycosylase activities on a G:U

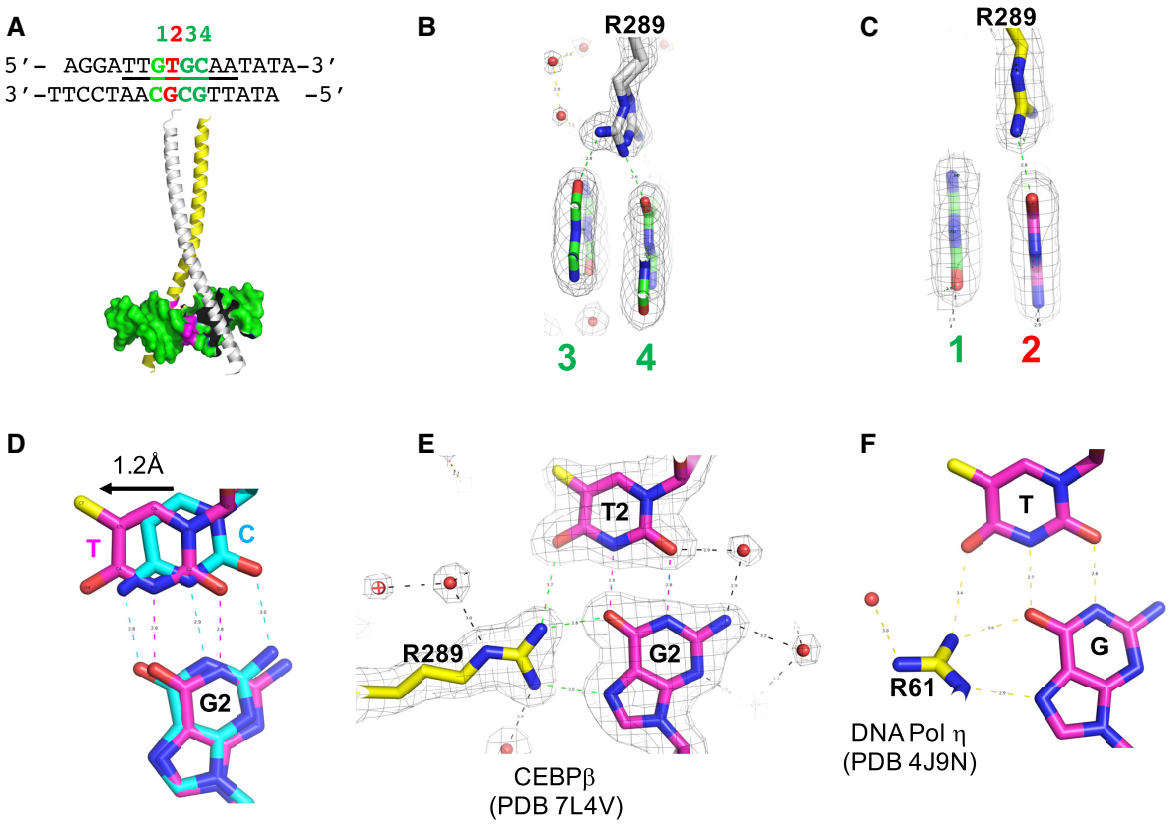


Figure 3. Structure of CEBPβ DNA binding domain in complex with G:T mispair (colored in magenta). (**A**) Each CEBPβ monomer (yellow and grey) recognizes one half-site. (**B**) Arg289 of one monomer bridges between two guanines at positions 3 and 4 of an unmutated half site. (**C**) Arg289 of the second monomer recognizes the mispaired T:G at position 2. (**D**) Comparison of T:G and C:G. The mispaired T of T:G moves away from the normal C position of a C:G pair (PDB 1GU4). (**E**) Interactions of Gua-Arg-Thy triad. The electron density (light grey) is contoured at 1.5σ above the mean. (**F**) R61 of human DNA pol η forms similar interactions with a T:G mismatch in the active site (PDB 4J9N).

mismatch within the CEBP element, by TDG, MBD4 and UDG, were also inhibited by the presence of CEBPβ (Figure 4C–E).

Next, we asked whether the mismatch is also protected if it occurs on the same DNA but outside of the CEBP element. We designed an additional T:G mismatch placed with 12-bp between it and the T:G at position 19, within the CEBP binding site (i.e. T:G at bp positions 6 and 19), 10-bp apart (T:G at positions 8 and 19), or 7-bp apart (T:G at positions 11 and 19) (Figure 5A). We evaluated two CEBPβ concentrations, differing by 2-fold, at which the higher concentration yield complete protection of a sole T:G at position 19 (within the CEBP binding site; lane 3 in Figure 5B and C). These experiments are done in the presence of a tenfold molar excess of DNA glycosylase (TDG or MBD4) over dsDNA probe. For both TDG and MBD4, T:G mismatches at positions 8 and 11 were partially protected by CEBPβ (lanes 5 and 7 in Figure 5B and C). Interestingly, the competition between the glycosylase and CEBPβ at position 11 (the closest position tested outside of the CEBP binding site) seems to allow a small portion of T:G at position 19 to be cleaved (lane 5 in Figure 5B and C). It may be that CEBPβ binding is marginally destabilized by the presence of the nearby mismatch. The largest difference between the two DNA glycosylases TDG and MBD4 is in the T:G at position 6, furthest away from the CEBP binding site – this mismatch was completely cleaved by TDG (lanes 8 and 9 in Figure 5B), but incompletely by MBD4 (i.e. still exhibiting partial protection by CEBPβ) (lanes 8 and 9 in Figure 5C). Nevertheless, the cleavage product of MBD4 is proportionally increased at T:G mismatches as the distance moves away from the CEBP binding site (cleavage at site 6 > site 8 > site 11 > site 19).

