Title: Co-opted transposons help perpetuate conserved higher-order chromosomal structures

Short title: Co-opted TEs maintain conserved 3D genome

One-sentence summary: Co-option of transposable elements maintains conserved 3D genome structures via CTCF binding site turnover in human and mouse.

Authors: Mayank NK Choudhary[1], Ryan Z Friedman[1], Julia T Wang[1], Hyo Sik Jang[1], Xiaoyu Zhuo PhD[1], Ting Wang PhD[1,*]

Affiliations: [1]The Edison Family Center for Genome Sciences & Systems Biology, Department of Genetics, Washington University, St. Louis, MO

[*]Corresponding author:
Ting Wang, PhD, Department of Genetics,
Washington University School of Medicine, 4515 McKinley Avenue
Campus Box 8510
St. Louis, MO 63110, USA.
Phone: (314) 286-0865
E-mail address: twang@genetics.wustl.edu

_2065_ Words
_0_ Table
_4_ Figures

ABSTRACT (125 words)

Transposable elements (TEs) shape genome regulation by harboring regulatory factor binding sites—including CTCF-binding sites known to tether chromatin loops and mark domain boundaries. The mechanisms underlying the emergence and evolution of these 3D genome structures remain unprobed. We find that TEs contribute extensively to emergence of species-specific loops in humans and mice by depositing novel anchor CTCF motifs, as well as to maintenance of conserved loops via CTCF binding site turnover. Deleting such TEs in human cells leads to collapse of conserved 3D structures. TEs have long been considered a source of genetic innovation; by comparing topologies, we show that TEs contribute regulatory plasticity that helps maintain conserved genome architecture—revealing a new paradigm for defining noncoding, regulatory conservation beyond classic DNA sequence preservation.

38 The 3D organization of various genomes has been mapped at high resolution
39 using a variety of methods (*1-5*). While genome folding is largely conserved in mammals
40 (*1,4*), the genetic forces shaping its emergence and evolution remain poorly understood.
41 Two distinct yet mutually non-exclusive models (*6*) have recently gained much traction:
42 that of phase separation (*7*) and of loop extrusion (*8,9*) by factors such as CTCF. In
43 relation to the latter, TEs are known to contain and disseminate functional regulatory
44 sequences (*10-13*) including that of CTCF. In contrast to relying on point mutations to
45 evolve a functional CTCF binding site, TE transposition presents an attractive model for
46 rapid regulatory sequence dissemination and regime building (*14-17*). Hence, we
47 hypothesized that TEs have been a rich source of sequence for the assembly and
48 tinkering of higher-order chromosomal structures. We studied the influence of all
49 repetitive elements (REs) in establishing higher-order chromosomal structures and, more
50 specifically, the role of TEs in their evolution.

51 We examined REs' contribution to loop anchor CTCF sites using published
52 genome-wide chromosomal conformation capture data from assays including ChIA-PET
53 (*2*) and Hi-C in human (GM12878, HeLa, HMEC, IMR90, K562, NHEK) and mouse (CH12-
54 LX) cell lines (*1*). We determined that 398 out of 3159 (12.6%) unique loop-anchoring
55 CTCF sites were derived from REs in the mouse lymphoblastoid cell line. These RE-
56 derived CTCF sites help establish 451 out of 2718 (16.6%) loops with discernible, unique
57 CTCF loop anchors (Fig 1A, B). In the corresponding human lymphoblastoid cell line,
58 REs contributed 935 out of 8324 (11.2%) unique loop-anchoring CTCF sites that help
59 establish 1244 out of 8007 (15.6%) loops. Overall, in humans, REs contributed 9-15%
60 of the anchor CTCF sites that result in 12-18% loops across a variety of cell lines (Fig
61 1A, B).

62 In both species, RE-derived loop-anchor CTCF sites were largely derived from
63 TEs (>95%) and their class of origin (SINE, LINE, LTR, DNA) showed a species-biased
64 distribution (Fig 1C). Using the highest resolution *in-situ* HiC maps in matched
65 lymphoblastoid cell types in mice (CH12-LX) and humans (GM12878), we compared the
66 composition of the RE-derived loop anchor CTCF sites. While the mouse lineage was
67 profoundly shaped by the SINEs (70%, 4x enrichment over background), the human

68 lineage was overrepresented by retroviral LTR elements and DNA transposons (36% and
69 22%, 2x and 3x enriched over the background respectively) (Fig 1D). At the family level,
70 the B2 SINEs in mice were 13-fold enriched over background and contributed 65% of
71 loop anchor CTCF sites. In humans, the hAT-Charlie family of DNA transposons
72 contributed 13% of loop anchor CTCF sites, a 4-fold enrichment over background (Fig
73 1E). These contributions are underestimates as we have yet to (i) uniquely identify all
74 loop anchor CTCF sites (especially in repetitive regions), and (ii) annotate all repetitive
75 elements, especially ancient TEs that have diverged far from their identity (*18*).

