# Title: CTCF directly binds G-quadruplex structures to regulate genome topology and gene expression

Daniela Samaniego-Castruita[1*], Isabella Han[1*], Roxroy C Morgan[1*], Samantha Carpenter[1], Bryce Williams[1], Abhijit Chakraborty[2], Ishwar Radhakrishnan[3], Ferhat Ay[2,4,5], Samuel A Myers[6,7], Anjana Rao[7#], Vipul Shukla[1,8,9#]

1 Department of Cell and Developmental Biology, Northwestern University, Chicago, USA 60611
2 Center for Autoimmunity and Inflammation, La Jolla Institute for Immunology, La Jolla, CA 92037
3 Department of Molecular Biosciences, Northwestern University, Evanston, 60201
4 Bioinformatics and Systems Biology PhD Program, University of California, San Diego, CA 92093
5 Department of Pediatrics, University of California San Diego, La Jolla, CA 92093
6 Laboratory for Immunochemical Circuits, La Jolla Institute for Immunology, La Jolla, USA, 92037
7 Division of Signaling and Gene Expression, La Jolla Institute for Immunology, La Jolla, USA, 92037
8 Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, USA 60611
9 Center for Human Immunobiology, Northwestern University, Chicago, USA 60611

* These authors contributed equally
# Co-Lead authors
Correspondence: vipul.shukla@northwestern.edu

**Abstract**

DNA G-quadruplexes (G4s) are non-B form secondary DNA structures that are highly conserved across evolution. G4 structures occupy key regulatory sites in the mammalian genomes and are implicated in several cellular processes. However, the mechanisms by which G4s contribute to distinct facets of genome function are not well understood. Here, we conduct a proteomics screen with G4s of diverse topologies to uncover novel G4 binding activities in genomic regulators of nucleosome remodeling, paraspeckle assembly, RNA splicing, and 3D genome organization. Among the most prominent hits, we identify the genomic architectural protein, CTCF, as one of the strongest G4 binders. Building on this discovery, we perform extensive biochemical validation of CTCF-G4 interaction and identify CTCF mutants, with pronounced affinity for G4s over its consensus DNA motif. By implementing well-established approaches and developing new G4 mapping tools, we define a comprehensive catalog of genomic G4s and investigate their association with CTCF binding. Using genetic reconstitution of mouse embryonic stem cells with a G4-specific CTCF mutant, we ascribe genomic functions to CTCF-G4 interactions in regulation of CTCF occupancy, chromatin looping and gene expression. Interestingly, our studies reveal a subset of G4-linked chromatin loop anchors that form "persistent loops" which are retained even upon CTCF depletion. Collectively, our work establishes the G4 binding activity of CTCF and provides new insights into the functional significance of G4 structures.

## Introduction

G-quadruplexes (G4s) are non-B-form DNA secondary structures that form at guanine-rich regions in the genome[1-7]. G4s form when four stretches of guanine (G) bases within G4 sequence motifs ($G_{\geq3}N_{1-7}G_{\geq3}N_{1-7}G_{\geq3}N_{1-7}G_{\geq3}$) come together in planar symmetry, stabilized by atypical Hoogsteen hydrogen bonding and monovalent cations (typically $K^+$ ions)[2-5,8]. Depending on the sequence context, relative orientation of the DNA strands, and the spacing within guanine stretches, G4s can adopt distinct topologies[5]. Notably, G4s are highly conserved across evolution[7,9], and putative G4-forming sequences (pG4s) are enriched at gene promoters[4,10], 5' untranslated regions[4,11], DNA replication origins[4,12] and telomeres[4-6], indicating that these structures likely have important functions in the genome. Consistent with this notion, studies have uncovered the association of G4 structures with regulation of gene expression[2,4,5], chromatin accessibility[13], DNA methylation[14,15], genomic stability[15-17], telomere homeostasis[4,5,18] and three-dimensional (3D) genomic architecture[19-21]. That said, the mechanisms by which G4s regulate these diverse genomic processes continues to be an area of active investigation. Recent affinity-based and proximity-based labelling approaches have identified G4 binding activity in transcription factors (SP1, FUS, E2F and NRF1)[22] as well as protein complexes involved in regulation of DNA methylation (DNMT1, UHRF1) and chromatin accessibility (SWItch sucrose non-fermentable, SWI/SNF complex)[23]. However, the biochemical basis of G4 interaction with these proteins, as well as the genomic contexts in which they occur, warrants additional investigation to functionally elucidate G4 biology.

Among the genomic functions of G4s, recent reports have proposed an intriguing link between G4 structures and higher-order genomic folding[19-21,24]. Mammalian genomes are segregated into chromosomal territories that are further partitioned into distinct compartments, separating gene-rich active regions (compartment A) from the gene-poor inactive regions (compartment B)[25-30]. At the scale of tens to hundreds of kilobases, the genome is further assembled into self-interacting regions known as topologically associating domains (TADs), with higher propensity of genomic interactions observed within, rather than, between TADs[25-29]. The TAD structures mediate enhancer-promoter interactions and are important for regulating cell-type specific gene expression programs[25-29]. Interestingly, G4 structures are shown to be enriched at TAD boundaries and are implicated in regulating insulation strength of TADs[19]. On a molecular level, mammalian TAD boundaries are frequently associated with binding of CCCTC-binding factor (CTCF) and the cohesin complex[28,30]. CTCF is a highly conserved Zinc-finger (ZF) containing DNA binding protein that is proposed to play a central role in a linear tracking model known as "loop-extrusion" for long-range chromatin interactions[26,31-34]. The loop-extrusion model posits that CTCF, bound to its DNA consensus motifs, particularly in convergent orientation, impedes the movement of the cohesin ring complex to define anchors of chromatin loops and TAD boundaries[30,33,35]. Though which convergent CTCF sites demarcate loop anchors or TAD boundaries and if additional determinants are required for specifying the boundary elements for chromatin looping are open questions.

The binding of CTCF to chromatin is highly dynamic with residence times in the range of minutes, although once associated to certain sites, CTCF can remain bound for tens of minutes[36,37]. CTCF binding in the genome is most well-characterized at a 12-15 base pair DNA consensus motifs (5--NCANNA-

G(G/A)N-GGC-(G/A)(C/G)(T/C)-3') which are present at ~80% of all CTCF binding sites[32,34,38,39]. CTCF harbors 11 cysteine2-histidine2 (C2H2)-type ZFs and structural studies have shown that the central ZFs 4-7 are critical for recognizing the consensus DNA motif[26,32,38,39]. It is estimated that in any given cell type roughly 30-60% of CTCF consensus motifs are occupied by CTCF[40]. It is worth mentioning that the affinity of CTCF for different genomic DNA motifs can differ greatly and the mechanisms for selective binding of CTCF to specific sites are not fully understood[39,41]. In this regard, a recent study suggested that G4 structures can enhance binding of CTCF to its consensus DNA sequence particularly at weaker motifs[20]. However, the authors in this study could not detect direct interaction of CTCF with G4s[20]. Moreover, the structural basis and the biophysical principle for the proposed model, which would require precise positioning of G4s relative to the CTCF consensus motifs, are unclear[20]. The relationship between CTCF and G4s is further complicated by another recent report, which proposed both direct and indirect roles of G4s in CTCF recruitment[24]. Besides its binding to DNA, CTCF is also reported to bind chromatin-associated RNA[42,43]. Taken together, these recent studies highlight that CTCF occupancy in the genome is more nuanced than previously appreciated.

Here, by performing a proteomics screen with G4s of distinct topologies, we uncover G4 binding activity in several known and novel proteins that play key roles in genome biology. We identify G4 binding activity in protein complexes involved in nucleosome remodeling, paraspeckle assembly, RNA splicing, and 3D genome organization. Among the major hits from the screen, we report robust G4 recognition by CTCF. We perform biochemical validation of CTCF-G4 interaction and characterize specific CTCF mutants with strong binding preference for G4s. Furthermore, we utilize murine embryonic stem cells (mESCs) with degron-tagged CTCF to acutely deplete endogenous CTCF and reconstitute with a G4-specific CTCF mutant. Using these experimental designs, we dissect the genomic roles of CTCF-G4 interaction for regulation of chromatin looping and gene expression. Together our work provides compelling evidence for CTCF-G4 interaction and contributes to the emerging roles of G4 structures in genome function and organization.

## Results

### A proteomics screen identifies CTCF as a G4 binding factor

To investigate the nuclear functions of DNA G4 structures, we used four biologically relevant oligonucleotides (oligos) which form G4s of parallel (Kit1 and Kit2), anti-parallel (Spb1) and mixed propellor/intermolecular (Telomere: Telo) topologies (**Fig.1A**). We minimally mutated the G-stretches in each of these oligos to generate respective non-G4 control oligos (**Fig.1A**). To control for potential differences in biotinylation that may in turn impact pulldown efficiencies, we extended the G4 and non-G4 oligos to carry a complementary DNA sequence, which was annealed to a common biotinylated adaptor oligo (**Fig.1A**). This oligo design also provides the double strand DNA (dsDNA)-G4 junction associated with G4s *in vivo* (**Fig.1A**). We prepared nuclear lysates from primary murine B cells stimulated with lipopolysaccharide (LPS) and interleukin-4 (IL-4) for 72 hours and used a B cell lymphoma cell line CH12F3 (CH12) cells. Since G4 formation is sensitive to salt concentrations, we used hypotonic nuclear lysis combined with dounce homogenization and performed buffer exchange to replace the lysis buffer

with an intracellular salt buffer (110 mM KCl and 10.5 mM NaCl), to mimic conditions conducive for G4 formation in cells (**Fig.1A**). For pulldown assays, we performed in-solution interaction of biotinylated G4 and non-G4 oligos with nuclear lysates and captured bound-fractions on streptavidin conjugated magnetic beads. For initial tests, we added G4-specific Flag-tagged single chain variable fragment BG4 (BG4) antibody[44] in nuclear lysates and observed robust enrichment of Flag-BG4 and a well-known G4 helicase, Bloom (BLM) with the G4 oligos compared to non-G4 oligos (**Fig.1B**).

To discover a catalog of G4-interacting proteins, we subjected the G4 and non-G4 pulldown fractions (two biological replicates from primary B and CH12 cells) to quantitative, mass spectrometry-based proteomics (**Fig.1C, Sup.Fig.1A-B**). Compared with non-G4 control oligos, our results showed G4-mediated enrichment of several protein complexes involved in important genomic functions, such as RNA splicing, mRNA cleavage, telomere homeostasis, nucleosome remodeling (SWItch-Sucrose Non-Fermentable or SWI/SNF complex), chromosome condensation (including CCCTC-binding factor or CTCF) and histone modifications (Polycomb repressor or PcG complex) (**Fig.1D, Sup.Fig.1C**). In fact, several subunits of the spliceosomal, mRNA cleavage, SWI/SNF, PcG and paraspeckle assembly were enriched in G4-bound fractions (**Sup.Fig.1C**). Importantly, many of the hits from our screen overlapped proteins that were recently identified using an analogous approach labeling G4 interacting proteins with click chemistry using a G4 ligand Pyridostatin (PDS)[23]; these common hits included proteins involved in nucleosome remodeling, RNA splicing and mRNA processing, among others (**Fig.1D**). Even though we used DNA G4s, we also observed enrichment of RNA processes such as translation initiation, where G4 structures have been previously implicated[4,45]. We also noted G4 binding activity in several proteins associated with paraspeckles and Cajal bodies, which are membrane-less nuclear bodies with biomolecular condensate-like features. We corroborated some of the top targets using western blotting and confirmed the enrichment of SWI/SNF subunits ARID1A and SMARCB1, heterochromatin associated protein TRIM28 (Tripartite motif-containing 28), the paraspeckle associated protein NONO (Non-POU domain-containing octamer-binding protein) (**Fig.1E**), and CTCF in G4-bound fraction compared with non-G4 control oligos (**Fig.1F**).

Since many of the identified proteins are subunits of multimeric protein complexes, we focused on characterizing the G4 binding activity of CTCF, which directly binds DNA and was recently shown to associate with G4 structures[20,24]. To begin, we tested the G4 binding specificity of CTCF to the four G4 oligos used in our screen. Compared to the respective non-G4 controls, CTCF showed strong binding to each of the four G4s of different topologies, highlighting that CTCF recognizes G4 structures rather than an underlying DNA sequence within the G4 oligos (**Fig.1G**). Altogether, the proteomics screen uncovered the ability of several important regulators of genomic function, including the genomic architectural protein, CTCF, to bind G4 structures specifically.

**Biochemical characterization of CTCF-G4 interaction**

For defining the biochemical basis of G4 recognition by CTCF, we established a system for *in vitro* transcription and translation of CTCF using T7 polymerase and rabbit reticulocyte lysate (**Fig.2A**). *In vitro* synthesized full length (FL) CTCF showed strong binding to biotinylated G4 oligos compared to the corresponding non-G4 oligos (**Fig.2A**). To further characterize the specific domains of CTCF directly

involved in G4 recognition, we performed deletional mutagenesis. CTCF is composed of 11 C2H2-type zinc fingers (ZFs) and unstructured regions at N- and C-termini. We found that the N- and C-terminal regions, as well as ZFs 1-3 and ZFs 8-11 are not essential for G4 recognition by CTCF (**Fig.2A**). However, akin to its binding to the consensus dsDNA motif, the central ZFs 4-7 (117 amino acids) were both essential and sufficient for G4 recognition (**Fig.2A**). To investigate this binding further, we generated the recombinant CTCF with central ZF4-7 fragment using bacterial expression (**Fig.2B**). The CTCF ZF4-7 fragment displayed clear binding to G4 DNA compared to non-G4 controls (**Fig.2C**), ruling out the possibility that CTCF ZF 4-7 interacts indirectly with G4s through a binding partner.

Since CTCF ZFs 4-7 mediates interaction with both G4 and dsDNA consensus motifs, we compared these binding activities in a competition assay. For these experiments, we mixed the biotinylated G4 oligo (*Kit2*) with either the (non-biotinylated) CTCF consensus dsDNA motif or a mutated version of the dsDNA consensus at different molar ratios (**Fig.2D**). Following in-solution binding of CTCF ZF4-7 with different oligo mixtures we recovered G4 bound fractions through pulldown with streptavidin conjugated beads. While the mutated dsDNA consensus motif did not affect G4 bound CTCF fraction, an equimolar ratio of the consensus dsDNA motif completely abolished CTCF binding to G4 (**Fig.2D**). Notably, the G4 bound CTCF fraction was progressively recovered with 2-fold molar excess of G4 and showed binding at saturation levels when 4-fold molar excess of G4 oligo was used (**Fig.2D**). Based on these results, we estimate that the binding affinity of CTCF ZF4-7 for G4 is approximately 2-4-fold lower than that for its dsDNA consensus motif.

To measure the binding affinities of CTCF ZF4-7 for G4 and consensus DNA, we then developed a binding assay in which CTCF interaction with different DNA substrates was measured under native (intracellular) salt concentrations (**Fig.2E**). For this, CTCF was incubated with biotinylated DNA substrates in solution, followed by streptavidin capture and on-bead fluorescence measurement using particle-based flow cytometry (**Fig.2E**). It is important to note that unlike electromobility shift assays (EMSAs) or other similar assays, this method allows measurement of binding affinities under more relevant salt concentrations in the absence of EDTA, which is a potent inhibitor for DNA binding activity of zinc finger proteins[46], such as CTCF. As expected, CTCF ZF4-7 domains exhibited robust binding to G4 DNA (*Kit2*) with a dissociation constant ($K_d$) of ~63nM, whereas non-G4 oligo did not show any appreciable binding (**Fig.2F**). In comparison, CTCF ZF4-7 bound more strongly to its dsDNA consensus motif with $K_d$ of ~34nM, which is comparable to previous estimates using fluorescence polarization (**Fig.2E**)[32]. Therefore, the binding affinity of CTCF ZF4-7 for its consensus DNA motif is approximately two-fold higher than that for G4 DNA, which is consistent with estimates from the competition assays (**Fig.2D**).

To further compare the CTCF-G4 interaction to CTCF binding to the consensus DNA motif, we generated a single point mutation of histidine 373 to arginine (H373R), the second histidine in C2H2 motif in ZF4, which is critical for recognition of CTCF consensus motif (**Sup.Fig.1C**). As expected, the CTCF ZF4-7 H373R exhibited major deficits in its ability to recognize the CTCF consensus motif, whereas in contrast, the H373R mutant showed no defects in G4 recognition (**Fig.2H-J**). Though the same region of CTCF provides specificity for consensus motif and G4 DNA, these results suggest that CTCF likely recognizes

196 G4 and linear DNA substrates through distinct mechanisms. To ask whether the Zn finger structures of
197 CTCF are required for G4 recognition, we performed binding measurements of CTCF to G4 DNA in the
198 presence of varying concentrations of a $Zn^{2+}$ chelator EDTA (**Fig.2K**) and observed a concentration-
199 dependent decrease in CTCF-G4 interaction with increasing amounts of EDTA (**Fig.2K, Sup.Fig.1D**).
200 These findings suggest that the ZF domain structure of CTCF is indeed important for G4 recognition.

201 A recent study from Wulfridge et.al. proposed that G4s facilitate enhanced binding of CTCF to its
202 consensus motif, partly because they did not detect direct interaction of CTCF with G4s[20]. To address this,
203 we compared CTCF-G4 interaction in the intracellular salt buffer used in our binding assays to the binding
204 buffer used for EMSA performed in this study[20] (**Fig. 2L-M**). Notably, the binding of CTCF ZF4-7 to
205 G4s was greatly reduced in the EMSA binding buffer (from Wulfridge et.al.[20]), which contained EDTA,
206 and removal of EDTA substantially recovered G4 binding (**Fig. 2L-M**). Based on these results we reason
207 that the failure to detect CTCF-G4 interaction is likely due to suboptimal binding conditions used in
208 previous studies[20,22].

209 **CTCF ZF4 mutant (H373R) associates with G-rich genomic regions**

210 Next, to investigate the G4-dependent functions of CTCF on a genome-wide level, we used a previously
211 described mouse embryonic stem cell (mESC) system in which GFP and auxin inducible degron (AID)
212 tags are knocked into the endogenous *Ctcf* locus (AID-CTCF)[35]. These AID-tagged CTCF mESCs also
213 harbors a the *tir* transgene, which encodes the Tir E3 ubiquitin ligase that ubiquitinates and degrades the
214 endogenous AID-CTCF in the presence of the plant hormone, indole 3-acetic acid (IAA or Auxin), within
215 hours (**Fig.3A**)[35]. We transduced the AID-CTCF mESCs with lentiviral expression plasmids for FLAG-
216 and V5-tagged full-length (FL) CTCF, FL H373R mutant CTCF, or empty vector (EV) control, carrying
217 mCherry as a reporter. The use of this experimental design avoids the problem that long-term CTCF
218 depletion is detrimental to most cell types and allows us to develop a system in which CTCF-depleted
219 mESCs could be reconstituted lentivirally to express FL CTCF or H373R mutant proteins.

220 We sort-purified transduced cells based on mCherry reporter (**Sup.Fig.2A**, *left*) and confirmed
221 degradation of the 130 kDa CTCF-AID-GFP fusion protein expressed from the endogenous *Ctcf* locus
222 was degraded at 8 or 24 hours of auxin treatment, while the reconstituted FL CTCF and H373R mutant
223 proteins (detected by immunoblotting with anti-V5) were not degraded (**Fig. 3B, Sup.Fig.2B**). Since our
224 biochemical studies were performed with CTCF ZF4-7 fragments, using in vitro translation assays, we
225 confirmed that both FL and the H373R mutant CTCF binds G4 in the full-length context (**Sup.Fig.2C**,
226 *top panel*). Moreover, as with the ZF4-7 fragment, the full length CTCF with H373R mutation also showed
227 reduced binding to the CTCF consensus DNA motif (**Sup.Fig.2C**, *bottom panel*).

228 To determine whether the expression of the H373R mutant was deleterious to cells, we characterized the
229 AID-CTCF mESCs post-auxin treatment. There were no significant differences in cell numbers
230 (**Sup.Fig.2E,** *left panels*) of cells expressing empty vector or WT and mutant CTCF, or spontaneous
231 differentiation to SSEA1-negative fibroblasts (**Sup.Fig.2F-G**) at 24 hours. However, by 72 hours of auxin
232 treatment, CTCF depletion upon auxin treatment led to a significant reduction in cell numbers of the
233 parental CTCF-AID-GFP-expressing cells and cells transduced with empty vector, but this loss of viability

234 was rescued to similar extents by reconstitution with FL CTCF or H373R mutant (**Sup.Fig.2E,** *right*
235 *panels*). Notably, the FL mutant CTCF H373R was expressed at lower levels than FL wildtype CTCF
236 (**Fig. 3B**, *top and middle panels*; **Sup. Fig 2C**), even in cells sorted for equivalent levels of mCherry
237 expression (**Sup.Fig.2A,** *right*) suggesting that the mutant CTCF protein may be less stable compared to
238 the wildtype FL CTCF. Together these experiments indicate that the H373R mutant is capable of
239 supporting the survival of CTCF-AID-GFP-depleted cells as efficiently as wildtype CTCF.