Finally, the inhibition of DNA glycosylase activity is unique to the mismatches occurring within or near the CEBP element, because CEBPβ does not inhibit TDG cleavage activity when the mismatch is within a random sequence, nor is it inhibited by the DNA binding domain of another TF, TCF4 (Figure 6A, B). There were also more general inhibitory effects on MBD4 at the higher molar ratio of TFs (CEBPβ or TCF4) to DNA probe, which could be the result of nonspecific DNA binding (Figure 6C, D; the highest protein concentration used was 0.5 μM).

DISCUSSION

Our understanding of the effects of DNA methylation on mutation varies widely. The risk of spontaneous oxidative deamination of 5mC to T (or C to U) is relatively well understood, and has been documented to occur in organisms ranging from bacteria (46) to humans (47). However, our focus here has been on the effects of 5mC methylation on

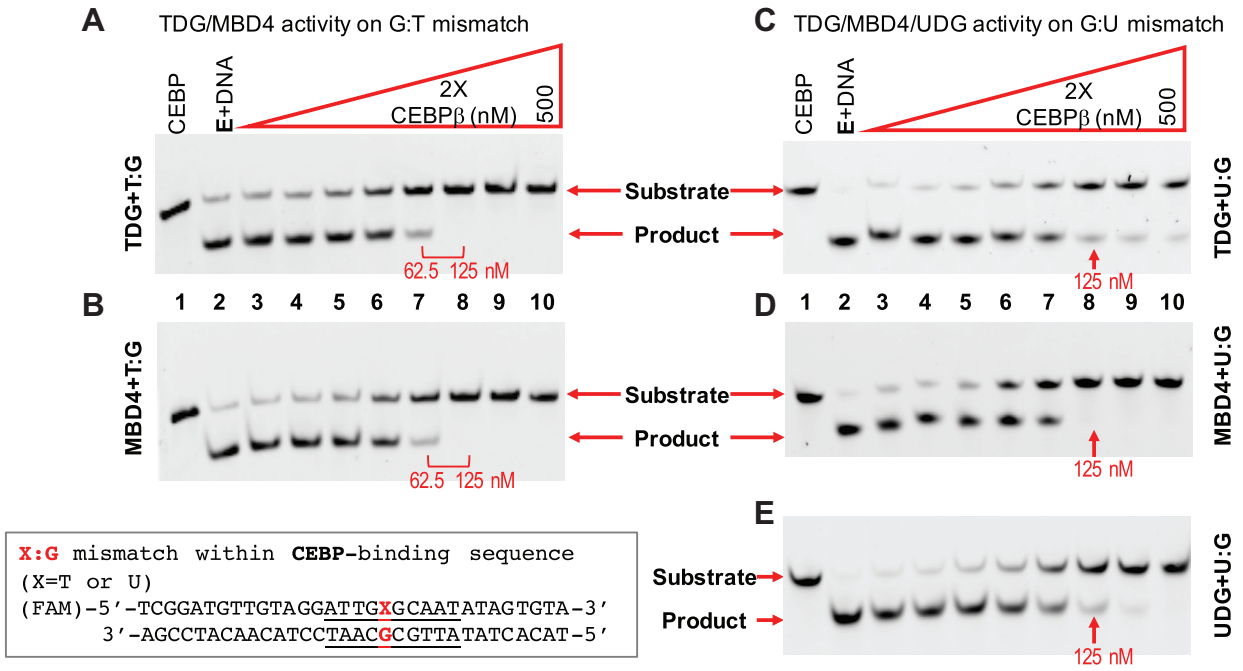


Figure 4. CEBPβ binding prevents G:T mismatch repair at CEBP binding sites. (**A**, **B**) TDG and MBD4 glycosylase activities on T:G mismatch within the CEBP binding site are inhibited by the incubation of CEBPβ with an increasing concentration (maximum of 0.5 μM at lane 10 of each panel, followed by serial 2-fold dilutions from lane 10 to lane 3, which thus has ∼4 nM). (**C**–**E**) The glycosylase activities of TDG, MBD4 and UDG, on a U:G mismatch within the CEBP binding site, are inhibited by the incubation of CEBPβ. However, the inhibition is incomplete at 125 nM for TDG and UDG.

the binding of specific proteins, and on how that protein binding modulates the mutation rate. We provided biological and bioinformatic, *in vitro* biochemical, and structural evidence for a plausible mechanism, by which accumulation of C→T mutation, following deamination of methylated cytosine, is modulated by binding of a 5mC-reader protein that has even higher affinity for T:G mismatches.