76 To study the evolution of chromatin loops, we compared their conservation
77 (Supplementary Methods) in matched human and mouse cell-types. Briefly, we used the
78 liftOver tool (*19*) to compare loops across species and required exactly one reciprocal
79 match (reciprocal best hit) to designate conserved loops. We found that 48% of all
80 mouse loops (1596 out of 3331) had a loop call in the corresponding syntenic region in
81 humans (Table S1.1). Our observation is in close agreement with prior studies (*1,4*) that
82 show about half of all higher-order chromosomal structures to be conserved. We then
83 sought to characterize the contribution of TEs to various classes of loops based on their
84 orthology.

85 We compared the origin of anchor CTCF sites of orthologous loops in mouse and
86 human. We found that out of 1596 orthologous loops, 142 (8.9%) in mouse and 108
87 (6.7%) in human had at least one TE-derived anchor CTCF site (Fig 2A). In addition to
88 orthologous loops, TE-derived anchor CTCF sites also gave rise to 24% (409 out of 1735)
89 and 15% (1136 out of 6716) non-orthologous (species-specific) loops in mouse and
90 humans, respectively (Fig 2B), consistent with the appreciable role of TEs in genome
91 innovation (*14-16,20,21*). Overall, the majority of TE-derived loop anchors in mouse were
92 established by a handful of young TE subfamilies (B3, B2_Mm2, B3A, B2_Mm1t) that
93 expanded in the rodent lineage (*22*). In contrast, multiple TE subfamilies of varying
94 evolutionary ages contributed diffusely to CTCF loop anchors in humans. Altogether,
95 TEs in humans contributed to fewer orthologous loops and distributed over more TE
96 subfamilies than in mouse (Fig 2C).

97 Intriguingly, 123/142 (87%) TE-derived orthologous loops in mouse were
98 discordant for TEs in humans (Table S1.2). As in: while the loops in humans were
99 anchored at the ancestral CTCF binding sites, the syntenic ancestral CTCF motifs were
100 either degraded or deleted in mouse, where the loops were now anchored at CTCF sites
101 derived from nearby, co-opted TEs instead. One such example is an orthologous loop
102 (as well as domain border) at the 3' end of the Gclc gene (Fig 2D) maintained in mouse
103 by a B3 SINE insertion ~100bp downstream of the degraded ancestral motif which was
104 well conserved in most non-rodent mammals (Figure S1). The degradation of the
105 ancestral CTCF motif incapacitates CTCF binding as evidenced by the CTCF-ChIP track
106 (Fig 2E); in contrast, the B3 element harbored strong CTCF binding, providing an
107 anchoring site to maintain the conserved loop in mouse. Similarly, we find that 89/108
108 (82%) TE-derived orthologous loops in human GM12878 cells were discordant for TEs
109 in mouse (Table S1.3). We hypothesized that TEs mediated binding site turnover for
110 CTCF and contributed to conserved genome folding events between human and mouse.
111 Since the mouse genome is replete with repeat-derived CTCF sites (*22*) that could
112 interfere with the targeted study of a specific TE candidate, we decided to validate these
113 hypotheses in human cell lines.
114 Here we examine two candidate TEs, both of which maintain conserved higher-
115 order chromosomal structures in humans: one belonging to the L1M3f subfamily of
116 LINEs, and the other belonging to the LTR41 subfamily of endogenous-retrovirus-
117 derived long terminal repeat (LTR). The former TE replaces the function of a lost ancestral
118 CTCF site (Figure S2), while the latter is functionally redundant for an ancestral CTCF
119 site still present in humans (Figure S3). Using CRISPR-Cas9, we obtained clones of
120 GM12878 cells bearing homozygous deletions of the L1M3f and LTR41 elements,
121 respectively (Figure S4, Table S2.4). We then performed HYbrid-Capture on the *in situ*
122 Hi-C library (Hi-C^2) to examine the effect of the TE deletion on the local 3D structure (*8*)
123 (Table S2.1, S2.2, S2.3).
124 The L1M3f-derived CTCF site was positioned at a conserved domain border and
125 anchored three chromatin loops (Figure S2). Upon deletion of this L1M3f, the conserved
126 local chromosomal structure collapsed as evidenced by (i) the loss of focal enrichment

127 in the homozygous TE knockout (KO) contact map in comparison to the wild-type (WT)
128 contact map, and (ii) the fusion of two neighboring domains (Hi-C^2 results: Fig 3A, Hi-C
129 results: Figure S5). The Virtual 4C plot anchored at the region surrounding the L1M3f
130 element showed three distinct peaks (corresponding to the three loops in the WT cell
131 line), which were lost in the KO (ΔL1M3f) cell line. We also found that cross-domain
132 interactions significantly increased from 8% in WT to 19% in KO cell lines (~2.4x, Welch's
133 t-test p-value<1.5×10^{-16}, Table S2.5) across the L1M3f-established domain boundary, a
134 change specific to the targeted domain and not seen in a control domain from a nearby
135 region (Fig 3C). Thus, the L1M3f element is necessary for maintaining the conserved
136 loops and domain boundary in humans. It represents a novel class of binding site
137 turnover (*23-26*) for CTCF leading to conservation in terms of function via establishment
138 of long-range interactions and potentially the underlying gene regulation, but not in
139 primary local sequence.

