

CLASH (Chromatin Loop Across-sample Score Harmonizer) quantifies a unified model of genetic and epigenetic variation, CTCF occupancy, and chromatin loop strength

Supplementary Material:

<https://docs.google.com/document/d/1xWTbcHb2UkOQnY-Kva7IiXpgrqyUs1bDhMwKSHG8Wzw/edit?usp=sharing>

Link to Figure folder:

https://drive.google.com/drive/folders/1cl1GZs2Sx5oiyhH-hcWyQuBqpJJXhY9K?usp=share_link

Valmik Ranparia¹

The Human Genome Structural Variation Consortium,

Geoffrey Fudenberg¹

Mark JP Chaisson^{1,2}

1. Department of Quantitative and Computational Biology, University of Southern California, CA, USA.

2. Corresponding author: mchaisso@usc.edu .

Abstract

Chromatin loops are central regulators of gene expression, yet how naturally occurring genetic and epigenetic variation influences loop formation across individuals remains unclear. Although CTCF motifs are frequently altered by SNPs, structural variation, and 5-methylcytosine (m^5C) CpG methylation, existing approaches lack the methodology needed to elucidate the relationship between these molecular changes, CTCF occupancy, and loop strength. Here, we combined high-resolution Hi-C, Fiber-seq, near telomere-to-telomere phased assemblies, and m^5C methylation maps across five lymphoblastoid cell lines to quantify how genetic and epigenetic variation shapes chromatin loop formation. We identified 381 differential contacts and showed that sequence variation, chromatin accessibility, and m^5C CpG methylation are each significantly associated with differential chromatin contacts and strongly correlate with CTCF occupancy. To study differential loop formation, we developed CLASH (Chromatin Loop Across-sample Score Harmonizer), a method that enables robust, quantitative comparisons of loop strengths across individuals to address inconsistent calls across samples from existing methods. CLASH substantially improves biological concordance with Hi-C signal and reveals a significant relationship between CTCF occupancy and loop strength. We found that 63% of

PWM-associated and 59% of m⁵C-associated effects on loop variation act through CTCF occupancy, establishing a unified model in which genetic and epigenetic variation modulate CTCF occupancy, and occupancy affects loop formation strength.

Introduction

In eukaryotic cells, the three-dimensional spatial organization of the genome plays a critical role in intricately regulating vital cellular processes such as transcription and replication. There are three primary structures within chromatin architecture: A/B compartments, topologically associating domains (TADs), and chromatin loops (Lieberman-Aiden et al. 2009; Xu et al. 2024; Holwerda and de Laat 2012). Chromatin loops are local topological features that influence gene expression by bringing linearly distant DNA segments into proximity and positioning enhancers/silencers near promoters. Chromatin loop topology differs across cell types (Bond et al. 2023; Grubert et al. 2020; Burren et al. 2017), developmental stages (Li et al. 2019; Siersbæk et al. 2017), and between individuals at the same cell state (Greenwald et al. 2019; Li et al. 2024). This difference between individuals has been implicated in a variety of diseases, including Cornelia de Lange Syndrome (Panarotto et al. 2022), polydactyly (Paliou et al. 2019), and some cancers (Yoon et al. 2024; Glushakow-Smith and Tothova 2025).

A detailed understanding of how chromatin architecture and loop formation differs between individuals, in particular within the same tissue, can help improve interpretation of how genetic and epigenetic variation influence traits and disease. The landscape of loop formation and chromatin architecture reflects dynamic interactions of DNA binding proteins including CTCF and cohesin (Rowley and Corces 2018) that are in turn modified by methylation, chromatin availability, and genetic variation. The relationship between each of these features and genome structure has been studied extensively, though largely as independent factors. Early studies on diverse tissues quantified the effects of methylation (Wang et al. 2012) and sequence variation (Maurano et al. 2012) on CTCF binding. Generally, differential CTCF binding is attributed more to epigenetic variation than sequence variation (Grubert et al. 2020; Monteagudo-Sánchez et al. 2024; Pękowska et al. 2018). Sequence variation associated with differential loop formation (interaction quantitative trait loci; iQTLs) was characterized in 30 individuals using Hi-C sequencing from chromatin immunoprecipitation (HiChIP), where 2,292 loops were associated with sequence variation (Bhattacharyya and Ay 2024).

The generation of a resource that profiles sequence variation, methylation, CTCF occupancy, and contact maps can help provide a uniform measure of the relative contribution of each class of variation to differential chromatin conformation along with mechanistic insight. Such a resource has been made possible by applications of long-read single-molecule sequencing (LRS); near telomere-to-telomere (T2T) genomes assembled using LRS comprehensively resolve variation (Logsdon et al. 2025), m⁵C may be measured directly from native LRS data (Simpson et al. 2017), and Fiber-seq profiles CTCF and histone occupancy (Stergachis et al.

2020), all with at (or near) base-pair and haplotype resolution. Here, we leveraged these technologies to produce a multi-modal genetic and epigenetic map of genomes to study the impact on chromatin structure either through modulation of CTCF occupancy or other indirect factors.

The effect of large-scale structural variation on chromatin structure was previously studied using high-quality references assembled by the Human Genome Structural Variation Consortium (HGSVC) (Li et al. 2024). This study found 185 structural variants (SVs, insertions, deletions, and rearrangements at least 50 bases) associated with modified topologically associating domains (TADs). Because the majority of SVs identified by the HGSVC are under 500 bases (Ebert et al. 2021), we reasoned that differential contacts not identified by this study reflected smaller, local changes to domain structure, and in particular differential chromatin loop formation, and focused on identifying contact differences using high-resolution contact maps. Furthermore, subtle changes in loop strength are associated with molecular phenotypes (Greenwald et al. 2019). To enable profiling variation in loop formation across populations, we developed CLASH (Chromatin Loop Across-sample Score Harmonizer), a method that assigns harmonized, quantitative loop-strength scores to candidate loops and enables robust cross-sample comparison of loop strength. Using CLASH-quantified chromatin loops from high-resolution Hi-C contact maps together with phased sequence variation, m⁵C methylation profiles, and single-molecule CTCF-occupancy measurements from Fiber-seq, we were able to determine the relative contribution of genetic and epigenetic variation to differential loop formation. We found that, while consistent with previous studies where the majority of differences that have an associated variant are in epigenetic changes, genetic variation accounted for roughly 30% of differential loop formation.

Results.

To enable the joint analysis of chromatin architecture, DNA methylation, and protein occupancy across individuals, we generated and integrated Hi-C, haplotype-resolved N⁶-methyladenosine (m⁶A) annotations of chromatin accessibility through Fiber-seq, haplotype-resolved Oxford Nanopore (ONT) m⁵C methylation tracks, and near-telomere-to-telomere (T2T) assembly datasets for five lymphoblastoid cell lines, GM19317, GM19347, HG01457, HG02666, and HG03248 (Logsdon et al. 2025); Fig. 1a). The five selected genomes contained an average of 4.0M SNVs, 833k indels < 50 bases, and 23.4k SVs ≥ 50 bases per haplotype, determined using the Phased Assembly Caller (Ebert et al. 2021). Hi-C samples yielded an average of ~163.5x genomic coverage with equalized proportions of read-pair orientations at distances between 1–2 kb, supporting downstream analyses at 2 kb resolution (Table S1). Fiber-seq data were generated using PacBio HiFi sequencing with a mean coverage of 33.5× per genome (Table S2). The full computational pipeline used is summarized in Fig. 1b.

All samples exhibited concordance of A/B compartment structures (average MSE = 0.024 ± 0.006, null MSE = 0.371, concordance = 94.7% ± 0.9%; Supplementary Figure 1) and

consistent large-scale structure from E1 values (Supplementary Figures 2 and 3). Additionally, all samples displayed strong agreement in accessibility profiles (average MSE of 1.731 ± 0.988 , null MSE = 7.929, concordance = $83.1\% \pm 2.2\%$; Supplementary Figure 1). Together, these confirm that all samples are in the same cell state and indicate that compartment-level differences do not confound subsequent analyses of contacts and loop formation.

We first investigated whether genetic and epigenetic mechanisms influence chromatin interactions genome-wide. Using DiffHiC to search for stringent pixelwise fold-change differences across the 5 samples, we identified 381 significantly different intra-chromosomal pixels between 446 unique bins that form 123 differential clusters with $\log_{10}CPM > -4$ and distance ranges from 10 kb to 1 Mb at 5 kb resolution by comparing each sample to a set formed of the remaining four samples (5 kb resolution Bonferroni $p < 2.3 \times 10^{-8}$; Supplementary Figures 4-8). We then assessed the various data types to measure how sequence variation within and between contact bins, as well as m^5C and chromatin accessibility rates, correlate with contact strength.

In loci where one sample exhibited a differential number of bases changed between contact bins, 60.7% of samples exhibiting differential interactions matched the sample containing the sequence change ($n = 107$; $p = 3.95 \times 10^{-20}$, binomial test; Supplementary Figure 9). Similarly, in loci where one sample exhibited a differential number of bases changed in contact bins, 61.9% of samples exhibiting differential interactions matched the sample containing the sequence change ($n = 118$; $p = 4.25 \times 10^{-23}$, binomial test; Supplementary Figure 10). These enrichments demonstrate that both large-scale structural rearrangements and smaller sequence changes in interaction-pair bins are strongly associated with altered chromatin contact strength across individuals, consistent with previous findings (Gorkin et al. 2019; Li et al. 2024). Similarly, for epigenetic mechanisms, differentially decreased contacts were also enriched in the samples with higher methylation levels with a match rate of 37.3% for m^6A methylation ($n = 75$; $p = 8.62 \times 10^{-4}$, binomial test) and 41.9% for m^5C CpG methylation ($n = 74$; $p = 1.58 \times 10^{-5}$, binomial test; Supplementary Figures 11 and 12). These enrichments demonstrate that methylation in interaction-pair bins is strongly associated with decreased chromatin interaction strength across individuals. This is expected for m^5C (Monteagudo-Sánchez et al. 2024), whereas no comparable evidence exists for chromatin structure identified through Fiber-seq profiling. Repeating these calculations with relaxed p-value thresholds above the Bonferroni p-value confirmed that a significant amount of subthreshold differential events exist and also align with mechanistic variation, although the overall proportion matched generally decreased, indicating the increased presence of false positives (Supplementary Figure 13).

After accounting for each of the four mechanisms as covariates, all mechanisms remained enriched among differential interactions (Fig. 2a). The adjusted match rates were 52.3% for bases changed between interaction-pair bins ($n = 65$, $p = 7.39 \times 10^{-9}$, binomial test), 55.4% for bases changed within interaction-pair bins ($n = 74$, $p = 1.97 \times 10^{-11}$, binomial test), 41.3% for

m^6A accessibility ($n = 46$, $p = 0.0012$, binomial test), and 44.4% for m^5C methylation ($n = 45$, $p = 2.20 \times 10^{-4}$, binomial test). Representative examples of differential interactions associated with each mechanism are shown in annotated Hi-C heatmaps (Figs. 2b-e). Across the 243 differential interactions where at least one mode of differences was present, 62.6% could be attributed to one or more of these four sources of variation (Fig. 2f). Together, these results show that genetic variation, differences in chromatin accessibility and protein occupancy, and CpG methylation collectively contribute to inter-individual variation in chromatin contact strength.

The majority of bins found to have at least one differential contact among samples appeared only once (Supplementary Figure 14). Furthermore, 39% of identified differential pixels were within complete homozygous deletions of at least one anchor bin and 15% were within 10 kb of chromatin loops (Supplementary Figure 15). Both of these percentages are significantly higher than expected by the null, confirming that deletions and chromatin loops account for a substantial fraction of genome-wide differential interaction patterns. Additionally, we calculated the number of clusters associated with deletions and loops, stratified by number of differential pixels within clusters (Supplementary Figure 16). The majority of unexplained pixels don't form clusters, consistent with diffHic's ability to find local differential interactions not associated with any broader form of genome structure. As cluster size increases, the majority of clusters can be explained by deletions and loop-like structures. We did not explicitly compute TAD boundaries on this dataset, and it is likely that many of the larger clusters associated with loops are also part of TAD boundaries.

The influence of genetic and epigenetic variation on chromatin structure is known to be, in part, mediated by CTCF binding (Rowley and Corces 2018). Thus, we measured the impact of genetic variation by associating global motif strength with occupancy across 262,463 CTCF sites (from 10 haplotypes), observing a positive association between PWM score and binding level ($r = 0.38$; $p < 2.2 \times 10^{-308}$; Pearson r ; Fig. 3a). This was confirmed by assessing the number of bases altered within each motif as a complementary metric, which showed a consistent negative association with occupancy ($r = -0.21$; $p < 2.2 \times 10^{-308}$; Pearson r ; Supplementary Figure 17). We next measured how epigenetic variation is associated with differential CTCF binding. Averaged m^5C methylation across CpGs within each motif was negatively correlated with occupancy ($n = 90,672$; $r = -0.41$; $p < 2.2 \times 10^{-308}$; Pearson correlation; Fig. 3b), and this trend was independently supported by comparing occupancy across hypo-, hyper-, and mixed-methylation states (hypomethylated $n = 58,094$ sites; hypermethylated $n = 140$ sites; mixed $n = 14$ sites; $r = -0.64$; $p = 4.4 \times 10^{-39}$; Pearson r ; Supplementary Figure 18). We also measured how chromatin accessibility is associated with differential CTCF binding. Chromatin accessibility was determined from averaged m^6A methylation rates from the 2 kb upstream and downstream of each 1 kb bin containing a CTCF site. Averaged m^6A methylation rates are weakly globally correlated with occupancy ($n = 262,280$, Pearson $r = 0.11$, $p < 2.2 \times 10^{-308}$; Supplementary Figure 19).

An ordinary least squares association model combining PWM score and m⁵C methylation explained more variance in CTCF occupancy than either factor alone (n = 90,672; PWM β = +0.406; m⁵C β = -0.085; Pearson r between observed and fitted values = 0.48; R² = 0.23; p = 1.1 × 10⁻¹⁶; Fig. 3c), indicating complementary effects. This outcome aligns with prior results showing that SNPs and variation in m⁵C at CTCF sites can affect CTCF binding (Wang et al. 2012); (Zeng et al. 2023). Repeating this analysis while including accessibility as a predictor led to a 0.02 increase in R² (Supplementary Figure 20). Due to the modest observed correlation between accessibility and CTCF occupancy, we did not consider accessibility in any downstream analyses.

We next sought to quantify the relationship between CTCF occupancy and chromatin loop formation. Loops were identified on all five of our samples using Mustache (Roayaei Ardakany et al. 2020) and HiCExplorer (Wolff et al. 2022) at 2 kb resolution. Consistent with Mustache's prior analyses of GM12878, 28.1% of our pooled loopcall set were loci anchored by two unique CTCF sites in at least one haplotype (Supplementary Figure 21). Additionally, reproducibility across samples and callers was low (Figs. 4a, b, Supplementary Figures 22-24), and many loci showed similar contact profiles but inconsistent loop detection (Fig. 4c). These observations demonstrated that conventional (binary) loop-calling approaches alone could not robustly capture biological variation or reliably represent cross-sample loop presence (Chowdhury et al. 2024; Forcato et al. 2017).

To overcome these limitations, we developed CLASH, a method that aggregates loop calls across tools and quantifies loop strength directly from the underlying Hi-C matrices as a continuous score (Fig. 4d). This approach assigns a standardized loop-strength score to every locus in every individual, enabling harmonized, quantitative comparison of loop strength across samples. CLASH substantially improved loop calls across global and per-locus metrics relative to binary callers, including correlation to biological signal ($\Delta r = +0.30$; $p < 2.2 \times 10^{-308}$, Fisher's r-to-z test) and MSE (Average Δ MSE per locus = -0.35; $p < 2.2 \times 10^{-308}$; Wilcoxon signed-rank test; Figs. 4e, f, Supplementary Figures 25 and 26), providing a more reliable basis for downstream differential analyses of how genetic and epigenetic mechanisms modulate chromatin loop formation.

We also compared CLASH's loop-scoring performance to "LC scores", a continuous score-like intermediary that was derived from the Mustache algorithm. While these scores were not designed to act as loop scores, no other existing methods exist to score loops and these LC scores are among the closest analogs that can be used for benchmarking. We comprehensively show CLASH scores outperform LC scores both globally and within-locus across a variety of Hi-C-derived and orthogonal metrics (Supplementary Figures 27-31). Harmonizing calls across samples through CLASH improves consistency; 67% of loops are present in two or more samples (Supplementary Figure 32).

With the 26,750 CLASH-scored loops across samples derived from the 5,314 loops that formed between two CTCF sites, we quantified the association between CTCF occupancy and loop strength. Across all samples, occupancy showed a positive genome-wide correlation with loop strength ($n = 17,439$; $r = 0.25$; $p = 5.8 \times 10^{-252}$, Pearson correlation; Fig. 4g). To assess locus-specific effects, we examined loci with high inter-sample variability in CLASH score (standard deviation ≥ 0.29 ; see Supplementary Figure 31 for a threshold sweep). Among these, the mean within-locus correlation was $r = 0.26$ ($n = 21$; binomial sign test, $p = 3.6 \times 10^{-3}$; Fig. 4h, Supplementary Figure 33), indicating a directionally consistent association between higher occupancy and increased loop strength at variable loci. This association is supported by prior work showing that binary changes in CTCF binding coincide with corresponding gains or losses of chromatin loops (Pugacheva et al. 2020). However, this study expands on this observation to the more biologically relevant model of a continual state.

We also tested whether CTCF occupancy affects higher-order chromatin organization at TAD boundaries. Using insulation profiles from Cooltools (Open2C et al. 2024), CTCF sites on chromosome 1 were classified as boundary or non-boundary sites. Although no significant global or per-locus correlation was observed across samples, within-sample analyses showed a positive association between occupancy and insulation strength, with stronger associations at boundary sites than non-boundary sites (boundary $n = 740$, boundary $r = 0.2617$, non-boundary $n = 3,063$, non-boundary $r = 0.0876$, Pearson r ; $\Delta r = 0.1741$; $p = 0.00326$, Welch's t test; Supplementary Figures 34 and 35). A similar relationship between CTCF occupancy and TAD strength has been reported previously although, like earlier, these studies only examined whether CTCF is present or absent and did not quantify occupancy as a continuous variable (Nora et al. 2017; Davidson et al. 2023).

We next sought to measure how genetic and epigenetic variation jointly affect loop strength independent of CTCF occupancy. Using loops formed between two CTCF sites, we quantified the association between sequence or methylation variation and loop strength, averaging PWM score and m⁵C methylation across both sites. PWM score showed a positive correlation with loop strength ($n = 26,557$; Pearson $r = 0.17$; $p = 8.8 \times 10^{-171}$; Fig. 5a, Supplementary Figure 36), and m⁵C methylation showed a negative correlation ($n = 15,932$; Pearson $r = -0.15$; $p = 2.1 \times 10^{-78}$; Fig. 5b, Supplementary Figure 37), as expected (Gorkin et al. 2019; Monteagudo-Sánchez et al. 2024). In both examples, the PWM- or methylation-affected CTCF site that lost loop formation also had the lowest CTCF occupancy across samples.

We also sought to determine whether CLASH identified putative iQTLs were enriched for eQTLs, and although statistical power was limited due to small sample size (iQTL $n = 11$; eQTL $n = 2$ at CLASH threshold ≥ 0.4), we observed a consistent increase in eQTL enrichment as loop-strength thresholds were made more stringent in agreement with previous results showing iQTLs are enriched in eQTLs (Supplementary Figure 38; (Bhattacharyya and Ay 2024)).

We next quantified how many differential loops could be associated with each mechanism. Among the 624 loci where CLASH loop strength varied by more than 0.6, 33.2% showed a correlation between occupancy and loop strength above $r \geq 0.25$, while 22.3% were associated with m⁵C methylation and 8.8% with PWM score. In total, these mechanisms collectively accounted for 52.6% of differential loops (Fig. 5c).

To determine if occupancy accounts for the observed mechanistic effect on loop strength, we first calculated the joint association between PWM score, m⁵C methylation, and occupancy with loop strength, as well as between PWM score and m⁵C methylation with loop strength. Including PWM score and m⁵C methylation together with occupancy yielded a lower correlation with loop strength than occupancy alone ($n = 4,489$, $\Delta r = -0.019$, $p = 0.042$; Steiger test), while excluding occupancy resulted in a significantly lower correlation with loop strength than occupancy alone ($n = 4,489$; $\Delta r = -0.061$; $p = 1.5 \times 10^{-4}$; Steiger test; Fig. 5d). We also directly quantified the extent to which variation in loop strength associated with PWM score or m⁵C methylation was statistically mediated by occupancy by applying a product-of-coefficients mediation analysis averaged across samples, and found that 63% of PWM's and 59% of m⁵C's effects on loop strength act through CTCF occupancy (Fig. 5e). As the majority of both mechanisms' effects on loop score were transmitted through CTCF occupancy, these results further support CTCF occupancy's role as a primary mechanism that affects loop strength. A summary of these relationships is shown in Fig. 5f.

Because the occupancy-based measurements did not incorporate SVs that insert or delete CTCF sites, we separately evaluated whether such variants were associated with differences in loop strength. We took advantage of the complete assemblies that provided sequence-resolved variants (Logsdon et al. 2025), where the annotation of sequence context determined if inserted or deleted variants added or removed CTCF binding sites. For each locus where SVs contained one or more loop-associated CTCF sites, we compared average loop scores between samples carrying the variant and those without it. Among 70 loops containing CTCF insertions (present in at least one but not all haplotypes), the mean loop-score difference was 0.0041 ± 0.1876 , and among 65 loops containing deletions (present in at least one but not all haplotypes), the mean difference was -0.0111 ± 0.1971 (Supplementary Figure 39). Contrary to the expectation that creating or removing loop-associated CTCF sites would consistently alter loop strength, these results indicate that variants showed minimal average effects at the genome-wide level (Li et al. 2024). Despite the lack of a consistent genome-wide effect, several individual loci exhibited clear loop-strength changes associated with variant CTCF sites (Supplementary Figures 40 and 41).

Discussion

Here, we sought to quantify fine-grained differences in three-dimensional genome structure across five diverse genomes with haplotype-resolved, telomere-to-telomere assemblies (Logsdon et al. 2025). The effect of SVs on genome structure had previously been characterized on these genomes (Li et al. 2024) using modest resolution contact maps (~30-fold coverage), where the average length of deletion and insertion SVs associated with changes in TADs was 8.9 kb and 1.8 kb, respectively, where insulation differences were calculated from pooled contacts within 100 kb of topological boundaries. In contrast, machine learning methods predict differential localized contacts from sequence features (Fudenberg et al. 2020) that are challenging to observe at low resolution.

We generated high-resolution contact maps to analyze the fine-scale changes in genome structure caused by smaller variation than the Li *et al.* study. We focused on quantifying differential loop formation because of the interpretable impacts on cellular regulation. There were two initial challenges: first, the methods to call loops were less consistent than what the high degree of conservation of chromatin structure predicts, and manual inspection of differences revealed missing calls where there was clear support in one or more samples. Second, differential loops were present where there was no genetic variation. The completeness of the near-T2T assemblies leaves no “(genetic variant) stone unturned”, so that differential loops unexplained by sequence variation were likely of epigenetic or stochastic origin.

The inclusion of phased epigenetic (m^5C) and CTCF occupancy measurements provides a more complete understanding of differential loop formation. Crucially, there is a 2.5-fold increase in differential loop formation due to methylation compared to sequence variation, and a 3.8-fold increase where differential CTCF binding is observed. Furthermore, the co-occurrence of differential methylation and sequence variation is minimal (1.8% of all differential sites). Although some of the epigenetic changes may be due to random silencing as a cell-line artifact, that is independent of the aim of quantifying which molecular changes lead to differential chromatin structure. At a time where there is a sharp interest in developing virtual cells using machine learning (Tang 2025), this data indicates that personalized epigenomes at the resolution of individual bases should be considered alongside personalized references to predict cell state.

After showing that SNPs in CTCF sites showed measurable effects on loop-formation (iQTLs), we also sought to determine whether these putative iQTLs were enriched for expression quantitative trait loci (eQTLs). We intersected these putative iQTLs with GTEx Whole Blood v8 eQTLs and, although statistical power was limited due to small intersection size (iQTL n = 11; eQTL n = 2 at CLASH threshold ≥ 0.4), there was a consistent increase in eQTL enrichment as

loop-strength thresholds were made more stringent (Supplementary Figure NNN). Additionally, the genomic landscape of CTCF sites was relatively constant across individuals; there were 306 ± 58 novel CTCF binding sites per sample relative to GRCh38 (0.55% of total CTCF sites) due to SVs. We did not detect a large number of novel loops forming due to inserted CTCF binding sites, nor a pervasive effect of SV on loop formation in general. With larger sample sizes, both eQTL and novel loop-forming inserted sequences may be found.