Our *in vitro* study focuses on one particular pathway of generating C→T mutations via binding of deamination products of C or 5mC by one member of CEBP family (COSMIC mutational signatures SBS1 and SBS2). However, there are many other pathways for generating C→T mutation, such as defective DNA repair enzymes (SBS6 and SBS30), treatment with alkylating agents (SBS11), or exposure to ultraviolet light (SBS7a-c) (Supplementary Figure S2B).

Possible involvement of other proteins in equivalent pro-mutagenic mechanisms

A recent study on the UV-induced CPD lesion within and around 64 binding motifs of 48 individual TFs of 10 different families revealed increased mutation rates within the binding regions—with variation among families of TFs—ascribed to TF binding having interfered with repair efficiency (31). Our structural and biochemical study on CEBPβ is highly synergistic with that analysis. Likewise, our findings are probably relevant to other members of the bZIP family of TFs that contain a conserved arginine residue corresponding to Arg289 of CEBPβ (42). That list includes activator protein 1 (AP-1 or Fos/Jun), cAMP response element (CRE) binding protein 1 (CREB1), activating transcription factor (ATF) and musculoaponeurotic fibrosarcoma oncogene homolog (MAF).

The effects we report here with CEBP might only be seen with DNA binding proteins having relatively high affinities for their DNA binding sites (whether or not that is affected by methylation or mismatches), and thus having high residence times on the DNA. In contrast, the majority of eukaryote TFs appear to have relatively low DNA affinities, relying for their specificity on combinatorial binding and elevated local concentrations in specific regions of the nucleus (48). Nevertheless, competition has been well documented between different DNA-binding proteins [e.g. (49–53)], and competition with a repair enzyme has been adapted for use in footprinting assays (54), and between a bZIP TF (CREB1) and DNA repair glycosylase UNG2 for damaged G:U mismatches within CRE element both *in vitro* and *in vivo* (55). More significantly yet, for the purposes of this study, is evidence that binding by other TFs is associated with elevated mutation rates at their binding sites (3,4,29–31).

Amount of binding protein required to have a pro-mutagenic effect

One question about this model, for elevated mutation rate within CEBP binding sites, is whether the amount of CEBP protein in the nucleus is sufficient to significantly affect repair enzyme accessibility of the roughly 85 000 CEBP sites in the human genome (so 170 000 due to diploidy). This is a difficult question to address; the question is not relevant just