140 Our second example was a species-specific LTR41-derived CTCF site ("c") that
141 is functionally interchangeable for an ancestral CTCF site derived from a much older TE
142 called MER82 ("d"), which had reduced CTCF binding strength (Figure S3B) and thus
143 was mostly decommissioned. In the WT contact map, we observed a bright focal
144 enrichment corresponding to CTCF binding sites a-c suggesting a looping interaction.
145 In contrast, there was little focal enrichment corresponding to a-d (Fig 3D, top row).
146 Additionally, in the WT Virtual 4C track anchored on "a", we observed a clear peak
147 corresponding to LTR41 ("c") suggesting an a-c loop (Fig 3E). Upon deletion of LTR41,
148 the conserved loop's anchor is offset to the MER82-derived CTCF site ("d") downstream
149 of the LTR41 as evidenced by the shift in the focal enrichment in the KO contact map
150 (Fig 3D, bottom row) and an increase in the KO Virtual 4C peak corresponding to the
151 MER82-derived CTCF site (i.e., a-d loop) (Fig 3E). Upon anchoring the Virtual 4C on a 5-
152 kb window containing LTR41 (c), we observed a peak loss at "a" corresponding to the
153 loss of the a-c loop in the KO, an interaction that existed in the WT cells (Fig 3F). With
154 the ~39kb shift of the anchor site, the half-megabase scale chromosomal structure
155 around the anchor region remained largely preserved (Figure S3C). Upon deletion of this
156 TE candidate, the local sequence configuration most likely resembled that of the pre TE-

157 insertion, ancestral genome. This example therefore illustrates a potential path by which
158 the local 3D genome evolved upon insertion of the LTR41 element as well as the
159 plasticity TEs (like LTR41 and MER82 in this case) bring to their host genomes.

160 These results support the hypothesis that TEs are able to contribute regulatory
161 robustness and strengthen conserved regulatory architecture as redundant or "shadow"
162 loop anchors. One can imagine that the mouse genome that underwent a lineage-
163 specific expansion of SINE B2s (*22*), which carry a CTCF binding site, to be saturated
164 with such events.

165 TEs are typically silenced by host repressive machineries including DNA and
166 histone methylation (*27-29*). However, a small fraction of TEs escape epigenetic silencing
167 and provide functional regulatory elements for the host in a process termed exaptation
168 (*30-33*). Since CTCF is a methylation sensitive chromatin factor and only binds to
169 unmethylated DNA (*34,35*), we examined the DNA methylation levels of anchor CTCF
170 sites of orthologous loops. We found that TE-derived CTCF sites were marked by
171 reduced DNA methylation, similar to their non-TE derived genomic counterparts (Fig 4A).
172 To understand the DNA methylation dynamics through evolution, we took advantage of
173 the differential mutation rate of 5-methylcytosine (5mC) to Thymine (T) (*36*).
174 Unmethylated cytosines (C) mutate to T at a lower rate than 5mC; thus, methylated DNA
175 exhibits higher frequency of C to T mutations (*37*). We found that TEs involved in turnover
176 events had a significantly lower frequency of methylation-associated C-to-T and G-to-A
177 mutations compared to an identically sampled background of TEs not involved in looping
178 (1000 simulations), but no difference in all other combined substitutions (summarized
179 human results: Fig 4B; full human and mouse results: Figure S6, S7, Table S3). These
180 results suggest that TEs providing CTCF turnover were hypomethylated over
181 evolutionary time to maintain their functional role, compared to other TE copies (Fig 4C).

182 Taken together, our findings reveal a formerly uncharacterized role that TEs have
183 played in the evolution of higher-order chromosomal structures in mammals. TEs have
184 contributed a substantial number of loop anchors in mouse and human 3D genomes, a
185 fraction of which were co-opted to help maintain conserved higher-order chromosomal
186 structures. TE transposition provides a novel method for CTCF binding site turnover to

maintain regulatory conservation (defined here as the preservation of long-range chromosomal interactions), by compensating for the loss of local primary sequence—local sequence that would have otherwise allowed the assessment of purifying selection. Deletion of these TEs in human cell lines eliminated the chromatin loops that they anchor and even resulted in collapse of conserved chromatin structure, as expected by our hypothesis. More strikingly, we demonstrate that in another case the loop anchor shifted to an alternative TE-derived CTCF site nearby, resulting in largely unchanged chromatin structure, underscoring the dynamic nature and robustness of the 3D genome. These TEs that maintain conserved chromatin loops via turnover are hypomethylated through evolutionary time, an observation that highlights the intimate interplay between genome, epigenome, and 3D genome in evolution. This research provides a foundation to study the impact of TEs and expand our understanding of chromosomal folding—its emergence, maintenance and transformation—in the context of evolving genomes. Ultimately, our study reveals how selfish genetic elements, regardless of their origins, can be repurposed to maintain latent genome sanctity and regulatory fidelity by conserving 3D structure.