The effect that genetic and epigenetic variation have on differential chromatin interactions between individuals reaffirms the importance of accounting for them in any functional genomics study. Previously, genetic and epigenetic factors have been shown to separately affect gene expression (Lappalainen et al. 2013; Banovich et al. 2014) and chromatin accessibility/transcription factor binding (Chen et al. 2024; Degner et al. 2012).

CLASH:

The loop-strength variation highlighted throughout this study could not be quantified by traditional binary loop callers that classify loci as either “loop” or “no loop.” Previous studies (Greenwald et al. 2019) have recognized the need to quantify loop-strength across samples, however only inter-individual comparisons rather than across loci comparisons were performed. This prevents global correlations between loop-strength and genetic, epigenetic, and occupancy variation from being computed. CLASH improves upon this method by converting the underlying Hi-C signal at the same locus in each sample into a single scalar value, accounting for contact counts surrounding the loop call as well as the enrichment of the contacts relative to the background contact counts at the same genomic distance. This shift from detection to measurement enables direct comparisons of subtle loop strength differences across individuals at matched genomic coordinates, which is essential for population-scale analyses and makes it uniquely suited for applications in identifying loci where subtle changes in loop-strength across samples/individuals are still significant, as in this project. CLASH could be used in future work to explore whether variation in loop strength contributes to disease-associated regulatory differences by testing whether loop strength at specific functional loci may serve as a predictive biomarker in precision-medicine contexts.

A key advantage of CLASH is that it generates reproducible scores directly from the Hi-C matrix, rather than relying on any particular loop-calling algorithm. This design makes CLASH compatible with any set of genomic loci provided by the user, resolving the inconsistencies among current loop-calling methods and ensuring that it remains broadly useful even as loop-calling methods improve. Furthermore, CLASH can serve as an orthogonal

benchmark for assessing the robustness of new loop-calling algorithms by providing a continuous, caller-independent measure of signal strength.

While CLASH's current implementation heavily involves empirically-derived parameters specified for 2 kb resolution, we aim to improve upon this in future work by transitioning the scoring parameters into a machine learning implementation that is calibrated to support multi-resolution scoring, to broaden CLASH's applicability.

Summary:

Our unified model provides a mechanistic basis for interpreting how genetic and epigenetic diversity contributes to chromatin architecture across individuals, while CLASH offers a quantitative framework for detecting these effects. Together, these analyses reveal a unified view of inter-individual variation in 3D genome structure: molecular variation primarily affects the potential for CTCF occupancy, which in turn affects chromatin loop-strength and potentially downstream gene regulation. The methods and findings described here lay the groundwork for future work to predict chromatin loop strength from additional biological factors as well as determine whether changes in chromatin loop strength may have functional and health-related consequences.

Methods:

Data processing:

All data used corresponded to five lymphoblastoid cell lines, GM19317, GM19347, HG01457, HG02666, and HG03248 (Logsdon et al. 2025). Hi-C data for all samples were generated, aligned to GRCh38 (GRCh38), and processed using the distiller-sm pipeline (<https://github.com/open2c/distiller-sm>) to resolutions as high as 1 kb. Phased SNP and structural variant calls for all samples were obtained from the HGSVC Phase 3 dataset (Ebert et al. 2021) generated from telomere-to-telomere haplotype assemblies aligned to GRCh38, and processed with dipcall ((Ebert et al. 2021; Li et al. 2018) and WhatsHap (Martin et al. 2023) to annotate CTCF-overlapping insertions, deletions, and SNPs. Single-molecule Fiber-seq reads, generated with PacBio HiFi for all samples, were aligned to GRCh38, phased with WhatsHap, and processed with pbmm2 (Li 2018), whatshap, and fibertools (Jha et al. 2024) to determine CTCF occupancy, with occupancy defined as the fraction of fibers showing a footprint at each motif. Haplotype-specific CpG methylation tracks were downloaded from HGSVC via globus, aligned

to GRCh38 with minimap2, phased with WhatsHap, and summarized per genomic bin and per CTCF site.

AB compartment analysation:

A/B compartments were computed from ICE-balanced Hi-C matrices at 100 kb resolution using Cooltools eigs_cis, with GC content used to orient the first eigenvector such that positive values correspond to active (A) chromatin. Compartment similarity across individuals was assessed using pairwise eigenvector mean squared error (MSE) and sign concordance.

Large-scale accessibility track analysis:

For all five samples, Fiber-seq reads across both haplotypes were partitioned into 100 kb genomic bins and the proportion of methylated adenines/total adenines within each bin was computed. For each pair of samples, we computed the average MSE between mean-centered m⁶A levels at the same genomic bin across samples and sign concordance as the percentage of bins in which both samples showed deviations in the same direction (positive or negative relative to their respective means).

DiffHic analysis:

Genome-wide differential chromatin interactions were identified using diffHic (Lun and Smyth 2015) by following the protocol detailed in the “diffHic User’s Guide”. We generated differential interaction lists from raw Hi-C contact matrices at 1 kb, 2 kb, and 5 kb resolutions by filtering out non-relevant pixels, normalizing for bias with normOffsets(), estimating dispersion, and applying quasi-likelihood tests across all five samples. We retained differential pixels that exhibited sufficient coverage ($\log\text{CPM} > -4$) and that passed Bonferroni multiple-test correction at genomic distances between 10 kb and 1 Mb. Each differential interaction was then annotated for four potential mechanisms – large structural variants, small sequence changes, m⁶A methylation, and m⁵C CpG methylation – using outlier frameworks to identify the sample or haplotype exhibiting the most extreme deviation for each feature. For each mechanism, we quantified whether the sample with the strongest molecular deviation corresponded to the sample with the largest Hi-C log-fold change and assessed significance using binomial tests. Loci explained by multiple mechanisms were then removed to confirm that each mechanism remained significantly enriched.

CTCF site determination and PWM score calculation:

Because 85% of chromatin loops form at CCCTC-binding factor (CTCF; (Rao et al. 2014), CTCF sites were identified in each haplotype by running FIMO (Grant et al. 2011) on haplotype-resolved assemblies and mapping the resulting coordinates back to GRCh38 using the Long Read Aligner software (Grant et al. 2011; Ren and Chaisson 2021), yielding ~50,000 sites

per haplotype. This matches expected CTCF site counts per haplotype (The ENCODE Project Consortium 2012).

For each motif, the underlying 19 bp sequence was scored using a position weight matrix (PWM) derived from the JASPAR MA0139.1 probability-odds matrix (Khan et al. 2017), producing a continuous measure of motif strength that reflects the impact of SNPs and small indels. PWM scores were computed by summing position-specific probability weights and normalizing by motif length to facilitate comparison across haplotypes and individuals. Sequence mismatches relative to the canonical 15 bp CTCF consensus motif were enumerated by aligning each haplotype sequence to the consensus and counting non-matching bases.

CTCF site m⁵C CpG methylation:

CpG methylation levels were quantified at single-base resolution from phased ONT methylation calls, with per-position methylation fractions computed as the proportion of reads supporting a methylated cytosine. For each haplotype-resolved CTCF motif, all CpG positions within the motif boundaries were extracted and averaged to obtain a site-level m⁵C methylation value. Additionally, a genome-wide hidden Markov model was used to classify each CpG as hypomethylated or hypermethylated.

CTCF site accessibility:

M⁶A rates (the number of m⁶A adenines / the total number of adenines) were calculated within each 1 kb bin for each sample. The average m⁶A rate of the four bins (two upstream, two downstream) surrounding each bin that contained a CTCF site was calculated. To account for technical biases from sequencing, the m⁶A rates were Z-scored within sample and sigmoid-transformed into a score between zero and one, representing how open the chromatin near each CTCF site was compared to the other CTCF sites in the sample. Z-scores were calculated within CTCF sites rather than genome-wide as downstream analysis was focused only on CTCF sites.

Quantifying correlations between genetic and epigenetic factors and occupancy:

Each CTCF site across all ten haplotypes was annotated with its PWM score, number of sequence mismatches, CpG methylation level and state, accessibility, and CTCF occupancy value. Global pairwise Pearson correlations were computed to assess how each genetic or epigenetic feature individually relates to CTCF occupancy. To evaluate their combined contributions, an ordinary least squares model was constructed using PWM scores, accessibility, and CpG methylation as joint predictors of occupancy after appropriate normalization. Because

the goal was to quantify association rather than perform prediction, the model was fit to the full dataset and significance was evaluated using standard linear regression statistics.

Calling loops with Mustache and HiCExplorer:

Chromatin loops were initially identified using both Mustache (Roayaei Ardakany et al. 2020) and HiCExplorer (Wolff et al. 2022) across 1 kb, 2 kb, and 5 kb resolutions, filtering for p-values of 0.1 and 0.01. To assess agreement between callers and across individuals, we compared loop sets using Jaccard indices after standardizing loop coordinates within 10 kb. Additionally, we computed the variance in Hi-C signal versus the number of samples with a called loop.

CLASH:

We developed CLASH, a quantitative loop-scoring framework that re-evaluates pooled loop calls across samples by extracting an adaptively sized Hi-C submatrix around each candidate loop and normalizing interaction strength to a distance-matched background. For each sample, CLASH identifies a refined loop center, adjusts the extraction window based on local matrix structure, scores the local matrix, and normalizes scores across samples within-locus. The method integrates center enrichment, hierarchical decay patterns, and matrix coherence into a unified 0–1 score that reflects the Hi-C signal of each loop. These continuous scores replace inconsistent binary loop calls and enable robust cross-sample comparison of loop strength. Full algorithmic details, scoring functions, and parameter choices are provided in the Supplementary Methods.

CLASH validation:

CLASH scores were compared to initial pooled Mustache/HiCExplorer binary loop calls using multiple quantitative metrics. We compared their global correlation with local Hi-C signal, assessed improvements in mutual information via bootstrap resampling, and computed per-locus MSE between predicted and observed Hi-C signal across samples. We also compared CLASH scores to Mustache LC scores using Hi-C-derived and orthogonal methods.

Calculating loop conservation:

Loop presence was defined as loops exhibiting a CLASH score of >0.3 . For each locus, we identified the number of individuals (max $N = 5$) in which a loop was present (k). The corresponding k-of-N conservation curve was generated and smoothed using cubic interpolation. The area under the curve (AUC) of the portion of the graph with $k \geq 2$ corresponds to the conditional conservation probability.

Quantifying the correlation between CTCF occupancy and CLASH loop score:

To assess how CTCF protein binding relates to chromatin loop strength, we restricted the analysis to loops whose two interaction bins each contained a CTCF site within 10 kb and computed a per-loop occupancy value by averaging the Fiber-seq-derived occupancy of its two interacting CTCF sites. These occupancy values were paired with CLASH loop scores across all samples to quantify the global relationship between CTCF binding and loop intensity.

To evaluate locus-specific effects, we also quantified the distribution of Pearson correlations between CTCF occupancy and CLASH loop strength between samples at each locus. Loci with fewer than four samples containing valid occupancy measurements and with ≤ 0.29 loop score standard deviation across samples were excluded. A binomial sign test was used to calculate statistical significance.

Quantifying the correlation between CTCF occupancy and CTCF insulation capability:

To evaluate how CTCF occupancy affects a CTCF site's ability to function as an insulator, insulating CTCF sites on chromosome 1 for each sample were identified using cooltools insulation. Identified sites were classified as either boundary sites, and all the other CTCF sites were classified as non-boundary sites. For each CTCF site, we computed a quantitative insulation score by extracting a 200 kb Hi-C window around the site and calculating an insulation score, based on an existing function (Crane et al. 2015), comparing upstream/downstream interactions versus cross-boundary interactions. Within each sample, Pearson correlations between occupancy and insulation score were computed across all CTCF sites, and the difference between boundary and non-boundary correlations was assessed using a two-sided Welch's t-test.

Quantifying the correlation between PWM scores, m⁵C methylation, CTCF occupancy, and CLASH loop scores:

Per-loop PWM and CpG methylation values were obtained by averaging interaction bin measurements and pairing them with CLASH loop scores and occupancy. We quantified individual and joint associations using Pearson correlations and Steiger's test for dependent correlations, enabling direct evaluation of whether PWM and m⁵C methylation contributed explanatory power beyond occupancy alone, including in a model excluding occupancy.

Mediation analysis:

To quantify how much of the effect of PWM scores and CpG methylation on loop strength is transmitted through CTCF occupancy, we performed a product-of-coefficients mediation analysis within each sample. Uncertainty was measured using a pooled nonparametric bootstrap.

eQTL analysis:

To assess whether SNPs within loop-associated CTCF sites that modulate loop strength act as interaction QTLs (iQTLs), we identified SNP–loop pairs and tested whether allele differences were associated with changes in CLASH loop scores across individuals. We then compared these iQTL candidates to the GTEx v8 Whole Blood significant eQTL catalog (The GTEx Consortium 2020) to determine whether loop-modulating variants were enriched for known expression QTLs. Enrichment across increasing loop strength difference thresholds was quantified by computing the log odds ratio of eQTL overlap.

Structural variation analysis:

To assess whether structural variants (SVs) that add or remove CTCF motifs influence chromatin loop formation, we identified insertions and deletions overlapping loop-associated CTCF sites and quantified the resulting gain or loss of motif instances. For each loop and individual, we compared CLASH loop scores between samples carrying a CTCF-altering SV and those without, and evaluated the direction and magnitude of loop strength changes across the genome.

Supplemental Methods:

Data processing

Hi-C data were generated for five male lymphoblastoid cell lines (GM19317, GM19347, HG01457, HG02666, and HG03248), each sequenced in four independent runs by Phase Genomics (Seattle, WA). Raw FASTQ files were aligned to the GRCh38 reference genome and processed using the distiller-sm pipeline, excluding unplaced contigs. PCR duplicates, self-circles, and dangling ends were removed following standard Hi-C quality-control steps.

The lower-limit of contact resolution imposed by molecular byproducts for all five samples was 1-2 kb, as summarized using MultiQC (v1.20; (Ewels et al. 2016)). To determine if the sequencing depth of our samples supported analysis at this high resolution, we adapted the Juicer (Durand et al. 2016) script provided by

https://github.com/aidenlab/juicer/blob/main/misc/calculate_map_resolution.sh to guide plausible Hi-C resolution analysis for our samples, which yielded map resolutions ranging from 2.45 kb – 2.95 kb (Table S1). These values are reasonable given prior work that has shown that a read depth of ~5 billion read pairs is necessary for Hi-C analysis at 1 kb resolution (Rao et al. 2015), ~1 billion read pairs allows for slightly underpowered Hi-C analysis at 2 kb (Rao et al. 2015; Lee et al. 2022), and ~500 million read pairs is sufficient for 5 kb analysis (Rao et al. 2015). Thus, we generated contact matrices at resolutions as high as 1 kb, but with a focus on 2

kb and 5 kb resolutions, for downstream analysis using the cooler (v0.10.3) framework, and applied iterative correction (ICE) for normalization via cooler balance. Across the five samples, cooler balance assigned NaN weights to an average of $12.84\% \pm 0.26\%$ of bins at 2 kb resolution (1,544,155 bins total) and $13.12\% \pm 0.30\%$ of bins at 5 kb resolution (617,669 bins total).

Phased variant call files (VCFs) containing both single-nucleotide polymorphisms and structural variants were obtained from the HGSCV Phase 3 dataset, and were generated from the pav2 pipeline (docker://becklab/pav:latest; (Ebert et al. 2021)). These VCFs were produced by aligning telomere-to-telomere haplotype assemblies generated from PacBio HiFi and Oxford Nanopore reads to GRCh38, calling variants with dipcall (v0.3) and phasing with whatshap (v2.8). The resulting phased VCFs were used to annotate deletions, insertions, and SNPs overlapping CTCF motifs and chromatin loops.

Single-molecule chromatin Fiber-seq data were produced using PacBio Revio HiFi sequencing (30 h movie time per SMRT Cell) at the University of Washington. Each sample achieved a mean genome-wide coverage of 33.5x, an average read length of 20.7 kb, HiFi yields of ~103 Gb, and read-quality scores ranging from Q30–Q33 (Supplementary Table S2). Reads containing pre-annotated N⁶-methyladenine (m⁶A) modifications were aligned to GRCh38 using pbmm2 (v1.10.0). The resulting BAM files were phased using whatshap phase guided by corresponding variant calls. Processed reads were analyzed using the standard fibertools command suite (v0.6.4) pipeline. M⁶A bases and nucleosomes were identified using the following command:

```
ft extract GM19317/GM19317_m6_nuc.bam \ --reference \ -q \ -a GM19317_all.tsv \
-n GM19317_nucleosomes.bed \ --m6a GM19317/GM19317_m6a.bed \ -t 8
```

Footprinting was performed against known CTCF motif coordinates (JASPAR MA0139.1) using the following command:

```
ft footprint "GM19317_m6_nuc.bam" \
--bed "GM19317H1_CTCF_COORDINATES_clean.bed" \
--yaml "ctcf.yaml" \
--out "GM19317H1_ctcf_footprint.bed"
```

For each motif, both the total number of fibers spanning the site and the number containing a CTCF-sized footprint were counted. The ratio of footprinted to total fibers defined the CTCF occupancy frequency for that site.

Per-base 5mC CpG methylation calls for all five samples were obtained from the Human Genome Structural Variation Consortium (HGSCV). Base-called reads were aligned to each

sample's telomere-to-telomere assembly using minimap2 (v2.30-r1287). Methylation profiles were phased using whatshap phase (v2.8) guided by dipcall-derived VCFs, yielding haplotype-specific methylation tracks. For downstream integration with Hi-C and Fiber-seq, the number of methylated CpG sites per fiber per genomic bin and per CTCF site was used as a proxy for total methylation signal.

AB compartment analysation

A/B compartments were computed using the Cooltools (v0.7.0) eigs_cis function applied to ICE-balanced Hi-C matrices at 100 kb resolution using the following command:

```
eigvals, eig_df = cooltools.eigs_cis(  
    GM19317_100000.cool,  
    phasing_track=bins[['chrom', 'start', 'end', 'frac_gc']],  
    n_eigs=1  
)
```

GC content was used as the phasing track to orient the first eigenvector (E1), such that positive values correspond to GC-rich, transcriptionally active (A) compartments. The resulting E1 eigenvalues for each genomic bin were compared between samples using pairwise MSE and sign concordance to quantify compartment similarity. A null expectation for MSE was estimated as the genome-wide variance of E1 values pooled across all samples, corresponding to the expected mean squared difference under random alignment of compartment eigenvectors. For visualization, E1 profiles across a representative 35 Mb region were plotted for all samples. Hi-C maps showing compartment state and inter-sample compartment changes were generated using matplotlib from sample cooler files. For all AB compartment, TAD domain, chromatin loop, and chromatin interaction-related analyses in this manuscript, we highlight representative loci and regions that illustrate the general trends observed genome-wide.

Chromatin accessibility analysis with Fiber-seq

For all five samples, Fiber-seq reads across both haplotypes were partitioned into 100 kb genomic bins (only relevant fragments of each read were kept in the case of reads that overlap multiple bins). For each read or fragment of a read, the total number of adenines and m⁶A methylated adenines were determined and summed across every read for every bin. The proportion of methylated adenines/total adenines within each bin was computed. For between-sample comparisons, these percentages were mean-centered within each sample to remove global shifts in modification levels and focus analyses on relative spatial variation along the genome. For each pair of samples, we computed the average MSE between mean-centered m⁶A levels at the same genomic bin across samples and sign concordance as the percentage of bins in which both samples showed deviations in the same direction (positive or negative relative to their respective means), excluding bins with 0 deviation. Final statistics reported averaged

MSE and sign concordance across each pair of samples. A permutation-based null model was constructed where for each sample pair, m⁶A values from one sample were randomly permuted across bins 1,000 times while preserving the empirical value distribution and MSE was recomputed to generate a null distribution of similarity values.

Identification of differential interactions using diffHic

Genome-wide differential chromatin interactions between the five samples were identified using the diffHic package (v1.38.0) following the protocol detailed in the “diffHic User’s Guide” (<https://bioconductor posit.co/packages-devel/bioc/vignettes/diffHic/inst/doc/diffHicUsersGuide.pdf>). The steps performed were:

1. Raw Hi-C contact matrices were generated for the 5 samples at 1 kb, 2 kb, and 5 kb resolution using the cooler dump() function. These matrices were used as input for diffHic.
2. Filter out centromeres, telomeres, interactions < 10 kb, interactions > 1 Mb, and interactions where across all 5 samples, the sum of contacts was < 2.
3. Normalize for bias using the normOffsets() function, which allows for cross-sample pixel comparisons. Following the documentation, the correctedContact() was not implemented as ICE normalization is not necessary for cross-sample pixel comparisons.
4. Estimate dispersion and fitting the quasi-likelihood negative binomial GLM model to the data. This step was repeated five times, with each iteration comparing one sample to the remaining four. This generated five sets of differential interaction lists, one for each sample. Our setup kept deletions in the Hi-C dataset as 0 counts, allowing them to be identified as differential between samples.

The resulting interaction lists were filtered to retain pixels with logCPM > -4 and interaction distances between 10 kb and 1 Mb. The logCPM threshold was selected based on the distributions of logCPM and logFC values , optimizing retention of informative contacts while preserving the expected unimodal logFC distribution centered around zero at 1 kb resolution (Supplemental Figures 4-8). Across filtered loci, logFC standard deviations followed a negative binomial distribution with a peak near $\sigma \approx 0.2$, reflecting that most interactions are conserved and a minority are differential. Significant interactions were identified using Bonferroni correction at each resolution, and genome-wide logFC profiles were visualized for each sample. The set of 381 significant interactions at 5 kb was used for mechanistic analyses.

Mechanistic analysis of diffHic differential interactions

Four mechanisms were evaluated for their association with diffHic-identified differential Hi-C signal:

1. Large-scale sequence differences between interaction bins
2. Small-scale sequence differences within interaction bins
3. m⁶A methylation
4. m⁵C methylation

For genetic variation, a leave-one-out (LOO) approach was used to identify the sample with the most extreme value at each differential pixel. For each interaction locus, the sample with a differential number of total bases changed (via SNPs, insertions, and deletions) both between interaction anchor bins and in interaction anchor bins was determined, as well as the sample with a differential logFC value. Samples were considered differential if they exhibited a median absolute deviation (MAD)-based Z-score ≥ 2 and their LOO t-test p-value fell below a Bonferroni-adjusted 0.02 threshold. This was chosen as the threshold because it corresponds to strong outlier deviation in a MAD-based framework while maintaining sensitivity given the sample size ($n = 5$). For bins where both (i) a differential genetic alteration sample and (ii) a differential Hi-C logFC sample were identified, we quantified the match rate between the two. Statistical significance was assessed using a two-sided binomial test under a null probability of 0.2 (reflecting the 1-in-5 chance of a match by random expectation).

M⁶A methylation rates were computed by determining, for each fiber and each bin, the number of methylated adenines divided by the total adenines. Rates were averaged across fibers and across both interacting bins for each differential Hi-C contact. To normalize baseline differences between individuals, genome-wide expected methylation ratios were computed and compared to the observed ratios. Bins with observed log-ratios deviating from expectation by >0.2 were flagged as differential, and the direction of the deviation was recorded. Because increased m⁶A methylation is expected to correlate with reduced chromatin contacts, loci were classified into four categories (+m⁶A, +logFC; +m⁶A, -logFC; -m⁶A, +logFC; -m⁶A, -logFC) based on whether the differential-methylation sample aligned with the minimum or maximum logFC value. The match rate of each set was calculated and statistical significance was assessed using a two-sided binomial test under a null probability of 0.2.

CpG methylation levels were obtained from phased ONT reads. For each resolution, per-bin methylation fractions were computed as the number of methylated cytosines divided by the total number of cytosines within that bin, separately for each haplotype (H1, H2). The methylation fractions of the two interaction bins were averaged to obtain a single value per haplotype per differential interaction. To identify whether a sample exhibited aberrant m⁵C levels at a locus, we applied a similar robust outlier framework as before. Across the 10 haplotypes, methylation values were converted into MAD Z-scores, and the haplotype with the largest absolute Z-score was considered a candidate differential sample. A leave-one-out (LOO) Z-score

was then computed by recalculating the median and MAD excluding the candidate haplotype sample. A haplotype was marked as differential if both the MAD Z-score and the LOO Z-score were ≥ 2 , and the direction of the deviation (“+” for higher methylation, “-” for lower) was recorded. For each differential m^5C event, we compared the direction of methylation deviation with the haplotype-sample’s Hi-C log fold-change (logFC). Because increased CpG methylation is expected to reduce chromatin contacts, loci were classified into four categories ($+m^5C$, $+logFC$; $+m^5C$, $-logFC$; $-m^5C$, $+logFC$; $-m^5C$, $-logFC$) based on whether the differential-methylation sample aligned with the minimum or maximum logFC value. For each category, we computed the proportion of loci in which the m^5C deviation correctly predicted the logFC extremum and assessed significance using a two-sided binomial test with null probability 0.2.