240 We then profiled the genomic distribution of CTCF using CUT-and-Tag (Cleavage Under Targets and
241 Tagmentation) in AID-CTCF mESCs not treated with auxin or reconstituted with EV, FL CTCF or H373R
242 mutant and treated with auxin for 24 hours. We noted a clear loss of CTCF binding in CTCF-depleted
243 (EV) mESCs treated with auxin, and CTCF binding was recovered in cells expressing FL or H373R
244 mutant (**Fig.3C**). The majority of peaks were shared between FL and H373R mutant (n=18,416) but the
245 H373R mutant showed fewer binding peaks (n=28,137) compared to FL CTCF (n=36,715) (**Fig.3C**). We
246 categorized all CTCF peaks from cells expressing FL and H373R mutant into sites preferentially bound
247 by FL CTCF (FL only, n=18,299), H373R CTCF (H373R only, n=9,721) and those that were similarly
248 enriched for FL and H373R binding (common, n=18,416) (**Fig.3D**). Genomic annotations of these regions
249 showed increased representation of promoters and 5' untranslated regions (5'UTRs) within common sites
250 bound by both FL and H373R CTCF in comparison to FL only, H373R only or randomly shuffled genomic
251 regions (shuffle) (**Fig.3E**).

252 We performed motif enrichment analysis on FL only, common and H373R only regions, and found the
253 canonical CTCF motif as the most enriched motif in FL only and common regions (**Fig.3F**). The number
254 of peaks with the canonical CTCF motif were reduced by two-fold in common regions as compared with
255 FL only regions. On the other hand, consistent with the diminished ability of the H373R mutant to
256 recognize the CTCF consensus motif, the CTCF motif was not among the top enriched motifs in H373R
257 peaks (**Fig.3F**). Besides CTCF motifs, the FL only, common and H373R only peaks displayed enrichment
258 of OCT4-SOX2-TCF-NANOG motifs, which is in line with previous studies demonstrating cooperativity
259 of these transcription factors with CTCF (**Fig.3F**)[47]. We also performed *de novo* motif discovery analysis
260 and observed enhanced GC content (**Sup.Fig.3A**) as well as enrichment of repetitive G4-like motifs
261 (**Fig.3G**), in common and H373R only peaks. Corroborating the biochemical studies, these analyses
262 highlight the reduced affinity of the H373R CTCF mutant to consensus CTCF motifs and reveal a
263 predisposition towards G-rich genomic sequences.

264 **Common sites bound by FL and ZF4 mutant CTCF are enriched for G4s**

265 Having observed an enrichment of G-rich sequences in regions bound by both FL and H373R CTCF, we
266 then mapped G4s. G4-specific antibodies, 1H6 and BG4, have been the mainstays for genomic G4
267 profiling[44,48]. Though these antibodies show good affinity for G4 DNA, they have also been reported to
268 cross-react with cytosine-rich and thymine-rich single-stranded (ss) DNA[49-51]. In our recent work[15], we
269 used a small molecule, N-methyl mesoporphyrin IX (NMM) as a detection probe for G4s in cells, which
270 binds G4 in 1:1 stoichiometry through a well-defined π-stacking mechanism (**Sup.Fig.3B**)[52,53].
271 Importantly, NMM is highly selective for G4s, particularly G4s of parallel topologies, which motivated

us to further develop it as a G4-specific pulldown probe (**Fig.4A**). Using propylphosphonic anhydride (T3P) as a coupling agent, we conjugated pegylated biotin amine with the carboxyl groups on NMM, which do not participate in G4 recognition by NMM (**Fig.4A**). We confirmed the generation of NMM-biotin complex (NMM-Bio) with mass spectrometry and purified NMM-Bio with chromatography. Upon titration, ~10μM NMM-Bio efficiently pulled down G4 DNA and showed no detectable affinity for non-G4 DNA (**Fig.4B, Sup.Fig.3C**).

We then adapted NMM-Bio in CUT-and-Tag workflow and profiled genomic G4s using both BG4 and NMM-Bio CUT-and-Tag and used unconjugated NMM combined with anti-biotin antibody as control (**Fig.4C**). We first asked whether the peaks identified by BG4 and NMM-Bio CUT-and-Tag in FL or H373R CTCF expressing mESCs overlapped with G4s identified using G4-seq in non-chromatinized mouse genomic DNA[7,54]. Genomic peaks identified by BG4 and NMM-Bio showed 53% and 50% overlap with regions marked by G4-seq, respectively (**Fig.4D**). The regions identified by G4-seq showed only a minor overlap (~10%) with G4 peaks identified by BG4 and NMM-Bio CUT-and-Tag, suggesting that BG4 and NMM-Bio identify distinct sets of genomic G4s in cells (**Fig.4D-E**). Comparison of BG4 and NMM-Bio CUT-and-Tag peaks overlapping G4-seq peaks showed a similar distribution of predicted G4s of distinct subtypes across methods (**Sup.Fig.4A**). Considering that BG4 and NMM-Bio methods reveal unique G4 regions, we decided to focus on the union of regions identified by either of these methods (i.e. 1989 peaks) for subsequent analysis.

To examine the association between genomic CTCF binding sites with G4 forming regions, we overlapped FL only, common, and H373R only CTCF sites with G4 regions identified using BG4 or NMM-Bio. In concordance with the G4 binding ability of FL and H373R, we observed most overlap of G4 peaks with the common CTCF peaks (36%, n=6,630), that were independently bound by both FL and H373R CTCF (**Fig.4F-G**). Even though the H373R mutant is expressed at a lower level and shows fewer statistically defined peaks, the proportion of H373R peaks overlapping G4 peaks was slightly higher than FL CTCF peaks (28% of H373R peaks vs 25% of FL peaks) (**Sup.Fig.4B**). We also analyzed ChIP-Seq data from a published study[39], where several ZFs of CTCF were individually mutated by histidine (H) to arginine (R) substitutions and overexpressed in murine B cells with a BioID tag (**Sup.Fig.4C**). We found enrichment of a strictly defined $G_3N_{1\text{-}7}G_3N_{1\text{-}7}G_3N_{1\text{-}7}G_3$ (3G-G4 motif) at CTCF (wildtype) binding sites (10% of peaks with 3G-G4 motif). Interestingly, the enrichment of 3G-G4 motif was further increased with mutations in ZFs 4 to 8 (from 10% to ~30-40%), suggesting a shift in CTCF binding to regions with a higher propensity to form G4s. Additionally, we found an increase in total number of CTCF binding sites (peaks) when G4s were stabilized in CH12 B cells using small molecule ligands pyridostatin (10μM PDS, 17,331 peaks) and 3,11-Difluoro-6,8,13-trimethyl-8H-quino[4,3,2,kl] acridinium methosulfate (2μM RHPS4, 17,458 peaks) when compared to DMSO (10,053 peaks) (**Sup. Fig. 4D**). Overall, these experiments reveal clear genome-wide associations of CTCF binding with G4 structures.

**CTCF directly interacts with genomic G4 structures**

To further dissect the CTCF-G4 interaction on a genomic level, we categorized CTCF peaks commonly bound by both FL and H373R CTCF into those with or without a canonical CTCF motif (Homer). As

310 expected, from the diminished ability of H373R mutant to bind CTCF consensus DNA, regions lacking
311 the canonical CTCF motif showed higher enrichment of G4 peaks (**Sup.Fig.4E,4F**). Within common
312 CTCF peaks (bound by FL and H373R CTCF) that overlapped G4 peaks, we calculated the distance
313 between the center of CTCF and G4 peaks (**Fig.5A**). Notably, a small number of regions with overlapping
314 CTCF and G4 peaks that harbored canonical CTCF motifs (n=1556), exhibited significantly higher
315 median distance between the centers of CTCF and G4 peaks, compared to regions which lacked a
316 canonical CTCF motif (n=5467) (**Fig.5A**). Though these results infer that G4 structures and CTCF binding
317 tightly coincide at sites lacking a defined canonical CTCF motif, the enrichment-based approaches such
318 as CUT-and-Tag are not ideal for interrogating the CTCF-G4 interaction at base-resolution. To this end,
319 we classified overlapping CTCF and G4 peaks into those that harbor only a single $G_3N_{1-7}G_3N_{1-7}G_3N_{1-7}G_3$
320 (3G-G4) motif in the with or without a canonical CTCF motif. Interestingly, in regions (n=279) which
321 carried single CTCF and 3G-G4 motifs, the average distance between the two motifs was >300 bp (**Fig.**
322 **5B**). Moreover, the average distance of the 3G-G4 motif to the nearest CTCF motif was ~5kb at
323 overlapping CTCF and G4 peaks that otherwise lacked a canonical CTCF motif (**Fig. 5C**). Together, these
324 analyses highlight two important points: 1) at several places in the genome CTCF binding directly
325 coincides with G4 sequences, and 2) even at CTCF binding sites where G4 and CTCF motifs co-exist, on
326 average these motifs are separated by a few hundred base pairs.

327 To assess this more carefully, we examined the footprints of FL and H373R CTCF binding at overlapping
328 CTCF (common regions) and G4 peaks. For these analyses we plotted the signal for CTCF binding relative
329 to the center of the 3G-G4 motif in regions categorized based on the presence or absence of a canonical
330 CTCF motif. Remarkably, the binding footprint of CTCF precisely matched the center of the G4 motif,
331 specially at sites lacking a canonical CTCF motif (**Fig.5D-E**). Furthermore, at regions that harbored
332 canonical CTCF motifs, two distinct CTCF footprints were evident (**Fig.5D-E**). One of these CTCF
333 footprints directly overlaid the center of G4 motifs and the other was located ~215 bp (mean distance)
334 downstream to the G4 motif (**Fig.5D-E**). The CTCF footprint at the center of G4 motifs was distinct in its
335 appearance (narrower) when compared to the downstream CTCF footprint, which we presume to be
336 mediated by the canonical CTCF motif (**Fig.5D-E**). Both FL CTCF and H373R mutant exhibited these
337 unique binding features (**Fig.5D-E**). However, compared to FL CTCF, the signal for H373R binding was
338 visibly reduced in regions harboring the canonical CTCF motif and enhanced at regions that lacked the
339 canonical CTCF motif (**Fig.5D-E**). From these results we propose that CTCF directly binds G4 structures
340 at least at a subset of genomic sites.

341 **CTCF-G4 interaction mediates chromatin looping and is associated with changes in gene expression**

342 To elucidate the role of CTCF-G4 interaction in regulation of gene expression, we performed RNA-seq
343 in cells expressing empty vector (EV) control with or without auxin and mESCs reconstituted with FL or
344 H373R mutant 24 hours post auxin mediated CTCF depletion. Principal component analysis showed
345 clustering of all replicates, with PC1 accounting for 73% of the variance and segregating the samples
346 based on auxin treatment (**Fig.6A**). Notably, PC2 further stratified samples based on CTCF expression,
347 highlighting the distinct transcriptional responses upon CTCF reconstitution. Differential gene expression
348 analysis in EV expressing samples treated with or without auxin revealed 978 differentially expressed

349 genes (DEGs; false discovery rate ≤0.05, log2 fold change > ±0.58) with 520 up- and 458 down-regulated
350 genes (**Sup.Fig.5A**). The comparison between FL and EV expressing cells treated with auxin identified
351 799 DEGs (361 up and 438 down), whereas H373R versus EV expressing cells showed 431 DEGs (189
352 up and 242 down) upon auxin treatment (**Fig.6B**). Of these DEGs, 233 DEGs (97 up and 136 down) were
353 shared in pairwise comparisons of FL and H373R to EV controls (**Fig.6C-E**). We reasoned that given the
354 common binding activity of FL and H373R CTCF towards G4s, these 233 DEGs might be enriched for
355 G4 signal. Indeed, the 233 DEGs exhibited enrichment of CTCF, BG4, and NMM-Bio signals at their
356 transcription start sites (TSSs) (**Fig.6F**). 38% of these genes (89/233) also harbored statistically defined
357 CTCF and G4 peaks in their vicinity (± 10kb of the gene), suggesting that these are likely G4-regulated
358 genes. Gene ontology analysis of the shared DEGs between FL and H373R showed that upregulated genes
359 were enriched for pathways related to embryonic development, cell fate commitment and vitamin
360 metabolism, whereas downregulated genes were associated with developmental programs linked to
361 organogenesis (**Fig.6G**). These data suggest that both FL and H373R CTCF are essential for maintaining
362 commitment to an embryonic stem cell program and repressing sporadic activation of organ/lineage
363 specific developmental pathways.

364 We also performed clustering of transcriptomics data with time-course sequencing (TC-seq) analysis to
365 study the relationship between CTCF binding, G4 formation and dynamic changes in gene expression
366 patterns in different experimental conditions. TC-seq generated 9 unique clusters, C1 through C9
367 (**Sup.Fig.5B-C**). We narrowed these clusters based on expression changes in response to CTCF depletion
368 with auxin and further sorted the clusters considering similar effects of FL and H373R reconstitution on
369 restoration of gene expression patterns. Among the interesting clusters, cluster C1 was upregulated by
370 CTCF, whereas cluster C2 was seemingly repressed by CTCF (**Sup.Fig.5B-C**). Moreover, clusters C1
371 and C2 were enriched in pathways related to immune response and neurogenesis, respectively. Notably,
372 in clusters C1 and C2, reconstitution with FL CTCF restored expression profiles resembling the untreated
373 (UT) EV controls (**Sup.Fig.5B-C**). Reconstitution with H373R mutant also restored the gene expression
374 patterns, albeit the effects were less prominent, which could be due to the lower expression levels of
375 H373R compared to FL (**Sup.Fig.5B-C; Fig.3B**). Nonetheless, the CTCF, BG4, and NMM-Bio signals
376 within ±200bp of the TSS of genes in clusters C1 and C2 also showed an observable enrichment when
377 compared to the non-auxin responsive cluster C4 as control (**Sup.Fig.5D**).

378 Finally, to link the changes in gene expression to CTCF-mediated long-range chromatin looping, we
379 analyzed a previously published high throughput chromosome conformation capture (Hi-C) dataset which
380 was generated in the AID-tagged CTCF mESCs, the same parent line used in our studies. We generated
381 chromatin contact maps reporting significant interactions with (auxin) or without (untreated, UT) auxin
382 treatment for 48 hours, as well as 48 hours post-auxin washout (washout). Using FitHiC, an algorithm for
383 identifying significant chromatin interactions, we identified a total of 318,770 chromatin interactions in
384 untreated cells at 20kb resolution, and upon auxin treatment a large proportion of these interactions were
385 lost (n=66,073 remaining post-auxin) (**Sup.Fig.5E**). Importantly, the majority of interactions were
386 recovered post-auxin washout (n=328,783) (**Sup.Fig.5E**). Using these interactions, we first asked if the
387 presence of G4s affects the strength of chromatin contacts (**Fig.6H**). Indeed, consistent with previous

388 reports[19,20], interactions that overlapped with a G4 peak showed a significantly lower adjusted p-value,
389 when compared to randomly surveyed interactions without G4s (**Fig.6H**). These data suggest that
390 chromatin loops that are associated with a G4 may be functionally distinct from the ones that do not carry
391 a G4. To assess this further, we compared the G4 distribution in all reported Hi-C interactions in untreated
392 mESCs and found that ~35% interactions were associated with G4 peaks (**Fig. 6I, (-)auxin**). Remarkably,
393 the G4 peaks were even more enriched (~66%) at chromatin loops that persisted upon CTCF depletion
394 with auxin treatment (**Fig.6I, (+)auxin**), we refer to these as "persistent loops". To rule out the possibility
395 that CTCF binding is retained at these sites (persistent loops) even upon auxin treatment, we plotted the
396 CTCF signal in (CTCF) peaks that overlapped Hi-C interaction windows in untreated conditions or those
397 that persisted upon auxin treatment in our datasets. As expected, the CTCF signal was similarly depleted
398 upon auxin treatment (EV AUX) in CTCF peaks associated with chromatin loops from untreated or auxin
399 conditions (**Sup.Fig.5F;** grey lines). We noted several examples of these "persistent loops" in the genome,
400 including a region around the *Pde6d* and *Cldn3* gene loci, where CTCF binding overlapped G4 containing
401 regions and while some chromatin contact loop anchors were lost upon auxin treatment, others remained
402 intact (**Fig.6J, Sup.Fig.6A**). Importantly, many of these sites were associated with gene expression
403 changes that were linked to FL and H373R CTCF binding. 197/233 DEGs co-regulated by CTCF FL and
404 H373R mutant were associated with Hi-C interactions with CTCF and G4 peaks, and intriguingly, 107/233
405 DEGs were associated with persistent loops, displaying a highly significant correlation (<0.00001, Chi-
406 square test) (**Sup.Fig.5G**). In contrast to the loops that persisted upon auxin treatment, several other
407 chromatin contacts were totally obliterated upon auxin treatment at certain genomic loci, such as the
408 *Cox15* gene locus, that also showed association with CTCF binding and G4s at a distant regulatory region
409 (**Sup.Fig.6B**). In conclusion, our findings suggest that G4s, in conjunction with CTCF binding, play a
410 critical role in regulation of gene expression possibly by influencing long-range chromatin interactions.

411 **Discussion**

412 Our study expands the understanding of DNA G4 structures as critical regulators of genomic function,
413 with particular focus on their direct interaction with the genomic architectural protein, CTCF. By
414 integrating proteomic screens and biochemical assays, we provide strong evidence for the specific
415 recognition of G4s by the central ZF domains of CTCF. Through genome-wide analyses, we suggest that
416 CTCF recruitment may be directly mediated by G4s at 15-20% of genome-wide CTCF binding sites; this
417 is in addition to the ~80 to 85% of CTCF binding sites that contain consensus CTCF motifs[38,39] With these
418 findings, we propose a model in which CTCF-G4 interaction cooperates with directional CTCF binding
419 to its consensus motif for establishing chromatin loops and TAD boundaries to regulate gene expression
420 and possibly other genomic processes (**Fig.6K**). This hypothesis is supported by our finding that the
421 anchor points of chromatin loops associated with G4s appear to be highly persistent and are less prone to
422 CTCF depletion. This suggests that G4 structures and the associated DNA conformation by themselves
423 may function as efficient loop anchors, even when CTCF is depleted. Alternatively, the G4 structures may
424 recruit other proteins that may continue to serve a barrier function, even in the absence of CTCF. Taken
425 together, our studies highlight the architectural function of G4 structures in the genome and provide
426 fascinating insights into the paradigm of 3D genome organization.

427 The N-terminal region of CTCF is shown to specifically interact with the STAG2 (SA2) subunit of cohesin
428 ring complex to impede loop extrusion[55,56]. It is important to note, that the cohesin ring complex is
429 estimated to have a diameter of ~40 nm[57], whereas the approximate size of CTCF is 3-5 nm. Therefore,
430 from a physical perspective, how CTCF even when bound to its DNA motif in the right orientation, can
431 serve as an effective roadblock for cohesin loop extrusion which occurs at an estimated 1kb/second[37], is
432 unclear. In this regard, G4 structures and the looped out DNA strand with associated DNA:RNA hybrids
433 could function as larger physical impediments for movement of cohesin rings. Moreover, a recent report
434 has accentuated the importance of DNA tension as a key parameter for the barrier activity of CTCF[37]. By
435 partially unwinding DNA strands and creating torsional stress, G4s can modulate DNA tension and may
436 contribute to the barrier activity of CTCF by influencing this parameter[58]. It is also important to point out
437 that besides CTCF, another genomic architectural protein, Yin-Yang1 (YY1) has been shown to bind G4
438 structures[21]. When considered together, these studies suggest that G4s have perhaps co-evolved for for
439 interactions with DNA-binding proteins to regulate long-range DNA looping.

440 The interaction of CTCF with G4 structures has been controversial. Our work indicates different assay
441 conditions as the possible source of discrepancies reported in previous studies[20,22,24]. We show that similar
442 to other zinc finger containing proteins[46], the binding of CTCF to G4s is sensitive to EDTA. We therefore
443 reason that the inability to detect CTCF-G4 interaction in previous studies could be attributed to the
444 presence of EDTA in binding and assay buffers[20,22]. Studies by Wulfridge et.al. suggested that G4
445 structures facilitate stronger binding of CTCF to weaker genomic motifs[20]. While this could be relevant
446 for CTCF binding at certain genomic sites, this model requires precise positioning of G4s relative to the
447 CTCF DNA motifs. Further, CTCF binding is disrupted by DNA methylation within the consensus DNA
448 motifs[32,59], and G4 structures have been linked to inhibition of DNA methylation through direct binding
449 and inhibition of DNMT1[14]. Therefore, as an alternative, G4 structures may regulate CTCF binding by
450 promoting hypomethylation at consensus DNA motifs. Moreover, the structural basis by which a proximal
451 G4 could influence CTCF binding to its consensus motif remains to be elucidated. Our studies on the other
452 hand provide evidence that CTCF can directly bind G4 structures, corroborating findings from a previous
453 study[24]. We demonstrate that similar to CTCF binding to its consensus DNA motif, G4 recognition by
454 CTCF involves ZFs 4 to 7 (117 amino acids). However, in contrast, a single point mutation (H373R) in
455 the ZF4 of CTCF abolishes binding to the consensus motif while binding to G4s remains unaffected.
456 Moreover, we estimate that the affinity of CTCF for G4s is ~2-fold lower than its affinity for the consensus
457 motif. Even though our studies imply that the C2H2 motif in ZF4 domain is not required for G4
458 recognition, the precise structural basis for CTCF-G4 binding is unresolved, and we plan to follow up on
459 this work in future studies.