REFERENCES:

1. SSP Rao et al., *Cell* **159**, 1665-1680 (2014).
2. Z Tang et al., *Cell* **163**, 1611-1627 (2015).
3. T Sexton et al., *Cell* **148**, 458-472 (2012).
4. JR Dixon et al., *Nature*. **485**, 376-380 (2012).
5. DU Gorkin et al., *Cell Stem Cell* **14**, 771-775 (2014).
6. W Schwarzer et al., *Nature* **551**, 51–56 (2017).
7. AR Strom et al., *Nature* **547**, 241–245 (2017).
8. AL Sanborn et al., *Proc Natl Acad Sci*. **112**, E6456-E6465. (2015).
9. G Fudenberg et al., *Cell Rep*. **15**, 2038–2049 (2016).
10. V Sundaram et al., *Genome Res*. **24**, 1963-1976 (2014).
11. G Bourque et al., *Genome Res*. **18**, 1752-1762 (2008).
12. G Kunarso et al. *Nat Genet*. **42**, 631-634 (2010).
13. PÉ Jacques et al., *PLoS Genet*. **9**(5) (2013).
14. T Wang et al., *Proc Natl Acad Sci*. **104**, 18613-18618 (2007).
15. EB Chuong et al., *Science* **351**, 1083-1087 (2016).
16. VJ Lynch et al., *Nat Genet* **43**, 1154-1159 (2011).
17. BJ Britten, EH Davidson. *Q Rev Biol*. **46**, 111-138 (1971).
18. APJ de Koning et al., *PLoS Genet*. **7** (2011).
19. AS Hinrichs, *Nucleic Acids Res*. **34**, 590-598 (2006).
20. C Feschotte, EJ Pritham. *Annu Rev Genet*. **41**, 331-368 (2007).
21. V Sundaram, T Wang. *BioEssays*. **40**, 1700155 (2018).
22. D Schmidt et al., *Cell* **148**, 335-348 (2012).
23. MZ Ludwig et al., *Development* **12**, 3325–3330 (1998).
24. AM Moses et al., *PLoS Comput Biol*. **2**, 130 (2006).
25. S Venkataram, JC Fay. *Genome Biol Evol*. **2**, 851–858 (2010).
26. D Villar et al., *Nat Rev Genet*. **15**, 221-233 (2014).
27. MA Matzke et al., *Plant Mol Biol*. **43**, 401-415 (2000).
28. JA Yoder et al., *Trends Genet*. **13**, 335-340 (1997).
29. RK Slotkin et al., *Nat Rev Genet*. **8**, 272-285 (2007).
30. A Huda et al., *Mob DNA* **1**, 2-12 (2010).
31. CB Lowe, D Haussler. *PLoS One* **7**, 43128 (2012).
32. G Bejerano et al., *Nature* **441**, 87–90 (2006).
33. MG Kidwell, DR Lisch. *Trends Ecol Evol*. **15**, 95-99 (2000).
34. C Kanduri et al., *Curr Biol*. **10**, 853-856 (2000).
35. S Kurukuti et al., *Proc Natl Acad Sci*. **103**, 10684-10689 (2006).
36. JC Shen et al., *Nucleic Acids Res*. **22**, 972-976 (1994).
37. AP Bird. *Nucleic Acids Res*. **8**, 1499–1504 (1980).
38. RW Meredith et al., *Science* **334**, 521-524 (2011).
39. RH Waterston et al., *Nature* **420**, 520-562 (2002).
40. M Haeussler et al., *Genome Biol.* **17**, 148 (2016).
41. MA Moreno-Mateos et al., *Nature Methods* **12**, 982-988 (2015).
42. W Bao et al., *Mob DNA* **6**, 11 (2015).
43. A Smit et al., RepeatMasker Open-4.0.6 2013-2015. (2017).

44. P Rice et al., *Trends Genet.* **16**, 276-277 (2000).

Acknowledgements We thank members of the Wang Lab for helpful discussions related to the project; Jessica Hoisington-López and Maria Lynn Jaeger from The Edison Family Center for Genome Sciences & Systems Biology for assistance with sequencing; Matthew Patana & Daniel Schweppe from the Siteman Flow Cytometry core for FACS expertise. M.N.K.C. was partly supported by the Precision Medicine Pathway, Washington University. T.W. is supported by NIH grants 5R01HG007354, 5R01HG007175, 5R01ES024992, 2U01CA200060, U24ES026699, U01HG009391 and American Cancer Society RSG-14-049-01-DMC.