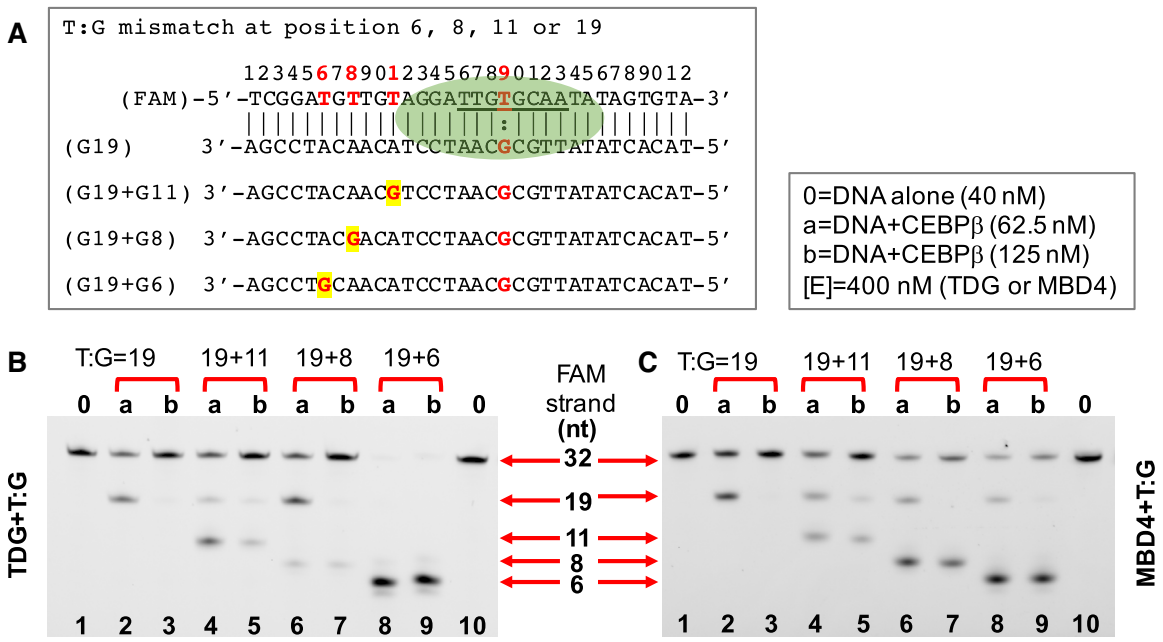


Figure 5. Partial protection of the T:G mismatch outside of the CEBP element. (**A**) Design of four DNA duplexes, all containing one T:G mismatch at bp position 19 (within the CEBP binding site), with or without a second T:G mismatch at positions 6, 8 or 11. The underlining indicates the CEBP binding site, while the green oval is meant to represent the area of DNA protected by CEBP binding (as illustrated in Figure 3). (**B**) Partial protection by CEBPβ from TDG cleavage at positions 11 and 8, and complete cleavage by TDG at position 6 (no protection). Note that TDG has thymine glycosylase activity in the following substrate preference order (low-to-high): TpT (position 8) < TpC < TpA (position 11) < TpG (positions 6 and 19) (64). (**C**) Partial protection by CEBPβ from MBD4 cleavage.

for CEBP, but applies to all other TFs that have increasingly been shown to boost mutation rates. Importantly, neither our model nor the data suggest that 100% of sites are protected by CEBP in a given lineage, just that CpGs within CEBP sites have a higher rate of mutation than expected based on the overall rate. Thus the issue is whether the levels of CEBP are high enough to protect *some* sites. We consider here a few possibilities for ways that the effective number of CEBP molecules might be higher than estimated.

First, concentrations of TFs (and repair enzymes) are likely to be inhomogeneous in at least two ways. Different regions of the nucleus have different concentrations of nuclear proteins (48), and the CEBP population is divided among six subtypes whose relative expression varies with tissue type and status (32) (see below). In particular, mutually-compensatory roles have been observed between CEBPβ and CEBPδ (56,57).

Second, the exclusion of repair enzymes does not have to be indefinite, but only until the next passage of a replication fork generates a mutated daughter duplex without a mismatch. So the amount of CEBP would not have to be so high in more-rapidly dividing cells (due to more frequent passage of replication forks).

Third, somatic mutational load in cancer genomes is correlated with topologically-associated chromatin domains: certain processes generate mutations in active chromatin domains, whereas others generate mutations in inactive domains (2). Only a small fraction of sites for a given TF are actually bound by that TF, and this binding depends in part on the surrounding 3D environment (58).

Finally, looking at levels of a single molecular species of CEBP may underestimate the effective concentration. The various subtypes vary by alternative initiation sites (there are at least three translationally regulated sub-isoforms of CEBPα and CEBPβ respectively; reviewed in reference (32)); though the DNA-binding domains are at the carboxyl termini, so all expressed isoforms are expected to have similar DNA-binding properties. Further, intra- and inter-family CEBP heterodimerizations, which effectively lowers the concentration of any single binding species, greatly expand the repertoire of DNA binding activities. Examples of such heterodimerization include CEBPβ and ATF4 (59).

CEBP expression in cancer cells

A broad-scale survey of mutation associated with oxidative stress in induced pluripotent stem cells suggests that closed chromatin contributes to exclusion of repair enzymes (60). Mutations across cancer genomes also vary with genome position, in association with intrinsic molecular organization of chromatin status, replication timing, DNA repair and transcription (61). Expression of CEBP family members is associated with patients having different cancer types. For example, CEBPβ protein expression level is positively associated with colorectal cancer but negatively associated with renal cancer (The Human Protein Atlas).

We examined gene expression levels of CEBP family members from the GEPIA database (62) in various tumors and matched normal tissues (match TCGA nor-