To assess joint explanatory power, differential bins were annotated with any mechanism for which a differential sample was detected, and match-rate analyses (total bases changed between interaction bins and in interaction bins; $+m^6A$, $-logFC$; $+m^5C$, $-logFC$) were calculated. This process was also repeated at lenient p-values greater than the Bonferroni to establish that subthreshold true differential events existed, although the stringent Bonferroni p-value was used for downstream analysis. To evaluate independence, match-rate analyses were recomputed after removing any differential interactions explained by two or more mechanisms, demonstrating that each mechanism retains significance when controlling for the others. A Venn diagram generated using the venn Python package (v0.1.3) summarized the overlap between the match rates of each mechanism in bins that exhibited at least one differential mechanism.

Next, we computed the frequency with which each unique genomic bin appeared across all differential interactions and quantified the proportion that fell within 10 kb of a chromatin loop in the 2D Hi-C matrix as well as within homozygous bin-eliminating deletions, calculating the null through circular permutation ($n = 1000$) by randomizing loop anchor positions at identical genomic distances and differential interaction anchors along the chromosome and testing against fixed homozygous deletion intervals with empirical permutation p-values. We also clustered differential pixels within $10\text{ kb} \times 10\text{ kb}$ of each other in the 2D Hi-C matrix and plotted the distribution of the number of differential pixels within each cluster, stratified by clusters within 10 kb of a loop or within homozygous deletions.

CTCF site determination and PWM score calculation

Because 85% of chromatin loops form at CTCF sites (Rao et al. 2015), CTCF motif locations in each sample were derived by running the FIMO software (v5.3.0) on haplotype-resolved assemblies for each sample using the following command:

```
fimo CTCF.meme GM19317.vrk-ps-sseq.asm-hap1.fasta
```

Results were mapped back to the reference genome GRCh38 using the Long Read Aligner software (v1.3.7.2). This process successfully mapped ~90% of CTCF sites for each sample back to the reference genome (averaging ~49,983 sites per haplotype). This matches expected CTCF site counts per haplotype.

The FIMO output also included the base sequence for each CTCF site. The position weight matrix (PWM) scores were calculated for each sequence for each haplotype. A PWM probability matrix with per-base probabilities at each position in the motif was constructed from JASPAR's 2018 MA0139.1 CTCF frequency matrix. For each haplotype-resolved CTCF site, the corresponding 19 bp sequence was scored by summing the position-specific probability weights associated with each base and dividing the total by the motif length to obtain a motif probability score, which enables direct comparison across the ten haplotypes in our dataset and integrates naturally with downstream quantitative models.

For each CTCF sequence, the number of bases mutated compared to the canonical 15 bp CTCF consensus sequence 5'-NCANNAGRNGGCRSY-3' (Hashimoto et al. 2017) was calculated. To accomplish this, the 15 bp window was aligned across each haplotype sequence to identify the highest-scoring local match. Within the best alignment, bases that violated the consensus rules were counted as mismatches, and the number of mismatches per sequence was recorded.

CTCF site m⁵C methylation

CpG methylation was quantified at single-base resolution using ONT-derived per-nucleotide methylation calls for each haplotype. This data was processed to determine, for each genomic position, the number of reads overlapping that cytosine and the number of reads supporting a methylated call for each sample and haplotype. For each position, a methylation fraction was computed as the number of methylated cytosines / the total number of cytosines. For every haplotype-resolved CTCF motif, all CpG positions falling between the start and end coordinates were retrieved. Multiple CpG methylation events on the same site were averaged, producing a site-level methylation profile for every motif and haplotype.

In parallel, a hidden Markov model (HMM) was applied genome-wide to classify each CpG position as hypomethylated or hypermethylated. These HMM state calls were merged into the CTCF annotations using the same coordinate-based procedure, yielding for each haplotype-specific CTCF site a categorical label of hypomethylated, hypermethylated, or mixed (if CpG positions within the motif belonged to different HMM states).

CTCF site accessibility

M⁶A rates (the number of m⁶A adenines / the total number of adenines) were calculated within each 1 kb bin for each sample. The average m⁶A rate of the four bins (two upstream, two downstream) surrounding each bin that contained a CTCF site was calculated. To account for technical biases from sequencing, the m⁶A rates were Z-scored within sample and

sigmoid-transformed into a score between zero and one, representing how open the chromatin near each CTCF site was compared to the other CTCF sites in the sample. Z-scores were calculated within CTCF sites rather than genome-wide as downstream analysis was focused only on CTCF sites.

Quantifying correlations between genetic and epigenetic factors and occupancy

Each CTCF site across all ten haplotypes was annotated with the computed PWM score, the number of sequence mutations relative to the consensus motif, the mean CpG methylation level, the CpG methylation state (hypomethylated, mixed, hypermethylated), the average accessibility rate, and the Fiber-seq-derived CTCF occupancy value. Global pairwise correlations between each feature and occupancy were computed to assess individual relationships. To quantify the joint association between sequence strength, accessibility, and CpG methylation and their relationship to CTCF occupancy, we fit an ordinary least squares (OLS) regression model using PWM score and methylation as predictors of occupancy and calculated the correlation between observed and fitted values. PWM scores were min–max normalized to the range [0,1], whereas methylation values were z-scored across all motif instances. Rows containing missing values for any of the predictors or occupancy were excluded. Because the goal of this analysis was association rather than prediction, the model was fit to the full dataset without a train–test split. Global significance of the association model was assessed using the F-test for linear regression with two predictors. All analyses were performed in Python using pandas, numpy, scikit-learn, scipy, matplotlib, and seaborn.

Loop discovery using Mustache and HiCExplorer

The following commands were used to discover loops, with appropriate length parameters:

```
hicDetectLoops -m GM19317_5000.cool -o GM19317_HICEX_5000_01.bedpe --maxLoopDistance 2000000 --windowSize 10  
--peakWidth 6 --pValuePreselection 0.1 --pValue 0.1
```

```
mustache -f GM19317.mcool -r 2kb -pt 0.1 -o GM19317_2000_01.tsv -p 50 -st 0.7
```

Chromatin loops were initially called on each resolution of each sample using both Mustache (v1.3.3; MUST) and HiCExplorer (v3.7.5; HICEX) software with p-values of both 0.1 and 0.01. Mustache loops called at 1 kb and 2 kb resolutions included an additional st parameter which was set at 0.7 following recommended settings. All HiCExplorer loops called with the hicDetectLoops() function included additional parameters of maxLoopDistance = 2000000, windowSize = 10, and peakWidth = 6. To compare loop sets, we computed Jaccard indices between callers and between samples after expanding each loop to a ±10 kb neighborhood to account for minor positional differences in peak localization. The same comparisons were also performed on two independent GM19317 technical replicates at 10 kb resolution.

We also calculated Jaccard index values with respect to varying sequencing depth by splitting the read pairs of unbalanced GM19317 cooler files at 1 kb, 2 kb, 5 kb, and 10 kb resolutions in half, balancing both output halves, and calling loops on both halves for all four resolutions using Mustache (HiCExplorer failed to call loops on either of the halves at any of the resolutions). Jaccard index values were then calculated between halves at each resolution, as they were above.

To determine the Hi-C variance at each sample, the Mustache and Hi-C Explorer loop-call sets for each sample at a resolution of 2 kb and p-value of 0.1 were first pooled. Multiple loops within a 50 kb window of each other were merged, and for each sample at each loop locus, the maximum contact count in the Hi-C contact map within a dynamic search window ranging from a 5×5 matrix (± 2 bins) for short-range loops (< 100 kb) up to an 11×11 matrix (± 5 bins) for long-range loops (> 200 kb) was selected as the refined center of the candidate loop. Then, for each sample at each locus, this center value was divided by the background mean for that sample, which was determined by calculating the average of the off-diagonal bins along the genomic distance band appropriate for loop size (± 9 – 14 bins for small loops ≤ 35 kb and ± 21 – 50 bins for larger loops) to give a simple proxy for Hi-C signal (see “CLASH Validation” for further discussion about this metric). Across each locus, the variance in Hi-C signal was then calculated.

CLASH

CLASH is a method that assigns a loop strength score to a provided Hi-C locus based on the structure of Hi-C signal at that locus, primarily accounting for the roughly radial symmetry of decaying contacts from the loop center (Eagen 2018). The input for CLASH is pooled loopcall sets of genomic coordinates where a previous method has called a chromatin loop. For each candidate loop, bins within 10 kb of the diagonal were excluded to avoid including windows with invalid ranges of separation and crossing over the diagonal of the Hi-C matrix. Then, for each called loop, the maximum contact count within a dynamic search window ranging from a 5×5 matrix (± 2 bins) for short-range loops (< 100 kb) up to an 11×11 matrix (± 5 bins) for long-range loops (> 200 kb) was selected as the refined center of the candidate loop. For each candidate loop center, background interaction levels were calculated from averaging off-diagonal bins along the genomic distance band appropriate for loop size (± 9 – 14 bins for small loops ≤ 35 kb and ± 21 – 50 bins for larger loops).

For each refined loop center, we extracted a local Hi-C submatrix with adaptive sizing. Small loops (< 35 kb) always used a 5×5 window (± 2 bins). For larger loops, we iteratively expanded the window from radius 2 up to a cap of radius 10, evaluating the outer “ring” at each step; expansion stopped when the mean ring intensity fell below an adaptive multiple of the

background (linearly decreasing from 1.5 \times at the inner rings toward 1.0 \times at the maximum radius). In sparse neighborhoods (any zero-valued neighbor in the immediate 3 \times 3 around the center), we fell back to a fixed 5 \times 5 window. To detect deletions near the loop center, we flagged submatrices with entire zero rows/columns that continue for ≥ 20 bins beyond the extracted window in the source matrix.

The five extracted sample matrices at a given genomic coordinate were then scored using an adaptive weighted average between two functions. The "division" function calculated the weighted decay of interaction strength from the loop center against each contact in the matrix, and the "hierarchical" function calculated the iterative shell-based decay for each pixel relative to the average of adjacent contacts in the previous shell. Matrices larger than the median size matrix at a given genomic coordinate were rewarded based on the increase in size. All scores were then robustly normalized using a median-absolute-deviation-based sigmoid transformation. The base score of each loop came from applying a weighted average of the two function scores based on the size of the matrix, with smaller matrices being influenced heavier by the hierarchical score and larger matrices being influenced heavier by the division score.

After computing the base scores, additional matrix terms were incorporated to refine the score. A coherence term quantified the variance among pixels within the 3 \times 3 region surrounding the loop center, with lower variance leading to rewards in the score. A contrast term captured the relative enrichment of the loop center compared to the local background, rewarding loops with sharply defined focal interactions. The support term, representing the proportion of nonzero terms, and context enrichment terms, representing the enrichment of the matrix surrounding the loop center, were used to penalize sparse or noisy matrices and to upweight locally enriched regions. These features were combined using a logistic scaling function to smooth transitions between high- and low-quality matrices and added to the base score to yield a score between 0-1 for each candidate loop that serves as a proxy for loop intensity at the given coordinates for each sample, thus allowing for inter-sample differential loop comparison.

A set of loop loci with diverse distributions of Hi-C signal at 2 kb resolution across samples were chosen to validate output scores during CLASH development. CLASH's threshold values, rewards, weights, and overall implementation were determined empirically based on this set of loci. Thus, this set of parameters is meant for 2 kb resolution analysis, which is the resolution we used for all downstream loop-related analysis.

For each matrix M of odd size N x N centered around the loop (L) in sample (S):

Hierarchical score:

- Set the center index of the matrix be $c = \lceil n/2 \rceil$
- Set $D_{\text{Hier}}(i,j)$ as the Chebyshev distance from the center; $D_{\text{Hier}}(i,j) = \max(|i - c|, |j - c|)$.
- For each $(i,j) \setminus (c,c)$, let the set of adjacent matrix elements in the layer closer to the center be $P(i,j) \in \{(p,q) : |p-i| \leq 1, |q-j| \leq 1, D_{\text{Hier}}(p,q) = D_{\text{Hier}}(i,j) - 1\}$ and the reference value $R(i,j) = \text{mean}(\{M(p,q) : (p,q) \in P(i,j)\})$

- For each $(i,j) \setminus (c,c)$, let the hierarchical decay $H(i,j) = 1 - (M(i,j)/(R(i,j) + \varepsilon))$ where $\varepsilon \ll 1$.
- If $H(i,j) < 0$, let enrichment $E(i,j) = (M(i,j)/(R(i,j))$ and $s(i,j) = (E(i,j) - 1)/(2.5)$. Set:

$$s'(i,j) = \min(1, \max(0, s(i,j))),$$

$$H(i,j) = -0.05 - 0.45 (s'(i,j))$$
- For each M , the hierarchical score $H_M = \text{mean } (H(i,j): (i,j) \in N^2, (i,j) \setminus (c,c))$

Division Score:

- Set the center index of the matrix $c = \lceil n/2 \rceil$ and $C = \text{the contact value at } M(c,c)$, the Euclidean radial distance of each matrix element from the center be defined as $D_{\text{div}}(i,j) = \sqrt{(i-c)^2 + (j-c)^2}$, and a corresponding weight for each distance $W(i,j) = 1/\sqrt(D_{\text{div}}(i,j))$
- Let the radial decay of each matrix element be defined as $V(i,j) = (1 - (M(i,j)/(C)) * W(i,j)$
- For each M , the division score $V_M = (\text{mean } (V(i,j): (i,j) \in N^2, (i,j) \setminus (c, c))) / (\sum W(i,j): (i,j) \in N^2, (i,j) \setminus (c, c))$

After H_M , and V_M were calculated for each of the 5 samples at a given locus, they were normalized within the locus across the 5 samples. For a given score type where $x \in (H_M, V_M)$ and x_{LS} is the matrix score for a given locus L in sample S, the median absolute deviation was calculated: $\text{MAD}_L = \text{median}(|x_{LS} - \text{median}(x_{LS})|)$. The Z-score for each sample was then calculated as $z_{LS} = (x_{LS} - \text{median}(x_{LS}))/\text{MAD}_L$. Each Z-score was then sigmoid transformed. This process generated within-locus normalized values H_{LS} and V_{LS} for every locus for every sample.

For all loops, the base score was calculated. $BS_M(n, \text{size}) =$

- $0.65H_{LS} + 0.35V_{LS}, 5 \leq n \leq 7$
- $0.5H_{LS} + 0.5V_{LS}, 7 < n \leq 11$
- $0.35H_{LS} + 0.65V_{LS}, n > 11$

Additional terms. Let $U = \text{mean } (M(i,j): (i,j) \in N^2)$, $U_{3x3} = \text{mean } (M(i,j): i,j \in (-1, 0, 1))$, $\sigma_{3x3}^2 = \text{Var}(M(i,j): i,j \in (0,1))$:

- The contrast score measures how strongly the center of the loop stands out from the background compared to the rest of the matrix. Contrast score = $1/(1 + \exp(-1.3 * (\min(8, \max((C/B + \varepsilon)/(U/B + \varepsilon), 0)) - 2.5)))$
- The coherence score measures how smooth the Hi-C signal in the matrix is. Coherence score = $1 / (1 + (\sigma_{3x3}^2/U_{3x3})^{0.7})$.
- The structural score SS = 0.5(contrast score) + 0.5(coherence score).

- The support fraction is the fraction of the matrix that has counts of non-zero.
 $F_{\text{sup}} = (\# \text{ of nonzero matrix elements}) / N^2$.
The zero fraction = $1 - F_{\text{sup}}$.
- The context enrichment measures the enrichment of the center 3×3 matrix elements compared to the background. $E_{\text{ctx}} = U_{3 \times 3} / B + \epsilon$
- A center dominant threshold was implemented to calculate if the center of the matrix was extremely enriched compared to the surrounding 3×3 ring of matrix elements. If $C / (\text{mean}(N_{3 \times 3}) + \epsilon) > 1.8$, then `center_dominant` = 1.
- An off-center threshold was implemented to calculate whether there is a signal in the matrix apart from the center. If $U/B < 1.2$, then `off_center` = 1
- The proportion of matrix elements with values above the background was calculated.
Fraction above background = $(\# M(i,j) > B) / N^2$
- A high neighbor threshold was implemented to determine if at least one non-center matrix element had a high value. This is true if there exists a $M(i,j) \in N^2$, $(i,j) \neq (c,c)$ such that $M(i,j) \geq 0.5C$.
- Severity aggregates the effects of some of the prior terms:
 - If `center_dominant` = 1, $S += 0.55$
 - If `off_center` = 1, then $S += 0.35$
 - If coherence < 0.65, then $sev += (0.65 - \text{coherence}) * 0.9$
 - If $F_{\text{sup}} < 0.75$, then $sev += (0.75 - \text{support_frac}) * 1.1$
 - If $E_{\text{ctx}} < 1.05$, then $sev += (1.05 - E_{\text{ctx}}) * (2.8 + (0.8 - \text{coherence}) * 1.1)$
- A penalty term for severity was defined as $P_{\text{Sev}} = 0.45 + (0.55 / (1 + e^{3 * (sev - 1.1)}))$
- A reward term was defined for matrices that exhibit strong signal-to-background contrast and coherence. If contrast > 2.0 and coherence > 0.346, then $R = 0.25 * \tanh(\text{contrast} / 3.8)$. Else $R = 0$.
- Another term W was implemented to handle additional matrix cases:
 - For matrices with both low center contact counts (`center_count < 5`) and low surrounding counts (`high_neighbors = false`), $W = 0.5$. However, matrices with low center counts but 75% of the matrix values were above the background value, $W = 1.5$. For other matrices where the `center_count < 5`, $W = \max(0.3, (\text{center count} / 5)^2)$
 - For matrices with $n < 11$ and `zero_fraction >= 0.33`, $W = 0.5$
 - For all other matrices where the `center_count > 5`, $W = 1$
- Another reward term `Boost` was added to slightly larger-than-median sized matrices (relative to the potential loops at the same loci across samples): $\text{Boost} = 0.02 * \max(0, \text{size} - \text{median_size}) * (0.8 + 0.4 * \text{coherence})$.

The final CLASH score was $C_M = \min((1.15 * W * (((BS_M + SS_M) * P_{\text{Sev}}) + R)) + \text{Boost}, 1.0)$

CLASH validation

A primary metric we used to validate CLASH scores in the absence of a ground truth was $\log(\text{center}/\text{background})$, where center is the Hi-C contact count of the loop center and background is the average local distance-matched contact count, as described above. Many existing loop-calling methods use some form of enrichment as calculated by center/background with various definitions of what constitutes the background to call loops, and Mustache and FitHiC specifically use distance-matched contacts as the background (Durand et al. 2016; Open2C et al. 2024; Wolff et al. 2022; Roayaei Ardakany et al. 2020; Kaul et al. 2020). Additionally, this metric is mostly independent from CLASH’s scoring function; while CLASH does calculate the local distance-matched background contact count, this value is primarily used in determining the size of the extracted matrix as matrix expansion ends when the average of the next added layer would approach this background value. Although this necessitates that the “division” function score does calculate a similar center/outer ring value, the relative contribution of this term to the final CLASH score is minimal due to the weighting of the center/outer ring value compared to the center/inner rings and due to the weighting of the final division score in determining the final loop score.

Using this metric, we compared CLASH’s loop-scoring performance to the initial pooled binary calls from Mustache and Hi-C Explorer. First, the correlation between loop score and Hi-C signal (calculated as $\log(\text{center}/\text{background})$) was compared for both datasets. Next, mutual information (MI) between each predictor and Hi-C signal was estimated using `mutual_info_regression`, and the difference in MI was assessed using a 200-iteration bootstrap test. Finally, per-locus MSE between predicted signal (initial or CLASH) and observed Hi-C signal across the five samples was computed, the fraction of loci showing MSE improvement under CLASH was recorded, and paired differences were tested using a Wilcoxon signed-rank test. All analyses were performed in Python using Pandas, Numpy, Scipy, Scikit-learn, Seaborn, and Matplotlib.

We also compared CLASH’s loop-scoring performance to “LC scores”, a continuous score-like intermediary that was derived from the Mustache algorithm (Roayaei Ardakany et al. 2020). Briefly, these scores were calculated as the Difference of Gaussians on z-scores of identified loops normalized to distance-matched contacts between consecutive scales. For each pixel, this Difference of Gaussians is calculated for each set of consecutive scales and the largest difference is chosen as the LC score, with higher scores being “stronger” loops, and higher scales being used to identify broader loops. We calculated these LC scores for each loop locus called by Hi-C Explorer and Mustache across all five samples, and resulting values ranged from -10^{-3} to 10^{-3} .

To compare CLASH and LC scores, we first calculated Pearson correlations comparing within-locus and globally pooled CLASH scores and LC scores with $\log(\text{center}/\text{background})$. To

investigate why LC scores performed well at ranking samples within locus (similarly to CLASH) but correlations broke down globally, we rank-normalized both the CLASH and LC scores globally and computed the variance in log(center/background) within each score bin, as we reasoned that similarly scoring loops should have low variance in Hi-C signal. We reasoned that global LC score discrepancies arise from locus-specific scale selection: different scales correspond to different kernels that output different scores, and taking the maximum over these scale-specific scores introduces an upward bias precluding score comparisons between different scales. To test this, we computed the same variance calculation as above, but decomposed the variance into within-scale variance and between-scale variance for LC scores. To match this procedure for CLASH scores, we also computed within-matrix-size and between-matrix-size variance.

We also performed orthogonal tests to externally validate CLASH's loop-scoring performance that do not rely on Hi-C signal by computing Pearson correlations between CLASH and LC scores with CTCF occupancy globally and within-locus across samples. Finally, as an independent validation of CLASH, we computed the enrichment of eQTLs in CLASH-derived putative iQTLs, as described below.

Quantifying the correlation between CTCF occupancy and CLASH loop-score

To quantify how CTCF protein occupancy globally relates to chromatin loop strength, CLASH-scored loops were first filtered to retain only those in which both loop-associated interaction bins contained a CTCF site within 10 kb. For each loop, the CTCF occupancy values of its two CTCF sites were averaged to generate a single occupancy value. If CTCF occupancy was missing, the other site's CTCF occupancy value was used, and if both were missing then the datapoint was excluded from analysis. This per-loop occupancy metric was then paired with the corresponding CLASH loop score, and Pearson correlations were computed across all loci and samples to assess the global relationship between CTCF binding and loop strength.

We also quantified the distribution of Pearson correlations between CTCF occupancy and CLASH loop-strength between samples at each locus. Loci with fewer than four samples containing both valid occupancy and loop score measurements were excluded. To ensure sufficient dynamic range for meaningful regression, loci were further filtered to retain only those with loop-score standard deviation greater than 0.29 (chosen after computing correlations and statistical significance across different standard deviation values) across samples as loci with minimal across-sample variation do not contain sufficient dynamic range to support reliable Pearson correlation. We collected all per-locus correlations and performed a binomial sign test to calculate statistical significance. All analyses were conducted in Python using Pandas, Numpy, Scipy, Seaborn, and Matplotlib.

Quantifying the correlation between CTCF occupancy and CTCF insulation capability

To investigate the effect of CTCF occupancy on CTCF's role as an insulator, we first used cooltools (v0.7.0) insulation to call insulating bins in our 5 kb resolution cooler files with the following command:

```
cooltools insulation -p 8 -o "$outfile" --view keep_chroms_filtered.bed --min-frac-valid-pixels 0.3 GM19317_5000.cool 100000
```

This process only successfully completed chromosome 1 for each sample, but provided sufficient data points (~5132 bins per sample) for analysis. We used this list of insulating bins to classify all of the CTCF sites within chromosome 1 for each sample as either a "boundary CTCF site", responsible for creating TAD domains, or a "non-boundary CTCF site" which does not form TAD domains. For each of the insulating bins, a boundary score was provided by cooltools insulation. For every CTCF site, boundary and non-boundary, a custom log₂ratio score was implemented based on an existing function from (Crane et al. 2015), to quantify each CTCF site's ability to act as an insulator. At 5 kb resolution, each CTCF coordinate was rounded to the nearest bin and a 200 kb (± 100 kb) window of the Hi-C contact matrix was extracted. The upper triangular portion for each of these matrices was divided into 3 sections relative to the CTCF site's reference bin:

- x1 (−/− triangle): contacts upstream of the site.
- x2 (+/− and −/+ square): contacts between downstream–upstream bins.
- x3 (+/+ triangle): contacts downstream of the site.

The log₂ratio score was defined as $\log_2((x_1 + x_3) / (x_2))$ and was similar to (Crane et al. 2015). Log₂ratio scores were calculated for each CTCF site. Each CTCF site also contained CTCF occupancy data, as previously determined. Pearson's correlation coefficients were calculated within each sample to compare protein occupancy levels with the log₂ratio insulation score across both boundary and non-boundary CTCF sites and averaged across samples. The significance of the difference between the boundary and non-boundary values was determined using a two-sided Welch's t-test on the per-sample correlation coefficients.