Supplementary Materials

Materials and Methods

Table S1 – S3

Figure S1 – S7

Author Contributions M.N.K.C. and T.W. conceived and designed this study; M.N.K.C. analyzed the data, performed experiments, generated sequencing libraries, and wrote the manuscript with inputs from T.W.; R.Z.F., J.T.W., and X.Z. contributed text and revised the manuscript; H.S.J. contributed reagents and resources; R.Z.F performed mutation frequency simulations along with M.N.K.C.; X.Z. generated TE ancestral sequences and TE alignments; T.W. supervised the project. All authors subsequently edited and approved the final manuscript.

Competing Interests

Authors declare no competing interests.

Corresponding Author

Correspondence and requests for materials should be addressed to Ting Wang.

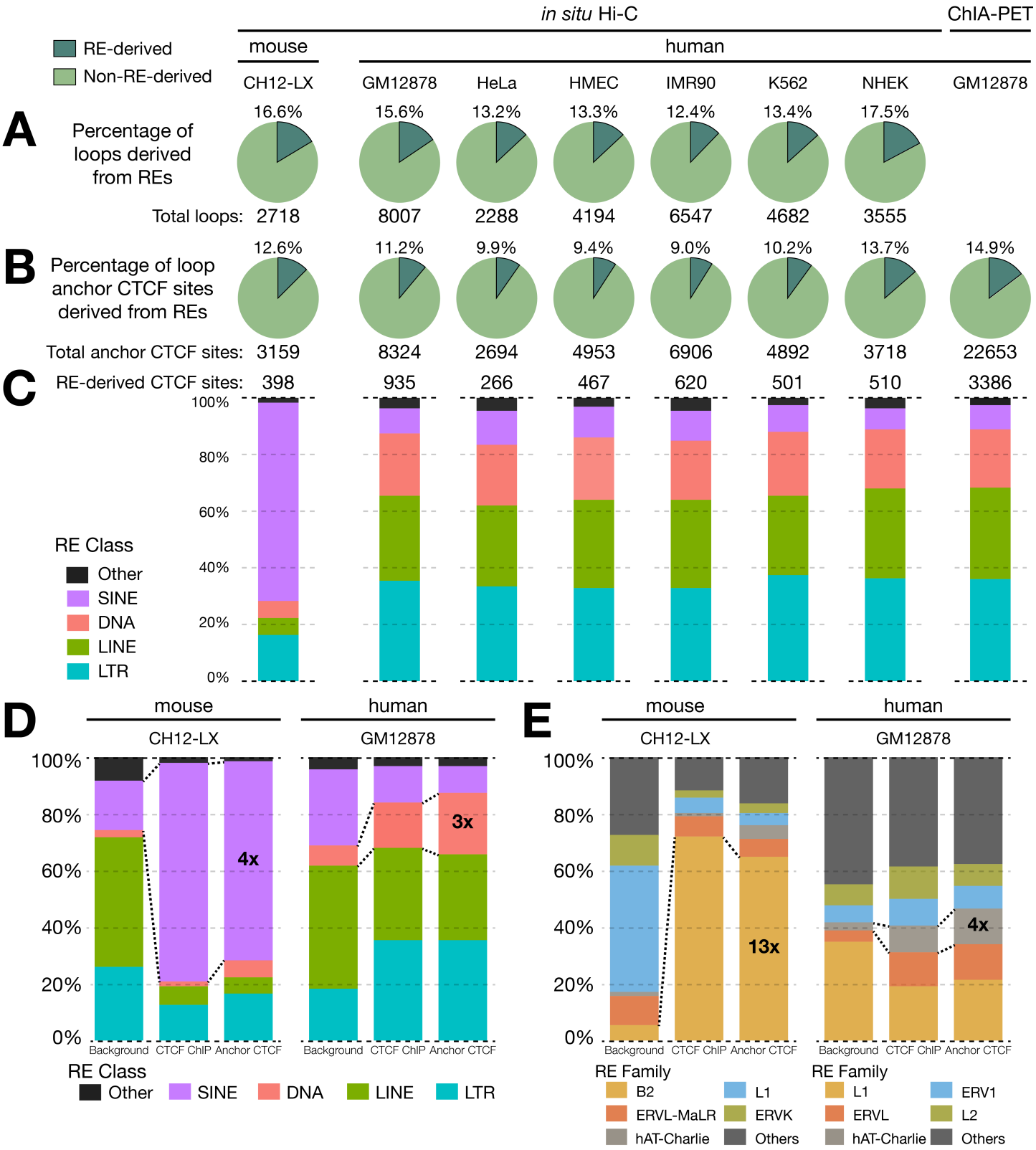

Figure 1: Contribution of repetitive elements (REs) to chromatin loops in humans and mouse. (**A**) Pie charts representing percentage of loops and (**B**) unique loop anchor CTCF sites derived from REs in a variety of human and mouse cell types. (**C**) Stacked bar plots showing the distribution of RE-derived anchor CTCF across major RE classes in the various human and mouse cell types. Stacked bar plots showcasing the distribution of RE-derived anchor CTCF vs. background and CTCF ChIP peaks across (**D**) major RE classes and (**E**) major RE families in matched blood lymphoblastoid cell line (mouse = CH12-LX; human = GM12878).