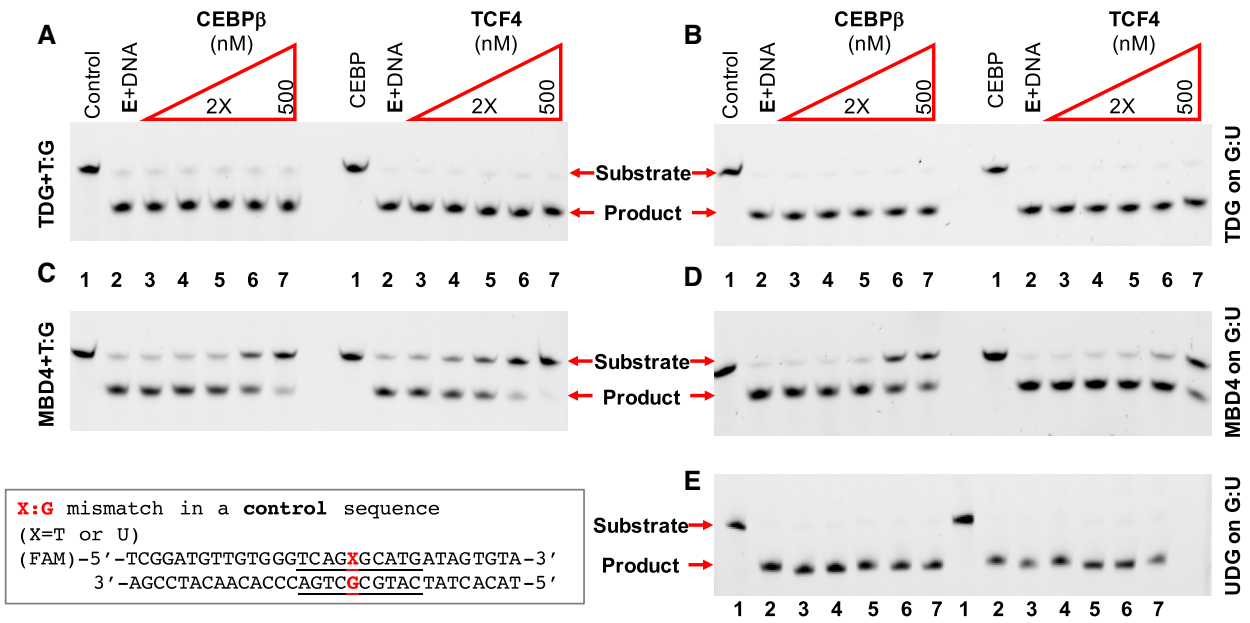


Figure 6. CEBPβ incubation with a non-specific DNA sequence, or TCF4 incubation with a CEBP binding sequence, does not inhibit TDG activity on T:G or U:G mismatches (**A**, **B**). (**C**, **D**) Non-specific DNA binding by CEBPβ or TCF4 at maximum concentration of 0.5 μM (lane 7 of each panel) inhibits MBD4 activity. (**E**) Non-specific DNA binding by CEBPβ or TCF4 does not inhibit UDG activity.

mal) (Supplementary Figure S3). There are a few relevant observations to make. First, in multiple tumors (acute myeloid leukemia, brain lower grade glioma, glioblastoma multiforme, rectum adenocarcinoma and ovarian serous cystadenocarcinoma), CEBPα has elevated levels, together with one another isoform. Second, three isoforms (CEBPβ, γ and ζ) are produced in higher amounts in pancreatic adenocarcinoma. Third, opposing regulatory patterns are seen: β, δ and ε isoforms are down while γ and ζ isoforms are up, in lymphoid neoplasm diffuse large B-cell lymphoma. Similarly, CEBPε levels are depressed while those of the γ and ζ isoforms are elevated in thymoma. The synergy of compensatory roles has been observed between CEBPβ and CEBPδ in induction of proinflammatory cytokines (56) and in mice that show defective adipocyte differentiation (57).

To summarize, we provide structural, biochemical and bioinformatic evidence for a mechanism to explain a pattern of locally-elevated C-to-T mutation. This mechanism involves TF binding that occludes a mismatch, and reduces its accessibility to mismatch repair proteins. Further support for this mechanism may require a complex analysis that links ChIP-seq, to determine binding site occupancy in a given cell type, with the mutation rate at those specific sites. Nevertheless, our results are consistent with those of a number of other studies, focusing on other TFs (Supplementary Figure S4) (63), and provide them with a likely molecular basis.

DATA AVAILABILITY

The X-ray coordinates and structure factor file have been submitted to PDB under accession number 7L4V (CEBPβ-T:G mismatch).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Ms. Yu Cao for technical assistance. We thank Dr. Hideharu Hashimoto for his previous involvement in protein purifications of MBD4 and TDG.
Authors contributions: J.Y. performed assays of DNA binding, base excision repair, and crystallization of CEBPβ-DNA. J.R.H. performed structural determinations of CEBPβ in complex with DNA. K.C.A, J.L, Y.H. analyzed somatic mutation data. J.K. analyzed expression data. R.M.B. participated in discussion throughout and assisted in preparing the manuscript. X.Z and X.C organized and designed the experimental scope of the study.

FUNDING

U.S. National Institutes of Health [R35GM134744 to X.C., R21GM138824, R01HL134780, R01HL146852 to J.L. and Y.H.]; Cancer Prevention and Research Institute of Texas (CPRIT) [RR160029]; American Cancer Society [RSG-18–043-01-LIB to J.L. and Y.H.]; X.C. who is a CPRIT Scholar in Cancer Research. The open access publication charge for this paper has been waived by Oxford University Press – *NAR* Editorial Board members are entitled to one free paper per year in recognition of their work on behalf of the journal.
Conflict of interest statement. None declared.