Quantifying the correlation between PWM scores, m⁵C methylation, CTCF occupancy, and CLASH loop scores

For each chromatin loop in each sample, PWM scores and CpG methylation levels from the two loop-associated CTCF sites were averaged to generate per-loop measures of motif strength and methylation. These values were paired with CLASH loop scores and occupancy values from earlier, and global Pearson correlations were computed to quantify the individual relationships between motif strength, methylation, and loop intensity. To test whether PWM and methylation provided explanatory power beyond CTCF occupancy, we compared the correlation between occupancy and loop-strength to the multiple correlation of a combined association

model incorporating occupancy, PWM scores, and CpG methylation using Steiger's test for dependent correlations, which accounts for shared outcomes and predictor non-independence. This analysis provided a direct statistical assessment of whether adding genetic (PWM) or epigenetic (m^5C) features improved explanatory power relative to occupancy alone. We performed an analogous comparison using PWM and methylation as the only predictors to evaluate their joint contribution independent of occupancy.

Mediation analysis

To quantify how much of the effect of CTCF motif strength (PWM score) and CpG methylation on loop strength is transmitted through CTCF occupancy, we performed a product-of-coefficients mediation analysis within each sample. We treated PWM scores or CpG methylation as the predictor, CTCF occupancy as the mediator, and CLASH loop strength as the outcome. For each predictor, mediated effects were estimated using linear regression to obtain the product of the predictor-occupancy (a) and occupancy-loop (b) paths, along with the corresponding direct (c) and total ($a \times b + c$) effects and the proportion of the total effect mediated. The percent mediated was calculated as $(a \times b) / (a \times b + c)$ and averaged across all 5 samples. Uncertainty was measured using a pooled nonparametric bootstrap in which loci were resampled within each sample and per-sample indirect effects were recomputed, averaged across samples for each bootstrap iteration (3,000 iterations), and compared to the mean total effect to derive bootstrap confidence intervals.

eQTL analysis

To test whether SNPs within loop-associated CTCF sites that modulate loop strength (putative interaction quantitative trait loci, iQTLs) are enriched for known expression QTLs (eQTLs), we obtained the GTEx v8 Whole Blood significant variant–gene pair dataset. For each loop containing a SNP within either loop-associated CTCF site, we recorded whether the variant was present in the GTEx eQTL list. Separately, we quantified whether each variant significantly altered loop strength by comparing the mean CLASH loop scores between individuals carrying the reference versus alternate allele and evaluating a series of loop strength difference thresholds. For each threshold (0.2–0.6), we computed the proportion of SNP–loop pairs classified as iQTLs and the proportion overlapping GTEx eQTLs. Enrichment was assessed by calculating the log odds ratio of eQTL overlap across increasing loop-strength thresholds.

Structural variation

We next investigated whether large SVs (insertions and deletions) that introduce or remove CTCF binding sites lead to measurable changes in chromatin loop formation. Structural variant calls for each sample haplotype, aligned to the GRCh38 reference genome, were filtered to retain only those altering at least 19 bp (the length of the CTCF consensus motif). FIMO motif scanning was then performed on each SV sequence to identify the number and positions of

CTCF motifs gained or lost. These SV-associated motif changes were subsequently intersected with CLASH-derived loop scores to assess whether the addition or removal of CTCF sites corresponded to loop strengthening or weakening.

For each loop locus, and for each sample, we then determined whether either haplotype carried an SV that added at least one new CTCF site (insertion) or removed at least one CTCF site (deletion) within either loop-associated interaction bin. Across the five individuals, this yielded two groups of loop strength measurements at each locus: (i) samples whose haplotypes contained a CTCF-altering SV at the interaction bin, and (ii) samples without such an SV. For each locus and for each SV class, we computed the change in loop strength as the difference between the mean CLASH loop score of the SV group and the non-SV group. Statistical significance of these differences was assessed using a two-sided Mann–Whitney U test comparing loop scores between SV and non-SV samples at each locus. Genome-wide distributions of Δ loop score values for CTCF-adding insertions and CTCF-removing deletions were then summarized and visualized to quantify the directional impact of SVs on chromatin loop formation.

Figures.

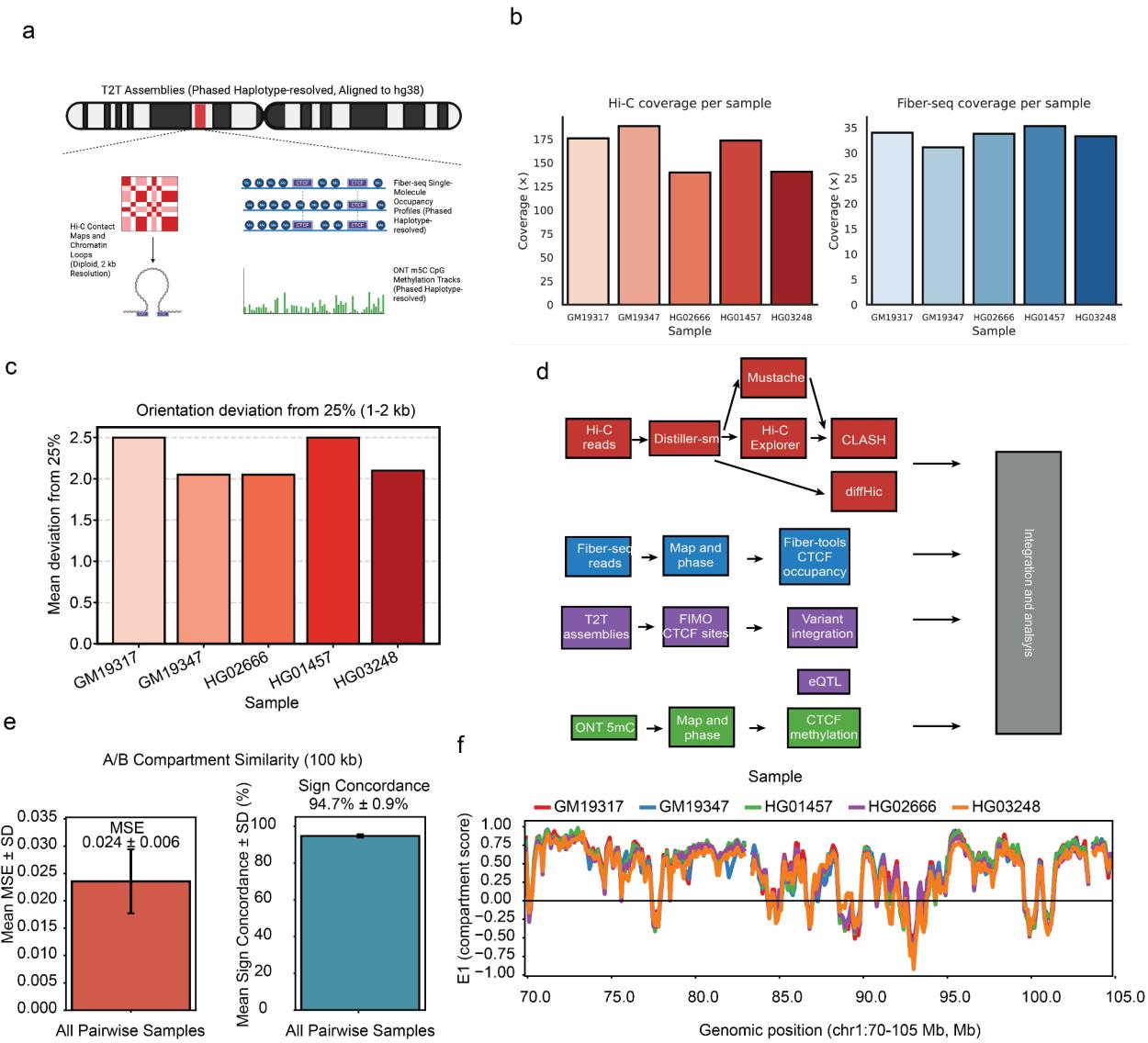


Figure 1. Multi-omic data generation, processing, and quality assessment across five lymphoblastoid samples. **a**, Overview of all datasets generated and integrated in this study, including Hi-C contact maps, Fiber-seq m^6A -based chromatin accessibility and CTCF occupancy, ONT-derived m^5C CpG methylation, and haplotype-resolved variant calls for GM19317, GM19347, HG01457, HG02666, and HG03248. All non-Hi-C data modalities are both phased and haplotype-aware. Hi-C contact matrices were aligned to the GRCh38 reference genome, and Fiber-seq, m^5C methylation, and variant data were aligned or lifted over to GRCh38 to enable joint analysis. **b**, Schematic overview of the computational workflows used in this study, including Hi-C processing and loop calling (distiller-sm, DiffHiC, Mustache, HiCEexplorer, and CLASH), Fiber-seq alignment and CTCF occupancy calling (whatshap and fibertools), m^5C CpG methylation phasing, structural variant integration, and CTCF motif identification using FIMO.

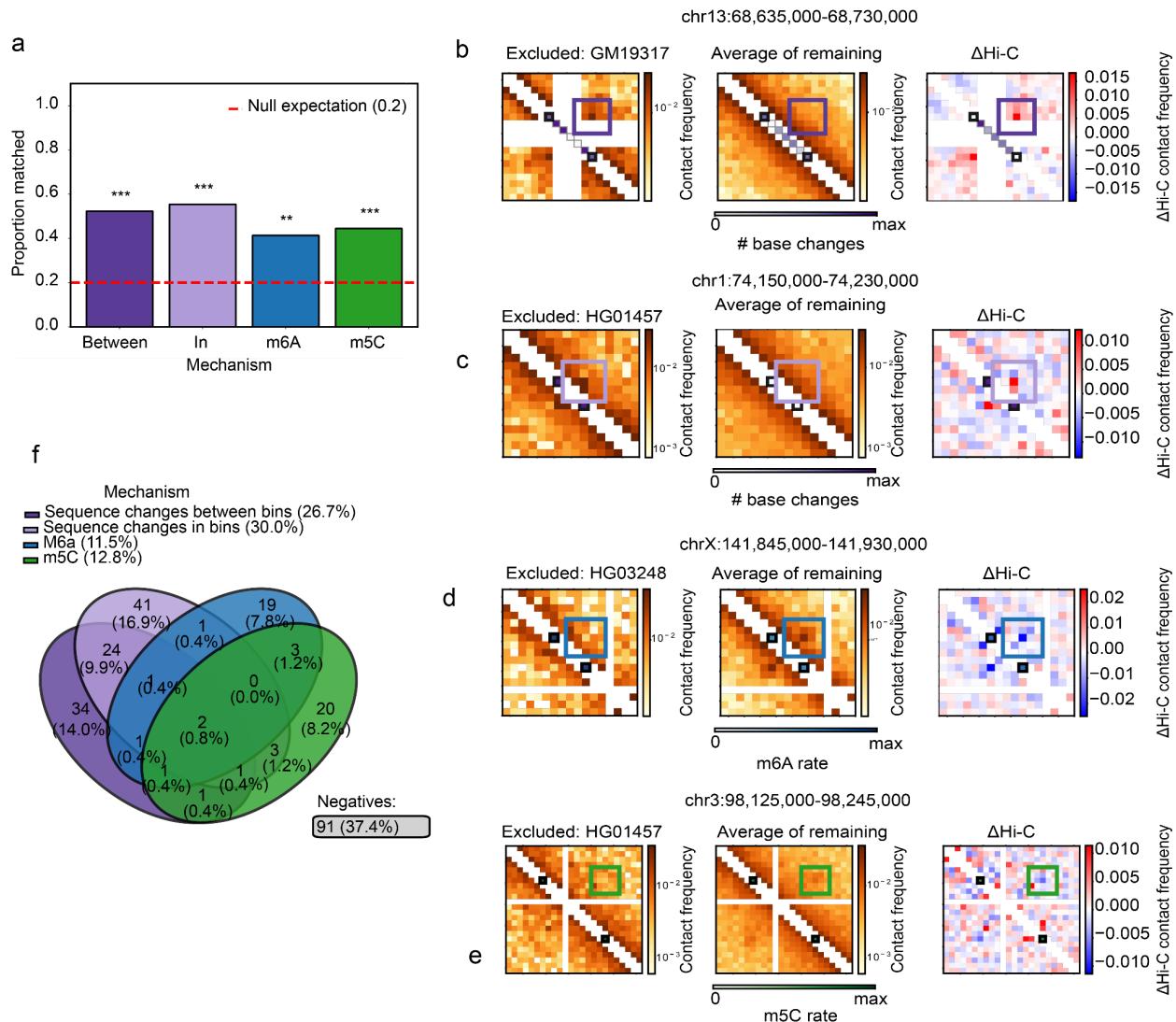


Figure 2. Mechanisms affecting diffHiC-identified differential chromatin interactions. **a,** Match rates for four mechanisms – sequence changes between interaction bins (dark purple), sequence changes within interaction bins (light purple), m⁶A methylation (blue), and m⁵C methylation (green) – after accounting for the influence of the other three mechanisms. All mechanisms remain significantly associated with differential chromatin contacts (between interaction bins: n = 65, p = 7.39×10^{-9} , 52.3%; in interaction bins: n = 74, p = 1.97×10^{-11} 55.4%; m⁶A: n = 46, p = 0.0012, 41.3%; m⁵C: n = 45, p = 2.20×10^{-4} , 44.4%). Significance levels are denoted as p < 0.05 (*), p < 0.01 (**), and p < 0.001 (***)¹, and were calculated using the binomial test compared against the null match rate of 0.2. **b–e,** Example of a differential chromatin interaction driven by sequence changes between interaction bins. Left: Hi-C contact map for the excluded sample (GM19317) at chr13:68,635,000–68,730,000. Middle: Average Hi-C contact map of the remaining four samples. Right: ΔHi-C map (GM19317 – mean of others), highlighting the interaction that differs most strongly in GM19317. The purple diagonal

bins mark the total number of base changes between the two interacting bins of the differential interaction, and the purple rectangle centers around the focal contact whose strength deviates in GM19317. Compared to the other samples, GM19317 carries a greater number of sequence changes between interaction bins, corresponding to an increase in contact strength (red shift). **c**, Example of a differential chromatin interaction at chr1:74,150,000-74,320,000 driven by sequence changes in interaction bins. Shown in the same layout as panel b. Interaction bins are annotated in purple for the number of bases changed within the bins. The highlighted interaction shows increased contact strength in HG01457 (red shift). **d**, Example of a differential chromatin interaction at chrX:141,845,000-141,930,000 aligning with differential accessibility in interaction bin anchors. Shown in the same layout as panel b. Interaction bins are annotated in blue for the chromatin accessibility rate in each bin. The highlighted interaction shows decreased contact strength in HG03248 (blue shift). **e**, Example of a differential chromatin interaction at chr3:98,125,000-98,245,000 driven by m⁵C methylation rate in interaction bins. Shown in the same layout as panel b. Interaction bins are annotated in green for the m⁵C methylation rate in each bin. The highlighted interaction shows decreased contact strength in HG01457 (blue shift). **f**, Venn diagram summarizing the 243 differential bins exhibiting at least one differential mechanism. In total, 62.6% of these bins can be explained by one or more of the four mechanisms shown in panel a.

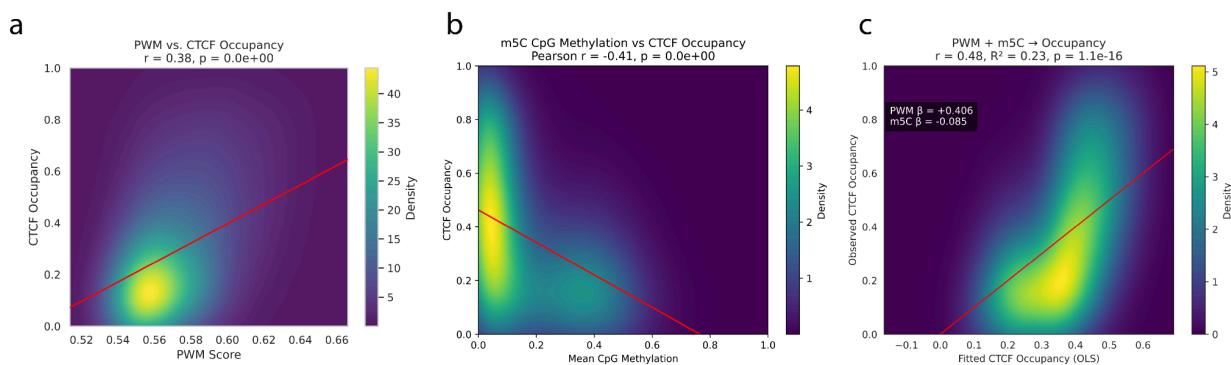


Figure 3. Quantification of genetic and epigenetic mechanisms influencing CTCF occupancy. **a**, Pearson correlation between CTCF motif strength and occupancy. Position weight matrix (PWM) scores were computed for a total of 509,826 CTCF motif instances, aggregated across 10 haplotypes, and correlated with Fiber-seq–derived CTCF occupancy (n = 262,463 sites; Pearson r = 0.38; p < 2.2 × 10⁻³⁰⁸). **b**, Pearson correlation between m⁵C CpG methylation and CTCF occupancy. Average m⁵C methylation percentage across CpG sites within each motif was correlated with occupancy (n = 90,672 sites; Pearson r = −0.41; p < 2.2 × 10⁻³⁰⁸). **c**, Linear

model estimating the combined effects of PWM score and m⁵C methylation on CTCF occupancy using ordinary least squares regression ($n = 90,672$ sites). Both predictors showed significant associations with occupancy ($\text{PWM } \beta = +0.406$; $\text{m}^5\text{C } \beta = -0.085$), and the model explained a moderate fraction of variance (Pearson r between observed and fitted values = 0.48; $R^2 = 0.23$; $p = 1.1 \times 10^{-16}$).

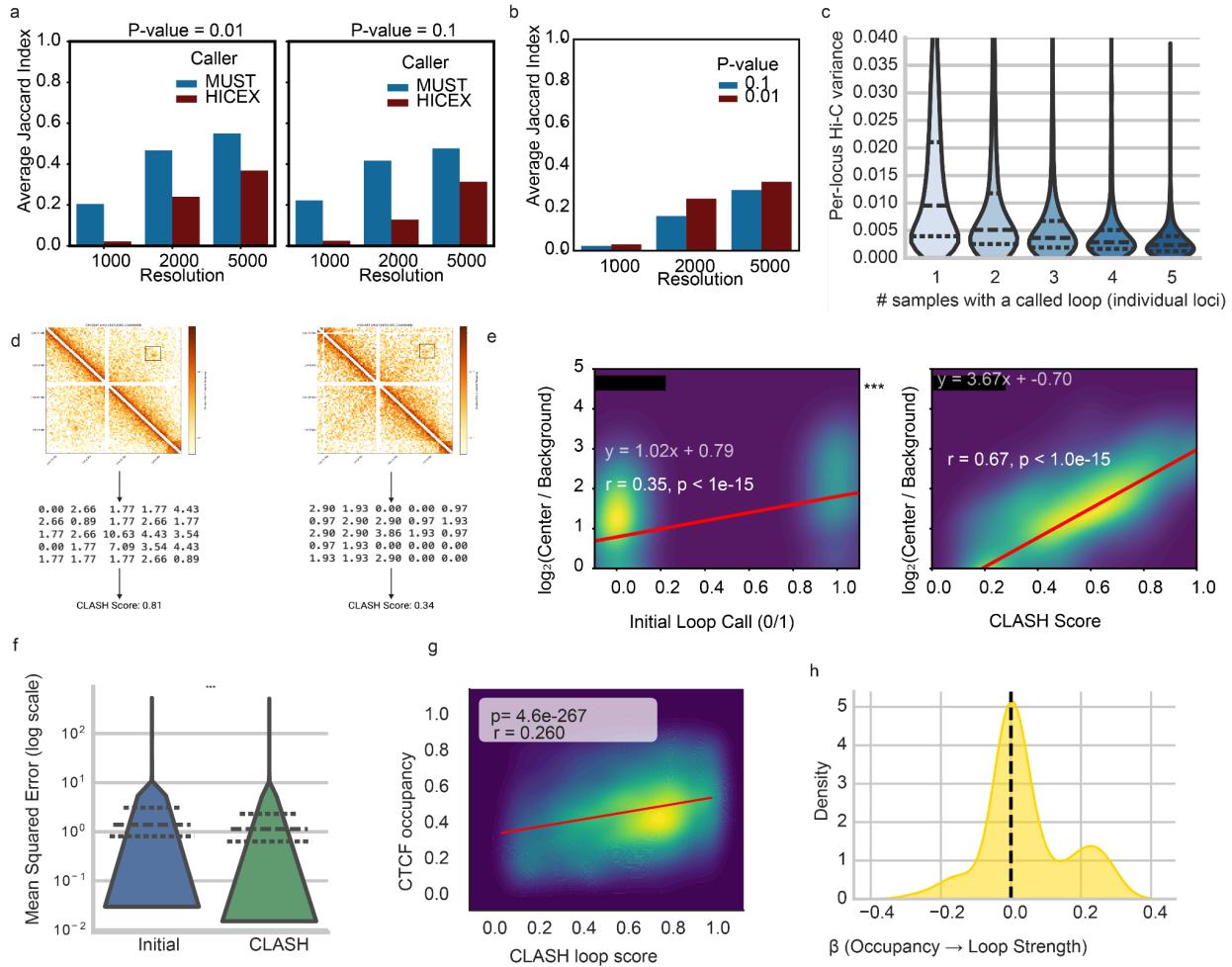


Figure 4. CLASH unification of loop calls and relation to CTCF occupancy, loop strength, and TAD organization. **a**, Average pairwise Jaccard index of loop calls between samples with a tolerance of 10 kb, shown separately for Mustache (MUST; blue) and Hi-C-Explorer (HICEX; red), across all resolutions. Values are displayed for loops filtered at significance thresholds of $P = 0.1$ (left) and $P = 0.01$ (right). **b**, Average Jaccard index of loop calls within each sample between Mustache and Hi-C-Explorer at matched resolutions, shown for $p = 0.1$ (blue) and $p = 0.01$ (red). **c**, Hi-C contact variance, measured by center contact count / local background count, at each locus across samples versus the number of samples with initial loop calls. Calls are pooled across Mustache and Hi-C-Explorer. **d**, Schematic illustration of how CLASH scores loops at the same locus (representative region of chr2:134741565-134936468 for samples GM19347 and HG01457) by extracting the matrix around the local maximum count. **e**, Global

biological concordance between Hi-C signal and initial binary loop-scores (left), and between Hi-C signal vs CLASH scores (right). CLASH nearly doubles the initial Pearson correlation from $r = 0.36$ to $r = 0.66$ ($\Delta r = +0.30$; $p < 2.2 \times 10^{-308}$; Fisher's r-to-z test). **f**, The per-locus improvement by CLASH compared to initial binary callers quantified using mean squared error (Average Δ MSE per locus = -0.35; $p < 2.2 \times 10^{-308}$, Wilcoxon signed-rank test). **g**, Global correlation between CTCF occupancy and loop strength across all samples ($n = 17,439$, Pearson $r = 0.25$, $p = 5.8 \times 10^{-252}$). **h**, Distribution of Pearson correlation values between CTF occupancy and loop strength across loci with CLASH score standard deviation ≥ 0.29 , showing that the mean r is 0.26 (Binomial sign test $p = 3.6 \times 10^{-3}$, $n = 21$). Significance levels are denoted as $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***)�.