460 In another useful contribution, we have developed a novel G4 mapping method that utilizes a highly
461 selective small molecule G4 ligand, NMM[53]. This adds to a growing catalog of G4-mapping methods that
462 each offer unique advantages but may have specific limitations. Notably, G4 mapping using the NMM-
463 Bio method appears to define a set of G4 peaks distinct from those defined by the BG4 antibody. We
464 adapted the NMM-Bio method in the CUT-and-Tag workflow to use with limiting cell numbers. The

465 NMM-Bio G4 mapping strategy can be used to complement existing BG4-based G4 mapping to generate
466 more comprehensive G4 maps for exploring G4 biology.

467 In addition to the CTCF-G4 interaction, our screen revealed a diverse array of both previously known and
468 novel G4-interacting proteins. Key features of our proteomics screen, distinct from previous approaches,
469 are the use of a mixture of DNA G4s of different topologies and the capture of proteins under native
470 intracellular salt concentrations. These conditions allowed us to detect proteins (such as CTCF) that were
471 not identified by previous G4 proteomics screens. Our work implicates G4 structures in regulation of
472 several genomic processes including RNA splicing, mRNA cleavage, nucleosomal remodeling, histone
473 modifications among others. While many of the hits from our screen were components of multi-subunit
474 protein complexes, the mechanisms by which they interact with G4s remain unknown.

475 An intriguing category of proteins highlighted in our screen were those associated with membrane-less
476 nuclear bodies with biomolecular condensate-like features such as Cajal bodies and paraspeckles. The
477 enrichment of condensate-associated proteins by G4s raises fascinating questions about the potential roles
478 of G4s in formation of biomolecular condensates, and this is consistent with recent work implicating RNA
479 G4s in liquid-liquid phase separation[60-62]. More broadly, from an evolutionary viewpoint, the number of
480 predicted G4 structures appear to have expanded during hominid evolution when compared across 37
481 different species[9]. This indicates that G4 structures may have been positively selected during human
482 evolution and co-opted in regulation of diverse genomic processes.

483 **Acknowledgements**. We thank Dr. Elphege Nora for sharing the mESCs carrying the AID-tagged CTCF.
484 We thank Drs. Xinkun Wang and Matthew Schipma and other members of the Northwestern University
485 Sequencing Core for providing next generation sequencing services, the Northwestern University
486 RHLCCC Flow Cytometry Facility team: Dr. Swaminathan S., Ostiguin C., Dias E., and others for their
487 help with cell sorting. Some figure schematics were created using Biorender. This work was supported by
488 the SPARK pilot award, R00 award from the National Cancer Institute (grant ID: CA248835) and
489 institutional startup funds from Northwestern University and the Lurie Cancer center to V.S.

490 **Author Contributions Statement**. DSC, IH, RCM and VS acquired, analyzed and interpreted the data.
491 DSC and VS performed the bioinformatics analysis. SC and BW performed characterization of ESCs. IR
492 helped with designing biochemical assays for CTCF-G4 interaction. AC and FA helped with Hi-C data
493 analysis. SAM performed the MASS-Spec for G4 proteomics screen. AR supervised the initial studies for
494 G4 proteomics screen and biochemical characterization of CTCF-G4 interaction. VS supervised the
495 studies, conceptualized the experiments and helped with data interpretation. DSC, IH, RCM and VS wrote,
496 and all members of the Shukla lab reviewed the manuscript. All authors were involved in reviewing and
497 editing the manuscript.

498 **Competing Interests Statement**. The authors declare no competing financial interest in this work.

499 **Data Availability.** All genome-wide sequencing datasets have been deposited to Gene Expression
500 Omnibus (GEO) repository, accession number GSE288430 and GSE288516. Any data and reagents will
501 be made available upon request.

502 **Code Availability**. The code used to process the next-generation sequencing datasets are deposited to
503 github (https://github.com/dsamanie7/G4_CTCF).

504

505 **Materials and Methods**

506 **Cell culture**

507 Murine B cells from C57Bl/6, mice were isolated using EasySep Mouse naive B cell isolation kit (#19484
508 Stem Cell Technology, Canada) from splenocytes. Primary B cells and CH12F3 (CH12) cells were
509 cultured at 37°C, 5% $CO_2$ in RPMI 1640 media supplemented with 10% FBS, 1x MEM non-essential
510 amino acids, 10 mM HEPES (pH 7.4), 2 mM Glutamax, 1 mM sodium pyruvate, 55 µM 2-
511 mercaptoethanol (all from Life technologies). B cells ($5x10^5$-$1x10^6$ cells/mL) were activated with 25
512 µg/mL LPS from *E. coli* O55:B5 (Sigma, St. Louis, MO) and 10 ng/mL rmIL-4. For pyridostatin (PDS)
513 (#SML2690, Sigma-Aldrich) and RHPS4 (Tocris) treatment, CH12 cells were incubated in drugs for 24
514 hours before analysis. All the cytokines used above were from Peprotech (Rocky Hill, NJ). The CH12F3
515 cell line was developed in Dr. T. Honjo's laboratory at Kyoto University and was independently validated
516 using cell stimulation and Ig class-switching experiments.

517 E14 and EN52.9.1 (CTCF-AID) ESCs were obtained from Dr. Elphège Nora (University of California,
518 San Francisco). mESCs were cultured at 37°C with 5% $CO_2$ in DMEM supplemented with 15% ESC-
519 grade fetal bovine serum (ES cell grade FBS from ThermoFisher), 2 mM GlutaMAX, 1X MEM non-
520 essential amino acids, 1 mM sodium pyruvate, 55 µM β-mercaptoethanol (all from Life Technologies),
521 and 1000 U/mL ESGRO murine leukemia inhibitory factor (mLIF Cat# ESG1106, Millipore Sigma). For
522 auxin-induced degradation of CTCF, media was supplemented with 500 µM auxin analog indole-3-acetic
523 acid sodium salt (IAA, Cat# I5148-2G, Millipore Sigma). Auxin washout was performed by washing cells
524 twice with PBS or DMEM.

525 **Detection of G-quadruplex binding proteins**

526 B cells activated in presence of 25 µg/ml LPS and 10 ng/ml IL-4 or CH12F3 cells were used to collect
527 nuclear extracts with Dignam extraction. Cells were resuspended in 10 mL Dignam buffer A (10 mM KCl,
528 1.5 mM $MgCl_2$) supplemented with protease inhibitor cocktail (Thermofisher) and incubated on ice for at
529 least 10 min. The cells were then homogenized and lysed with a dounce homogenizer (10 strokes) and
530 nuclei were collected by centrifugation at 600 g for 5 min and washed again with buffer A. Cells were
531 then resuspended in 1ml of Dignam buffer C (0.2 mM EDTA, 25% glycerol (v/v), 20 mM HEPES-KOH
532 (pH 7.9), 0.42 M NaCl, 1.5 mM $MgCl_2$) supplemented with protease inhibitor cocktail and rotated at 4°C
533 for 1 hour. The lysates were collected and passed through a PD-10 buffer exchange column (GE
534 Healthcare) equilibrated with Intracellular Buffer (25 mM HEPES pH 7.5, 10.5 mM NaCl, 110 mM KCl,
535 130 nM $CaCl_2$, 1 mM $MgCl_2$). The G4 and non-G4 oligonucleotides (**Fig.1A**) were folded in the presence
536 of 10mM HEPES (pH 7.4) and 150mM KCl, annealed with a complementary biotinylated adaptor
537 denatured at 95°C for 5 minutes with temperature ramp down to 25°C at 0.1°C/second. The annealed
538 oligos were then captured on streptavidin conjugated magnetic beads and incubated with nuclear lysates
539 to identify G4-binding proteins. For initial studies, the nuclear lysates collected in Intracellular Buffer
540 were spike-in with 1 µg FLAG-tagged BG4 scFv antibody (Cat# MABE917, Millipore Sigma), pre-
541 cleared with MyOne T1 streptavidin-conjugated DynaBeads (Cat# 65-601, Invitrogen) for 1 hour and
542 incubated with streptavidin beads conjugated with a mixture of G4 or non-G4 forming oligonucleotides
543 for 4 hours. Samples were then washed with Intracellular Buffer with 0.05% Tween-20 for a total of 5

544 washes and denatured using the LDS-PAGE sample buffer (Cat# NP0008, Invitrogen) for
545 immunoblotting. The oligo sequences are included in **supplementary table 1** and immunoblotting
546 antibodies in **supplementary table 2**.

547 **MASS-Spectrometry for Proteomics screen**

548 For the proteomics screen, G4 and non-G4 control pull-down fractions were loaded on denaturing gel and
549 stained with silver stain (Silver stain kit from Pierce). Streptavidin beads from the pulldown were
550 resuspended in 50 uL of 50 mM Tris pH 8, 150 mM NaCl, frozen and stored at -80C until processing. For
551 processing, dry urea was added to a final concentration of 8 M and incubated at room temperature for 15
552 min. Urea was diluted to 4 M with the same buffer and 20 units of LysC (Wako) was added and incubated
553 at 37°C for 60 min. The supernatant was transferred to a new tube at 37°C and the beads were digested
554 further with 250 ng trypsin (Promega) for 60 min. The trypsin digest was combined with the LysC digest,
555 urea was diluted to < 2 M and allowed to digest overnight. Peptide digests were desalted, TMT (Thermo)
556 labeled "on-column", mixed, and fractionated into six basic reversed phase fractions before being
557 concatenated back into three before analysis by liquid chromatography tandem mass spectrometry (ref:
558 PMC6495249). Liquid chromatography was performed using a Proxeon EASY-nLC 1000 at a flow rate
559 of 200 nl/min. Peptides were separated at 50°C using a 75 micron i.d. PicoFit (New Objective) column
560 packed in house with 1.9 um AQ-C18 material (Dr. Maisch) to 20 cm in length over an 84 min effective
561 gradient. Mass spectrometry was performed on a Thermo Scientific Q Exactive Plus. After a precursor
562 scan from 300 to 2,000 m/z at 70,000 resolution the top 12 most intense multiply charged precursors were
563 selected for HCD at a resolution of 35,000. Data were searched with Spectrum Mill (Agilent) using the
564 Uniprot Mouse database with 261 common laboratory contaminants. A fixed modification of
565 carbamidomethylation of cysteines and variable modifications of N-terminal protein acetylation,
566 oxidation of methionine, and TMT 11-plex labels were searched. The enzyme specificity was set to LysC/
567 trypsin and a maximum of three missed cleavages was used for searching. The maximum precursor-ion
568 charge state was set to 6. The precursor mass tolerance and MS/MS tolerance were set to 20 ppm. The
569 peptide and protein false discovery rates were set to 1%.

570 For statistical enrichment analysis, we performed a moderated T-test followed by Benjamini-Hochberg
571 correction for false discovery rate estimation. The enriched samples were the numerator, and the non-G4
572 DNA sequences were used at the denominator. The positive hits from MASS-Spec are included in
573 **supplementary table 3 and 4.**

574 **Immunoblotting**

575 Protein lysates were resolved using a NuPAGE 4-12% Bis-Tris gel (Cat# NP0321BOX, Invitrogen) and
576 transferred to PVDF membrane (Cat# IB23001, Invitrogen) using the iBlot 2 Semi-Dry Gel Transfer
577 system (Cat# IB21001, Invitrogen) at 25 V for 7 min. Membrane was blocked with 5% nonfat dry milk in
578 TBS-T buffer (1 M Tris-HCl pH 7.4, 5 M NaCl, 2.5 M KCl, 0.5 M EDTA, 0.05% Tween-20) and
579 incubated with an appropriate primary antibody (**supplementary table 2**), followed by incubation with
580 horseradish peroxidase (HRP)-conjugated secondary antibody, diluted in 5% milk solution. Each antibody
581 incubation was followed by three 10-min washes with TBSTE (50mM Tris-HCl pH 7.4, 150mM NaCl,
582 0.05% Tween-20, 1mM EDTA) buffer at room temperature. Signal was detected with enhanced
583 chemiluminescence reagents (Invitrogen) with either X-ray film or imaging with Chemidoc Luminescence
584 system (BioRad). The antibodies for immunoblotting are included in **supplementary table 2**.

585

586 **Site-directed mutagenesis of CTCF**

587 Site-directed mutagenesis to introduce H373R mutation in CTCF was performed following the Q5 Site-
588 Directed Mutagenesis Kit from New England BioLabs (NEB Cat# E0554S) with modifications. Forward
589 and reverse primers designed for the mutation were synthesized by Integrated DNA Technologies (IDT).
590 Template plasmid (10 µg) was mixed with forward and reverse primers (each 10 µM) and Q5 Hot Start
591 High-Fidelity 2X Master Mix (NEB Cat# M0494S) for amplification. Following the PCR amplification,
592 the product was subjected to a Kinase, Ligase, and Dpnl (KLD) reaction at room temperature for 5
593 minutes. Subsequently, the KLD reaction product was transformed into chemically competent *E. coli* cells
594 (NEB Cat# C2987H). Individual colonies were picked and cultured overnight in LB medium with
595 kanamycin at 37 °C, and plasmid DNA was extracted using PureLink Quick Plasmid Miniprep Kit
596 (Invitrogen Cat# K210010). The final plasmid DNA was sequenced to confirm the successful introduction
597 of mutation. Sequencing was performed by ACGT DNA Sequencing Services.

598 ***In vitro* translated CTCF**

599 *In vitro* transcription and translation of CTCF were performed using the TNT Coupled Reticulocyte Lysate
600 System (Promega Cat# L4610) according to the manufacturer's instructions. Briefly, 1 µg of plasmid
601 DNA template with CTCF open reading frame downstream to the T7 promoter was combined with TNT
602 reaction buffer (Promega Cat# L462A), TNT T7 polymerase (Promega Cat# L481A), amino acid mixture
603 (Promega Cat# L995A-L996A), TNT rabbit reticulocyte lysate (Promega Cat# L483A), and recombinant
604 RNase inhibitor (Takara Bio Cat# 2313A). The reaction mixture was incubated at 30°C for 90 minutes on
605 a thermocycler. A small portion of the final protein product (~2 µL) was kept for immunoblotting as input,
606 and the remaining product was directly incubated with annealed biotinylated G4 oligos to assess G4-CTCF
607 binding (please see below, CTCF-G4 binding assay).

608 **Recombinant CTCF purification**

609 Coding sequences of N-terminal Flag-tagged human CTCF fragment ZF4-7 WT and H373R mutant
610 (amino acid 350-466) were sub-cloned in frame with Glutathione-S-Transferase (GST) protein separated
611 by a PreScission protease cleavage site in the pGEX6p1 vector. BL21 (DE3) cells were transformed with
612 the pGEX6p1 vector with GST-CTCF ZF4-7 fusion. Single colonies were picked and grown overnight in
613 5 mL LB at 37°C before further expanding in 2 to 3 litres of LB (1:100) with shaking at 37°C for 2-3
614 hours until the OD at 600 nm reached 0.5-0.6 (log phase). $ZnCl_2$ was added to a final concentration of 25
615 µM, expression was induced by the addition of isopropyl β-D-1-thiogalactopyranoside (IPTG) to 0.5 mM,
616 and the cultures were incubated overnight at 16°C. Cells were harvested by centrifugation, resuspended
617 in lysis buffer containing 20 mM Tris-HCl (pH 7.5), 500 mM NaCl, 5% (v/v) glycerol, 0.5 mM Tris (2-
618 carboxyethyl)-phosphine hydrochloride (TCEP), and 25 µM $ZnCl_2$. The cell suspension was then lysed
619 by sonication for a total of 4 min (5 seconds on, 30 seconds off cycle) with a probe sonicator (Missonix)
620 followed by centrifuge at 10,000 g for 20 min, at 4°C. Lysates were mixed with polyethylenimine (Sigma)
621 at pH 7 to a final concentration of 0.3%–0.4% (w/v) before centrifugation at 16,500 rpm. Expression of
622 protein was confirmed by Coomassie staining on SDS-PAGE gel by comparing with an uninduced control
623 culture. The supernatant was then mixed with 2ml washed and equilibrated glutathione magnetic beads
624 (Cat# 78602, Pierce) and rotated for 4 hours overnight. The lysate beads mixture was then washed three
625 times with lysis buffer and then re-suspended in 2 mL elution buffer (100 mM Tris-HCl (pH 7.6), 10%
626 Glycerol, 25 µM $ZnCl_2$, 100 mM KCl, 1 mM DTT) and treated overnight with PreScission protease (Cat#
627 Z02799, Genescript). The supernatant was then collected, and the proteins were purified by gel filtration
628 on a Superdex 200 column (GE Healthcare). The protein concentrations were quantified by absorbance at

280 nm using Nanodrop. The purity of protein was additionally confirmed by Coomassie staining of samples on an SDS-PAGE gel.

## CTCF binding to G4 and consensus motif

To measure CTCF-G4 binding, *in vitro*-translated or recombinant CTCF protein were incubated with either a mixture of three different biotinylated G4 (*Kit2, Kit1 and Spb1*) and non-G4 oligos, all of which showed similar binding to CTCF (**Fig. 1G**). For quantitative binding assays, the *Kit2* G4 or non-G4 control oligos were used. For CTCF binding to consensus DNA motif, the double stranded CTCF consensus or mutated consensus oligos were annealed to the same biotinylated adaptor also used for G4 binding assays. The CTCF incubation with different oligos was performed in Intracellular Buffer (25 mM HEPES pH 7.5, 10.5 mM NaCl, 110 mM KCl, 130 nM $CaCl_2$, 1 mM $MgCl_2$) containing 0.125% BSA for 1 hour at 4°C with rotation. MyOne T1 streptavidin-conjugated DynaBeads (Cat# 65-601, Invitrogen) beads were pre-washed five times with Intracellular Buffer containing 0.125% BSA and were used to capture biotinylated oligos at 4°C for 1 hour with rotation. After incubation, samples were washed four times with Intracellular Buffer containing 0.05% Tween-20. To reveal the signal in some experiments, the binding was assessed by immunoblotting. In other cases, for quantitative binding measurements, signal was measured by on-bead fluorescence with flow cytometry (Cytek). For immunoblotting based measurements, the western blotting under denaturing conditions were performed as described above (**Antibodies, supplementary table 2**). For competition assays in **Fig. 2D,** the non-biotinylated double stranded CTCF consensus or mutated consensus oligos were mixed at different molar concentration with G4 or non-G4 oligos.

For quantitative binding assays, streptavidin-conjugated magnetic beads (Invitrogen, 65-601) with captured oligos were incubated with anti-Flag primary antibody (Cat# 14793S, 1:100 dilution, Cell Signaling) for 30 min at 4°C with rotation, followed by two washes and a 30-min incubation with a PE-conjugated anti-rabbit secondary antibody (Cat# 406421, 1:400 dilution, Biolegend) at 4°C with rotation. Beads were washed, resuspended in intracellular Buffer, and on-bead fluorescence was measured using flow cytometry (Cytek). For the detection of binding between CTCF and consensus motifs, the G4 or non-G4 oligos were replaced with CTCF oligos (**supplementary table 2**). For comparison with the conditions used in the study by Wulfridge et al.[20], Intracellular Buffer in this assay was replaced with Binding Buffer (50 mM TRIS pH 8.0, 100 mM NaCl, 1 mM DTT, 0.1 mM EDTA, 5% glycerol, 0.05% NP-40). Additional assays were performed under identical conditions with EDTA removed to assess its impact on CTCF-G4 interaction.

## NMM-Bio conjugation and purification

The NMM-Biotin conjugation was performed by a commercial vendor WuXi AppTec. 5mg NMM (8.61 μmol) was dissolved in anhydrous Dichloromethane, DCM (2 mL). Under nitrogen the NMM solution was added to 2 molar equivalents of Propylphosphonic anhydride (T3P, 13.21 mg, 20.75 μmol), 5 molar equivalents of N,N-Diisopropylethylamine (DIEA, 6.71 mg, 51.88 μmol) and Amine-PEG11-Biotin (8 mg, 10.38 μmol) at 20°C. The reaction was stirred at 20°C for 12 hours. After that, thin layer chromatography (TLC) with DCM/Methanol = 9/1, stained by visible light was performed and showed complete consumption of NMM (retention factor = 0.6) in the reaction. TLC (DCM/Methanol = 4/1, stained by visible light) also showed complete consumption of Amine-PEG11-Biotin (Retention factor= 0.3) in the reaction. Seven similar scale reactions were performed as described above and combined to generate a large batch of NMM-Bio. The reactions were concentrated under reduced pressure at 20 °C to remove extra DCM. After full wavelength scanning, the NMM-Bio complex showed maximum absorption at 395 nm, as also described earlier[53]. Following this, NMM-Bio was purified by a silica column under

395 nm wavelength with medium pressure liquid chromatography (MPLC) using a Welch Ultimate XB NH2 column with Heptane and Ethanol. The NMM-Bio conjugation was confirmed by 1H-NMR and high-resolution mass spectrometry (HRMS-TOF). Finally, the NMM-Bio complex was concentrated and lyophilized (21.5 mg, 10.3 μmol, 14.2% yield) as deep red color solid.