REFERENCES

1. Olivieri,M., Cho,T., Alvarez-Quilon,A., Li,K., Schellenberg,M.J., Zimmermann,M., Hustedt,N., Rossi,S.E., Adam,S., Melo,H. *et al.*

(2020) A genetic map of the response to DNA damage in human cells. *Cell*, **182**, 481–496.
2. Akdemir,K.C., Le,V.T., Kim,J.M., Killcoyne,S., King,D.A., Lin,Y.P., Tian,Y., Inoue,A., Amin,S.B., Robinson,F.S. *et al.* (2020) Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nat. Genet.*, **52**, 1178–1188.
3. Sabarinathan,R., Mularoni,L., Deu-Pons,J., Gonzalez-Perez,A. and Lopez-Bigas,N. (2016) Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature*, **532**, 264–267.
4. Perera,D., Poulos,R.C., Shah,A., Beck,D., Pimanda,J.E. and Wong,J.W. (2016) Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature*, **532**, 259–263.
5. Schuster-Bockler,B. and Lehner,B. (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, **488**, 504–507.
6. Alexandrov,L.B., Kim,J., Haradhvala,N.J., Huang,M.N., Tian Ng,A.W., Wu,Y., Boot,A., Covington,K.R., Gordenin,D.A., Bergstrom,E.N. *et al.* (2020) The repertoire of mutational signatures in human cancer. *Nature*, **578**, 94–101.
7. Consortium, I.T.P.-C.A.o.W.G. (2020) Pan-cancer analysis of whole genomes. *Nature*, **578**, 82–93.
8. Wijesinghe,P. and Bhagwat,A.S. (2012) Efficient deamination of 5-methylcytosines in DNA by human APOBEC3A, but not by AID or APOBEC3G. *Nucleic Acids Res.*, **40**, 9206–9217.
9. Burns,M.B., Lackey,L., Carpenter,M.A., Rathore,A., Land,A.M., Leonard,B., Refsland,E.W., Kotandeniya,D., Tretyakova,N., Nikas,J.B. *et al.* (2013) APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature*, **494**, 366–370.
10. Chan,K., Roberts,S.A., Klimczak,L.J., Sterling,J.F., Saini,N., Malc,E.P., Kim,J., Kwiatkowski,D.J., Fargo,D.C., Mieczkowski,P.A. *et al.* (2015) An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat. Genet.*, **47**, 1067–1072.
11. Stewart,J.A., Schauer,G. and Bhagwat,A.S. (2020) Visualization of uracils created by APOBEC3A using UdgX shows colocalization with RPA at stalled replication forks. *Nucleic Acids Res.*, **48**, e118.
12. Neddermann,P., Gallinari,P., Lettieri,T., Schmid,D., Truong,O., Hsuan,J.J., Wiebauer,K. and Jiricny,J. (1996) Cloning and expression of human G/T mismatch-specific thymine-DNA glycosylase. *J. Biol. Chem.*, **271**, 12767–12774.
13. Hendrich,B., Hardeland,U., Ng,H.H., Jiricny,J. and Bird,A. (1999) The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature*, **401**, 301–304.
14. Vierstra,J., Lazar,J., Sandstrom,R., Halow,J., Lee,K., Bates,D., Diegel,M., Dunn,D., Neri,F., Haugen,E. *et al.* (2020) Global reference mapping of human transcription factor footprints. *Nature*, **583**, 729–736.
15. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
16. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **172**, 650–665.
17. Yang,J., Horton,J.R., Wang,D., Ren,R., Li,J., Sun,D., Huang,Y., Zhang,X., Blumenthal,R.M. and Cheng,X. (2019) Structural basis for effects of CpA modifications on C/EBPbeta binding of DNA. *Nucleic Acids Res.*, **47**, 1774–1785.
18. Yang,J., Horton,J.R., Li,J., Huang,Y., Zhang,X., Blumenthal,R.M. and Cheng,X. (2019) Structural basis for preferential binding of human TCF4 to DNA containing 5-carboxylcytosine. *Nucleic Acids Res.*, **47**, 8375–8387.
19. Wu,P., Qiu,C., Sohail,A., Zhang,X., Bhagwat,A.S. and Cheng,X. (2003) Mismatch repair in methylated DNA. Structure and activity of the mismatch-specific thymine glycosylase domain of methyl-CpG-binding protein MBD4. *J. Biol. Chem.*, **278**, 5285–5291.
20. Hashimoto,H., Zhang,X. and Cheng,X. (2012) Excision of thymine and 5-hydroxymethyluracil by the MBD4 DNA glycosylase domain: structural basis and implications for active DNA demethylation. *Nucleic Acids Res.*, **40**, 8276–8284.
21. Hashimoto,H., Hong,S., Bhagwat,A.S., Zhang,X. and Cheng,X. (2012) Excision of 5-hydroxymethyluracil and 5-carboxylcytosine by the thymine DNA glycosylase domain: its structural basis and implications for active DNA demethylation. *Nucleic Acids Res.*, **40**, 10203–10214.