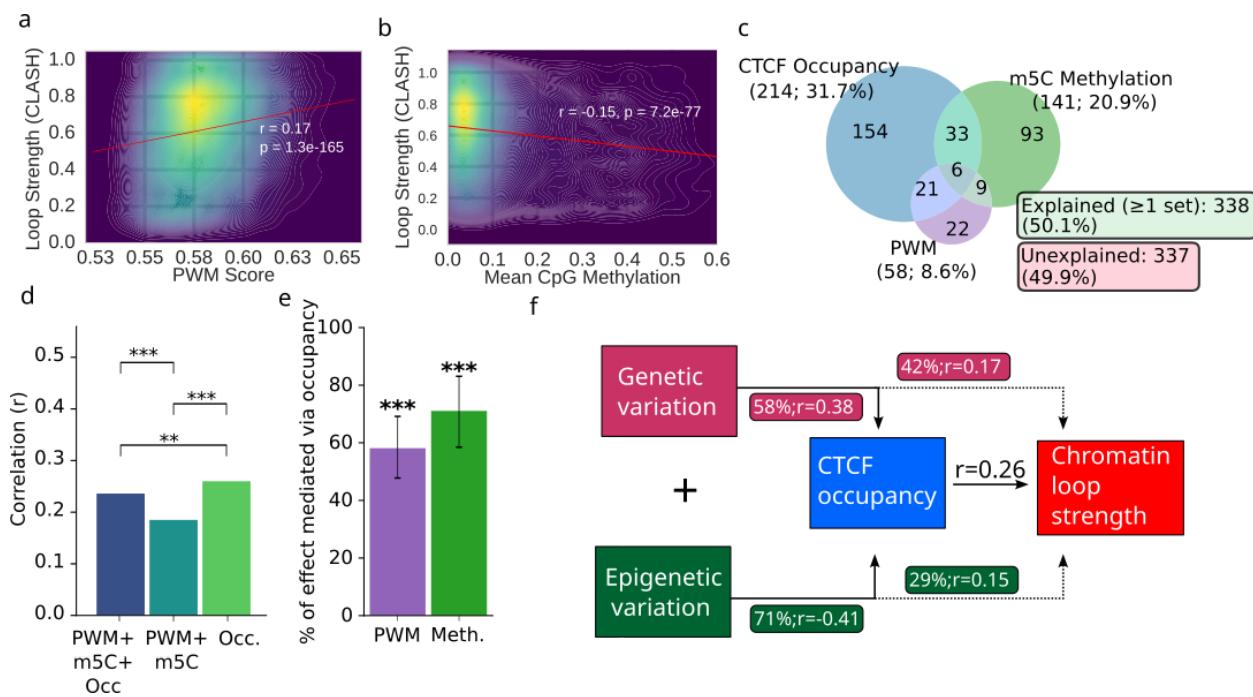


Figure 5. PWM score and CPG methylation correlations with loop scores, mediation analysis, and structural variation analysis. **a**, Correlation between PWM score (averaged across both haplotypes and both loop-associated CTCF sites) and loop strength ($n = 26,557$ loci, Pearson $r = 0.17$, $p = 8.8 \times 10^{-171}$). **b**, Correlation between m⁵C methylation level at loop-associated CTCF sites and loop strength ($n = 15,932$ loci, Pearson $r = -0.15$, $p = 2.1 \times 10^{-78}$). **c**, Proportion of differential loops ($n = 624$ loci with CLASH score range ≥ 0.6) explained by each mechanism using $r \geq 0.25$ as the explanatory threshold per-locus (CTCF occupancy = 33.2%; m⁵C methylation = 22.3%; PWM score = 8.8%; together explaining 52.1% of differential loops). **d**, Comparison of correlation of loop strength with combined genetic, epigenetic, and occupancy features (left) versus correlation of loop strength with combined genetic and epigenetic features (center) versus correlating loop strength with occupancy alone (right).

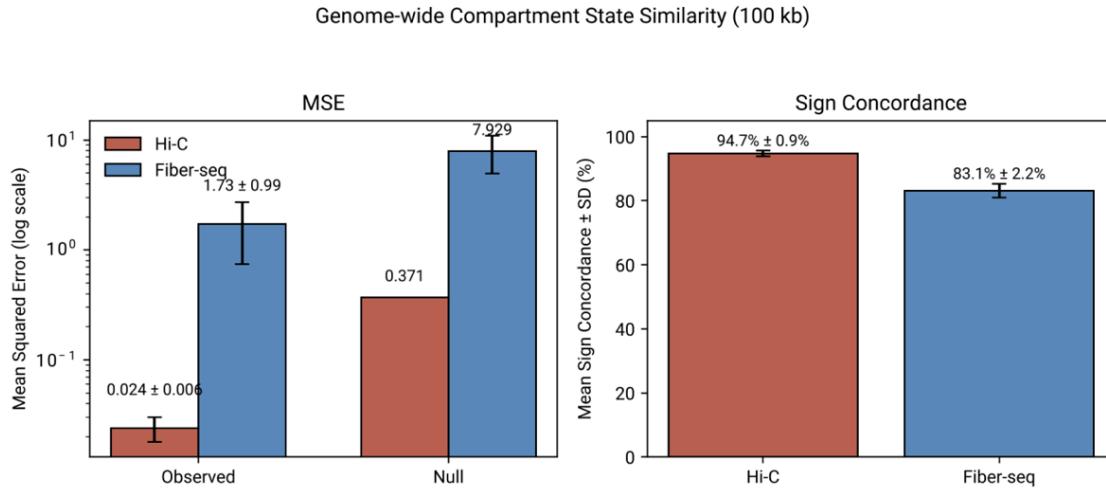
Including all three mechanisms together reduces the correlation ($n = 4,489$, $\Delta r = -0.019$, $p = 0.042$; Steiger test). Excluding occupancy substantially reduces the correlation ($n = 4,489$; $\Delta r = -0.061$; $p = 1.5 \times 10^{-4}$; Steiger test). **e**, Product-of-coefficient mediation analysis quantifying the per-sample proportion of PWM (63%) and m⁵C (59%) effects on loop strength that act through CTCF occupancy, averaged across samples. Black bars denote 95% bootstrap confidence intervals on the mediated proportion, estimated by pooled bootstrap of the indirect effect. **f**, Schematic summary of the unified model: genetic and epigenetic variation modulates CTCF occupancy, which in turn regulates loop strength. Significance levels are denoted as $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***)�.

Table S1. Hi-C sequencing and mapping summary statistics across the five samples.

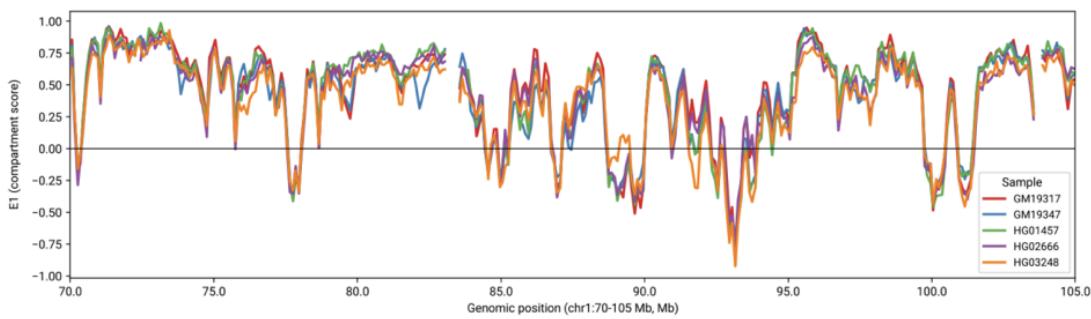
Sample	GM19317	GM19347	HG02666	HG01457	HG03248
Total read pairs (millions)	1813.6	1946.7	1440.9	1789.8	1449.8
% unmapped	31.40%	32.30%	28.50%	31.10%	30.10%
% one-sided	10.30%	10.20%	10.70%	9.90%	10.00%
% two-sided	58.30%	57.50%	60.80%	59.00%	59.90%
% duplicated	12.80%	14.20%	12.10%	19.20%	11.90%
Unique read pairs (millions)	824.9	842.8	701.1	712	695.9
Cis interactions (count)	658065376	648963062	576398312	596436231	616383635
Trans interactions (count)	166854257	193862696	124683354	115560791	79549255
FF (1-2kb)	24.60%	25.10%	24.90%	24.60%	24.70%
RF (1-2kb)	20.60%	20.90%	21.10%	20.60%	21.30%
FR (1-2kb)	30.00%	28.90%	29.10%	30.00%	29.20%
RR (1-2kb)	24.80%	25.10%	24.90%	24.80%	24.80%
Coverage (x)	175.76	188.65	139.64	173.44	140.50
Suggested map resolution (Rao, bp)	2500	2450	2950	2900	2950

Table S2. Fiber-seq summary statistics across the five samples.

Sample	GM19317	GM19347	HG02666	HG01457	HG03248
Coverage (x)	34	31.1	33.8	35.3	33.3
Mean read length (bp)	20850	22591	24462	21151	14236
Hifi yield (gbp)	105	96	105	109	103

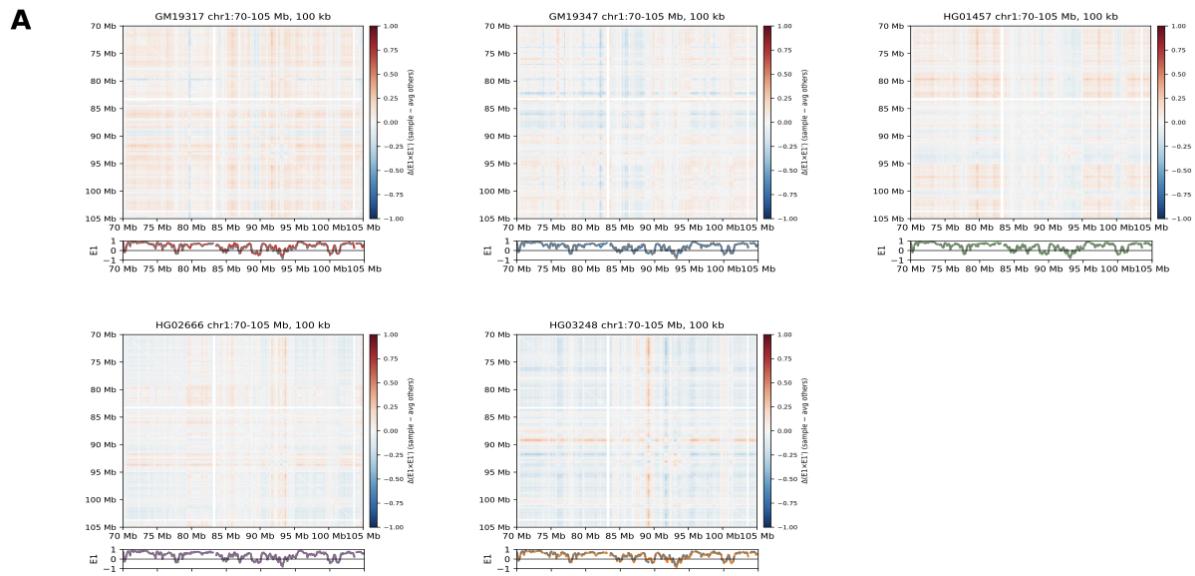


Supplementary Figure 1. Pairwise comparisons of A/B compartment eigenvector values (E1) and accessibility profiles across all five samples at 100 kb resolution. Hi-C A/B compartment profiles show high similarity with an average MSE of 0.024 ± 0.006 (red) and sign concordance, calculated as the fraction of bins with the same E1 sign, of $94.7\% \pm 0.9\%$ across sample pairs. The null MSE of 0.371 was defined by the expected MSE if A/B compartments between samples were not correlated and calculated as E1 variance. Similarly, Fiber-seq accessibility compartment profiles show high similarity, with an average MSE of 1.731 ± 0.988 (red) and sign concordance of $83.1\% \pm 2.2\%$ across sample pairs. MSE was calculated by mean-centered per-bin signals (m^6A methylated adenines/total adenines), and the sign of each compartment was determined by the direction of the m^6A deviation from the mean of that sample for each bin. The null MSE of 7.929 was defined by the expected MSE if compartments between samples were not correlated and calculated by randomly permuting mean-centered per-bin m^6A values across genomic bins.

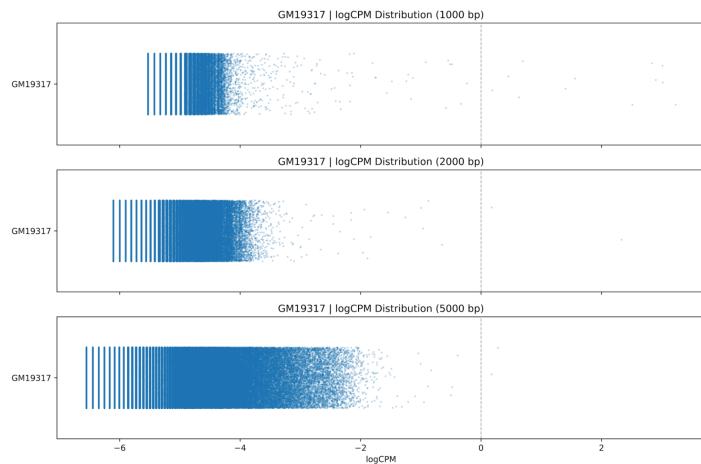


Supplementary Figure 2. E1 compartment eigenvector profiles at 100 kb resolution for all five samples plotted across a representative region of chromosome 1 (chr1:70–105 Mb).

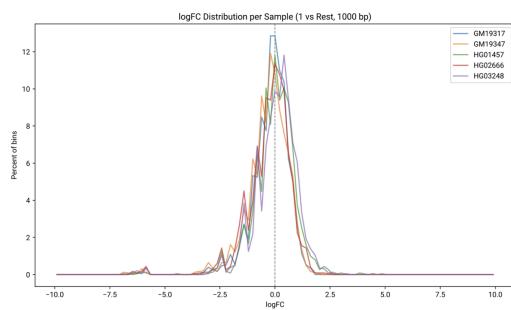
Colors indicate GM19317 (red), GM19347 (blue), HG01457 (green), HG02666 (purple), and HG03248 (orange), and samples exhibit largely similar eigenvector profiles. The region chr1:81,800,000-82,600,000 shows a decrease in GM19347 E1 scores compared to the other samples, and this can be attributed to GM19347 H1 containing $>5\times$ more total bases deleted in that region than the other 9 haplotypes, likely disrupting enhancers within that region.



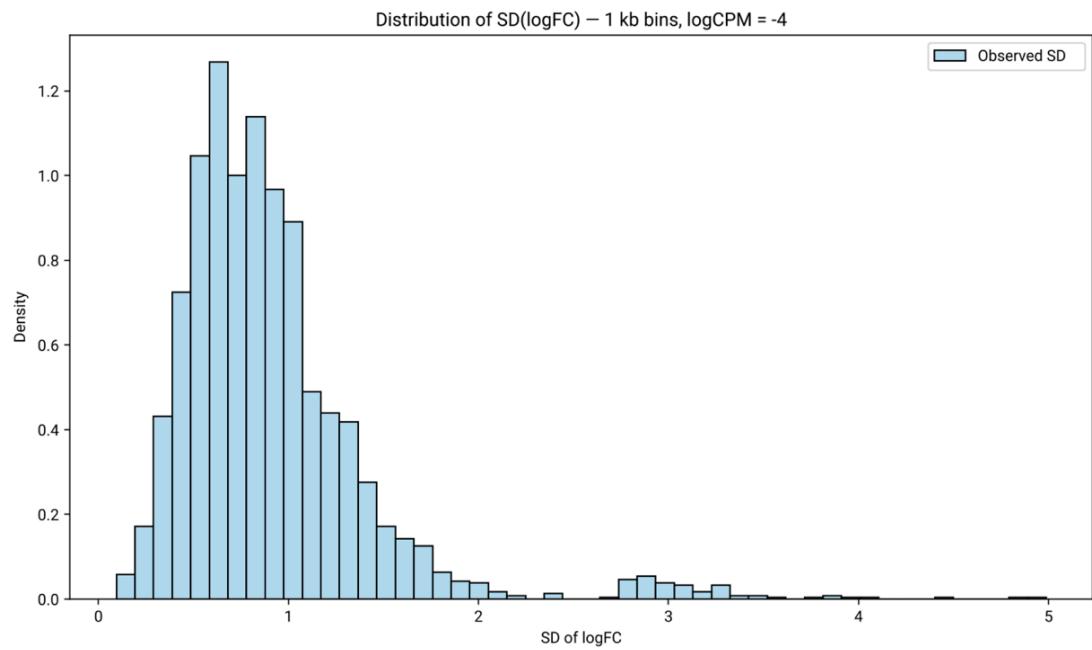
Supplementary Figure 3. Hi-C contact maps at 100 kb resolution for the representative region chr1:70–105 Mb for all five samples. For each individual, the mean contact map of the other four samples was subtracted to highlight sample-specific deviations. A/B compartments were computed using the first eigenvector (E1), with red indicating A compartments and blue indicating B compartments. The corresponding E1 tracks are shown below each heatmap and include both the sample-specific E1 profile and the average profile from the other four samples.



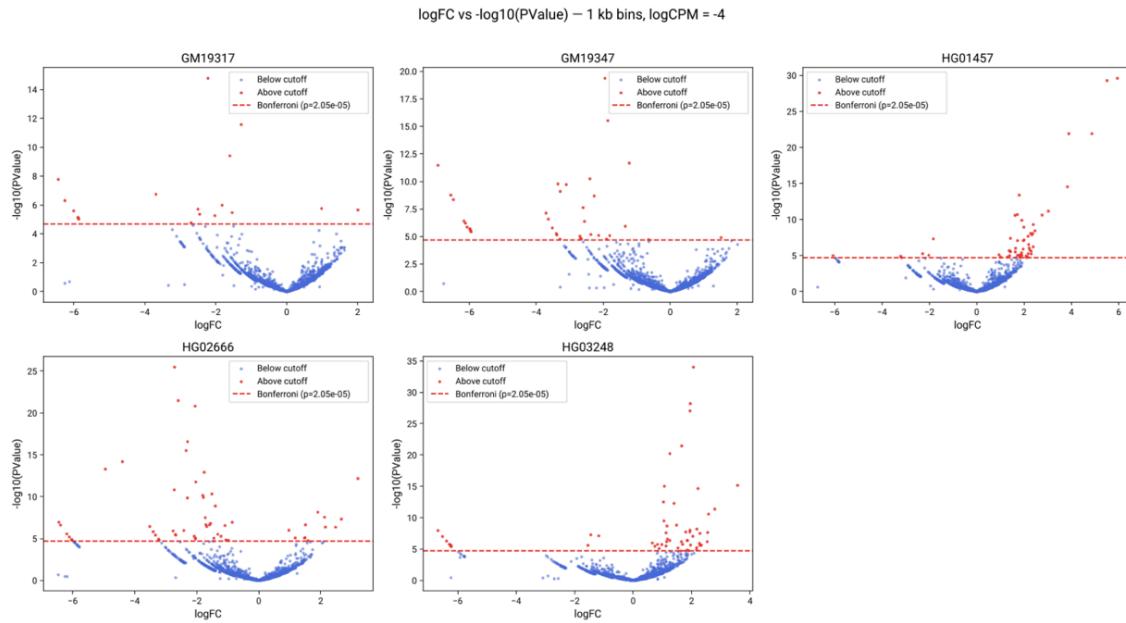
Supplementary Figure 4. Distribution of logCPM values across all loci for GM19317 at 1 kb, 2 kb, and 5 kb resolution, used to determine the filtering threshold ($\text{logCPM} > -4$) that retains interactions with sufficient statistical power.



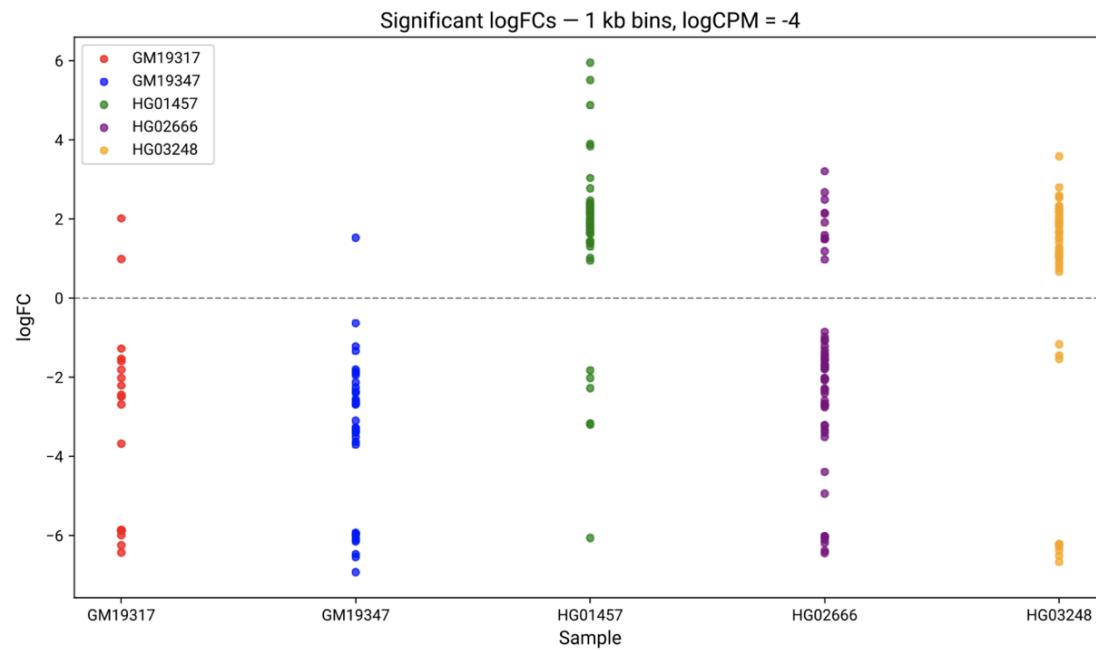
Supplementary Figure 5. Distribution of filtered logFC values ($\text{logCPM} > -4$) for each sample, showing the expected unimodal peak centered near zero at 1 kb resolution.



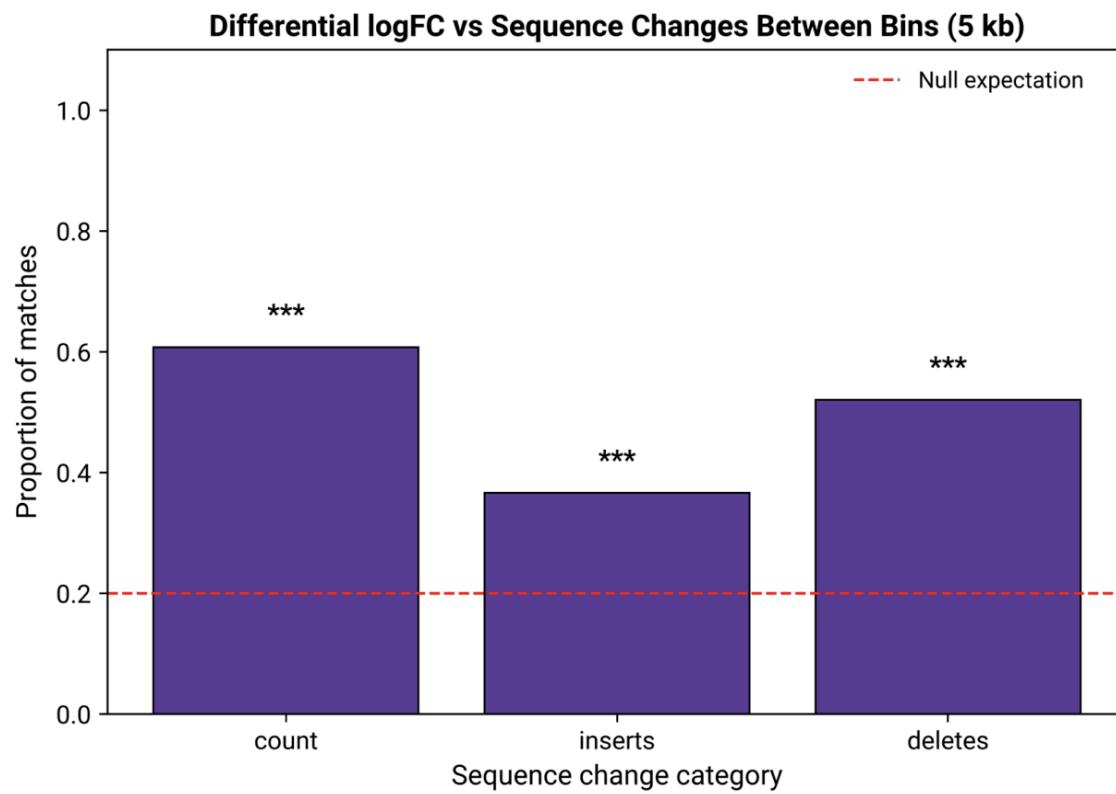
Supplementary Figure 6. Distribution of logFC standard deviations across filtered loci ($\log CPM > -4$) at 1 kb, following an approximately negative-binomial shape with most interactions conserved ($\sigma \approx 0.5$), and a minority showing strong differential signals.



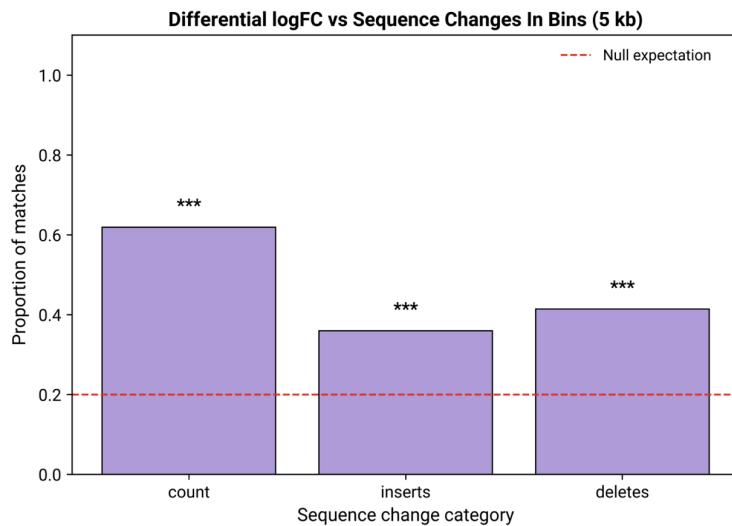
Supplementary Figure 7. LogFC versus log(p-value) for filtered interactions (logCPM >-4) at 1 kb resolution for all bins. Bonferroni multiple-test correction at 1 kb resolution was applied with Bonferroni $p = 2.05 \times 10^{-5}$, and pixels that passed the threshold are colored red.