**CUT-and-Tag library preparation, sequencing, and analysis**

The CUT-and-Tag protocol from EpiCypher was adapted with modifications. Approximately 300,000 ESCs were collected from 24 hour auxin treatment. The isolated nuclei from mESCs were mixed with activated Concanavalin A (ConA)-coated magnetic beads (Cat# 21-1401, Epicypher) and incubated for 10 minutes at room temperature. Upon nuclei extraction and binding nuclei to activated beads, incubation steps were completed in a modified Digitonin 150 buffer (20 mM HEPES pH 7.5, 150 mM KCl, 0.5 mM spermidine, 1x protease/phosphatase inhibitor) until the tagmentation step. This modification, in which NaCl was replaced by similar concentration of KCl, was to optimize stabilization of G4 structures during prolonged incubation steps, as shown previously[13]. Other buffers were prepared and used according to manufacturer's protocol. The nuclei-conjugated ConA bead complexes were incubated overnight at 4 °C with primary antibody (CTCF, BG4, NMM-Bio, or control NMM; **supplementary table 2**) in the modified Digitonin 150 buffer. After primary incubation, nuclei-bead complexes were incubated with secondary anti-biotin antibody (Cat# ab53494, Abcam) in the modified Digitonin 150 buffer for 1 hour at 25 °C, followed by tagmentation using CUTANA pAG-Tn5 (Cat# 15-1117, Epicypher). Targeted chromatin tagmented products were column purified using a DNA Clean & Concentrator Kit (Cat# D4013, Zymo Research). DNA libraries were amplified with the following thermocycling conditions: initial denaturation at 98°C for 30 seconds, 18 cycles of denaturation at 98 °C for 10 seconds, annealing at 63 °C for 30 seconds, and extension at 72 °C for 1 minute, followed by a final extension at 72 °C for 1 minutes. After PCR amplification, the libraries were column purified once again to be quantified and qualified before pooling for sequencing. Purified libraries were sequenced on Illumina NovaSeq 6000.

Paired-end reads (50bp) were mapped to the mouse genome mm10 GRCm38 (Dec. 2011) from UCSC and E.coli K-12 genome (NZ_CP132594.1) with Bowtie ("--p 8)[63]. Reads mapping to the *E.coli* genome were filtered out from sam files using 'grep' command in bash. The remaining mm10 mapped reads were marked for duplication and kept with picard-tools-2.21.4 MarkDuplicates (https://broadinstitute.github.io/picard/). Peaks were called using MACS2 (2.1.4) ("callpeak --keep-dup all" using IgG samples as input)[64]. The intersected peaks from the same condition were obtained using HOMER[65] mergePeaks and filtered with grep. Intersection between the different methods were performed with bedtools[66] "intersect" (bedtools/2.29.2) or HOMER mergePeaks. Deeptools[67] (3.5.5) was used to plot the signal at the TSS and center of the peaks.

**Lentiviral production and transduction of mESCs**

HEK293T cells were cultured to 80% confluence and transfected using the Lipofectamine® 3000 Transfection Kit (Invitrogen, Cat#L3000001), following manufacturer's protocol, with 2.5μg DNA (1.25μg lentiviral plasmid; 3-parts psPAX2, 1-part pMD2.G – and 1.25μg EV, FL or H373R plasmids). Media containing lentivirus secreted by HEK293T cells was collected after 48 hours and transfection was assessed by flow cytometry for mCherry. Media was aspirated from ESCs cultured in a 6-well plate and replaced with a 1:1 mix of lentiviral media and DMEM ESC media (0.5 multiplicity of infection) with polybrene (10ng/ml) and centrifuged at 2000 rpm for 90 minutes at 37°C to facilitate transduction. Following transduction, cells were rested for 4 hours before aspirating viral media and replacing with fresh DMEM. Transduced cells were sorted by fluorescence-associated cell sorting based on expression

717 of mCherry reporter. Sorted cells were expanded in culture for one week in DMEM ESC media and
718 assessed again for mCherry expression prior to conducting experiments.

719 **RNA-seq library preparation, sequencing, and analysis**

720 RNA was purified from cell lines treated with or without auxin using the RNeasy Plus Mini Kit (Qiagen)
721 following manufacturer's instructions. Purified RNA (200 ng) was subsequently utilized to prepare RNA
722 libraries with the NEBNext Ultra II Directional RNA Library Prep Kit (New England Biolabs, Cat#
723 E7760), following manufacturer's recommended protocol. Samples were sequenced on the Illumina
724 Novaseq 6000 platform. Paired-end (50 bp) reads were mapped to the mouse genome mm10/GRCm38
725 using STAR4 (2.7.10a)[68] (--genomeLoad LoadAndRemove --outFilterMismatchNmax 4 --
726 outFilterMultimapNmax 100 --winAnchorMultimapNmax 100). Counts were obtained with
727 featureCounts[69] (subread v2.0.3)5 -p -B -g gene_name -s 2 --countReadPairs). Differentially expressed
728 genes were calculated with DESeq2[70], filtering out genes that did not have any count in any
729 condition.DEGs were defined using an adjusted p-value≤0.05 and a log2 fold change ≥±0.58. Pathway
730 enrichment analysis of DEGs was performed using Metascape (https://metascape.org). TC-seq package[71]
731 was used on RNA-seq read counts to detect the temporal pattern of the differential gene expressions
732 (RPKM values) ("algo = 'cm', k = 9). DEGs were assigned to a cluster (C1- C9) representing a specific
733 temporal pattern of expression.

734 **Motif analysis**

735 For de novo and known motif scanning, we used HOMER "findMotifsGenome.pl" function to identify de
736 novo and known motifs in the different subsets of CTCF peaks. In parallel, we also performed motif
737 discovery analysis with MEME (5.5.2) with the following parameters: "-dna -nmotifs 20 -minw 6 -maxw
738 50 -revcomp".
739 For identifying putative G4 forming motifs, we used regular expressions in 'awk'commands to categorize
740 potential G4 forming mitifs from fasta sequences. The loop1–7 was defined with the following expression:
741 $G_3$ $N_{1–7}$ $G_3$ $N_{1–7}$ $G_3$ $N_{1–7}G_3$. The G4s with long loops were defined as sequences with any loop length >7
742 (up to 12 for any loop and 21 for the middle loop). The simple bulge was defined as sequences with a G4s
743 with a bulge of 1–7 bases in one G-run or multiple 1-base bulges. The two tetrad/ complex bulge was
744 defined as sequences with G4s with two G-bases per G-run with several bulges of 1–5 bases.
745

746 **Hi-C analysis**

747 We downloaded all available Hi-C replicates for Untreated, auxin-inducible degron (AID), and Washoff
748 conditions from the GEO repository (GSE98671) to investigate chromatin interactions under different
749 treatments. The raw Hi-C reads were processed using the HiC-Pro v2.8.0 pipeline[72], which involved
750 mapping the data to the mm10 genome assembly to generate genomewide interaction matrices. To identify
751 significant chromatin interactions, we employed the Fit-Hi-C v1.1.3 tool[73], configuring it with the options
752 "-U 5000000" and "-L 20000" to call interactions at a 20Kb resolution. This approach allowed us to
753 robustly detect and analyze significant genomic interactions across the various experimental conditions.
754 We further filtered the Hi-C interactions by selecting only those with a q-value < 0.05.
755

756 **References**

757 1 Rich, A. The molecular structure of polyinosinic acid. *Biochim Biophys Acta* **29**, 502-509 (1958).
758 https://doi.org/10.1016/0006-3002(58)90005-2

759 2 Bochman, M. L., Paeschke, K. & Zakian, V. A. DNA secondary structures: stability and function of G-
760 quadruplex structures. *Nat Rev Genet* **13**, 770-780 (2012). https://doi.org/10.1038/nrg3296
761 3 Miglietta, G., Russo, M. & Capranico, G. G-quadruplex-R-loop interactions and the mechanism of
762 anticancer G-quadruplex binders. *Nucleic Acids Res* **48**, 11942-11957 (2020).
763 https://doi.org/10.1093/nar/gkaa944
764 4 Rhodes, D. & Lipps, H. J. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res* **43**,
765 8627-8637 (2015). https://doi.org/10.1093/nar/gkv862
766 5 Spiegel, J., Adhikari, S. & Balasubramanian, S. The Structure and Function of DNA G-Quadruplexes.
767 *Trends Chem* **2**, 123-136 (2020). https://doi.org/10.1016/j.trechm.2019.07.002
768 6 Varshney, D., Spiegel, J., Zyner, K., Tannahill, D. & Balasubramanian, S. The regulation and functions of
769 DNA and RNA G-quadruplexes. *Nat Rev Mol Cell Biol* **21**, 459-474 (2020).
770 https://doi.org/10.1038/s41580-020-0236-x
771 7 Marsico, G. *et al.* Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic*
772 *Acids Res* **47**, 3862-3874 (2019). https://doi.org/10.1093/nar/gkz179
773 8 Bacolla, A. & Wells, R. D. Non-B DNA conformations, genomic rearrangements, and human disease. *J Biol*
774 *Chem* **279**, 47411-47414 (2004). https://doi.org/10.1074/jbc.R400028200
775 9 Wu, F. *et al.* Genome-wide analysis of DNA G-quadruplex motifs across 37 species provides insights into
776 G4 evolution. *Commun Biol* **4**, 98 (2021). https://doi.org/10.1038/s42003-020-01643-4
777 10 Gray, L. T., Vallur, A. C., Eddy, J. & Maizels, N. G quadruplexes are genomewide targets of transcriptional
778 helicases XPB and XPD. *Nat Chem Biol* **10**, 313-318 (2014). https://doi.org/10.1038/nchembio.1475
779 11 Bugaut, A. & Balasubramanian, S. 5'-UTR RNA G-quadruplexes: translation regulation and targeting.
780 *Nucleic Acids Res* **40**, 4727-4741 (2012). https://doi.org/10.1093/nar/gks068
781 12 Cayrou, C. *et al.* Genome-scale analysis of metazoan replication origins reveals their organization in
782 specific but flexible sites defined by conserved features. *Genome Res* **21**, 1438-1449 (2011).
783 https://doi.org/10.1101/gr.121830.111
784 13 Hansel-Hertsch, R. *et al.* G-quadruplex structures mark human regulatory chromatin. *Nat Genet* **48**,
785 1267-1272 (2016). https://doi.org/10.1038/ng.3662
786 14 Mao, S. Q. *et al.* DNA G-quadruplex structures mold the DNA methylome. *Nat Struct Mol Biol* **25**, 951-
787 957 (2018). https://doi.org/10.1038/s41594-018-0131-8
788 15 Shukla, V. *et al.* TET deficiency perturbs mature B cell homeostasis and promotes oncogenesis
789 associated with accumulation of G-quadruplex and R-loop structures. *Nat Immunol* **23**, 99-108 (2022).
790 https://doi.org/10.1038/s41590-021-01087-w
791 16 De, S. & Michor, F. DNA secondary structures and epigenetic determinants of cancer genome evolution.
792 *Nat Struct Mol Biol* **18**, 950-955 (2011). https://doi.org/10.1038/nsmb.2089
793 17 De Magis, A. *et al.* DNA damage and genome instability by G-quadruplex ligands are mediated by R loops
794 in human cancer cells. *Proc Natl Acad Sci U S A* **116**, 816-825 (2019).
795 https://doi.org/10.1073/pnas.1810409116
796 18 Salvati, E. *et al.* Telomere damage induced by the G-quadruplex ligand RHPS4 has an antitumor effect. *J*
797 *Clin Invest* **117**, 3236-3247 (2007). https://doi.org/10.1172/JCI32461
798 19 Hou, Y. *et al.* Integrative characterization of G-Quadruplexes in the three-dimensional chromatin
799 structure. *Epigenetics* **14**, 894-911 (2019). https://doi.org/10.1080/15592294.2019.1621140
800 20 Wulfridge, P. *et al.* G-quadruplexes associated with R-loops promote CTCF binding. *Mol Cell* **83**, 3064-
801 3079 e3065 (2023). https://doi.org/10.1016/j.molcel.2023.07.009
802 21 Li, L. *et al.* YY1 interacts with guanine quadruplexes to regulate DNA looping and gene expression. *Nat*
803 *Chem Biol* **17**, 161-168 (2021). https://doi.org/10.1038/s41589-020-00695-1
804 22 Spiegel, J. *et al.* G-quadruplexes are transcription factor binding hubs in human chromatin. *Genome Biol*
805 **22**, 117 (2021). https://doi.org/10.1186/s13059-021-02324-z

806 23 Zhang, X., Spiegel, J., Martinez Cuesta, S., Adhikari, S. & Balasubramanian, S. Chemical profiling of DNA
807 G-quadruplex-interacting proteins in live cells. *Nat Chem* **13**, 626-633 (2021).
808 https://doi.org/10.1038/s41557-021-00736-9
809 24 Tikhonova, P. *et al.* DNA G-Quadruplexes Contribute to CTCF Recruitment. *Int J Mol Sci* **22** (2021).
810 https://doi.org/10.3390/ijms22137090
811 25 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles
812 of the human genome. *Science* **326**, 289-293 (2009). https://doi.org/10.1126/science.1181369
813 26 Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137**, 1194-1211 (2009).
814 https://doi.org/10.1016/j.cell.2009.06.001
815 27 Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat Rev Genet* **19**,
816 789-800 (2018). https://doi.org/10.1038/s41576-018-0060-8
817 28 Szabo, Q., Bantignies, F. & Cavalli, G. Principles of genome folding into topologically associating domains.
818 *Sci Adv* **5**, eaaw1668 (2019). https://doi.org/10.1126/sciadv.aaw1668
819 29 Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* **164**, 1110-
820 1121 (2016). https://doi.org/10.1016/j.cell.2016.02.007
821 30 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin
822 looping. *Cell* **159**, 1665-1680 (2014). https://doi.org/10.1016/j.cell.2014.11.021
823 31 Filippova, G. N. *et al.* An exceptionally conserved transcriptional repressor, CTCF, employs different
824 combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc
825 oncogenes. *Mol Cell Biol* **16**, 2802-2813 (1996). https://doi.org/10.1128/MCB.16.6.2802
826 32 Hashimoto, H. *et al.* Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to
827 DNA. *Mol Cell* **66**, 711-720 e713 (2017). https://doi.org/10.1016/j.molcel.2017.05.004
828 33 Davidson, I. F. & Peters, J. M. Genome folding through loop extrusion by SMC complexes. *Nat Rev Mol*
829 *Cell Biol* **22**, 445-464 (2021). https://doi.org/10.1038/s41580-021-00349-7
830 34 Ghirlando, R. & Felsenfeld, G. CTCF: making the right connections. *Genes Dev* **30**, 881-891 (2016).
831 https://doi.org/10.1101/gad.277863.116
832 35 Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains
833 from Genomic Compartmentalization. *Cell* **169**, 930-944 e922 (2017).
834 https://doi.org/10.1016/j.cell.2017.05.004
835 36 Hansen, A. S., Pustova, I., Cattoglio, C., Tjian, R. & Darzacq, X. CTCF and cohesin regulate chromatin loop
836 stability with distinct dynamics. *Elife* **6** (2017). https://doi.org/10.7554/eLife.25776
837 37 Davidson, I. F. *et al.* CTCF is a DNA-tension-dependent barrier to cohesin-mediated loop extrusion.
838 *Nature* **616**, 822-827 (2023). https://doi.org/10.1038/s41586-023-05961-5
839 38 Kim, T. H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome.
840 *Cell* **128**, 1231-1245 (2007). https://doi.org/10.1016/j.cell.2006.12.048
841 39 Nakahashi, H. *et al.* A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep* **3**, 1678-
842 1689 (2013). https://doi.org/10.1016/j.celrep.2013.04.024
843 40 Ong, C. T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nat Rev*
844 *Genet* **15**, 234-246 (2014). https://doi.org/10.1038/nrg3663
845 41 Marina-Zarate, E., Rodriguez-Ronchel, A., Gomez, M. J., Sanchez-Cabo, F. & Ramiro, A. R. Low-affinity
846 CTCF binding drives transcriptional regulation whereas high-affinity binding encompasses architectural
847 functions. *iScience* **26**, 106106 (2023). https://doi.org/10.1016/j.isci.2023.106106
848 42 Saldana-Meyer, R. *et al.* RNA Interactions Are Essential for CTCF-Mediated Genome Organization. *Mol*
849 *Cell* **76**, 412-422 e415 (2019). https://doi.org/10.1016/j.molcel.2019.08.015
850 43 Hansen, A. S. *et al.* Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region
851 in CTCF. *Mol Cell* **76**, 395-411 e313 (2019). https://doi.org/10.1016/j.molcel.2019.07.039

852 44 Biffi, G., Di Antonio, M., Tannahill, D. & Balasubramanian, S. Visualization and selective chemical
853 targeting of RNA G-quadruplex structures in the cytoplasm of human cells. *Nat Chem* **6**, 75-80 (2014).
854 https://doi.org/10.1038/nchem.1805
855 45 Maizels, N. G4-associated human diseases. *EMBO Rep* **16**, 910-922 (2015).
856 https://doi.org/10.15252/embr.201540607
857 46 Nyborg, J. K. & Peersen, O. B. That zincing feeling: the effects of EDTA on the behaviour of zinc-binding
858 transcriptional regulators. *Biochem J* **381**, e3-4 (2004). https://doi.org/10.1042/BJ20041096
859 47 Levasseur, D. N., Wang, J., Dorschner, M. O., Stamatoyannopoulos, J. A. & Orkin, S. H. Oct4 dependence
860 of chromatin structure within the extended Nanog locus in ES cells. *Genes Dev* **22**, 575-580 (2008).
861 https://doi.org/10.1101/gad.1606308
862 48 Henderson, A. *et al.* Detection of G-quadruplex DNA in mammalian cells. *Nucleic Acids Res* **42**, 860-869
863 (2014). https://doi.org/10.1093/nar/gkt957
864 49 Kazemier, H. G., Paeschke, K. & Lansdorp, P. M. Guanine quadruplex monoclonal antibody 1H6 cross-
865 reacts with restrained thymidine-rich single stranded DNA. *Nucleic Acids Res* **45**, 5913-5919 (2017).
866 https://doi.org/10.1093/nar/gkx245
867 50 Ray, S. *et al.* Custom DNA Microarrays Reveal Diverse Binding Preferences of Proteins and Small
868 Molecules to Thousands of G-Quadruplexes. *ACS Chem Biol* **15**, 925-935 (2020).
869 https://doi.org/10.1021/acschembio.9b00934
870 51 Zeraati, M. *et al.* I-motif DNA structures are formed in the nuclei of human cells. *Nat Chem* **10**, 631-637
871 (2018). https://doi.org/10.1038/s41557-018-0046-3
872 52 Nicoludis, J. M. *et al.* Optimized end-stacking provides specificity of N-methyl mesoporphyrin IX for
873 human telomeric G-quadruplex DNA. *J Am Chem Soc* **134**, 20446-20456 (2012).
874 https://doi.org/10.1021/ja3088746
875 53 Sabharwal, N. C. *et al.* N-methylmesoporphyrin IX fluorescence as a reporter of strand orientation in
876 guanine quadruplexes. *FEBS J* **281**, 1726-1737 (2014). https://doi.org/10.1111/febs.12734
877 54 Chambers, V. S. *et al.* High-throughput sequencing of DNA G-quadruplex structures in the human
878 genome. *Nat Biotechnol* **33**, 877-881 (2015). https://doi.org/10.1038/nbt.3295
879 55 Li, Y. *et al.* The structural basis for cohesin-CTCF-anchored loops. *Nature* **578**, 472-476 (2020).
880 https://doi.org/10.1038/s41586-019-1910-z
881 56 Nora, E. P. *et al.* Molecular basis of CTCF binding polarity in genome folding. *Nat Commun* **11**, 5612
882 (2020). https://doi.org/10.1038/s41467-020-19283-x
883 57 Barrington, C., Finn, R. & Hadjur, S. Cohesin biology meets the loop extrusion model. *Chromosome Res*
884 **25**, 51-60 (2017). https://doi.org/10.1007/s10577-017-9550-3
885 58 Harteis, S. & Schneider, S. Making the bend: DNA tertiary structure and protein-DNA interactions. *Int J*
886 *Mol Sci* **15**, 12335-12363 (2014). https://doi.org/10.3390/ijms150712335
887 59 Hark, A. T. *et al.* CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus.
888 *Nature* **405**, 486-489 (2000). https://doi.org/10.1038/35013106
889 60 Asamitsu, S. *et al.* RNA G-quadruplex organizes stress granule assembly through DNAPTP6 in neurons.
890 *Sci Adv* **9**, eade2035 (2023). https://doi.org/10.1126/sciadv.ade2035
891 61 Luige, J., Armaos, A., Tartaglia, G. G. & Orom, U. A. V. Predicting nuclear G-quadruplex RNA-binding
892 proteins with roles in transcription and phase separation. *Nat Commun* **15**, 2585 (2024).
893 https://doi.org/10.1038/s41467-024-46731-9
894 62 Zhang, Y. *et al.* G-quadruplex structures trigger RNA phase separation. *Nucleic Acids Res* **47**, 11746-
895 11754 (2019). https://doi.org/10.1093/nar/gkz978
896 63 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short
897 DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009). https://doi.org/10.1186/gb-2009-
898 10-3-r25