22. Hashimoto,H., Zhang,X. and Cheng,X. (2013) Selective excision of 5-carboxylcytosine by a thymine DNA glycosylase mutant. *J. Mol. Biol.*, **425**, 971–976.
23. Hashimoto,H., Zhang,X. and Cheng,X. (2013) Activity and crystal structure of human thymine DNA glycosylase mutant N140A with 5-carboxylcytosine DNA at low pH. *DNA Repair (Amst.)*, **12**, 535–540.
24. Otwinowski,Z., Borek,D., Majewski,W. and Minor,W. (2003) Multiparametric scaling of diffraction intensities. *Acta Crystallogr. A*, **59**, 228–234.
25. McCoy,A.J., Grosse-Kunstleve,R.W., Adams,P.D., Winn,M.D., Storoni,L.C. and Read,R.J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr.*, **40**, 658–674.
26. Headd,J.J., Echols,N., Afonine,P.V., Grosse-Kunstleve,R.W., Chen,V.B., Moriarty,N.W., Richardson,D.C., Richardson,J.S. and Adams,P.D. (2012) Use of knowledge-based restraints in phenix.refine to improve macromolecular refinement at low resolution. *Acta Crystallogr. D. Biol. Crystallogr.*, **68**, 381–390.
27. Emsley,P. and Cowtan,K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr. D. Biol. Crystallogr.*, **60**, 2126–2132.
28. Read,R.J., Adams,P.D., Arendall,W.B. 3rd, Brunger,A.T., Emsley,P., Joosten,R.P., Kleywegt,G.J., Krissinel,E.B., Lutteke,T., Otwinowski,Z. *et al.* (2011) A new generation of crystallographic validation tools for the protein data bank. *Structure*, **19**, 1395–1412.
29. Gonzalez-Perez,A., Sabarinathan,R. and Lopez-Bigas,N. (2019) Local determinants of the mutational landscape of the human genome. *Cell*, **177**, 101–114.
30. Morova,T., McNeill,D.R., Lallous,N., Gonen,M., Dalal,K., Wilson,D.M. 3rd, Gursoy,A., Keskin,O. and Lack,N.A. (2020) Androgen receptor-binding sites are highly mutated in prostate cancer. *Nat. Commun.*, **11**, 832.
31. Frigola,J., Sabarinathan,R., Gonzalez-Perez,A. and Lopez-Bigas,N. (2021) Variable interplay of UV-induced DNA damage and repair at transcription factor binding sites. *Nucleic Acids Res.*, **49**, 891–901.
32. Tsukada,J., Yoshida,Y., Kominato,Y. and Auron,P.E. (2011) The CCAAT/enhancer (C/EBP) family of basic-leucine zipper (bZIP) transcription factors is a multifaceted highly-regulated system for gene regulation. *Cytokine*, **54**, 6–19.
33. Sun,C., Duan,P. and Luan,C. (2017) CEBP epigenetic dysregulation as a drug target for the treatment of hematologic and gynecologic malignancies. *Curr. Drug Targets*, **18**, 1142–1151.
34. Roe,J.S. and Vakoc,C.R. (2014) C/EBPalpha: critical at the origin of leukemic transformation. *J. Exp. Med.*, **211**, 1–4.
35. Tolomeo,M. and Grimaudo,S. (2020) The "Janus" role of C/EBPs family members in cancer progression. *Int. J. Mol. Sci.*, **21**, 4308.
36. Nerlov,C. (2007) The C/EBP family of transcription factors: a paradigm for interaction between gene expression and proliferation control. *Trends Cell Biol.*, **17**, 318–324.
37. Mao,P., Brown,A.J., Esaki,S., Lockwood,S., Poon,G.M.K., Smerdon,M.J., Roberts,S.A. and Wyrick,J.J. (2018) ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. *Nat. Commun.*, **9**, 2626.
38. Elliott,K., Bostrom,M., Filges,S., Lindberg,M., Van den Eynden,J., Stahlberg,A., Clausen,A.R. and Larsson,E. (2018) Elevated pyrimidine dimer formation at distinct genomic bases underlies promoter mutation hotspots in UV-exposed cancers. *PLos Genet.*, **14**, e1007849.
39. Barak,Y., Cohen-Fix,O. and Livneh,Z. (1995) Deamination of cytosine-containing pyrimidine photodimers in UV-irradiated DNA. Significance for UV light mutagenesis. *J. Biol. Chem.*, **270**, 24174–24179.
40. Rishi,V., Bhattacharya,P., Chatterjee,R., Rozenberg,J., Zhao,J., Glass,K., Fitzgerald,P. and Vinson,C. (2010) CpG methylation of half-CRE sequences creates C/EBPalpha binding sites that activate some tissue-specific genes. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 20311–20316.
41. Kribelbauer,J.F., Laptenko,O., Chen,S., Martini,G.D., Freed-Pastor,W.A., Prives,C., Mann,R.S. and Bussemaker,H.J. (2017) Quantitative analysis of the DNA methylation sensitivity of transcription factor complexes. *Cell Rep.*, **19**, 2383–2395.