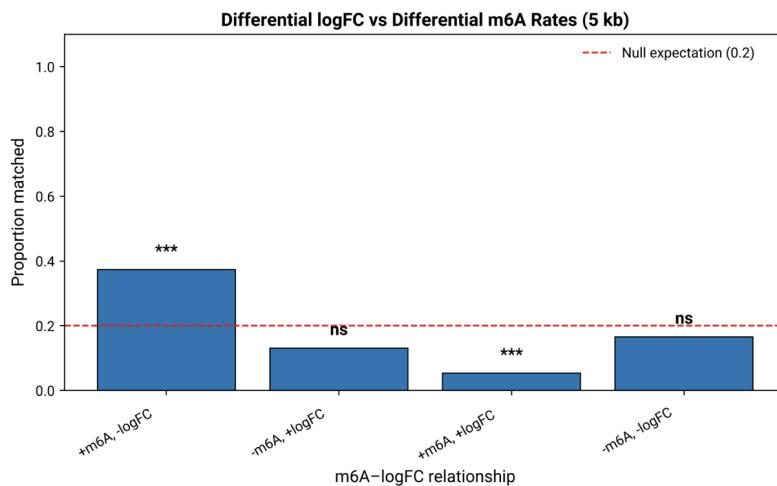
Supplementary Figure 8. LogFC values for each filtered differential interaction (logCPM >-4) for each sample at 1 kb resolution that pass the Bonferroni p-value.



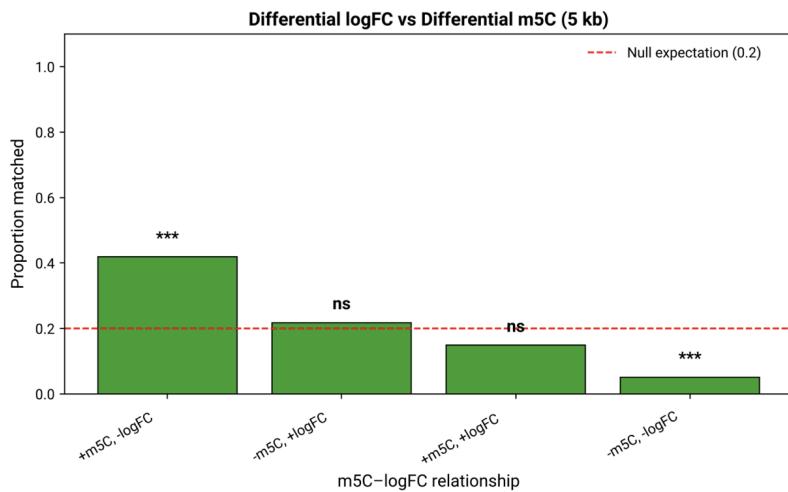
Supplementary Figure 9. Match rates of differential loci where the sample with a differential logFC value matches the sample with the greatest number of base sequence changes (total base changes, inserted bases, and deleted bases) between interaction bins compared to the null 0.2 under random matching at 5 kb resolution. The match rate for total base changes is 60.7% ($n = 107$, $p = 3.95 \times 10^{-20}$, binomial test). Significance levels are denoted as $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***)�.



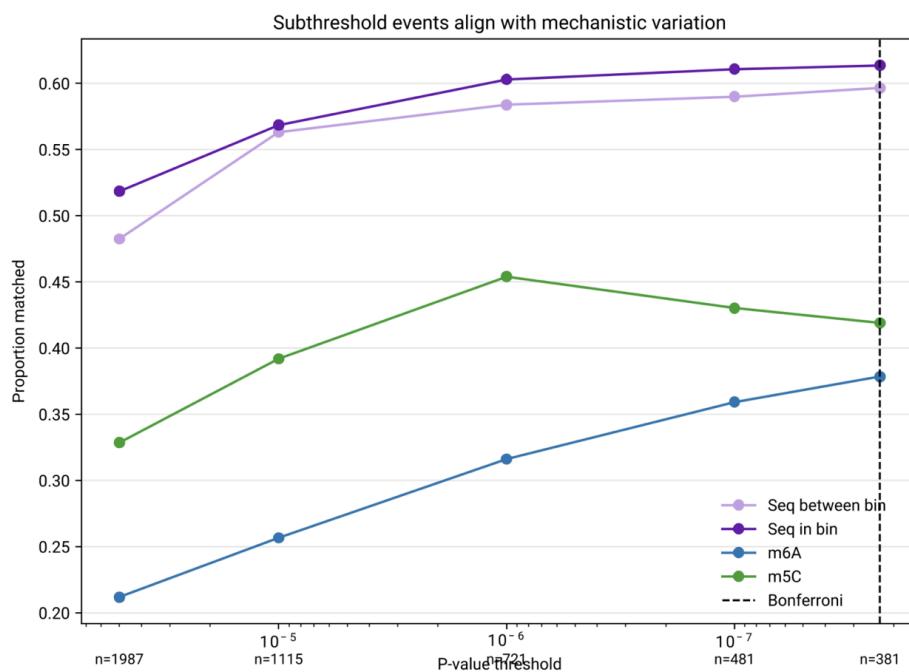
Supplementary Figure 10. Match rates of differential loci where the sample with a differential logFC value matches the sample with the greatest number of base sequence changes (total base changes, inserted bases, and deleted bases) within interaction bins compared to the null 0.2 under random matching at 5 kb resolution. The match rate for total base changes is 61.9% ($n = 118$, $p = 4.25 \times 10^{-23}$, binomial test). Significance levels are denoted as $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***)�.



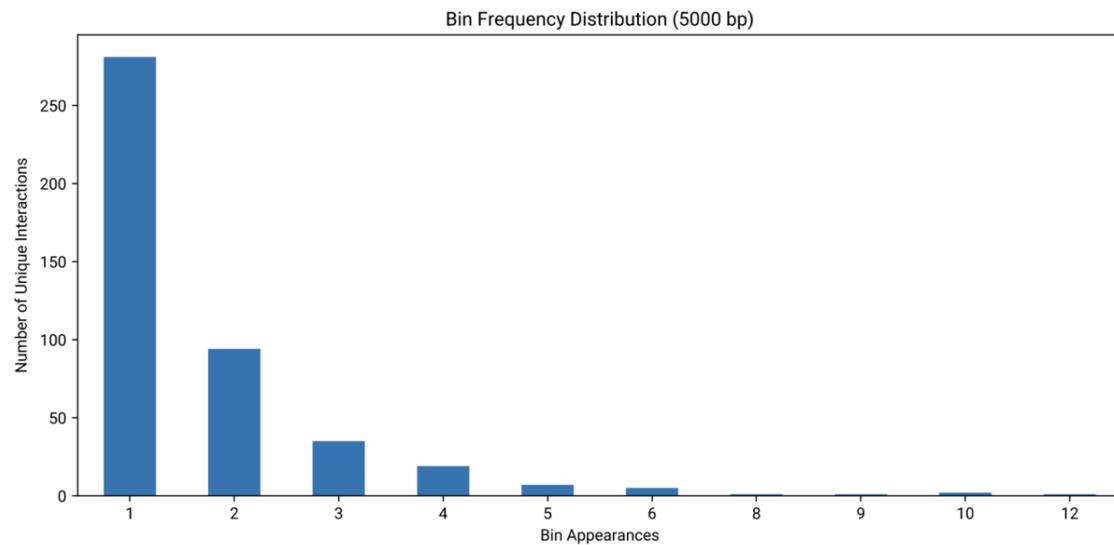
Supplementary Figure 11. Match rates of differential loci where the sample with the highest (+) or lowest (-) logFC value matches the sample with the highest (+) or lowest (-) m⁶A methylation level compared to the null 0.2 under random matching at 5 kb resolution. The match rate between the sample with the lowest logFC and highest m⁶A is 37.3% ($n = 75$, $p = 8.62 \times 10^{-4}$, binomial test). Significance levels are denoted as $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***)�.



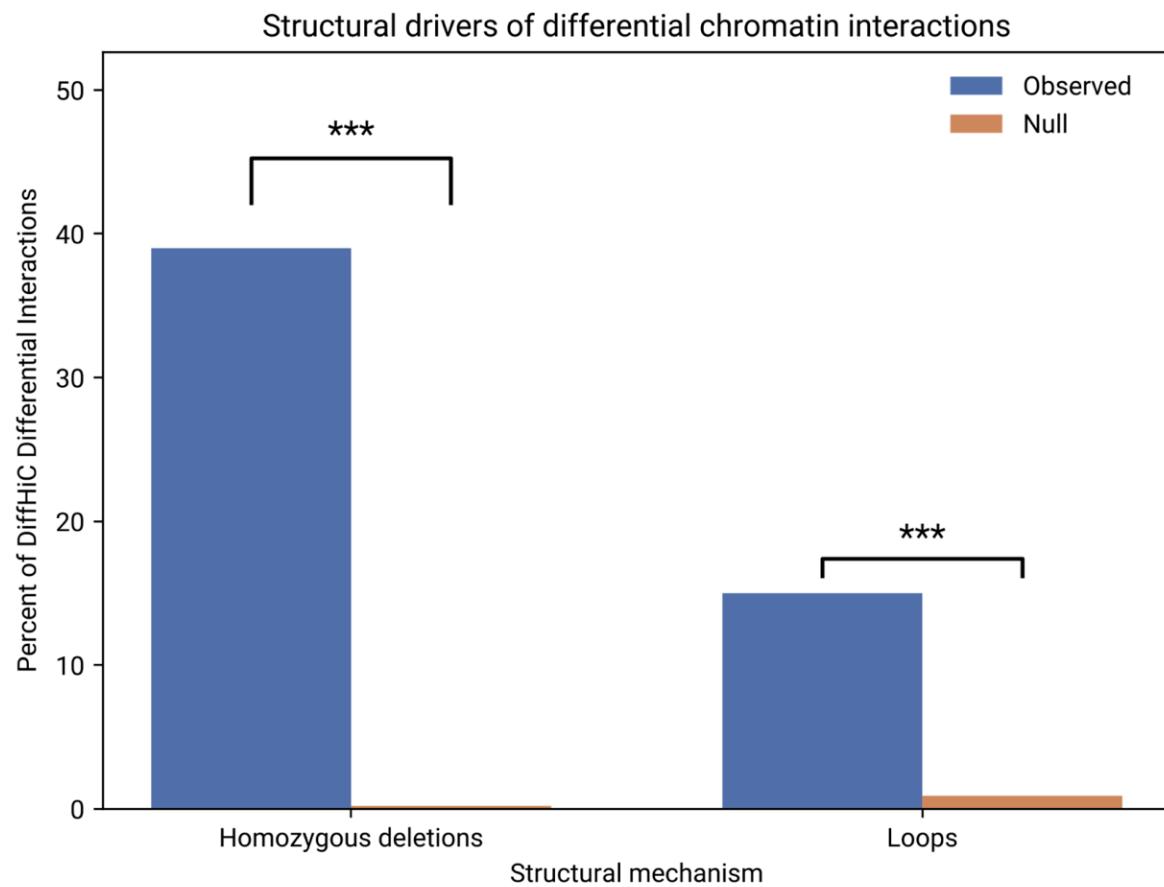
Supplementary Figure 12. Match rates of differential loci where the sample with the highest (+) or lowest (-) logFC value matches the sample with the highest (+) or lowest (-) m⁵C methylation level compared to the null 0.2 under random matching at 5 kb resolution. The match rate between the sample with the lowest logFC and highest m⁵C is 41.9% (n = 74, p = 1.58×10^{-5} , binomial test). Significance levels are denoted as p < 0.05 (*), p < 0.01 (**), and p < 0.001 (***)�.



Supplementary Figure 13. Proportion of matches across mechanisms (light purple, number of sequence changes between bins; dark purple, number of sequence changes in bins; blue, m⁶A methylation; green, m⁵C methylation) at less stringent p-value thresholds up to the Bonferroni p-value at 5 kb resolution. The number of total data points passing each p-value filter is included along the x-axis.

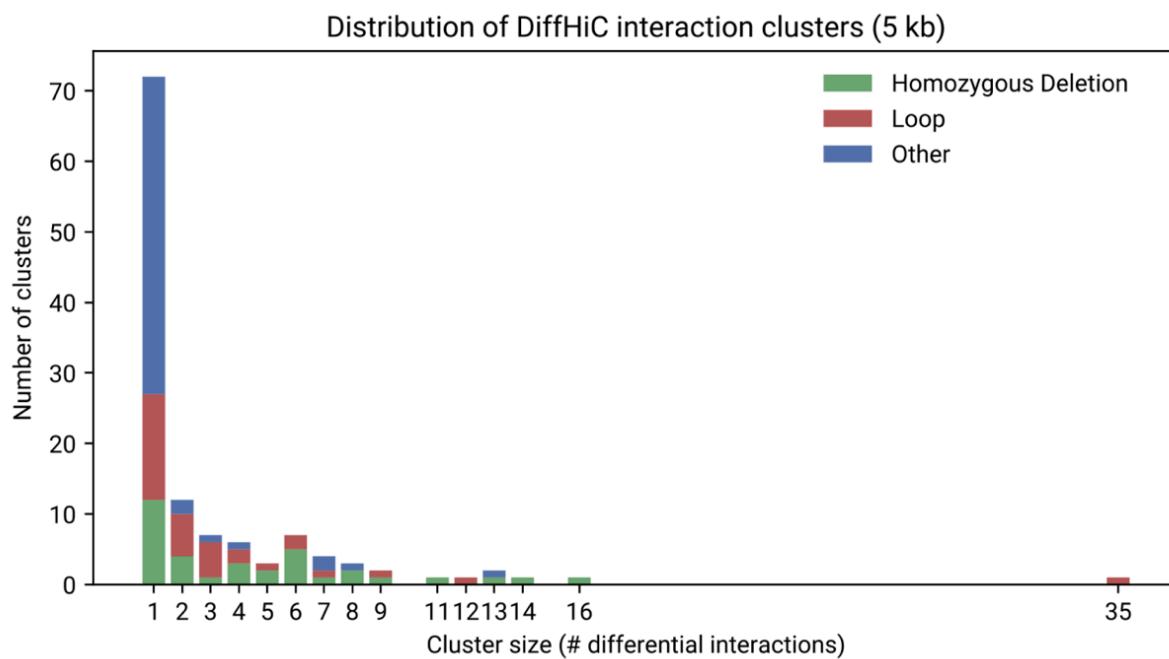


Supplementary Figure 14. Frequency distribution of diffHic identified bins among the 381 differential pixels identified at 5 kb resolution across the genome.

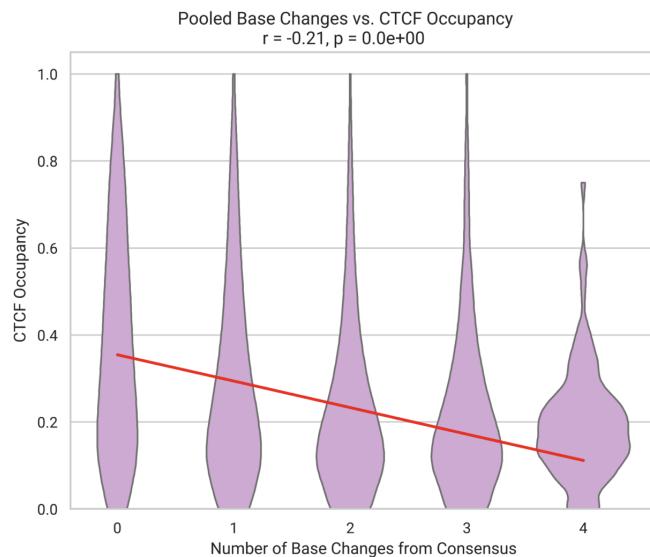


Supplementary Figure 15. Proportion (blue) of differential pixels located within homozygous deletions (39%) or within 10 kb of chromatin loops (15%) at 5 kb resolution.

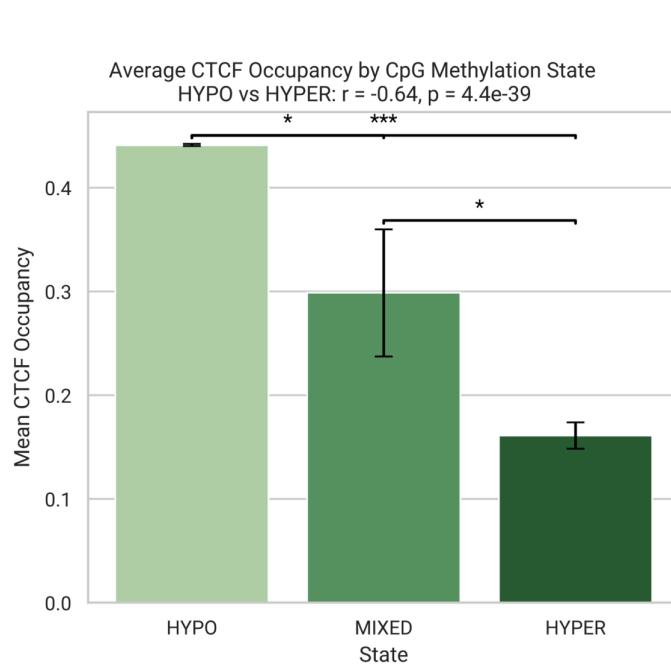
Null proportions (orange) for loops were calculated through circular permutation ($n = 1,000$) randomizing loop anchor positions at identical genomic distances with empirical permutation p-values. Null proportions for deletions ($n = 1,000$) were calculated by circularly permuting differential interaction anchors along each chromosome and testing against fixed homozygous deletion intervals. Significance levels are denoted as $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***)�.



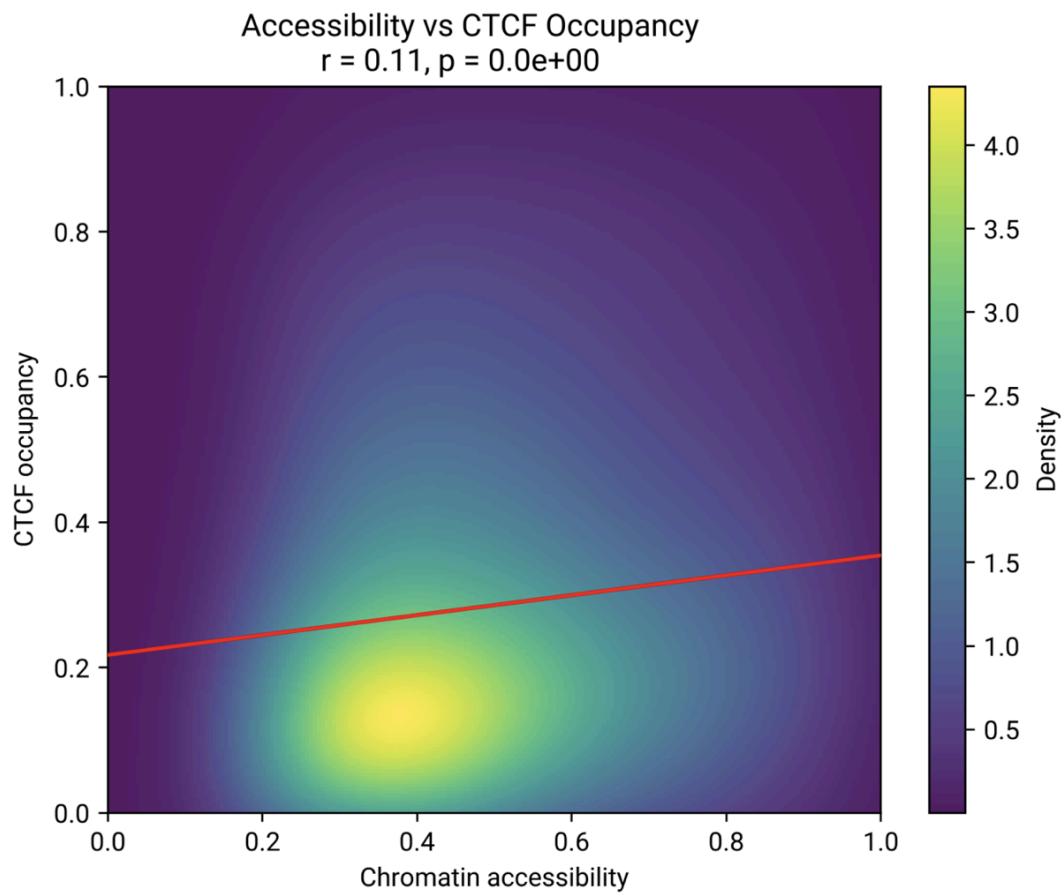
Supplementary Figure 16. Number of diffHic clusters associated with homozygous deletions (green), loops (red), and neither deletions nor loops (blue) at 5 kb resolution, stratified by number of differential pixels within clusters ($n = 381$ pixels, as 123 distinct clusters). Clusters were formed by grouping differential pixels located within a 10 kb \times 10 kb window of each other in the 2D Hi-C matrix.



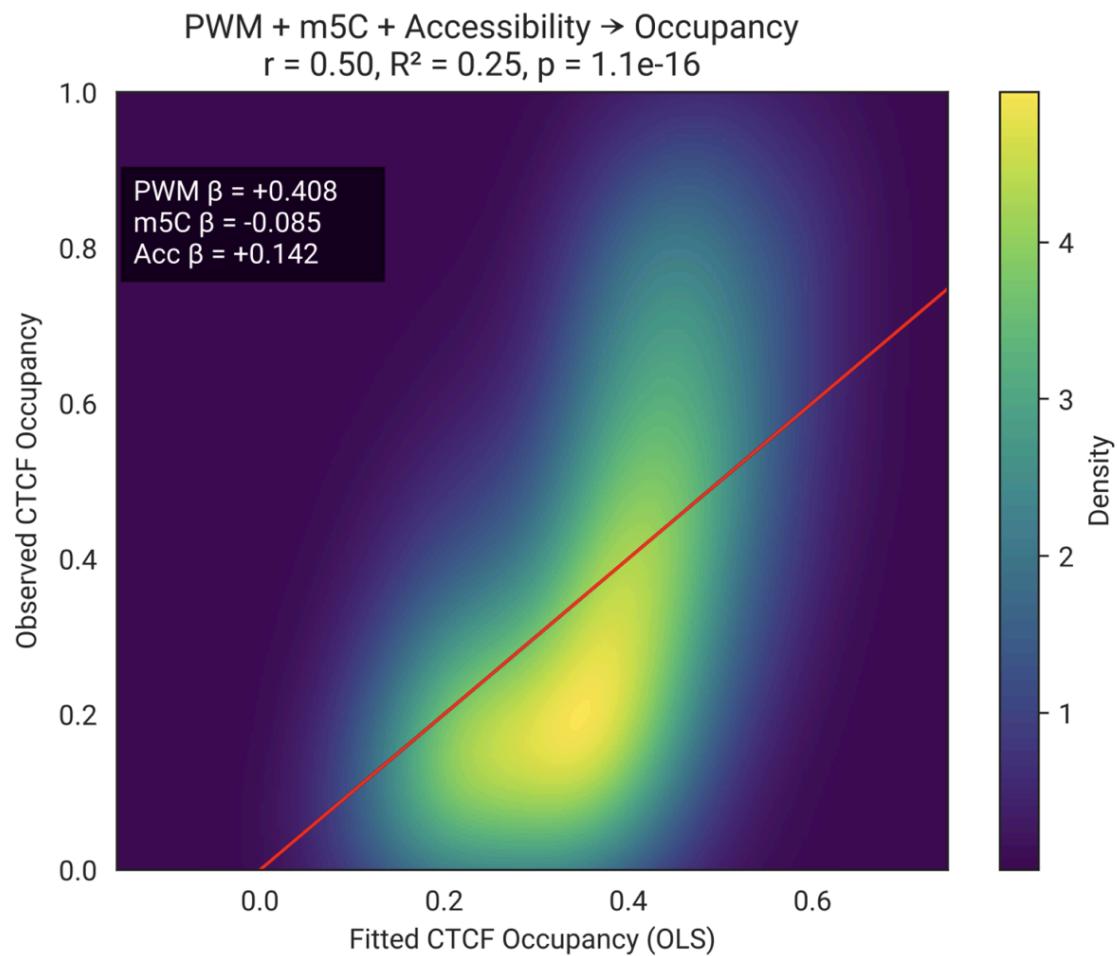
Supplementary Figure 17. Correlation between the number of base substitutions within each CTCF motif and CTCF occupancy. Pearson $r = -0.21$ ($n = 262,463$ sites; $p < 2.2 \times 10^{-308}$). Significance levels are denoted as $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***)�.