899 64 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
900 https://doi.org/10.1186/gb-2008-9-9-r137
901 65 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory
902 elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010).
903 https://doi.org/10.1016/j.molcel.2010.05.004
904 66 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features.
905 *Bioinformatics* **26**, 841-842 (2010). https://doi.org/10.1093/bioinformatics/btq033
906 67 Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: a flexible platform for exploring
907 deep-sequencing data. *Nucleic Acids Res* **42**, W187-191 (2014). https://doi.org/10.1093/nar/gku365
908 68 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
909 https://doi.org/10.1093/bioinformatics/bts635
910 69 Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-
911 and-vote. *Nucleic Acids Res* **41**, e108 (2013). https://doi.org/10.1093/nar/gkt214
912 70 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq
913 data with DESeq2. *Genome Biol* **15**, 550 (2014). https://doi.org/10.1186/s13059-014-0550-8
914 71 Spies, D., Renz, P. F., Beyer, T. A. & Ciaudo, C. Comparative analysis of differential gene expression tools
915 for RNA sequencing time course data. *Brief Bioinform* **20**, 288-298 (2019).
916 https://doi.org/10.1093/bib/bbx115
917 72 Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**,
918 259 (2015). https://doi.org/10.1186/s13059-015-0831-x
919 73 Ay, F., Bailey, T. L. & Noble, W. S. Statistical confidence estimation for Hi-C data reveals regulatory
920 chromatin contacts. *Genome Res* **24**, 999-1011 (2014). https://doi.org/10.1101/gr.160374.113

921

Figure 1

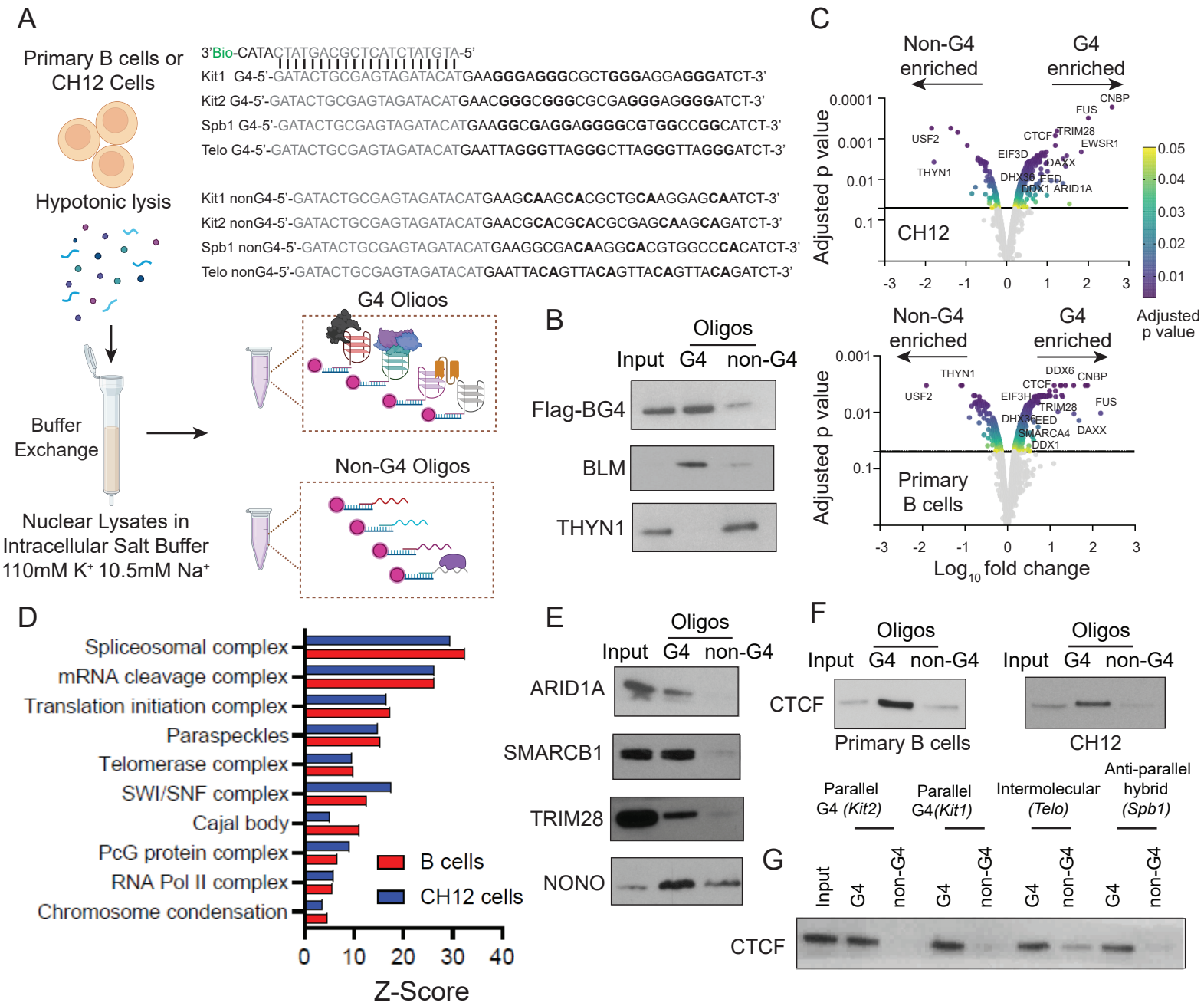

A
Primary B cells or CH12 Cells
Hypotonic lysis
Buffer Exchange
Nuclear Lysates in Intracellular Salt Buffer 110mM K+ 10.5mM Na+
3'Bio-CATACTATGACGCTCATCTATGTA-5'
Kit1 G4-5'-GATACTGCGAGTAGATACATGAAGGGAGGGCGCTGGGAGGAGGGATCT-3'
Kit2 G4-5'-GATACTGCGAGTAGATACATGAACGGGCGGGCGCGAGGGAGGGATCT-3'
Spb1 G4-5'-GATACTGCGAGTAGATACATGAAGGCGAGGAGGGGCGTGGCCGGCATCT-3'
Telo G4-5'-GATACTGCGAGTAGATACATGAATTAGGGTTAGGGCTTAGGGTTAGGGATCT-3'
Kit1 nonG4-5'-GATACTGCGAGTAGATACATGAAGCAAGCACGCTGCAAGGAGCAATCT-3'
Kit2 nonG4-5'-GATACTGCGAGTAGATACATGAACGCACGCACGCGAGCAAGCAGATCT-3'
Spb1 nonG4-5'-GATACTGCGAGTAGATACATGAAGGCGACAAGGCACGTGGCCCACATCT-3'
Telo nonG4-5'-GATACTGCGAGTAGATACATGAATTACAGTTACAGTTACAGTTACAGATCT-3'
G4 Oligos
Non-G4 Oligos
B
Oligos
Input G4 non-G4
Flag-BG4
BLM
THYN1
C
Non-G4 enriched
G4 enriched
Adjusted p value
CH12
Primary B cells
Log10 fold change
D
Spliceosomal complex
mRNA cleavage complex
Translation initiation complex
Paraspeckles
Telomerase complex
SWI/SNF complex
Cajal body
PcG protein complex
RNA Pol II complex
Chromosome condensation
B cells
CH12 cells
Z-Score
E
ARID1A
SMARCB1
TRIM28
NONO
F
CTCF
Primary B cells
CH12
G
Parallel G4 (Kit2)
Parallel G4(Kit1)
Intermolecular (Telo)
Anti-parallel hybrid (Spb1)
Input
G4
non-G4

Figure 1: Proteomics screen for G4 binding proteins

A. A schematic illustrating hypotonic lysis of primary B cells or CH12 cells combined with dounce homogenization and buffer exchange, followed by incubation with biotinylated G4 or non-G4 oligos. Top right: nucleotide sequence of each oligo used, gray indicating sequence complementary to the biotinylated (Bio) adaptor anchor, and bolded letters indicating guanine stretches (G4 oligo) for G4 formation or mutation of guanine stretches in non-G4 oligos.
B. Immunoblot showing the enrichment of Flag-BG4 (spike-in) and a G4 helicase BLM in nuclear lysates pulled down with G4 or non-G4 oligos. THYN1 is included as ssDNA binding protein used as negative control.
C. Enrichment profile of G4 protein pulldown fractions from CH12 (top) and primary B (bottom) cells quantified by Tandem Mass Tagging followed by Mass-spectrometry.
D. Top protein complexes enriched with G4s from the proteomics screen in primary B cells (red) or CH12 cells (blue). Protein complex enrichment was determined by Metascape analysis with proteins showing 2-fold increase over non-G4 controls and p-adjusted value >0.05. The X-axis represents the Z-score.
E. Immunoblot showing the enrichment of ARID1A, SMARCB1, NONO, and TRIM28 in G4 pulldown compared with non-G4 control oligos.
F. Immunoblot showing the enrichment of CTCF upon pull down with G4 oligos or non-G4 control oligos from primary B cells (left) or CH12 cells (right).
G. Immunoblot showing the enrichment of CTCF with G4 oligos of distinct topologies, including parallel (Kit2), parallel-propeller (Kit1), intermolecular (Telo), and anti-parallel hybrid (Spb1) G4. The relative enrichment is compared to respective control oligos.

Data is panels B, E, F and G are representative of at least three independent experiments.

Figure 2

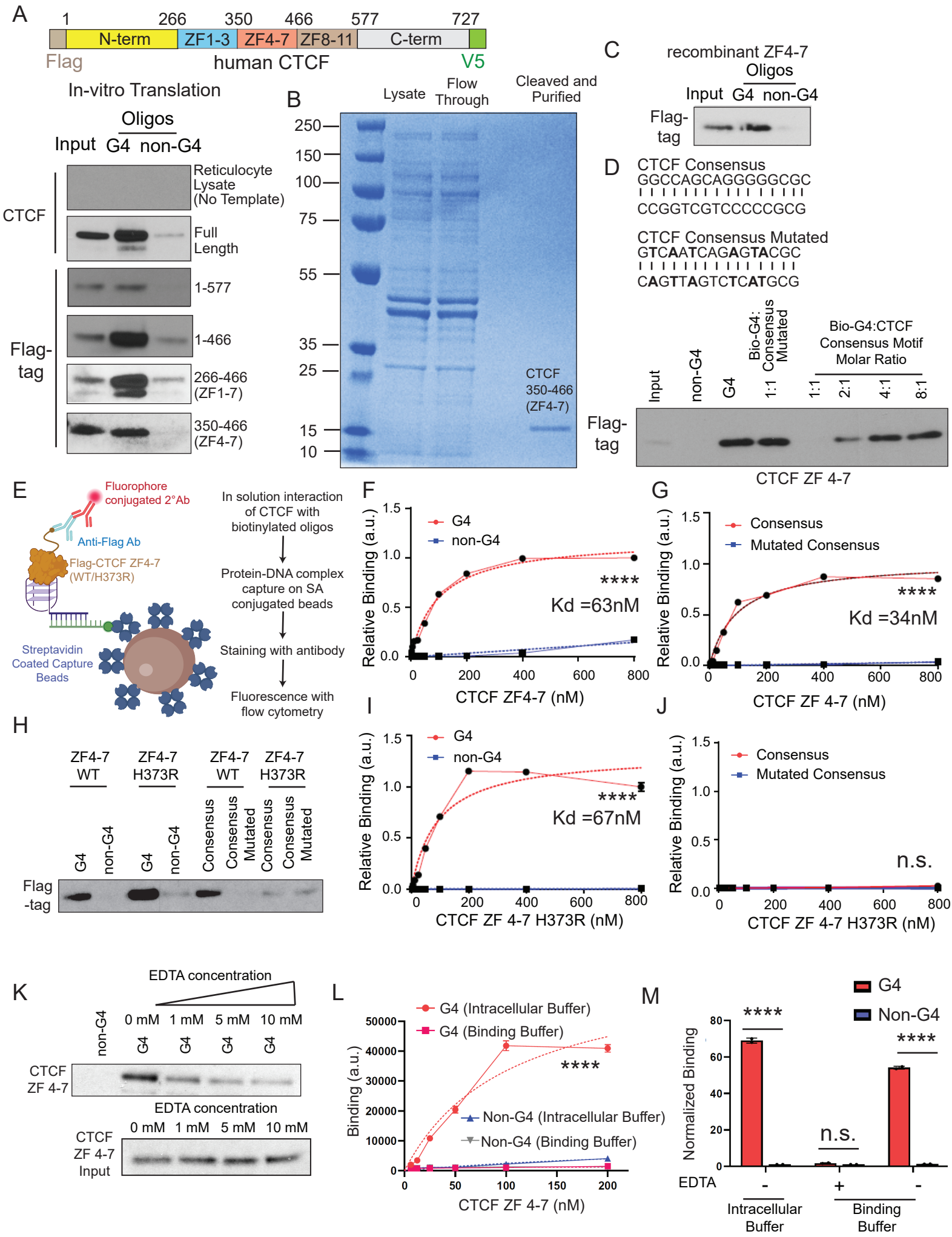

A
1
266
350
466
577
727
N-term
ZF1-3
ZF4-7
ZF8-11
C-term
Flag
human CTCF
V5
In-vitro Translation
Oligos
Input G4 non-G4
CTCF
Reticulocyte Lysate (No Template)
Full Length
Flag-tag
1-577
1-466
266-466 (ZF1-7)
350-466 (ZF4-7)
B
Lysate
Flow Through
Cleaved and Purified
250
150
100
75
55
35
25
15
10
CTCF 350-466 (ZF4-7)
C
recombinant ZF4-7
Oligos
Input G4 non-G4
Flag-tag
D
CTCF Consensus
GGCCAGCAGGGGGCGC
CCGGTCGTCCCCCGCG
CTCF Consensus Mutated
GTCAATCAGAGTACGC
CAGTTAGTCTCATGCG
Input
non-G4
G4
Bio-G4: Consensus Mutated
1:1
Bio-G4:CTCF Consensus Motif Molar Ratio
1:1
2:1
4:1
8:1
Flag-tag
CTCF ZF 4-7
E
Fluorophore conjugated 2°Ab
Anti-Flag Ab
Flag-CTCF ZF4-7 (WT/H373R)
Streptavidin Coated Capture Beads
In solution interaction of CTCF with biotinylated oligos
Protein-DNA complex capture on SA conjugated beads
Staining with antibody
Fluorescence with flow cytometry
F
Relative Binding (a.u.)
G4
non-G4
****
Kd =63nM
CTCF ZF4-7 (nM)
G
Consensus
Mutated Consensus
****
Kd =34nM
CTCF ZF 4-7 (nM)
H
ZF4-7 WT
ZF4-7 H373R
ZF4-7 WT
ZF4-7 H373R
G4
non-G4
G4
non-G4
Consensus
Consensus Mutated
Consensus
Consensus Mutated
Flag-tag
I
G4
non-G4
****
Kd =67nM
CTCF ZF 4-7 H373R (nM)
J
Consensus
Mutated Consensus
n.s.
CTCF ZF 4-7 H373R (nM)
K
EDTA concentration
0 mM 1 mM 5 mM 10 mM
non-G4
G4
G4
G4
G4
CTCF ZF 4-7
EDTA concentration
0 mM 1 mM 5 mM 10 mM
CTCF ZF 4-7 Input
L
G4 (Intracellular Buffer)
G4 (Binding Buffer)
****
Non-G4 (Intracellular Buffer)
Non-G4 (Binding Buffer)
Binding (a.u.)
CTCF ZF 4-7 (nM)
M
G4
Non-G4
****
****
n.s.
Normalized Binding
EDTA
-
+
-
Intracellular Buffer
Binding Buffer

Figure 2 Biochemical validation of CTCF-G4 interaction

A. Top: a schematic of the human CTCF with an N-terminal FLAG (beige) and C-terminal V5 tag (green) expressed using *in vitro* transcription and translation with rabbit reticulocyte lysates. Bottom: immunoblot showing the expression of *in vitro* transcribed and translated full length (FL) CTCF or mutants lacking specific domains. Pulldown was performed with biotinylated G4 oligos or non-G4 controls. Input without oligo pulldown was used as loading control, and reticulocyte lysate without template was used as negative control.
B. Coomassie stained gel validating the expression and purification of recombinant CTCF fragments containing the central ZF4-7 fragment (CTCF 350-466) expressed in *E. coli*.
C. Immunoblot showing the enrichment of recombinant CTCF ZF4-7 with G4 oligos, compared with non-G4 controls.
D. Top: oligonucleotide sequence of CTCF consensus and mutated consensus motifs used in oligo pulldown. Each mutated base is bolded in the mutated consensus motif sequence. Bottom: Immunoblot showing the enrichment of recombinant CTCF ZF4-7 with biotinylated (*Kit2)* G4 compared to non-G4 control oligos. The biotinylated G4 oligo was mixed with varying molar ratios of dsDNA consensus motif or with an equimolar concentration of mutated consensus.
E. Schematic of G4 binding assay performed using flow cytometry. Flag-tagged recombinant CTCF was mixed with different biotinylated oligos followed by capture on streptavidin (SA) magnetic beads. The specific interaction was quantified by on-bead fluorescence signal measured using an anti-Flag antibody.
F. Binding curve for interaction between wild-type (WT) recombinant CTCF ZF4-7 and *Kit2* G4 (red, circle) or non-G4 (blue, square) oligos. The relative binding was calculated by normalizing the fluorescence signal of bead only negative control, stained with the anti-Flag antibody. X-axis denotes CTCF protein concentrations in nanomolar (nM), Y-axis represents arbitrary units (a.u.) of binding. Curved dashed lines represent the inferred non-linear fit for the data.
G. Binding curve for interaction between WT recombinant CTCF ZF4-7 and consensus or mutated consensus oligos. X-axis denotes CTCF protein concentrations in nanomolar (nM), Y-axis represents arbitrary units (a.u.) of binding. Curved dashed lines represent the inferred non-linear fit for the data.
H. Immunoblot showing the pull down of recombinant WT or H373R CTCF ZF4-7 with biotinylated G4, non-G4, consensus, or mutated consensus oligos.
I. Binding curve showing the interaction between H373R recombinant CTCF ZF4-7 and *Kit2* G4 (red, circle) or non-G4 (blue, square) oligos. X-axis denotes CTCF protein concentrations in nanomolar (nM), Y-axis represents arbitrary units (a.u.) of binding. Curved dashed lines represent the inferred non-linear fit for the data.
J. Binding curve showing the interaction between H373R recombinant CTCF ZF4-7 protein and consensus (red, circle) or mutated consensus (blue, square) oligos. X-axis denotes CTCF protein concentrations in nanomolar (nM), Y-axis represents arbitrary units (a.u.) of binding. Curved dashed lines represent the inferred non-linear fit for the data.
K. Immunoblot showing the pull down of recombinant CTCF ZF4-7 with biotinylated G4 oligos in presence of varying concentrations of EDTA (0mM to 10mM). Input without oligo pulldown was used as loading control (bottom) and lysate from protein pulled down with non-G4 was used as negative control.
L. Binding curve showing the interaction of recombinant CTCF ZF4-7 with *Kit2* G4 or non-G4 oligos in either intracellular buffer (orange circle and blue triangle, respectively) or binding buffer from Wulfridge et.al. (red square and grey triangle, respectively). X-axis denotes CTCF protein

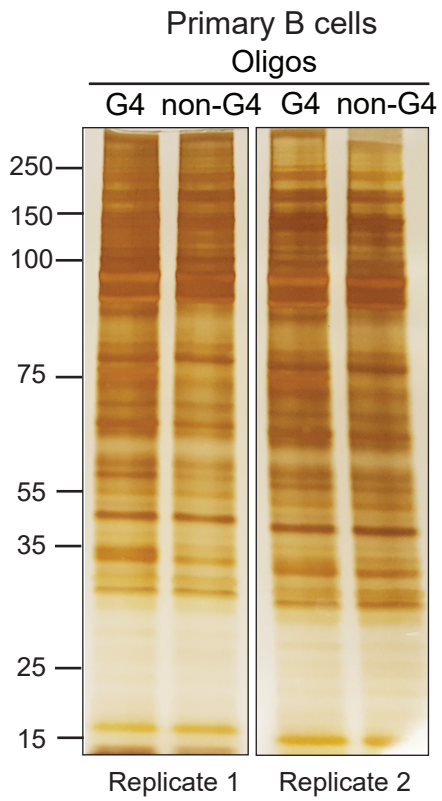
concentrations in nanomolar (nM), Y-axis represents arbitrary units (a.u.) of binding. Curved dashed lines represent the inferred non-linear fit for the data. ****p value ≤ 0.0001 calculated with 2-way ANOVA

M. The effect of EDTA on binding of the recombinant CTCF ZF4-7 protein to *Kit2* G4 (red) or non-G4 (blue) oligos in either intracellular or binding buffer. The binding was normalized to bead only negative controls. Data points represent the mean and ± standard error from three independent experiments. **** p value ≤ 0.0001 calculated with student T-test. n. s. Not significant.