42. Yang,J., Zhang,X., Blumenthal,R.M. and Cheng,X. (2020) Detection of DNA modifications by sequence-specific transcription factors. *J. Mol. Biol.*, **432**, 1661–1673.
43. Zhao,Y., Gregory,M.T., Biertumpfel,C., Hua,Y.J., Hanaoka,F. and Yang,W. (2013) Mechanism of somatic hypermutation at the WA motif by human DNA polymerase eta. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 8146–8151.
44. Lee,Y.S., Gao,Y. and Yang,W. (2015) How a homolog of high-fidelity replicases conducts mutagenic DNA synthesis. *Nat. Struct. Mol. Biol.*, **22**, 298–303.
45. Takata,K., Arana,M.E., Seki,M., Kunkel,T.A. and Wood,R.D. (2010) Evolutionary conservation of residues in vertebrate DNA polymerase N conferring low fidelity and bypass activity. *Nucleic Acids Res.*, **38**, 3233–3244.
46. Cherry,J.L. (2018) Methylation-induced hypermutation in natural populations of bacteria. *J. Bacteriol.*, **200**, e00371-18.
47. Zhou,Y., He,F., Pu,W., Gu,X., Wang,J. and Su,Z. (2020) The impact of DNA methylation dynamics on the mutation rate during human germline development. *G3 (Bethesda)*, **10**, 3337–3346.
48. Kribelbauer,J.F., Rastogi,C., Bussemaker,H.J. and Mann,R.S. (2019) Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. *Annu. Rev. Cell Dev. Biol.*, **35**, 357–379.
49. Zabet,N.R. and Adryan,B. (2013) The effects of transcription factor competition on gene regulation. *Front Genet*, **4**, 197.
50. Gottgens,B. (2015) Creating cellular diversity through transcription factor competition. *EMBO J.*, **34**, 691–693.
51. Sokolik,C., Liu,Y., Bauer,D., McPherson,J., Broeker,M., Heimberg,G., Qi,L.S., Sivak,D.A. and Thomson,M. (2015) Transcription factor competition allows embryonic stem cells to distinguish authentic signals from noise. *Cell Syst.*, **1**, 117–129.
52. d'Azzo,A. and Annunziata,I. (2020) Transcription factor competition regulates lysosomal biogenesis and autophagy. *Mol Cell Oncol*, **7**, 1685840.
53. Szentes,S., Zsibrita,N., Koncz,M., Zsigmond,E., Salamon,P., Pletl,Z. and Kiss,A. (2020) I-Block: a simple *Escherichia coli*-based assay for studying sequence-specific DNA binding of proteins. *Nucleic Acids Res.*, **48**, e28.
54. Devchand,P.R., McGhee,J.D. and van de Sande,J.H. (1993) Uracil-DNA glycosylase as a probe for protein–DNA interactions. *Nucleic Acids Res.*, **21**, 3437–3443.
55. Moore,S.P., Kruchten,J., Toomire,K.J. and Strauss,P.R. (2016) Transcription factors and DNA repair enzymes compete for damaged promoter sites. *J. Biol. Chem.*, **291**, 5452–5460.
56. Lu,Y.C., Kim,I., Lye,E., Shen,F., Suzuki,N., Suzuki,S., Gerondakis,S., Akira,S., Gaffen,S.L., Yeh,W.C. *et al.* (2009) Differential role for c-Rel and C/EBPbeta/delta in TLR-mediated induction of proinflammatory cytokines. *J. Immunol.*, **182**, 7212–7221.
57. Tanaka,T., Yoshida,N., Kishimoto,T. and Akira,S. (1997) Defective adipocyte differentiation in mice lacking the C/EBPbeta and/or C/EBPdelta gene. *EMBO J.*, **16**, 7432–7443.
58. Dror,I., Golan,T., Levy,C., Rohs,R. and Mandel-Gutfreund,Y. (2015) A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res.*, **25**, 1268–1280.
59. Podust,L.M., Krezel,A.M. and Kim,Y. (2001) Crystal structure of the CCAAT box/enhancer-binding protein beta activating transcription factor-4 basic leucine zipper heterodimer in the absence of DNA. *J. Biol. Chem.*, **276**, 505–513.
60. Yoshihara,M., Araki,R., Kasama,Y., Sunayama,M., Abe,M., Nishida,K., Kawaji,H., Hayashizaki,Y. and Murakawa,Y. (2017) Hotspots of de novo point mutations in induced pluripotent stem cells. *Cell Rep.*, **21**, 308–315.
61. Lim,B., Mun,J. and Kim,S.Y. (2017) Intrinsic molecular processes: impact on mutagenesis. *Trends Cancer*, **3**, 357–371.
62. Tang,Z., Li,C., Kang,B., Gao,G., Li,C. and Zhang,Z. (2017) GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.*, **45**, W98–W102.
63. Afek,A., Shi,H., Rangadurai,A., Sahay,H., Senitzki,A., Xhani,S., Fang,M., Salinas,R., Mielko,Z., Pufall,M.A. *et al.* (2020) DNA mismatches reveal conformational penalties in protein-DNA recognition. *Nature*, **587**, 291–296.
64. Scharer,O.D., Kawate,T., Gallinar,P., Jiricny,J. and Verdine,G.L. (1997) Investigation of the mechanisms of DNA binding of the human G/T glycosylase using designed inhibitors. *Proc. Natl. Acad. Sci. USA*, **94**, 7878–4883.