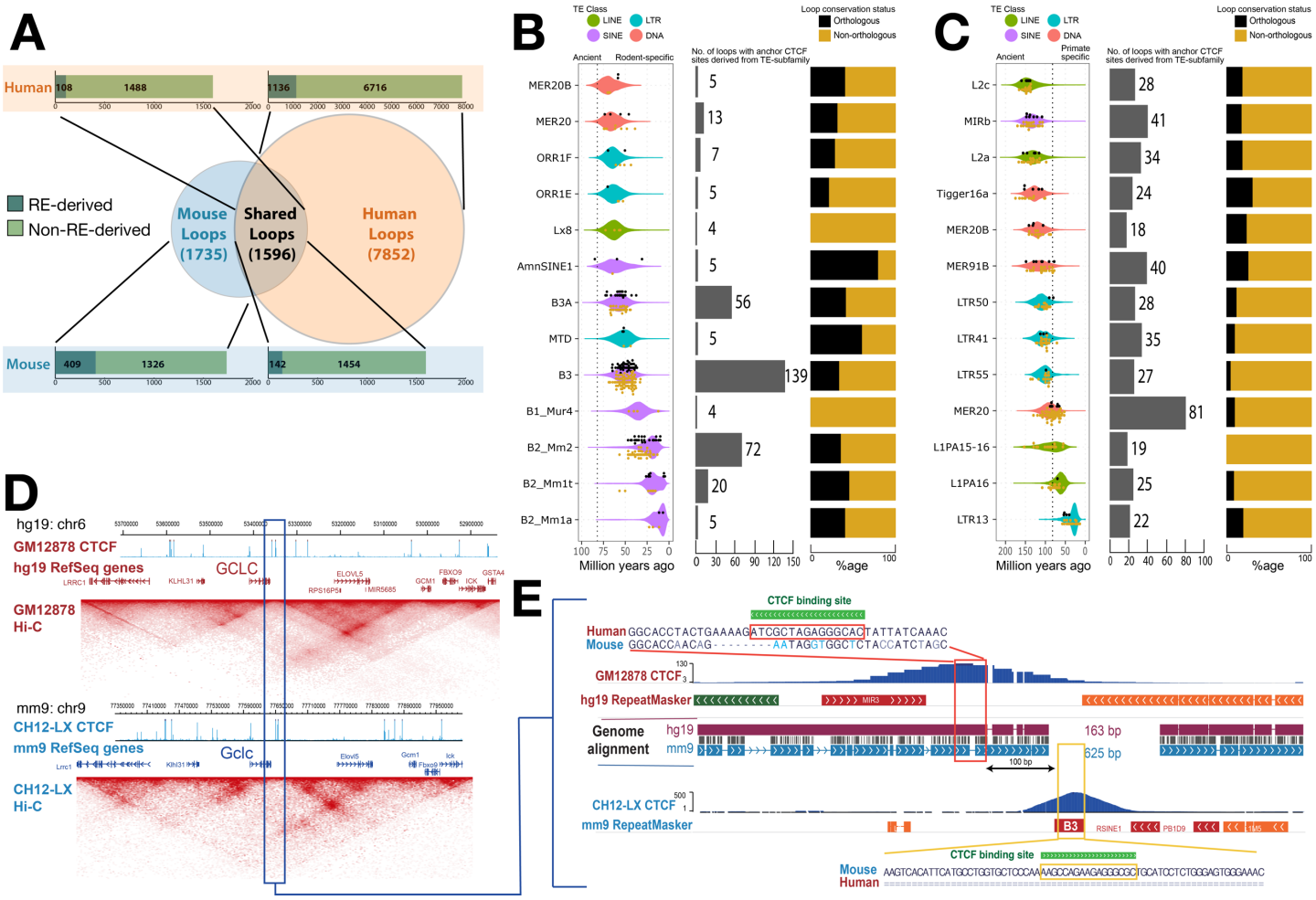

Figure 2: Contribution of TEs in the conservation landscape of human and mouse loops. (**A**) Venn diagram representing the various classes of chromatin loops based on their orthology and bar plots showing the contribution of REs to anchor CTCFs of each class of loops. (**B**) Age distribution (left), total contribution to loop anchor CTCF sites (middle), distribution of orthologous and non-orthologous loops (right) derived from the top 13 TE subfamilies in mouse and (**C**) humans. Estimated primate/rodent divergence time (82 million years ago) is from (Meredith et al, 2011). (**D**) Contact maps representing a conserved 3D structure in syntenic regions between human and mouse. (**E**) A B3 SINE transposon insertion helps in maintaining the conserved 3D structure via CTCF binding site turnover with remnants of the ancestral CTCF motif, well conserved in most non-rodent mammals (Extended Data Figure 1), still seen in the mouse genome. (Note the region is inverted in mouse hence the opposite direction of the motif.)