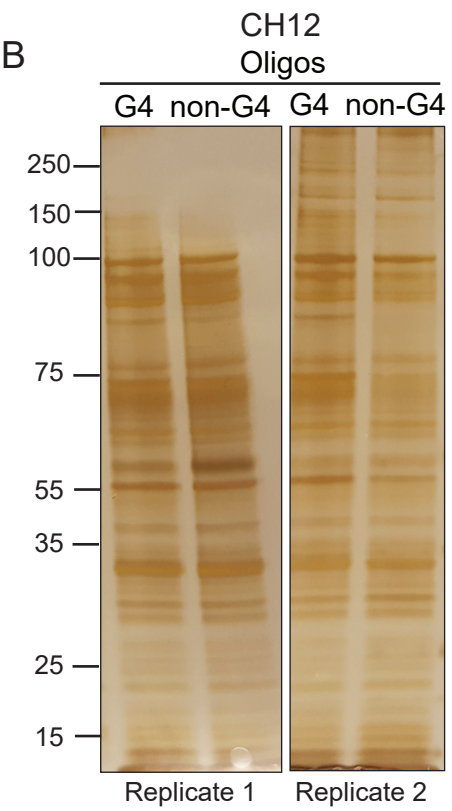
Data points in figures F, G, I and J represent the mean ± standard deviations from three independent experiments. **** p value ≤ 0.0001 calculated with 2-way ANOVA.

Supplementary figure 1

A



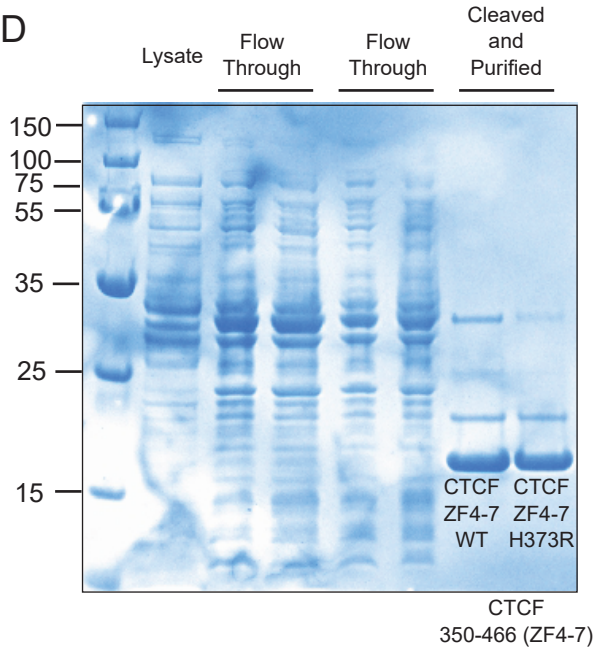

B



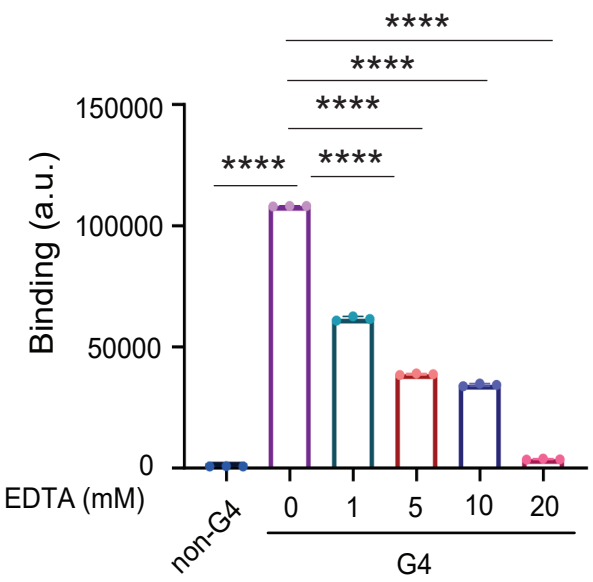

C

| Complexes | Primary B cells | CH12 cells |
|---|---|---|
| Spliceosomal complex | Cirbp,Ddx5,Hnrnpa1,Rbm3,Sf3a2,Srsf2,Snrpc,Snrnp70,Snrpb,Snrpb2,Snrpd1,Snrpe,U2af2,Sf1,Ik,Txnl4a,Hnrnpu,Snrpa,Sf3b6,Ppih,Sf3b5,Zcrb1,Rnpc3,Snrpd3,Sf3a1,Mfap1a,Snrnp48,Phf5a,Snrpa1,Prpf31,Snrpf,Prpf4,Cactin,Sugp1,Prpf3,Ddx42,Ddx41,Smu1,Sf3a3,Snrnp35,Smndc1,Snip1,Rbm17,Sf3b1,Sf3b3,Snrpd2,Sf3b4,U2af1,Ddx46,Hnrnpa3,Prpf38a,Sf3b2 | Cirbp,Hnrnpa1,Rbm3,Sf3a2,Srsf2,Snrpc,Snrnp70,Snrpb,Snrpb2,Snrpd1,Snrpe,Alyref,U2af2,Hnrnpu,Snrpa,Sf3b6,Sf3b5,Zcrb1,Rnpc3,Snrpd3,Sf3a1,Mfap1a,Snrnp48,Phf5a,Snrpa1,Snrpf,Sugp1,Prpf3,Sf3a3,Smndc1,Snip1,Rbm17,Sf3b1,Sf3b3,Snrpd2,Sf3b4,U2af1,Hnrnpa3,Sf3b2 |
| mRNA cleavage complex | Cpsf2,Cpsf4,Cpsf3,Fip1l1,Sympk,Nudt21,Wdr33,Pcf11,Cpsf1,Clp1,Cstf2,Cstf3 | Cpsf2,Cpsf4,Cpsf3,Fip1l1,Sympk,Wdr33,Pcf11,Cpsf1,Clp1,Cstf2,Cstf3 |
| Translation initiation complex | Eif3a,Eif4e,Eif3e,Eif3b,Eif3g,Eif3i,Eif3d,Eif3c,Eif3f,Eif3h,Eif3k,Eif3j1,Eif3m,Eif3l | Eif3a,Eif3e,Eif3b,Eif3g,Eif3i,Eif3d,Eif3c,Eif3f,Eif3h,Eif3k,Eif4b,Eif3j1,Eif3m,Eif3l |
| SWI/SNF complex | Dpf2,Smarca4,Smarcb1,Smarcc1,Smarce1,Smarcd1,Ss18,Arid1a | Dpf2,Smarca4,Smarcb1,Smarcc1,Smarce1,Smarcd1,Arid1a,Ss18 |
| PcG protein complex | Eed,Ezh2,Tfdp1,Suz12,Ogt,Rbbp7,Ep300 | Bmi1,Cbx4,Eed,Ezh2,Suz12,Phc2 |
| Paraspeckles | Nono,Pspc1,Nudt21,Sfpq,Alkbh5,Fus | Nono,Pspc1,Sfpq,Alkbh5,Fus |

D




E



Supplementary Figure 1

A. Silver stain of pulldown fractions from G4 and non-G4 control oligos in the nuclear lysates of LPS and IL-4-stimulated primary B cells.
B. Silver stain of pull-down fractions from G4 and non-G4 control oligos in the nuclear lysates of CH12 cells.
C. List of protein complex subunits enriched in G4 proteomics screen of primary B cells and CH12 B cells. Multiple subunits within each complex were enriched.
D. Expression and purification of recombinant CTCF ZF4-7 WT or H373R (ZF4-7) mutant from *E. coli*.
E. Binding assay showing the interaction of recombinant CTCF ZF4-7 protein with G4 (red, circle) or non-G4 (blue, square) oligos in the absence or presence of increasing concentrations of EDTA. X-axis denotes CTCF protein concentrations in nanomolar (nM), Y-axis represents arbitrary units (a.u.) of binding. **** p value $\leq 0.0001$ calculated with student T-test.

Figure 3

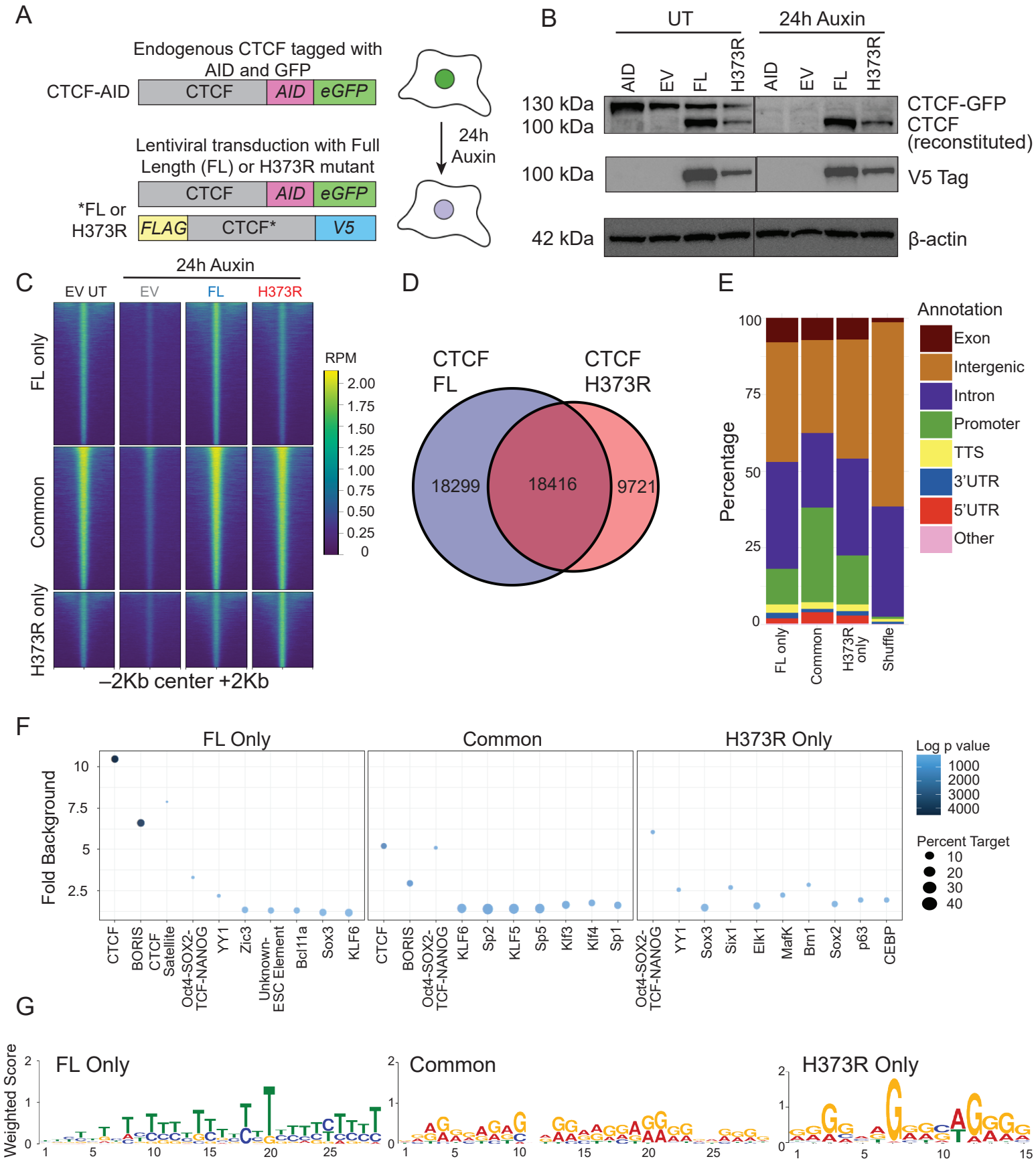
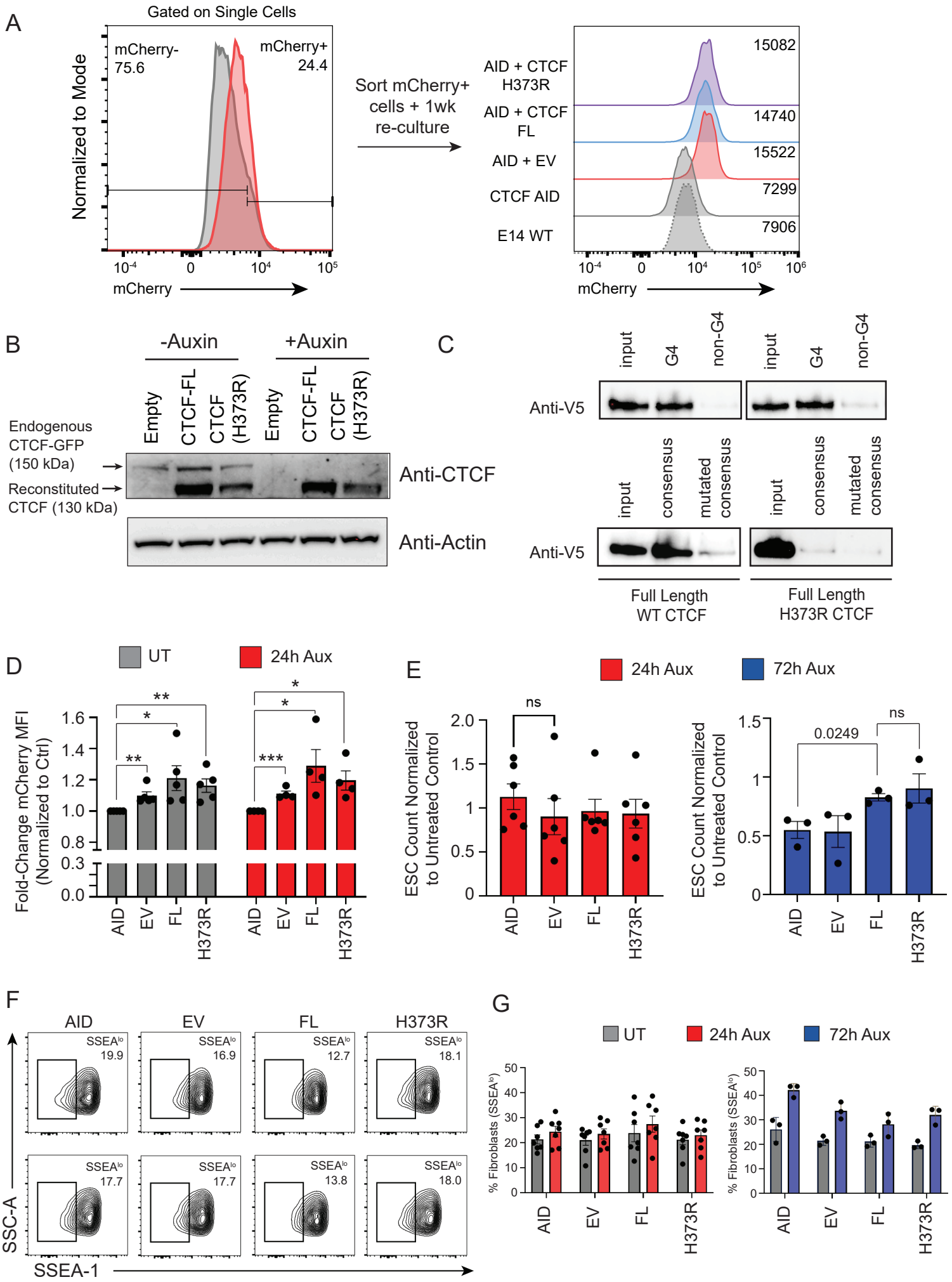

A
Endogenous CTCF tagged with AID and GFP
CTCF-AID
CTCF
AID
eGFP
Lentiviral transduction with Full Length (FL) or H373R mutant
*FL or H373R
CTCF
AID
eGFP
FLAG
CTCF*
V5
24h Auxin
B
UT
24h Auxin
AID
EV
FL
H373R
AID
EV
FL
H373R
130 kDa
100 kDa
CTCF-GFP
CTCF (reconstituted)
100 kDa
V5 Tag
42 kDa
β-actin
C
24h Auxin
EV UT
EV
FL
H373R
FL only
Common
H373R only
RPM
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0
–2Kb center +2Kb
D
CTCF FL
CTCF H373R
18299
18416
9721
E
Annotation
Exon
Intergenic
Intron
Promoter
TTS
3'UTR
5'UTR
Other
Percentage
100
75
50
25
0
FL only
Common
H373R only
Shuffle
F
FL Only
Common
H373R Only
Fold Background
10
7.5
5
2.5
CTCF
BORIS
CTCF Satellite
Oct4-SOX2-TCF-NANOG
YY1
Zic3
Unknown-ESC Element
Bcl11a
Sox3
KLF6
CTCF
BORIS
Oct4-SOX2-TCF-NANOG
KLF6
Sp2
KLF5
Sp5
Klf3
Klf4
Sp1
Oct4-SOX2-TCF-NANOG
YY1
Sox3
Six1
Elk1
MafK
Brn1
Sox2
p63
CEBP
Log p value
1000
2000
3000
4000
Percent Target
10
20
30
40
G
Weighted Score
FL Only
Common
H373R Only

Figure 3 Genome-wide profiling of FL and H373R CTCF protein

A. Schematic showing the endogenous CTCF tagged with Auxin Inducible Degron (AID) tag and GFP (top) and the exogenous Full Length (FL)* or H373R* CTCF construct tagged with FLAG (at N-terminus) and V5 (at C-terminus) (bottom). The cells were treated with auxin for 24 hours to monitor degradation of endogenous CTCF.
B. Immunoblot showing the expression of GFP-tagged endogenous (130 kDa) or V5-expressing reconstituted (100 kDa) CTCF (top) with or without auxin treatment for 24 hours. The expression of reconstituted CTCF was validated with immunoblotting using anti-V5 antibody (middle). AID represents the original mESCs without auxin treatment. EV, FL and H373R are cells transduced with empty vector, full length and H373R CTCF, respectively. Beta-actin expression is used as loading control (bottom).
   Heatmaps showing CUT-and-Tag signal intensity of CTCF binding across ±2 kb regions centered on peaks with or without auxin treatment (24 hours) in EV, FL and H373R expressing cells. Peaks were classified into FL-only (top), Common (middle), and H373R-only (bottom) regions. CUT-and-Tag signal intensity, normalized to Reads Per Million (RPM), is depicted by the color scale. CTCF CUT-and-Tag was performed as three independent replicates across different experimental conditions. The heatmaps are merged/averaged signals from three replicates.
C. Venn diagram showing the number of CTCF peaks identified using CUT-and-Tag in auxin-treated mESCs transduced with either FL (blue) or H373R (red) CTCF.
D. Genomic annotations of CTCF peaks categorized as FL only, Common, and H373R only, along with genomic shuffle as control regions.
E. Motif enrichment analysis of CTCF peaks categorized as FL-only (left), Common (middle), and H373R-only (right) regions. Y-axis represents the fold enrichment of motifs compared to the genomic background; X-axis denotes enriched motifs. Circle size indicates the percentage of peaks containing each motif, and the color intensity corresponds to the log-transformed p-value for motif enrichment performed using HOMER.
F. Top sequence motifs identified using *de novo* motif discovery in FL-only (left), Common (middle), and H373R-only (right) CTCF peaks. The Y-axis represents the weighted score, indicating the contribution of each nucleotide within the motif.