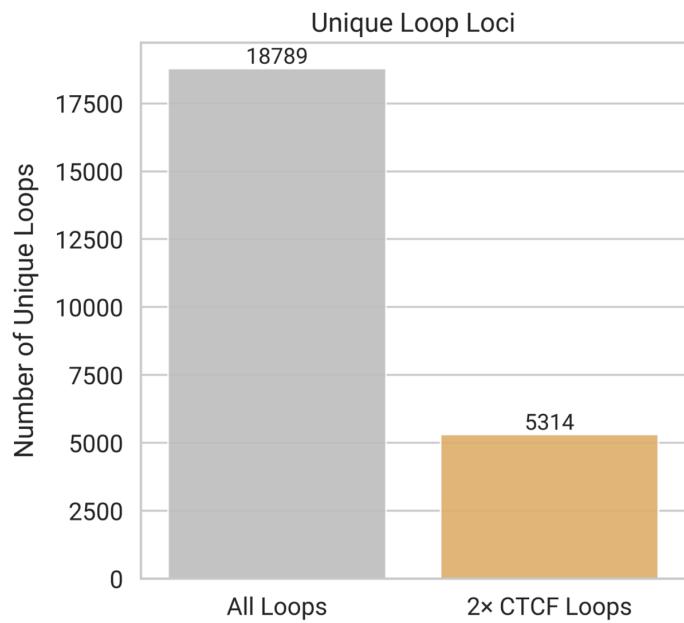
Supplementary Figure 18. Correlation between CTCF m⁵C CpG methylation state and CTCF occupancy. The pearson $r = -0.64$ between hypomethylated and hypermethylated states (hypomethylated $n = 58,094$ sites; hypermethylated $n = 140$ sites; mixed $n = 14$ sites; $p < 2.2 \times 10^{-308}$). Significance levels are denoted as $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***)�.



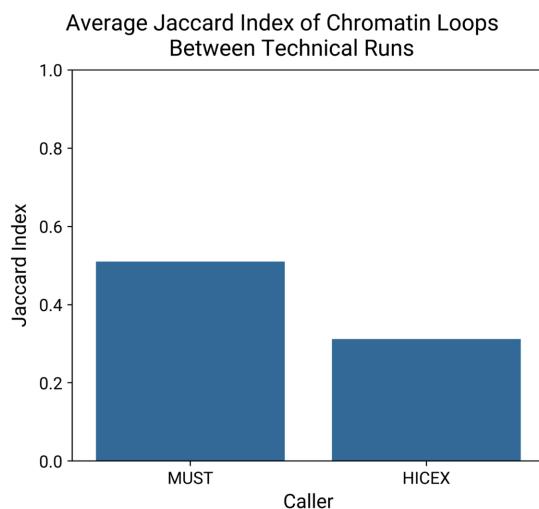
Supplementary Figure 19. Pearson correlation between local chromatin accessibility and CTCF occupancy ($n = 262,280$, Pearson $r = 0.11, p < 2.2 \times 10^{-308}$). Chromatin accessibility was determined from averaged m⁶A methylation rates from the 2 kb upstream and downstream of each 1 kb bin containing a CTCF site.



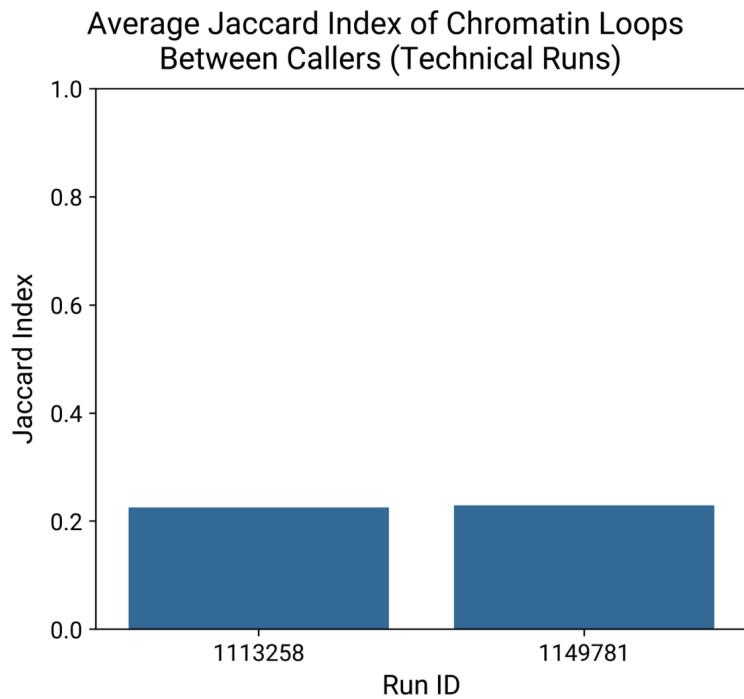
Supplementary Figure 20. Linear model estimating the combined effects of PWM score, m⁵C methylation, and local chromatin accessibility on CTCF occupancy using ordinary least squares regression (n = 90,672 sites). All predictors showed significant associations with occupancy (PWM $\beta = +0.408$; m⁵C $\beta = -0.085$, accessibility $\beta = +0.142$), and the model explained a moderate fraction of variance (Pearson r between observed and fitted values = 0.50; $R^2 = 0.25$; $p = 1.1 \times 10^{-16}$).



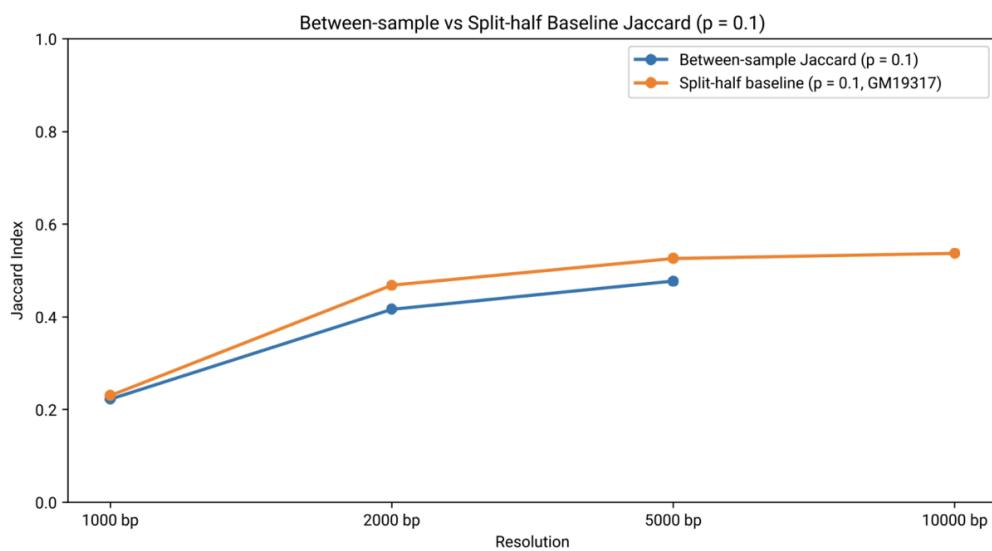
Supplementary Figure 21. Total number of unique loci with at least one loop call in the pooled callset, identified by Mustache and HiCExplorer across all five samples at 2 kb resolution using a p-value threshold of 0.1. The loop callset contained 18,789 loops (grey), and among these, 5,314 loops (yellow; 28.1%) were formed between two CTCF sites (where each loop-anchor had one unique CTCF site within 10 kb) in at least one haplotype. For reference, Mustache called 18,068 total loops in GM12878 at 5 kb resolution, with 5,318 loops (29.4%) forming between two CTCF anchors ([Roayaei Ardakany et al. 2020](#)).



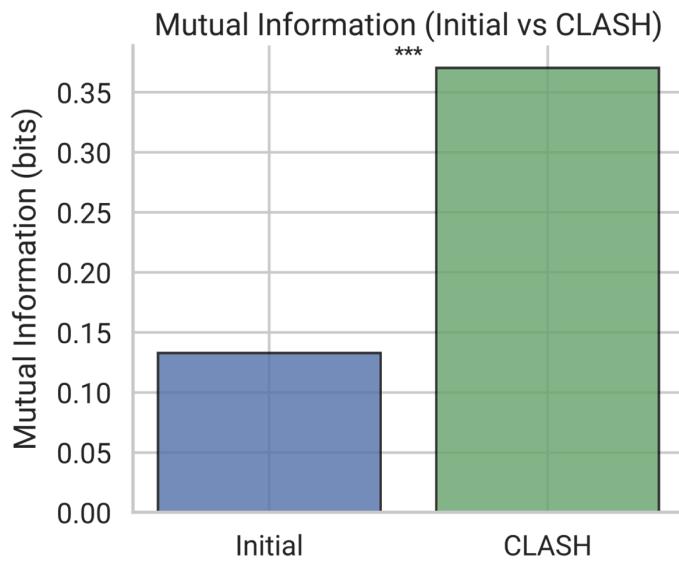
Supplementary Figure 22. Average Jaccard index of loop calls between technical runs for Mustache and HiCExplorer at 10 kb resolution. Technical run IDs 1113258 and 1149781 correspond to sample GM19317.



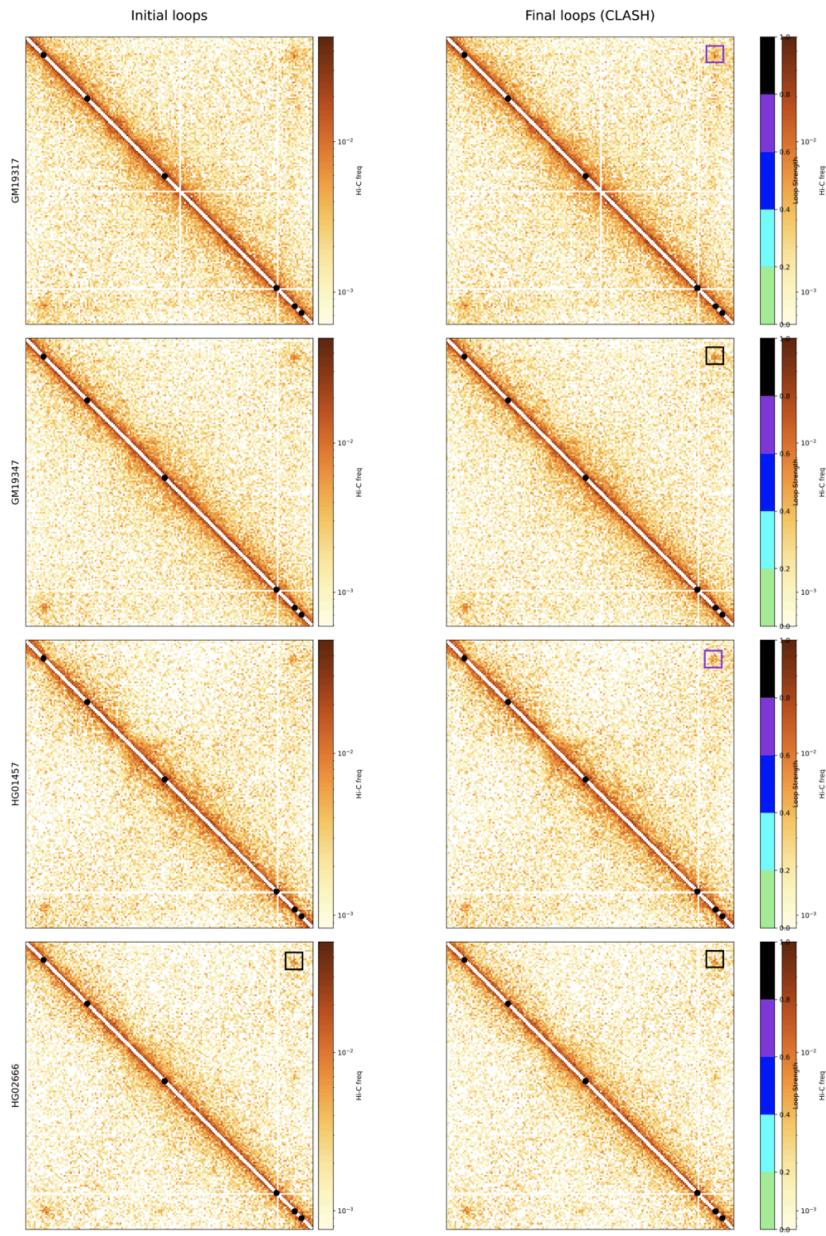
Supplementary Figure 23. Average Jaccard index of loop calls between Mustache and HiCExplorer across two technical runs at 10 kb resolution. Technical run IDs 1113258 and 1149781 correspond to sample GM19317.



Supplementary Figure 24. Average between-sample Jaccard index values at 1 kb, 2 kb, and 5 kb resolutions (blue) and maximum Jaccard index values expected due to sequencing depth for GM19317 at 1 kb, 2 kb, 5 kb, and 10 kb resolutions (orange). The expected Jaccard index given perfect loop-callers with perfect data would approach 1.0, but the observed baseline Jaccard index doesn't substantially increase from 2 kb to 10 kb resolutions, despite the dataset having more than sufficient sequencing depth power at 10 kb resolution. Thus, the between-sample Jaccard indices approach the limit set by Mustache's calling ability rather than by sequencing depth.

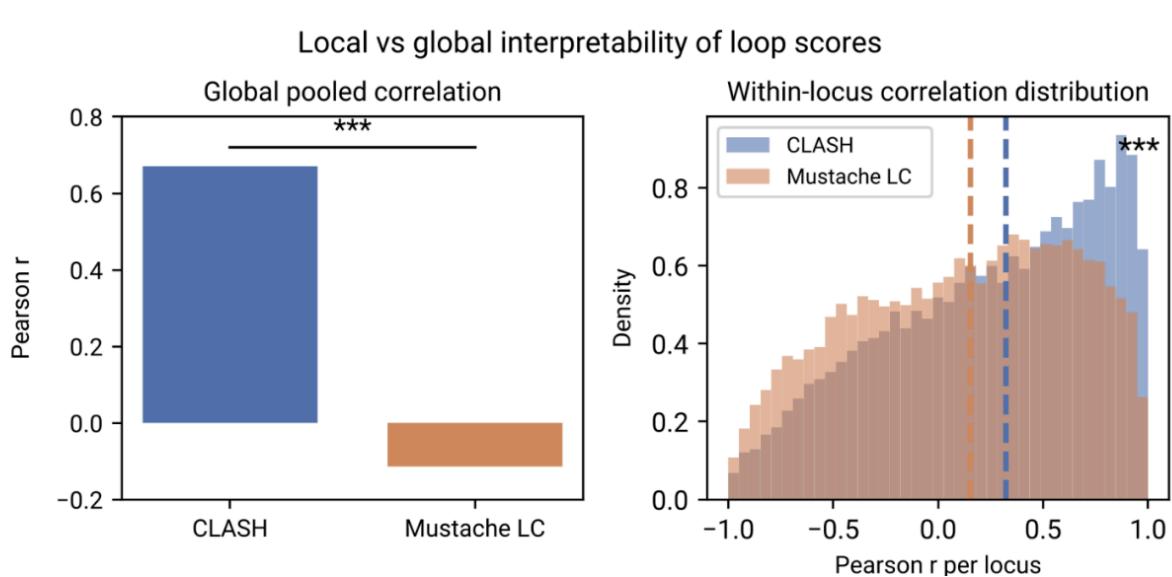


Supplementary Figure 25. Mutual information (bits) encoded by loop calls before (blue) and after CLASH harmonization (green).

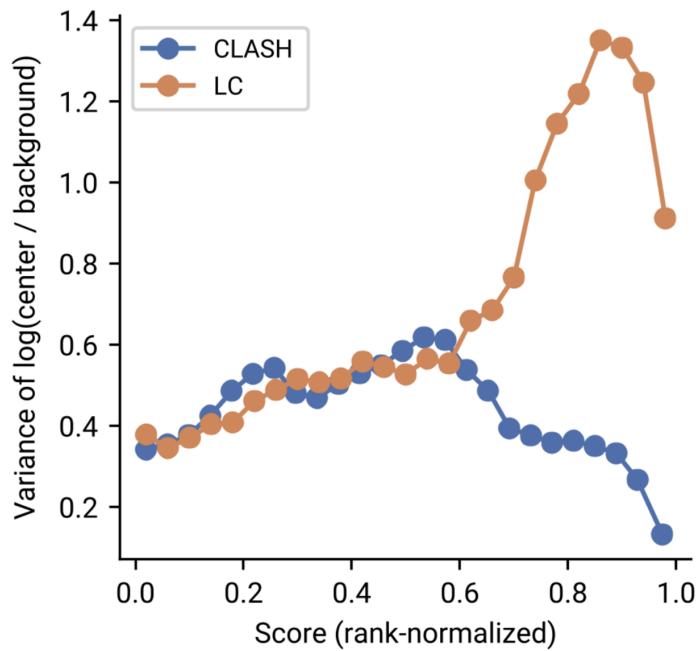


Supplementary Figure 26. Example chromatin loop before and after CLASH

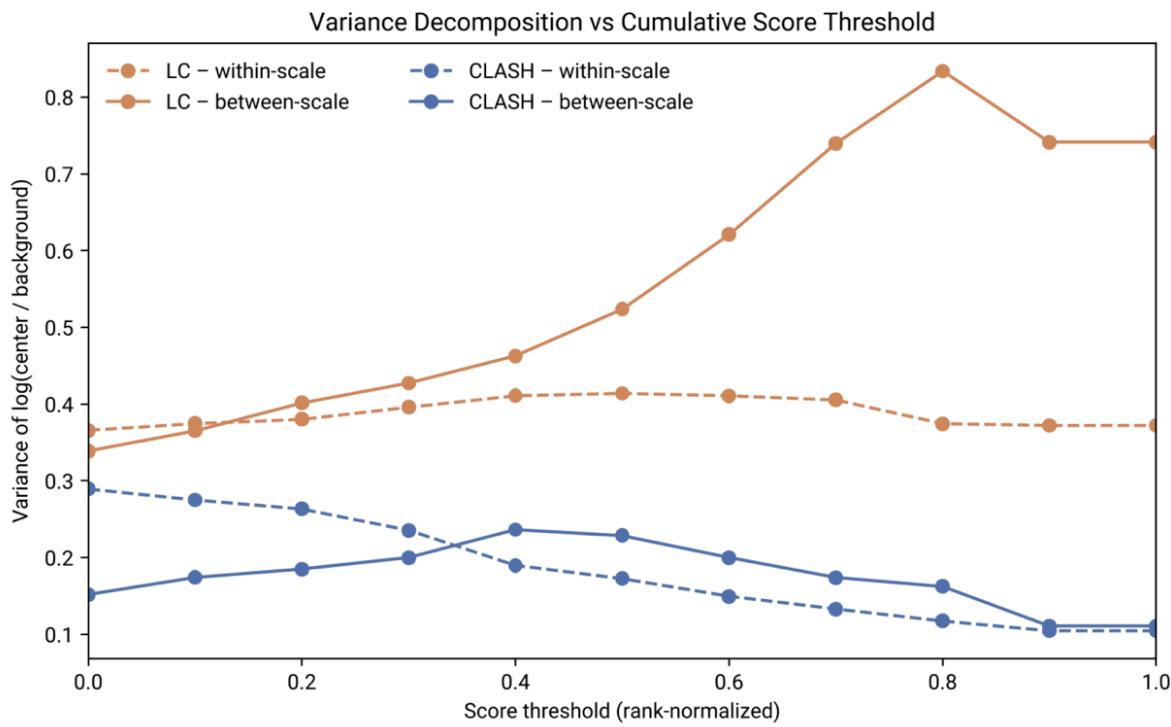
harmonization. The representative locus (chr14:76,052,000–76,352,000) illustrates loop-calling inconsistency across samples prior to CLASH harmonization (left), where Mustache and HiCExplorer fail to identify the same loop across individuals. They call the loop only in sample HG02666 (black rectangle) despite clear contact enrichment across all samples. After CLASH harmonization (right), the same locus is consistently scored as strong (purple rectangle) or very strong (black rectangle) across all five samples. CTCF sites are represented as black dots along the diagonal.



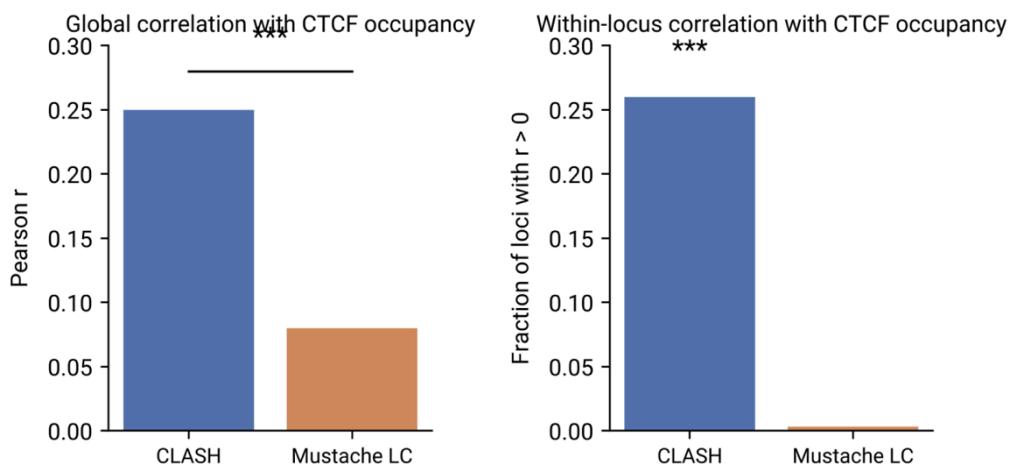
Supplementary Figure 27. Global pooled (left) and within-locus (right) Pearson r correlations of Mustache LC scores (orange) and CLASH scores (blue) with log(center contacts/background contacts). There are strong correlations between CLASH score and Hi-C signal both globally and per-locus, while LC scores have weak inverse correlations with Hi-C signal globally (CLASH Pearson $r = 0.67$, LC Pearson $r = -0.11$, $n = 91397$, Steiger $p < 2.2 \times 10^{-308}$) and a lower correlation distribution than CLASH within-locus (CLASH median pearson $r = 0.32$, LC median pearson $r = 0.16$, $n = 18281$, Wilcoxon signed-rank $p = 5.7 \times 10^{-156}$). Significance levels are denoted as $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***)�.



Supplementary Figure 28. Variance of log(center/background) values across rank-normalized CLASH (blue) and Mustache LC scores (orange). CLASH and LC scores perform similarly from 0.0 to approximately 0.6, at which point CLASH shows significantly less center/background variance for equivalent ranked scores and outperforms LC scores, which show significantly more. This indicates that CLASH is assigning high scores to structurally similar Hi-C matrices globally, while LC scores assign high scores to matrices with heterogeneous structures.

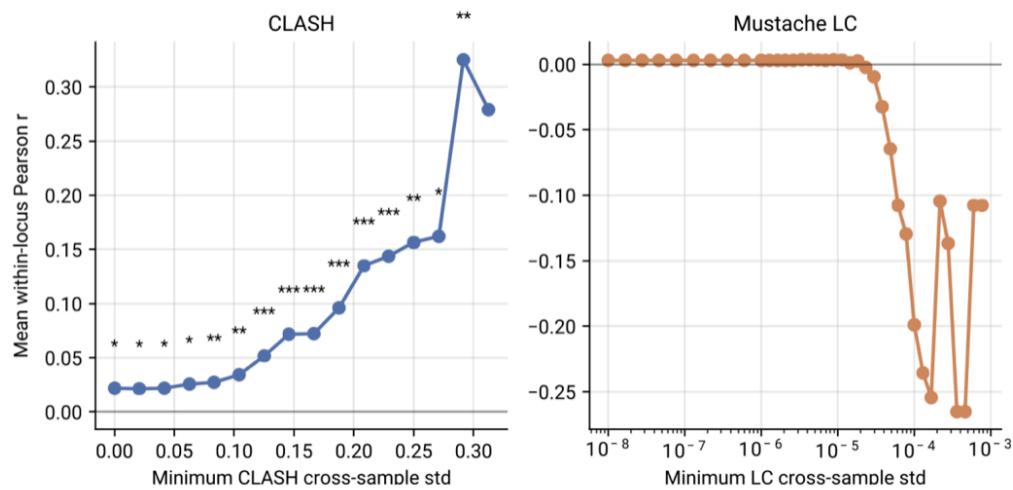


Supplementary Figure 29. Decomposition of within-scale (dashed) and between-scale (solid) variance for CLASH (blue) and Mustache LC (orange) across cumulative rank-normalized score thresholds. For LC scores, scale corresponds to different σ values, whereas for CLASH scores, scale corresponds to different matrix sizes. LC scores exhibit significantly more between-scale variance than within-scale variance at higher scores precluding meaningful global comparisons. Conversely, CLASH scores exhibit similar within-matrix size and between matrix-size variance, indicating that scores are consistent across matrix sizes and thus can be used for global comparisons.