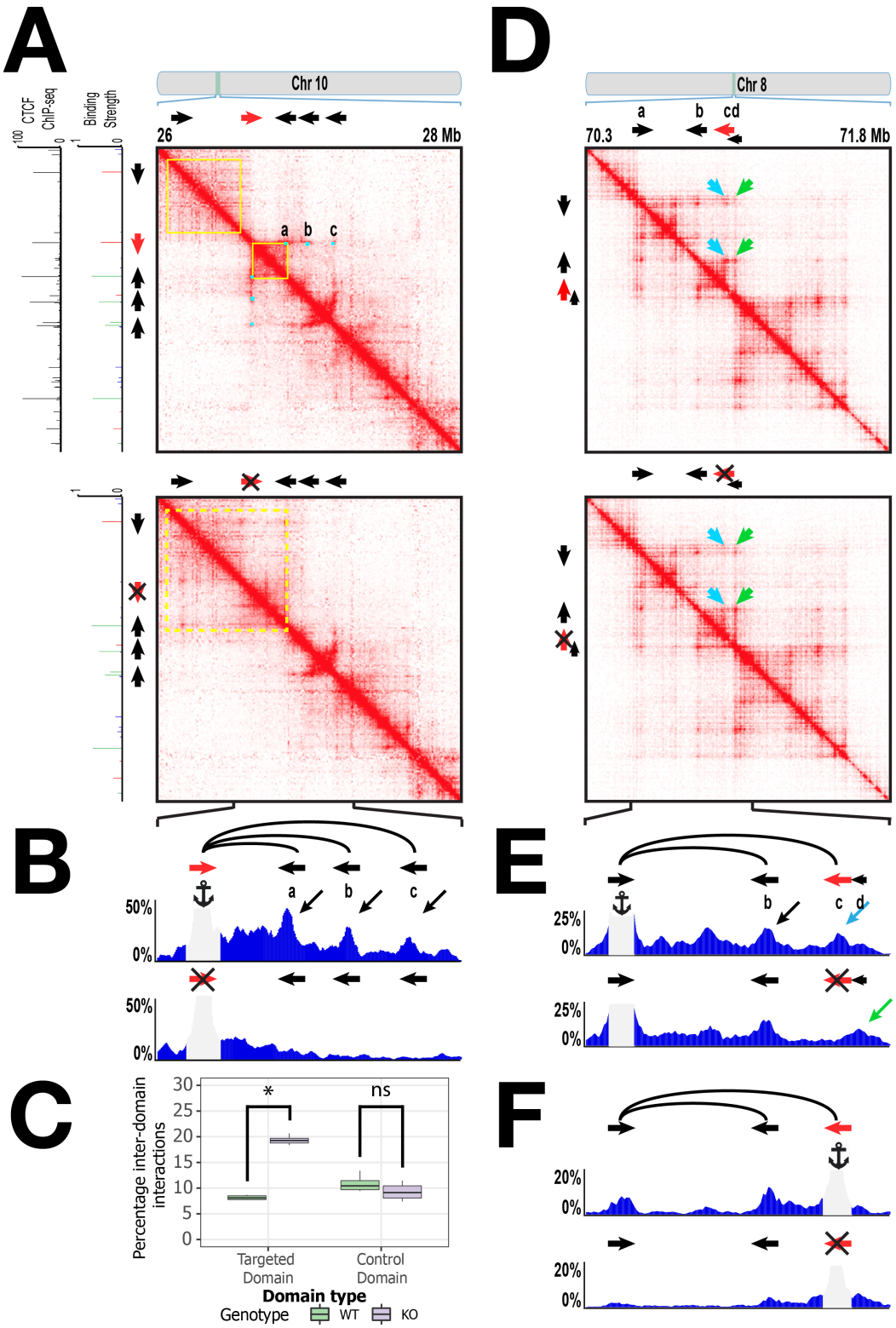

Figure 3: TEs are necessary for maintaining conserved higher-order chromosomal structures in humans. (**A**) Results of a CRISPR/Cas9-based deletion of an L1M3f element at chr10:26–28 Mb in GM187278 cells. Mega-contact maps (details in Methods) generated using Hi-C[2] technology for the (top) WT locus and (bottom) KO (ΔL1M3f) locus. (**B**) Virtual 4C plot displaying total percent interactions emanating from an anchor on a 5kb-window containing the L1M3f element. (**C**) Boxplot measuring the percent inter-domain interactions (Table S2.5) across the targeted domain and a control domain (boundaries unaffected by CRISPR edits) using subsampled contact maps (details in Methods)**.** (**D**) Results of CRISPR/Cas9-based deletion of an LTR41 element at chr8:70.3–71.8 Mb in GM187278 cells. Mega-contact maps generated in Hi-C[2] experiments for the (top) WT locus and (bottom) KO (ΔLTR41) locus. (**E**) Virtual 4C plot displaying total percent interactions emanating from an anchor on a 5kb-window containing the left anchor CTCF of the conserved loop, and (**F**) the LTR41 element.