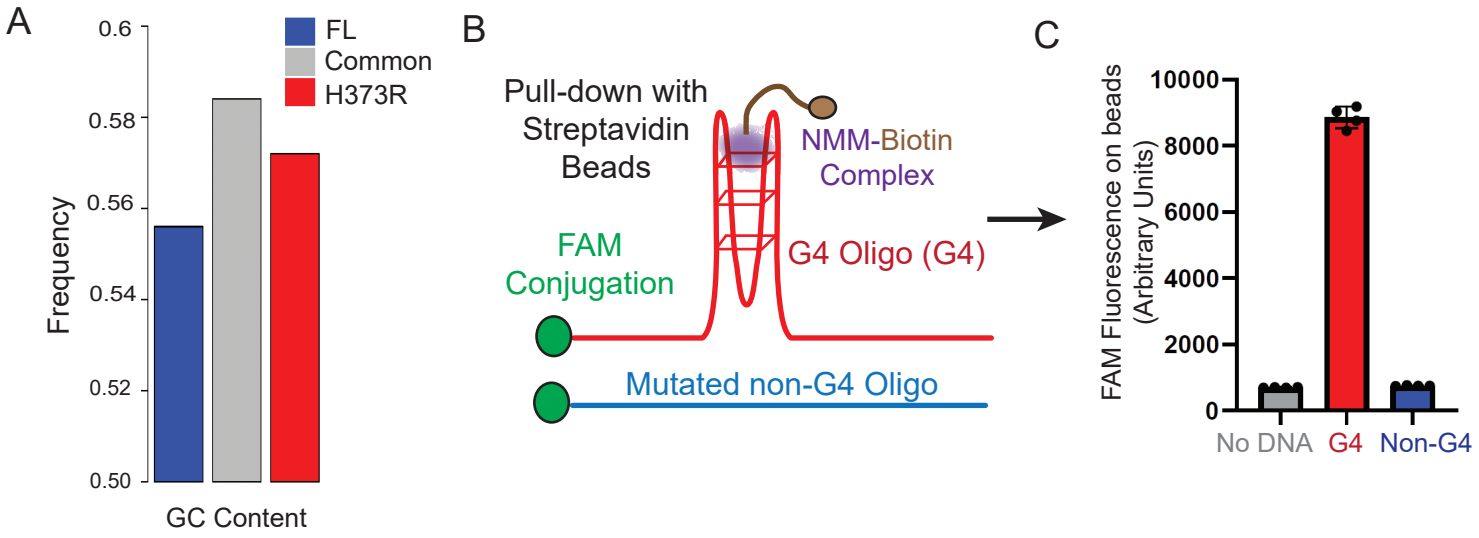
Supplementary figure 2



Supplementary Figure 2

A. Sorting strategy for mESCs harboring AID-tagged CTCF transduced with empty vector (EV), CTCF FL and CTCF H373R constructs expressing mCherry reporter. mCherry+ cells were sorted and expanded in culture for one week prior to CUT-and-Tag.
B. Western blot for CTCF expression in cell lysates derived from EV, FL and H373R expressing mESCs after 8 hours with or without auxin treatment. Beta-Actin was used as a loading control.
C. Immunoblot showing the enrichment of *in vitro* transcribed and translated full length or H373R CTCF pulled down with either G4, non-G4, CTCF consensus, or mutated consensus oligos.
D. Barplots quantifying the median fluorescence intensity (MFI) of mCherry in AID, EV, FL and H373R expressing mESCs after 24h with and without auxin treatment. mCherry MFI was normalized to the non-transduced AID-CTCF (AID) mESCs in the same experiment. * p value $\leq$ 0.05, ** p value $\leq$ 0.01, *** p value $\leq$ 0.001, calculated using Student T-test.
E. Barplots quantifying normalized cell counts in mESCs expressing EV, FL and H373R with auxin treatment for 24 and 72 hours. Cell counts were normalized to the respective untreated controls in each condition. The ratio below 1 indicates reduced cell numbers. Marked p values calculated using Student T-test.
F. Representative flow cytometry plots for SSEA-1 expression in AID, EV, FL and H373R mESCs 24 hours with or without auxin treatment.
G. Barplots quantifying the frequencies of fibroblasts ($SSEA^{lo}$) in AID, EV, FL and H373R mESCs at 24 and 72 hours with or without (untreated, UT) auxin treatment.

Supplementary figure 3



Supplementary Figure 3

A. Barplot showing GC content across FL only (blue), common (grey) and H373R only (red) CTCF peaks.
B. Schematic representation of the experimental design of NMM-Bio binding to a G4 oligo.
C. A fluorescein (FAM)-conjugated *Kit2* G4 oligo (red) or mutated non-G4 control (blue) was incubated with NMM-Bio and captured on streptavidin beads. The fluorescence signal was measured using flow cytometry. Data is representative of 3 independent replicates.

Figure 4

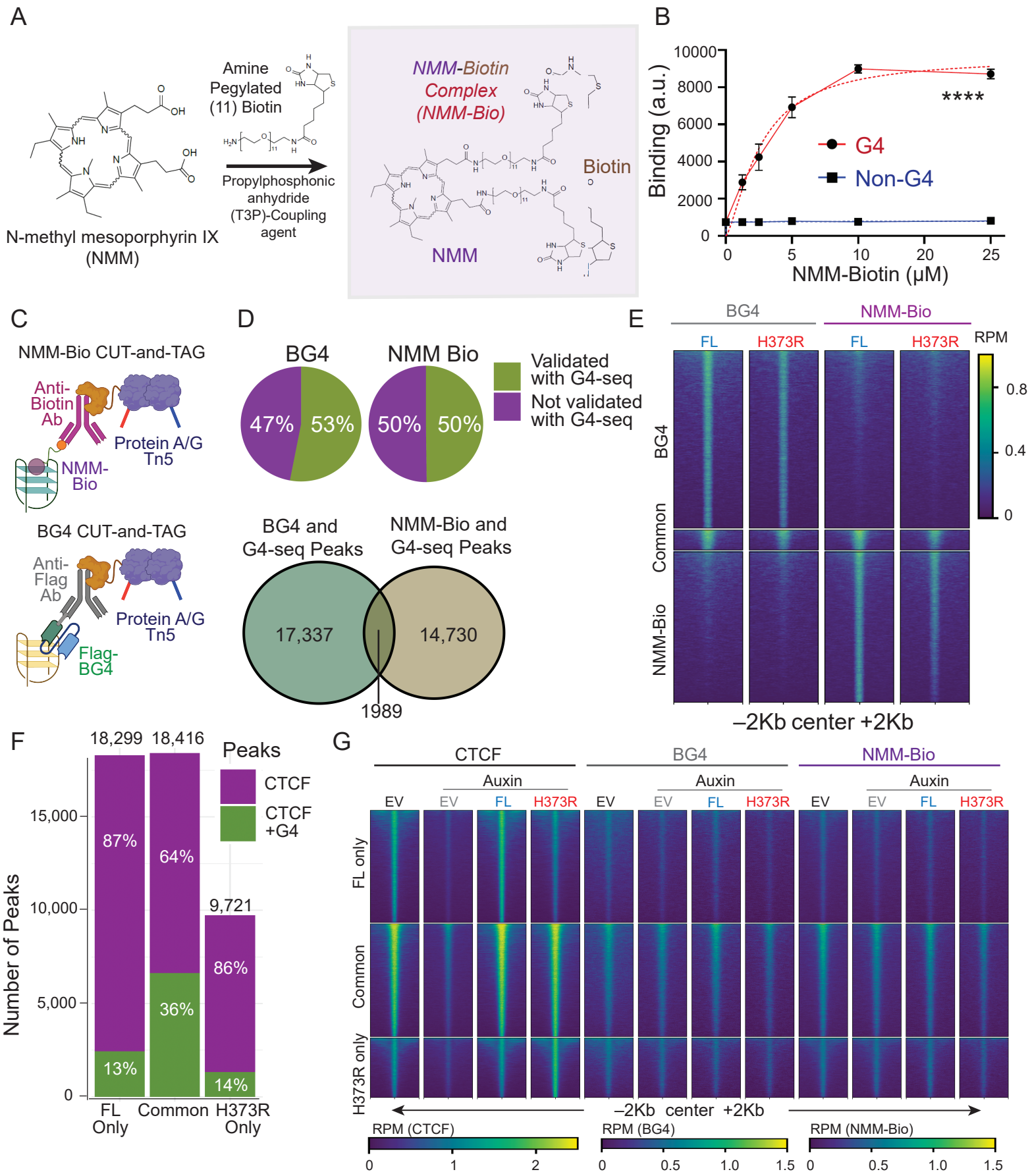

A
N-methyl mesoporphyrin IX (NMM)
Amine Pegylated (11) Biotin
Propylphosphonic anhydride (T3P)-Coupling agent
NMM-Biotin Complex (NMM-Bio)
Biotin
NMM
B
Binding (a.u.)
NMM-Biotin (μM)
G4
Non-G4
****
C
NMM-Bio CUT-and-TAG
Anti-Biotin Ab
Protein A/G Tn5
NMM-Bio
BG4 CUT-and-TAG
Anti-Flag Ab
Protein A/G Tn5
Flag-BG4
D
BG4
NMM Bio
47%
53%
50%
50%
Validated with G4-seq
Not validated with G4-seq
BG4 and G4-seq Peaks
NMM-Bio and G4-seq Peaks
17,337
14,730
1989
E
BG4
NMM-Bio
FL
H373R
RPM
BG4
Common
NMM-Bio
–2Kb center +2Kb
F
18,299
18,416
9,721
Peaks
CTCF
CTCF +G4
87%
13%
64%
36%
86%
14%
Number of Peaks
FL Only
Common
H373R Only
G
CTCF
BG4
NMM-Bio
Auxin
EV
FL
H373R
FL only
Common
H373R only
–2Kb center +2Kb
RPM (CTCF)
RPM (BG4)
RPM (NMM-Bio)

Figure 4 Genome-wide G4 profiling and its association to CTCF binding

A. Schematic for the generation of N-methyl mesoporphyrin IX (NMM) conjugated with biotin (NMM-Bio).
B. Binding curve of NMM-Bio to G4 (red) and non-G4 (blue) oligos. X-axis marks different NMM-Bio concentrations used, and Y-axis represents binding signal as arbitrary units (a.u.). Data points are mean ± standard error from four independent experiments. **** p value ≤ 0.0001 calculated with 2-way ANOVA.
C. Schematic representation of CUT-and-TAG for NMM-Bio and BG4 based G4 profiling. NMM-Bio was combined with anti-biotin antibody, while flag-tagged BG4 single chain variable fragment antibody was combined with anti-Flag antibody. In both methods, tagmentation was performed using Tn5 transposase fused with protein A/G.
D. Overlap of BG4 and NMM-Bio peaks with previously published G4-seq peaks identified in mouse genomic DNA. Top: pie charts show the percentage of BG4 and NMM-Bio peaks with or without G4-seq overlap. Bottom: Venn diagram highlights shared and unique peaks between BG4 and NMM-Bio methods at regions that are common with G4-seq regions in both conditions.
E. Heatmaps showing signal intensity for BG4 and NMM-bio peaks ±2 kb from the center. Peaks were classified into unique and shared regions identified by NMM-Bio and BG4 methods in mESCs expressing FL and H373R CTCF upon 24 hours of auxin treatment. CUT-and-Tag signal intensity normalized to Reads Per Million (RPM), is depicted by the color scale.
F. Bar plot quantifying the overlap of G4 peaks (identified by either BG4 or NMM-Bio method) with CTCF peaks categorized into FL-only, Common, and H373R-only regions.
G. Heatmaps showing signal intensity (RPM) of CTCF, BG4, and NMM-Bio signals across FL-only, Common, and H373R-only CTCF peaks. Regions were classified into FL-only, Common, and H373R-only peaks in mESCs expressing EV, FL and H373R CTCF with or without (EV control) auxin treatment.

BG4 and NMM-Bio CUT-and-Tag were performed as three independent replicates across different experimental conditions. Unconjugated NMM (NMM) was used as negative control, and the peaks were called using NMM signal as background. The heatmaps are merged signals from three replicates.

Figure 5

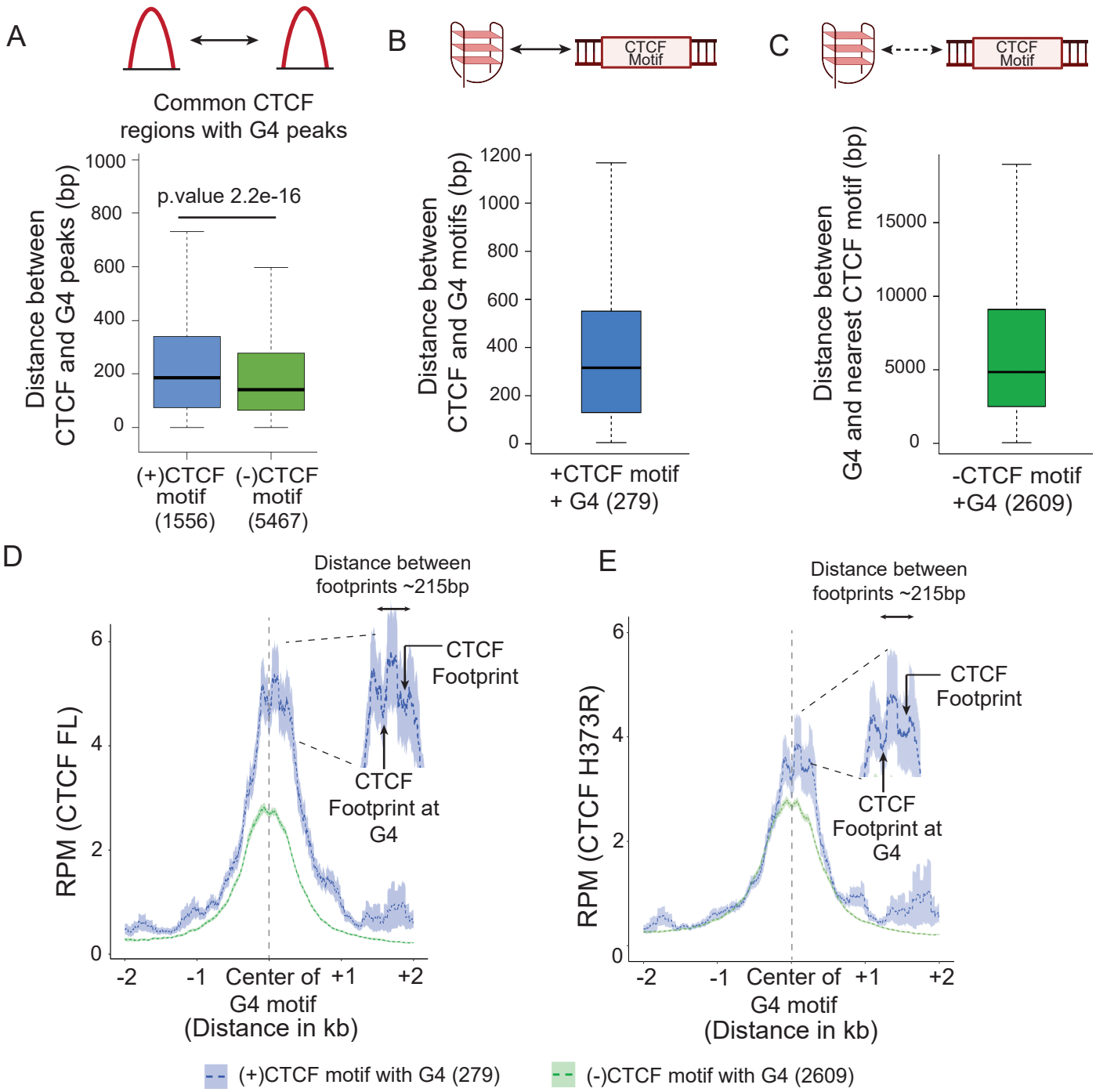
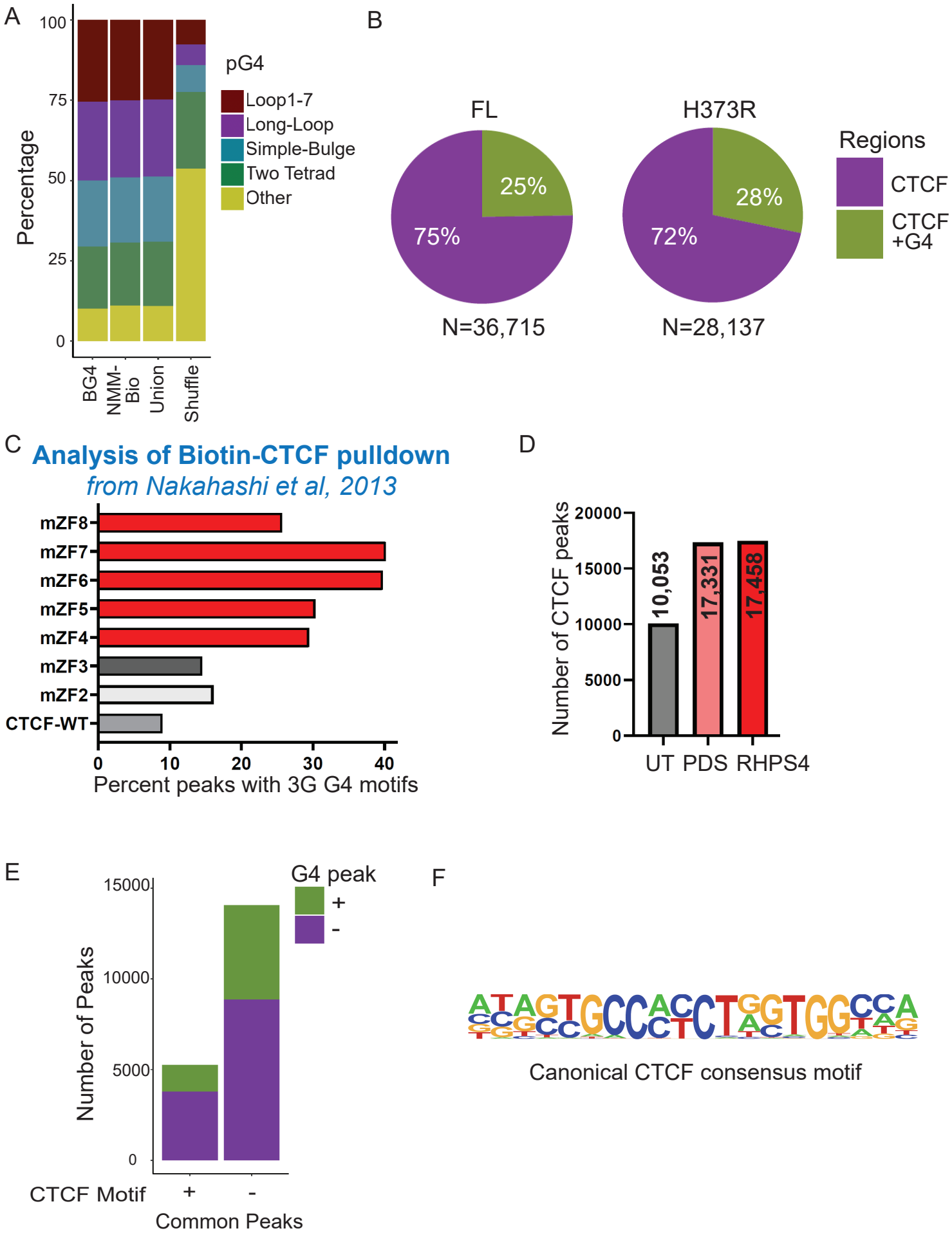

A
Common CTCF regions with G4 peaks
Distance between CTCF and G4 peaks (bp)
p.value 2.2e-16
1000
800
600
400
200
0
(+)CTCF motif (1556)
(-)CTCF motif (5467)
B
CTCF Motif
Distance between CTCF and G4 motifs (bp)
1200
1000
800
600
400
200
0
+CTCF motif + G4 (279)
C
CTCF Motif
Distance between G4 and nearest CTCF motif (bp)
15000
10000
5000
0
-CTCF motif +G4 (2609)
D
Distance between footprints ~215bp
CTCF Footprint
CTCF Footprint at G4
RPM (CTCF FL)
6
4
2
0
-2
-1
Center of G4 motif
+1
+2
(Distance in kb)
E
Distance between footprints ~215bp
CTCF Footprint
CTCF Footprint at G4
RPM (CTCF H373R)
6
4
2
0
-2
-1
Center of G4 motif
+1
+2
(Distance in kb)
(+)CTCF motif with G4 (279)
(-)CTCF motif with G4 (2609)

Figure 5 CTCF directly associated with genomic G4s

A. Box and whisker plot quantifying the distance between the centers of the overlapping CTCF and G4 peaks. The regions were further separated based on the presence (+CTCF motif) or absence (-CTCF motif) of a canonical CTCF motif within the peak. The statistical significance was calculated using Wilcoxon signed-rank test.
B. Box and whisker plot quantifying the distance between the center of the CTCF motifs (+CTCF motif) and the 3G-G4 motif within common (FL and H373R enriched) CTCF peaks that overlap a G4 peak.
C. Box and whisker plot quantifying the distance between the 3G-G4 motif to the nearest CTCF motif in common (FL and H373R) CTCF peaks that overlap a G4 peak.
D. Histograms displaying the signal from DNA footprinting analysis at overlapping CTCF and G4 peaks separated into regions with (blue) or without (green) a canonical CTCF motif. The FL CTCF footprints are centred at the 3G-G4 motif within G4 peaks. The zoomed in area is the center of G4 peak with canonical CTCF motif. The shaded area represents standard deviation.
E. Histograms displaying the signal from DNA footprinting analysis at overlapping CTCF and G4 peaks separated into regions with (blue) or without (green) a canonical CTCF motif. The H373R CTCF footprints are centred at the 3G-G4 motif within G4 peaks. The zoomed in area is the center of G4 peak with canonical CTCF motif. The shaded area represents standard deviation.

Supplementary figure 4



Supplementary Figure 4

A. Barplot showing different classes of predicted G4 motifs in BG4, NMM-Bio, or union of BG4 and NMM-Bio peaks that overlap regions identified in G4-seq. Genomic shuffle regions are used as controls. Please also see methods for details about the respective G4 motifs.
B. Pie-charts showing the percent of CTCF peaks that overlap with G4 peaks (green) identified in mESCs expressing FL or H373R CTCF.
C. Bargraph quantifying percent CTCF peaks with 3G-G4 motifs ($G_3N_{1-7}G_3N_{1-7}G_3N_{1-7}G_3$). CTCF ChIP-seq was generated with BirA-tagged wildtype (WT) and ZF mutant CTCF proteins overexpressed in B cells[39]. Each ZF from ZF2 to ZF8 was mutated at the second histidine in the C2H2 motif.
D. Quantification of the number of CTCF peaks identified using ChIP–seq in B cell line CH12 cells treated with or without G-quadruplex stabilizing ligands PDS (10μM) and RHPS4 (2μM) for 24 hours.
E. Bar plot showing the number of common (FL and H373R) CTCF peaks separated based on the presence (+) or absence (-) of a canonical CTCF motif. The peaks were further categorized based on their overlap with (+) or without (-) a G4 peak.
F. A weighted matrix of the canonical CTCF motif used in the analysis in panel E. and Figure 5.

Figure 6

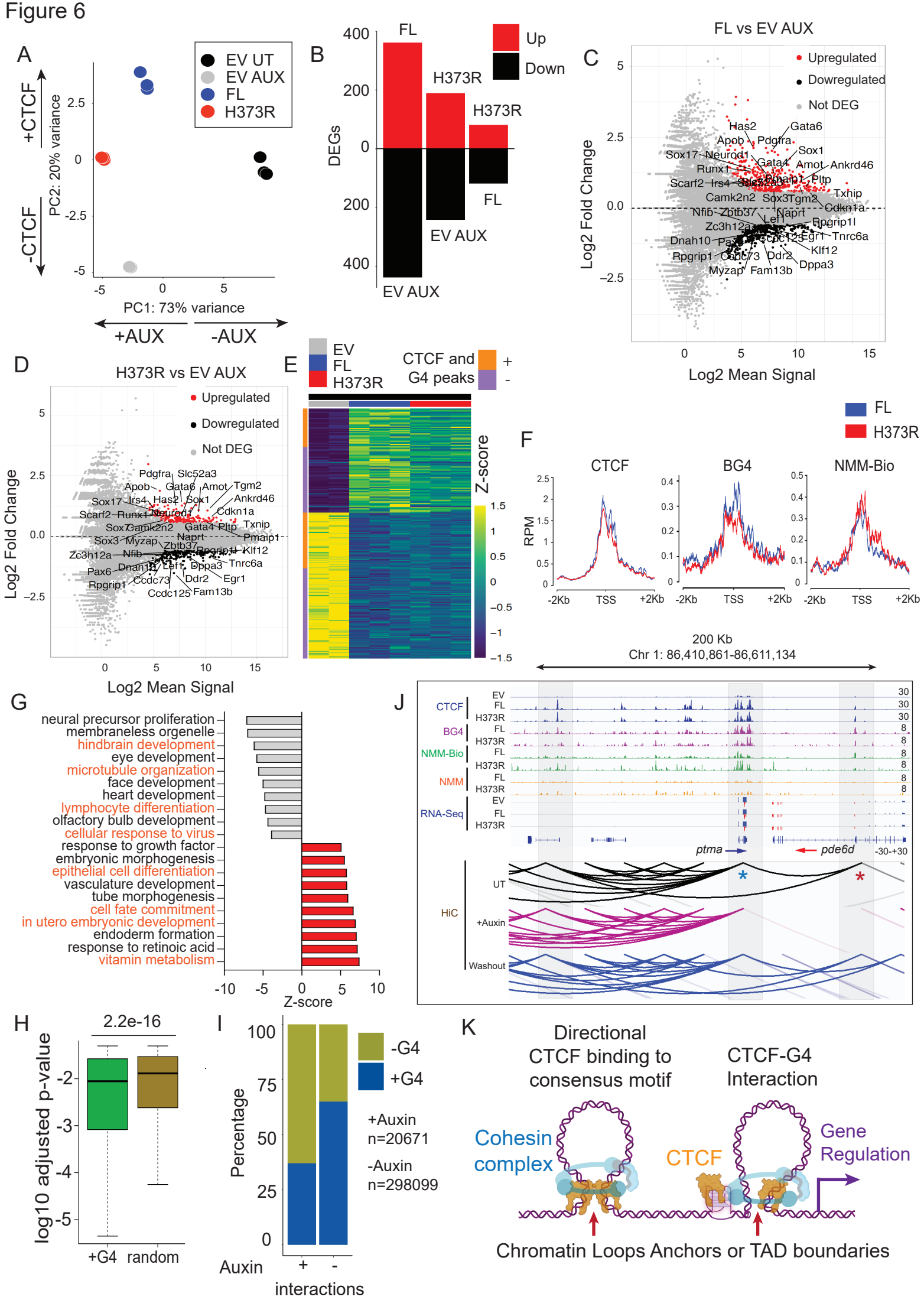
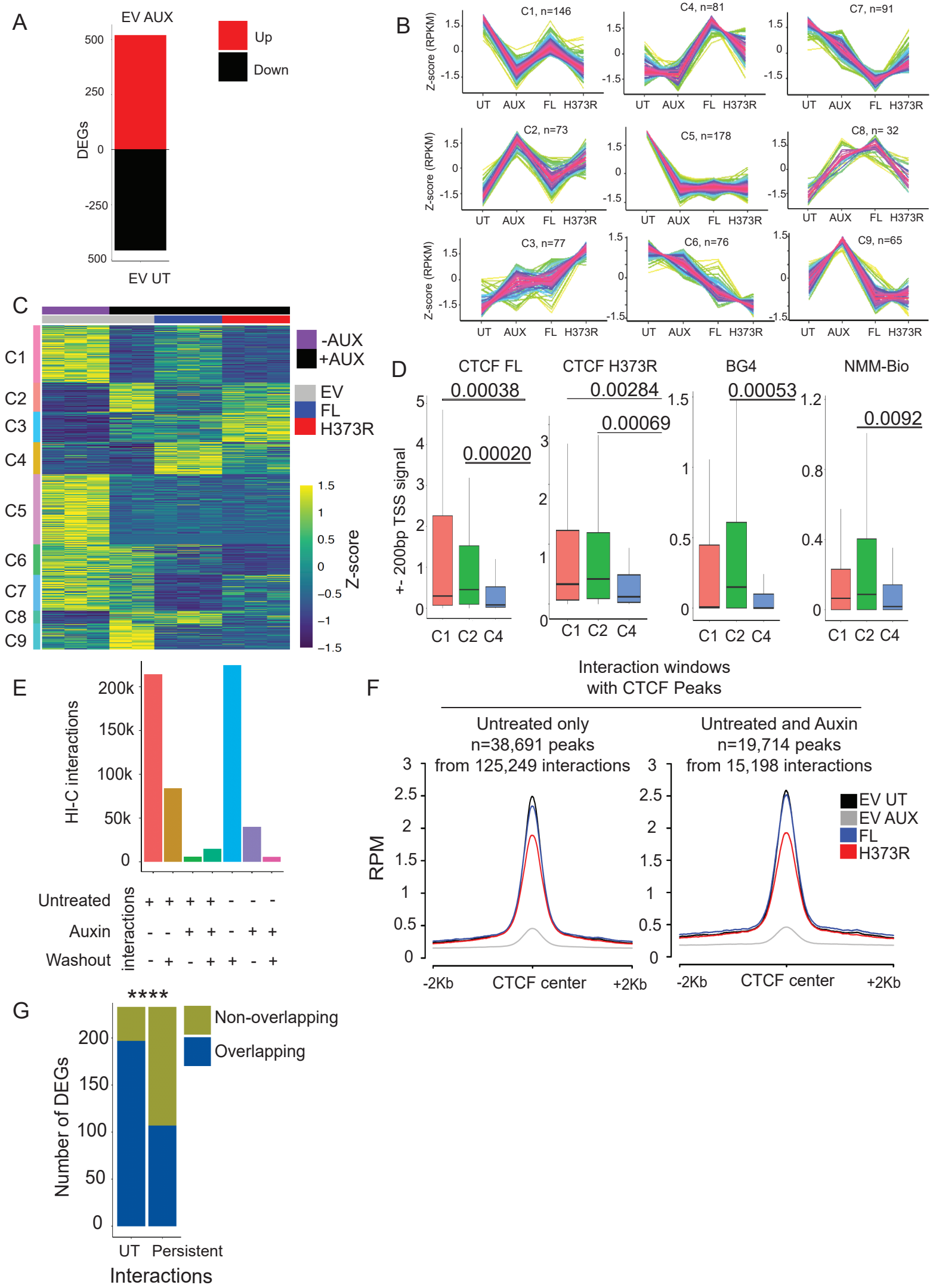

A
EV UT
EV AUX
FL
H373R
+CTCF
-CTCF
PC2: 20% variance
PC1: 73% variance
+AUX
-AUX
B
FL
H373R
H373R
FL
EV AUX
EV AUX
Up
Down
DEGs
C
FL vs EV AUX
Upregulated
Dowregulated
Not DEG
Log2 Fold Change
Log2 Mean Signal
D
H373R vs EV AUX
Upregulated
Dowregulated
Not DEG
Log2 Fold Change
Log2 Mean Signal
E
EV
FL
H373R
CTCF and G4 peaks
+
-
Z-score
F
FL
H373R
CTCF
BG4
NMM-Bio
RPM
-2Kb
TSS
+2Kb
200 Kb
Chr 1: 86,410,861-86,611,134
G
neural precursor proliferation
membraneless orgenelle
hindbrain development
eye development
microtubule organization
face development
heart development
lymphocyte differentiation
olfactory bulb development
cellular response to virus
response to growth factor
embryonic morphogenesis
epithelial cell differentiation
vasculature development
tube morphogenesis
cell fate commitment
in utero embryonic development
endoderm formation
response to retinoic acid
vitamin metabolism
Z-score
J
CTCF
BG4
NMM-Bio
NMM
RNA-Seq
EV
FL
H373R
ptma
pde6d
-30-+30
HiC
UT
+Auxin
Washout
H
2.2e-16
log10 adjusted p-value
+G4
random
I
Percentage
-G4
+G4
+Auxin
n=20671
-Auxin
n=298099
Auxin
+
-
interactions
K
Directional CTCF binding to consensus motif
CTCF-G4 Interaction
Cohesin complex
CTCF
Gene Regulation
Chromatin Loops Anchors or TAD boundaries

Figure 6 CTCF-G4 interactions mediate specific gene expression changes

A. Principal component analysis of RNA-seq data of mESCs with EV untreated (UT) (black), EV auxin (AUX) (gray) and FL auxin (blue) and H373R auxin (red). Percent variance for principal components 1 (PC1) and 2 (PC2) are shown on X and Y-axis respectively. Arrows in each axis represent the features explaining the variance: with or without auxin (PC1) and presence or absence of CTCF (PC2).
B. Barplot showing the numbers of differentially expressed genes (DEGs) in pairwise comparisons between full length (FL) vs Empty vector (EV)-AUX, H373R vs EV-AUX and H373R vs FL. Genes are separated into up (red) and downregulated (black)
C. MA plots showing the DEGs in FL CTCF vs EV-AUX. Red and black points are upregulated and downregulated genes respectively. X-Axis represents Log2 base mean signal and Y-axis denotes Log2 fold change in expression.
D. MA plots showing the DEGs in H373R CTCF vs EV-AUX. Red and black points are upregulated and downregulated genes respectively. X-Axis represents Log2 base mean signal and Y-axis denotes Log2 fold change in expression.
E. Heatmap showing the Z-score from the RPKM expression values of 233 common DEGs between FL vs EV-AUX and H373R vs EV-AUX comparisons.
F. Profile plot showing the CTCF, BG4 and NMM-Bio signal as Reads Per Million (RPM) at the transcription start sites (TSS) ±2 kb within the 233 common DEGs in FL vs EV-AUX and H373R vs EV-AUX.
G. Downregulated (gray) and upregulated (red) pathways in the 233 common DEGS in FL vs EV-AUX and H373R vs EV-AUX. X-axis represents Z-score. Pathways marked in orange are enriched for genes that show overlap with both CTCF and G4 peaks.
H. Box and whisker plot comparing log10 p-values from Hi-C interactions (untreated condition) that overlap G4 peaks (green; G4) or genomic shuffle regions (random; olive) as controls. The statistical significance was calculated using Wilcoxon signed rank test.
I. Barplot quantifying Hi-C interactions with or without auxin treatment (for 48 hours) in AID-tagged mESCs. The Y-axis represents the percent of interactions with (blue) or without (olive) an overlapping G4 peaks identified using NMM-Bio or BG4 in mESCs reconstituted with FL or H373R CTCF treated with auxin. The unique interactions that remained post-auxin treatment (n=20,671) were compared to unique interactions in untreated conditions (n=298, 099).
J. Representative genome browser view of CTCF (blue), BG4 (purple), NMM-Bio (Green), NMM (Orange) and RNA-seq (blue/red) at *Ptma* and *Pde6d* gene loci. The expression of *Pde6d* gene is positively regulated by FL CTCF and H373R mutant. Shaded areas define regions of interest with CTCF and G4 peaks. Bottom part is HiC contact maps from Nora et.al.[35] in the parent mESCs without auxin (untreated, UT), with auxin treatment for 2 days (+Auxin) and 2 days following auxin washout (Washout). The red and blue asterisks highlight the interaction anchor points that are lost or retained upon auxin treatment. The same parent mESCs were used in our studies.
K. A proposed model depicting the role of CTCF-G4 interaction in mediating long-range chromatin looping to define stable chromatin loops or boundaries of topologically associating domains (TADs).

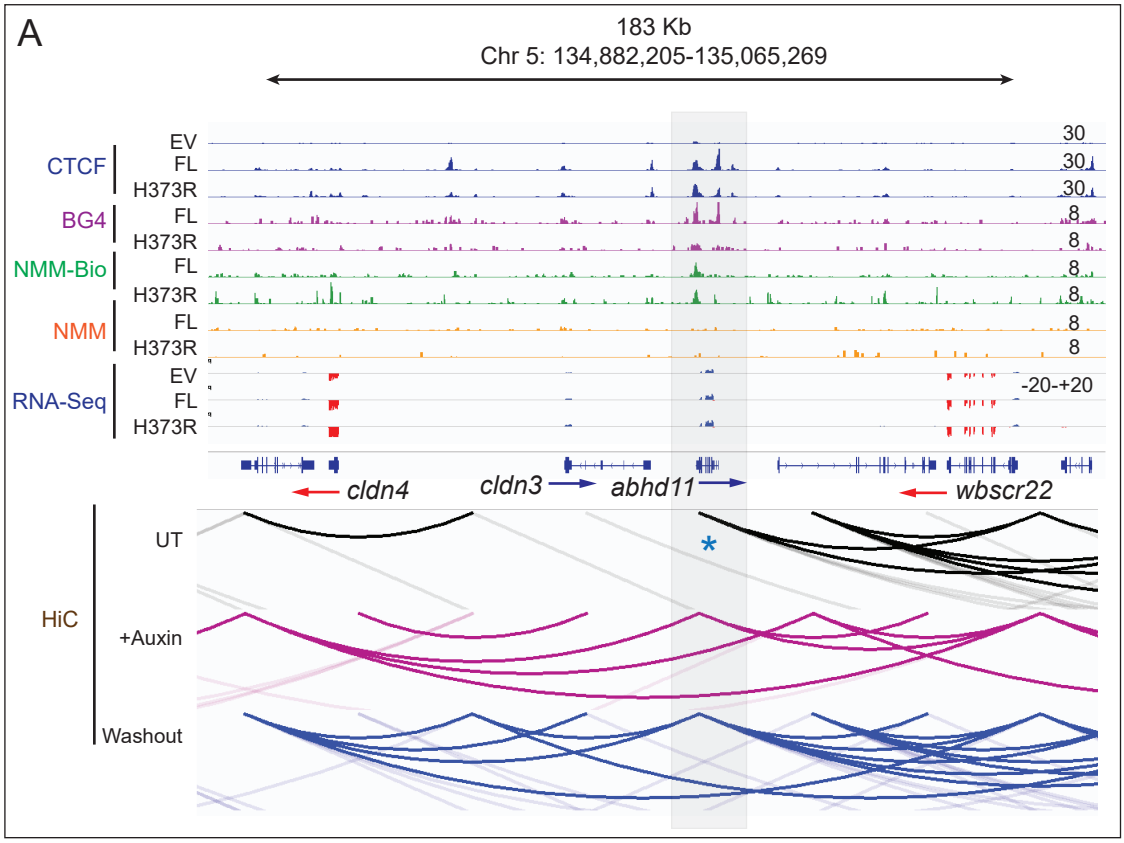
Supplementary figure 5

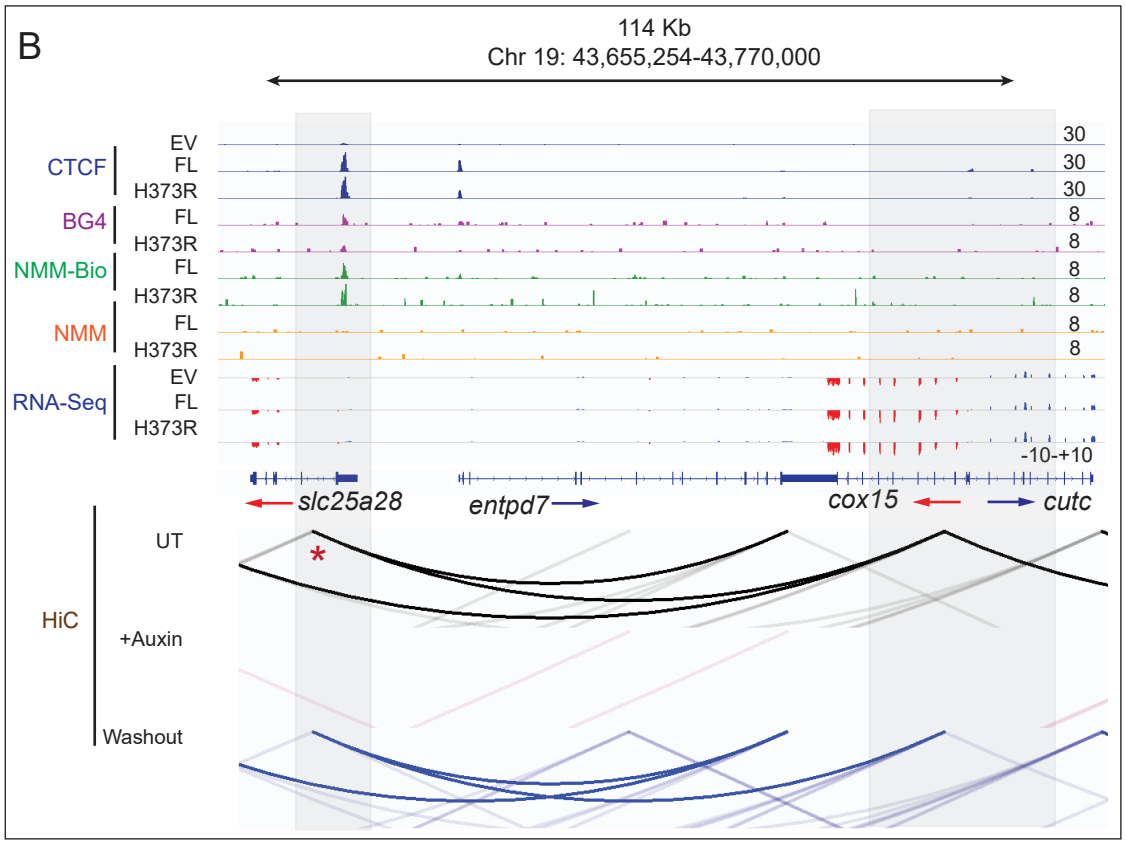


Supplementary Figure 5

A. Quantification of the differentially expressed genes in EV-AUX vs EV-UT pair-wise comparison. Genes are separated into up (red) and downregulated (black)
B. Analysis of temporal gene expression modules with TC-seq. Gene expression profiles were analyzed in EV-UT, EV-AUX, FL and H373R cells and genes were clustered based on their temporal expression patterns. Nine clusters (C1-C9), or expression patterns, were identified. Y-axis indicates the Z-score calculated using RPKM. X-axis indicates the sample condition.
C. Heatmap showing the Z-score from the RPKM expression of EV-UT, EV-AUX, FL and H373R in the nine clusters (C1-C9) identified in TC-seq.
D. Box and whisker plots comparing the signal of CTCF FL, CTCF H373R, BG4 FL and NMM-Bio within ± 200bp of the TSS of genes in clusters C1, C2, and C4. The statistical significance was calculated using Wilcoxon signed-rank test.
E. Barplot reporting the total number of Hi-C interactions in untreated (UT), auxin and washout conditions. The conditions are represented as + or – to highlight conditions in which the Hi-C interactions were identified.
F. Histograms displaying CTCF signal from EV-UT, EV-AUX, FL and H373R ESCs plotted in Hi-C interactions with windows overlapping CTCF peaks in untreated or persistent interactions in untreated and auxin treated conditions.
G. Barplot showing the total number of DEGs (promoters) that overlap (blue) or do not overlap (non-overlapping; olive) Hi-C interactions from untreated (UT) condition or those that persist (persistent) upon auxin treated conditions. The statistical significance is calculated using Chi-square test, **** p-value <0.00001.

Supplementary figure 6



Supplementary Figure 6

**A and B**. Representative genome browser view of CTCF (blue), BG4 (purple), NMM-Bio (Green), NMM (Orange) and RNA-seq (blue/red) at *Cldn3* (A) and *Cox15* (B) gene loci. Shaded areas define regions of interest with CTCF and G4 peaks. Bottom part is HiC contact maps in the parent mESCs (AID-tagged CTF) without auxin (untreated, UT), with auxin treatment for 2 days (+Auxin) and 2 days following auxin washout (Washout). The red and blue asterisks highlight the interaction anchor points that are lost or retained upon auxin treatment.