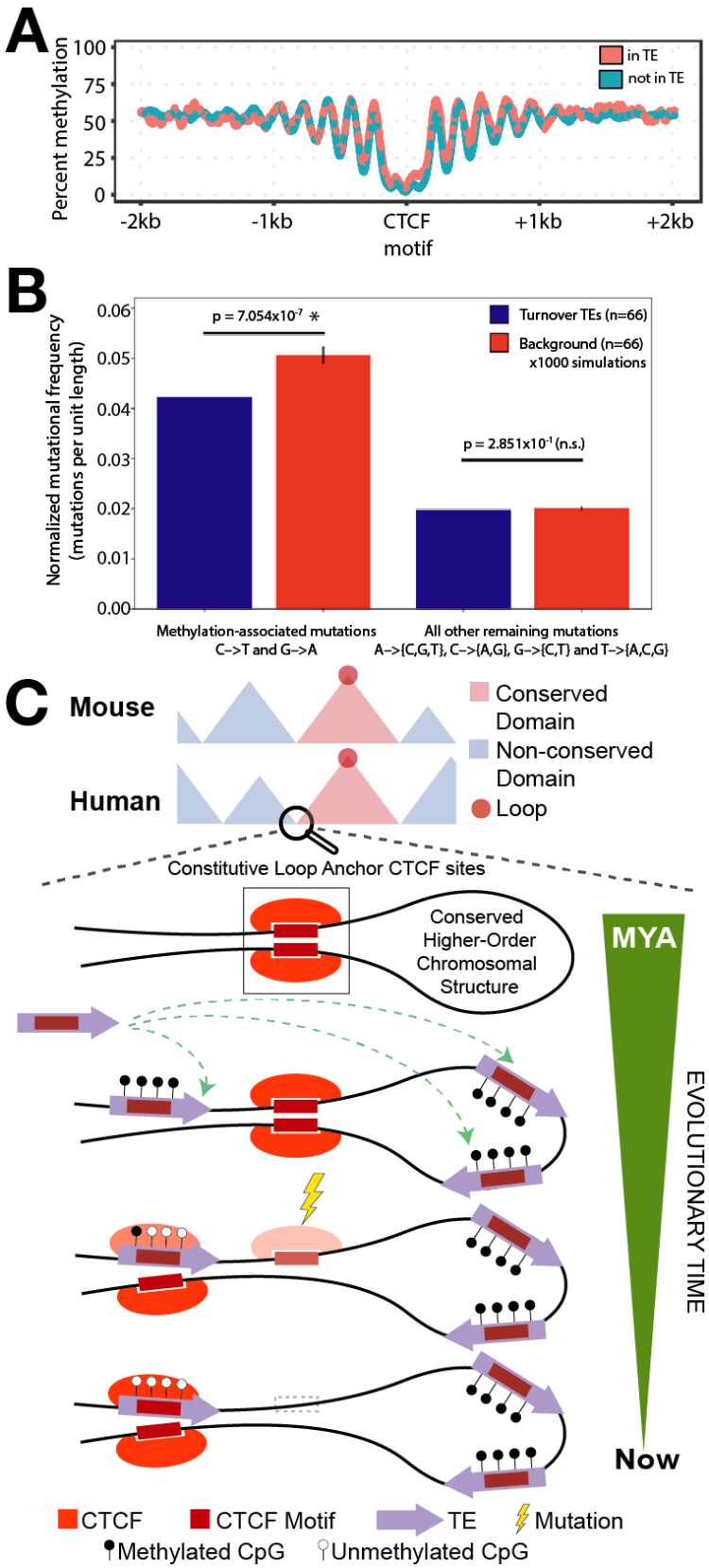

Figure 4: Turnover TEs are hypomethylated through evolutionary time. (**A**) Methylation signature ±2kb around CTCF sites that help maintain orthologous loops segmented by the origin of the anchor CTCF site (**B**) Methylation-associated and non-methylation mutational signature of individual TEs relative to its ancestral sequence in humans (mouse TE data available in Extended Data Figure 6). Alignments were performed using crossmatch (shown here) and Needle (details in Methods, results in Extended Data Figure 7). Error bars show one standard deviation of the means from 1000 simulations. (**C**) Schematic depicting the framework of TE-mediated CTCF binding site turnover that highlights the intimate reciprocity between the TE, genome and epigenome, to help maintain conserved 3D genome.