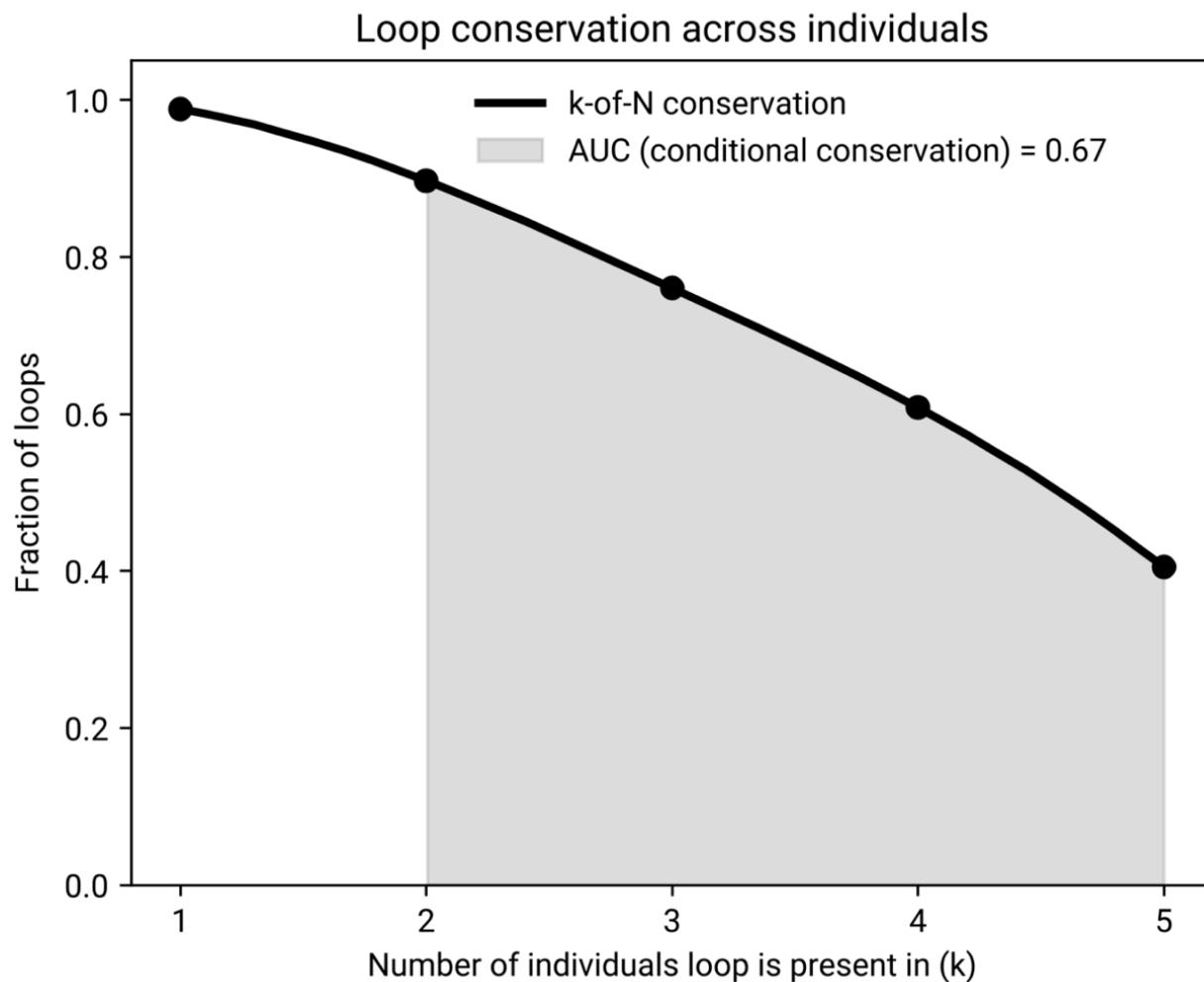
Supplementary Figure 30. Global Pearson correlation (left) and within-locus correlations (right) between CLASH (blue) and Mustache LC (orange) scores with CTCF occupancy.

Globally, CLASH scores show significantly more independent correlation with CTCF occupancy than LC scores (CLASH Pearson $r = 0.25$; LC Pearson $r = 0.08$; $n = 17,483$; Steiger $p < 2.2 \times 10^{-308}$). For CLASH per-locus comparisons, only loci that exhibited a CLASH score standard deviation of 0.29 across samples were considered, and for Mustache per-locus comparisons, only loci that exhibited a LC score standard deviation of 1×10^{-5} were considered. These values were chosen as the maximal r values that maintained statistical significance (for CLASH – no LC values reached statistical significance) based on the results of Supplementary Figure 31. For the within-locus analysis, CLASH-selected differential loci exhibited a significant enrichment for positive correlations with CTCF occupancy (CLASH Pearson $r = 0.26$; $n = 21$ loci; binomial sign test $p = 3.6 \times 10^{-3}$), whereas LC-selected differential loci did not (LC Pearson $r = 0.03$; $n = 2,590$ loci; binomial sign test $p = 0.38$). Significance levels are denoted as $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***)�.

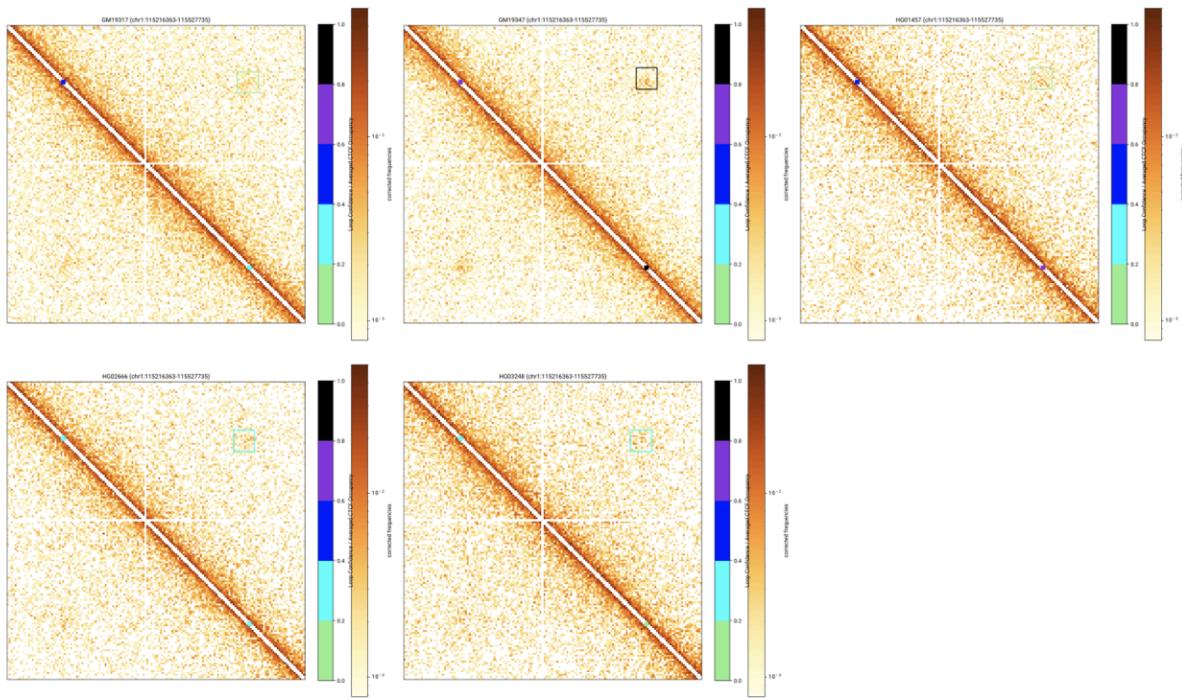


Supplementary Figure 31. Average within-locus Pearson correlations between CLASH (blue) and Mustache LC (orange) scores with CTCF occupancy across a range of

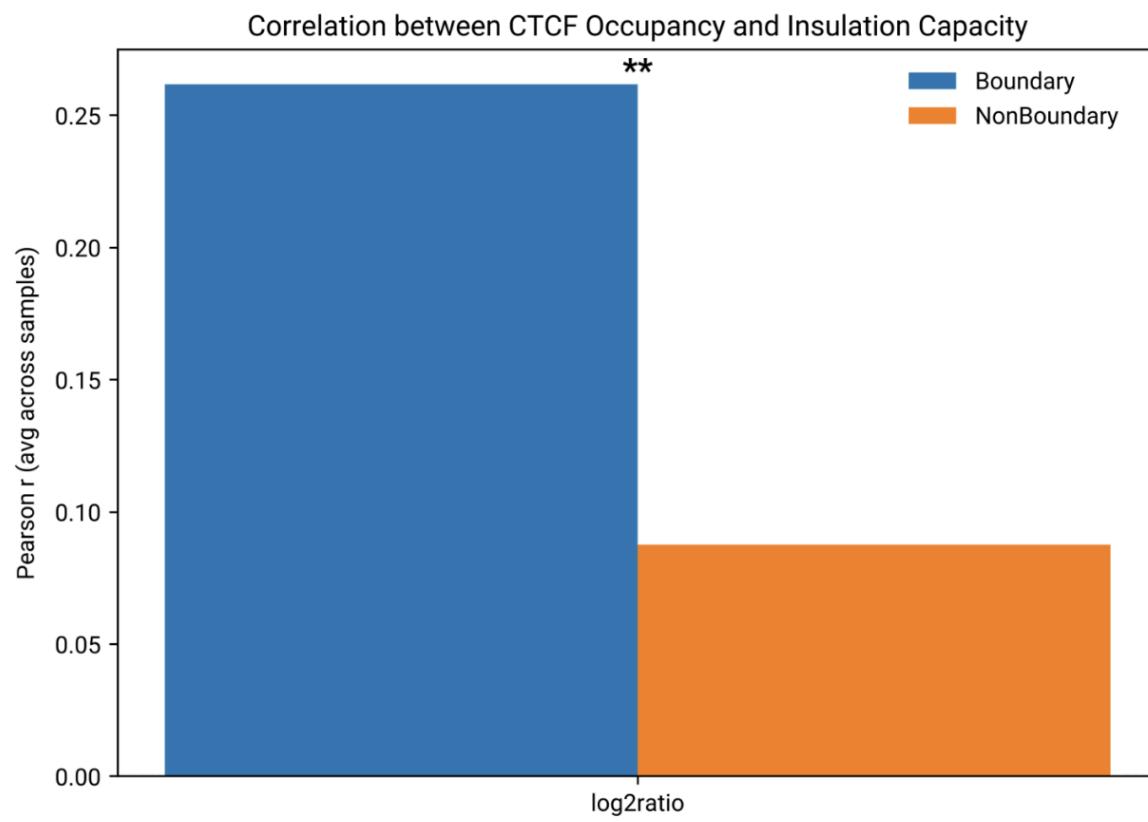
cross-sample standard deviation score thresholds. Significance levels are denoted as $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***)



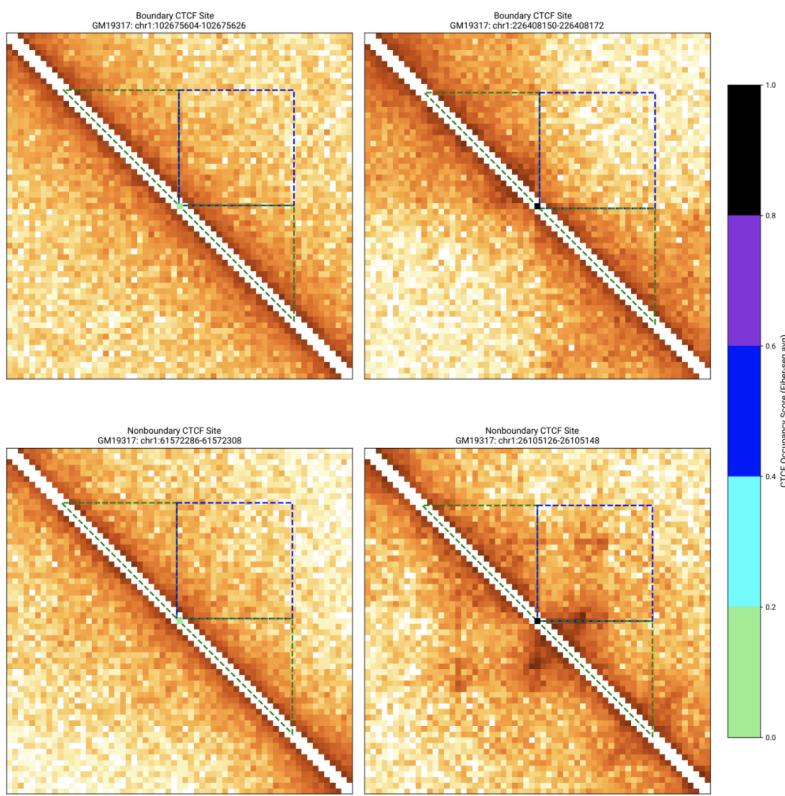
Supplementary Figure 32. Smoothed k-of-N conservation curve of binarized CLASH-scored loops. Loops were classified as present if they exhibited a CLASH score of > 0.3 . The area under the curve (AUC) of the portion of the graph with $k \geq 2$ corresponds to the conditional conservation probability.



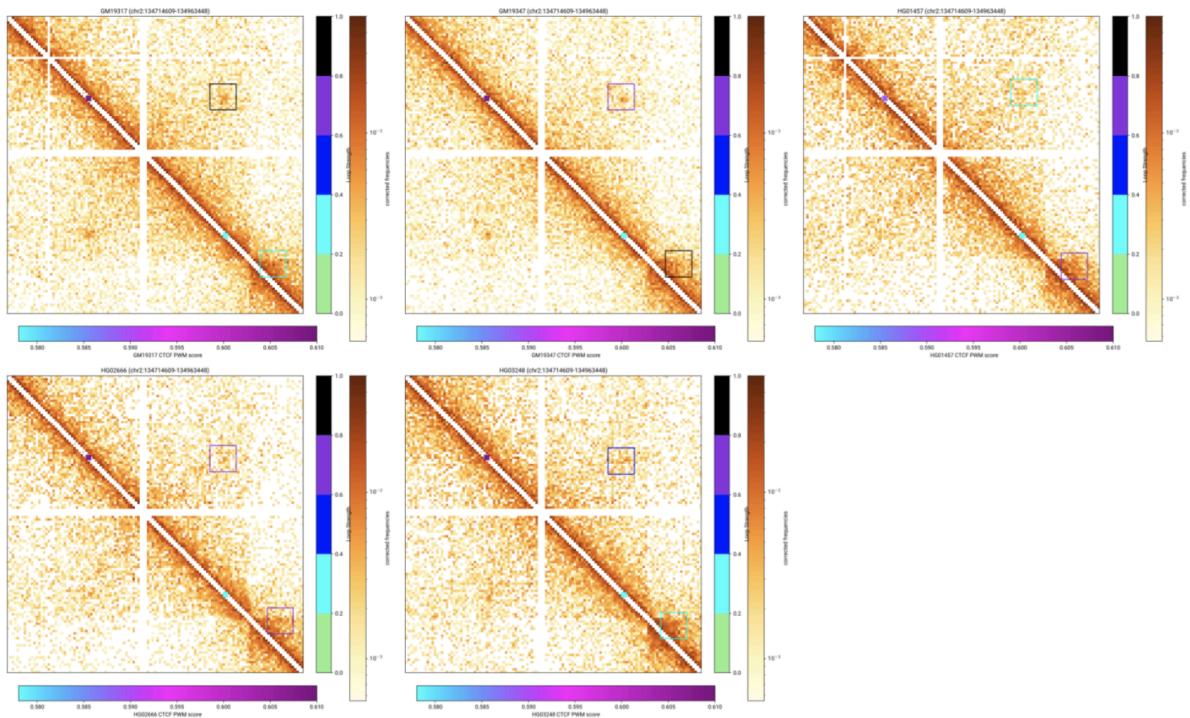
Supplementary Figure 33. Example locus (chr1:115216363-115527735) across all five samples, illustrating a positive relationship between CTCF occupancy and loop scores. Both CTCF occupancy and CLASH loop scores are annotated for each sample with light green = 0-0.2, light blue = 0.2-0.4, blue = 0.4-0.6, purple = 0.6-0.8, and black = 0.8-1.0. The loop score is annotated by the rectangle surrounding the loop, and CTCF occupancy of each CTCF site is annotated as rectangles along the diagonal. GM19347 has higher CTCF occupancy in both loop-associated CTCF sites than the other samples, corresponding to an increase in loop score.



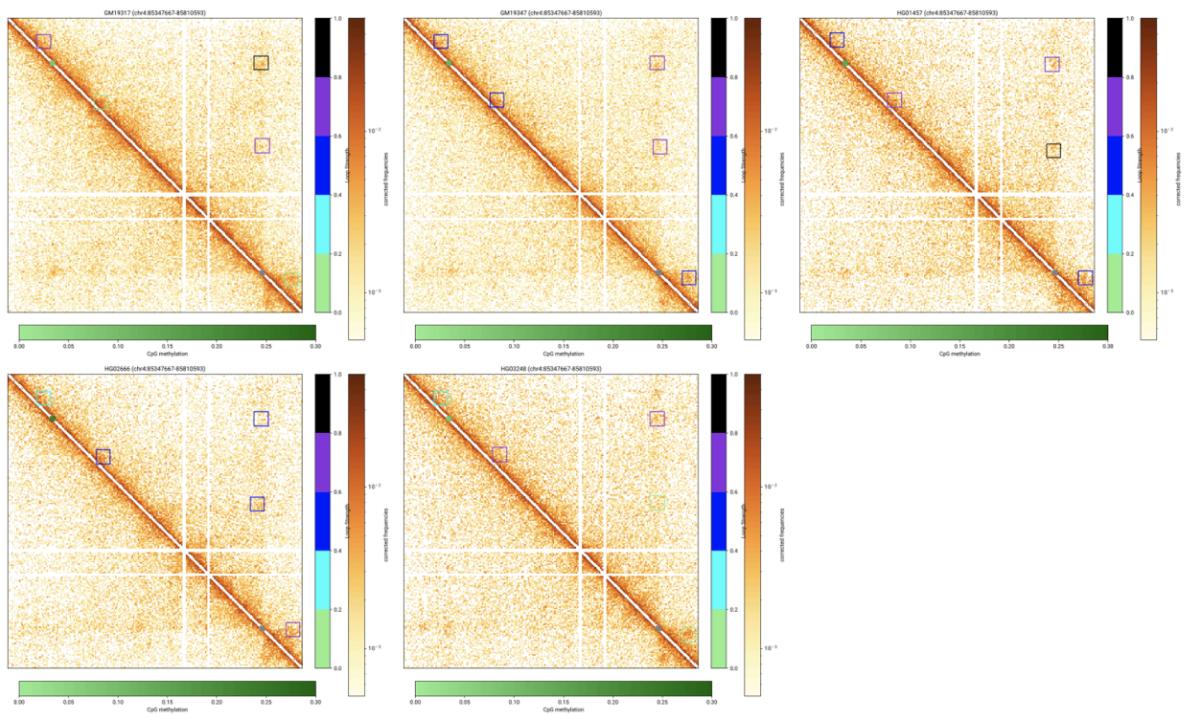
Supplementary Figure 34. Average within-sample correlations between CTCF occupancy and log₂ratio insulation scores (representing TAD strength) at boundary versus non-boundary sites on chromosome 1 (boundary n = 740, boundary Pearson r = 0.2617, non-boundary n = 3,063, non-boundary Pearson r = 0.0876; Δr = 0.1741; p = 0.00326, Welch's t-test).



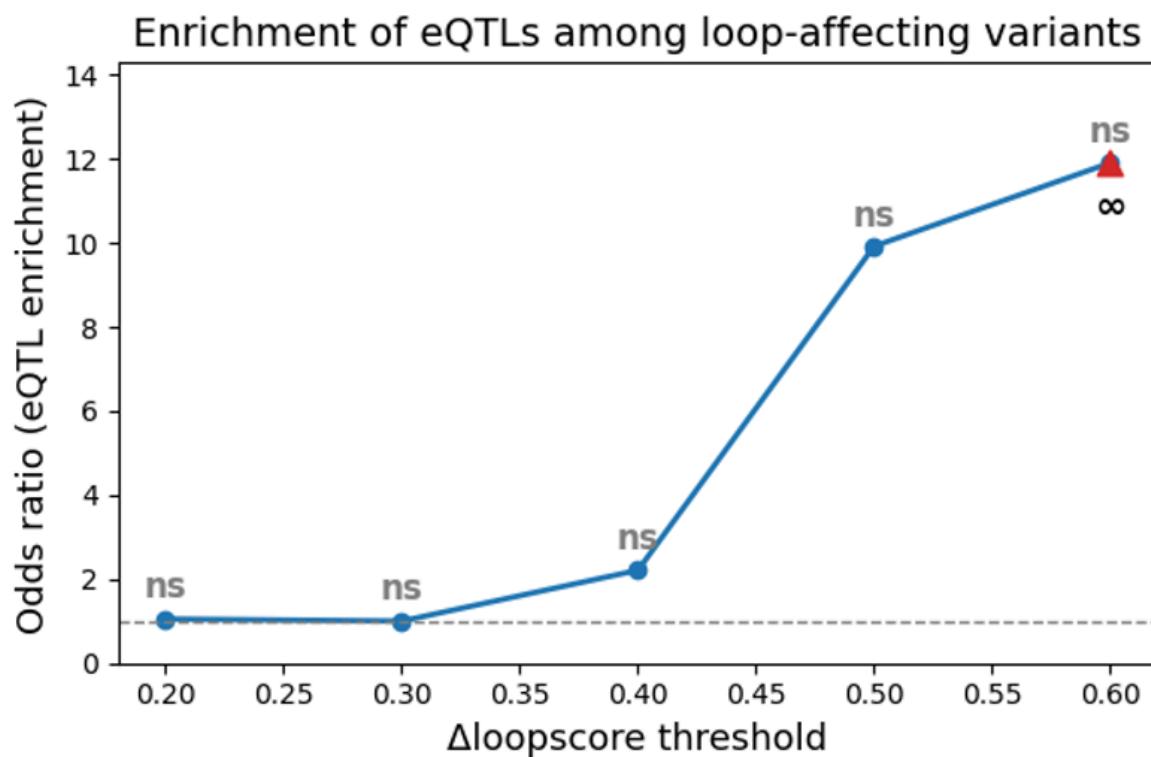
Supplementary Figure 35. Demonstrative example showing TAD analysis results for four loci in GM19317. CTCF sites classified as boundary sites are presented as the top two panels, whereas CTCF sites classified as non-boundary are shown in the bottom two panels. CTCF sites with low occupancy are presented on the left, while CTCF sites with high occupancy are shown on the right. Green triangles are overlaid on each Hi-C map representing cis CTCF site contacts considered, while the blue square represents trans CTCF site contacts considered.



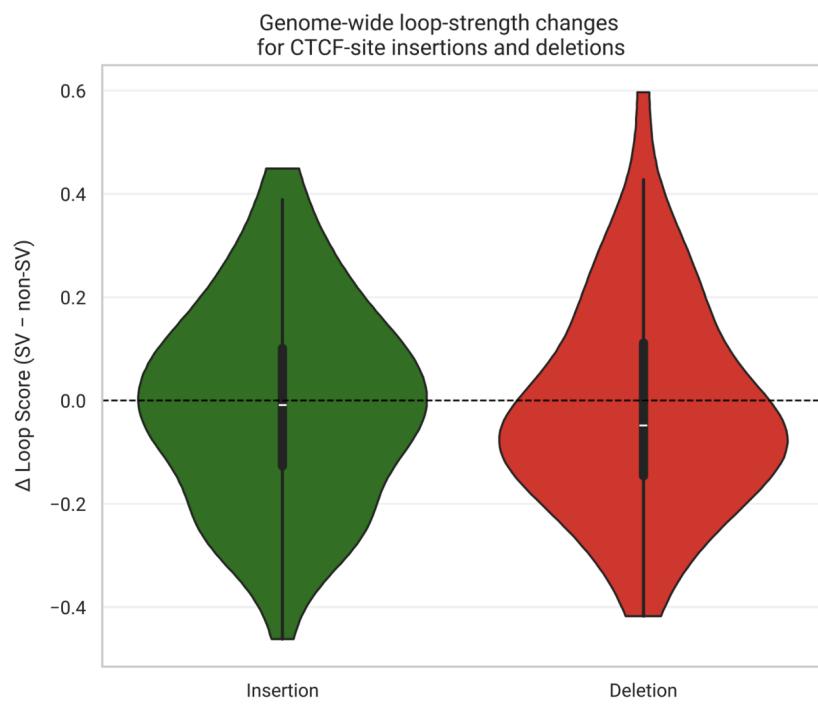
Supplementary Figure 36. Representative locus (chr2:134,714,609–134,963,448) illustrating the expected relationship between PWM scores and loop strength across samples. In HG01457, the upstream loop-associated CTCF site contains a heterozygous SNP at the 14th base position (G → A), resulting in a reduced PWM score (lighter purple). This nucleotide change corresponds to both the lowest CTCF occupancy at that site and the weakest loop formation across samples. For clarity, only relevant CTCF sites are shown.



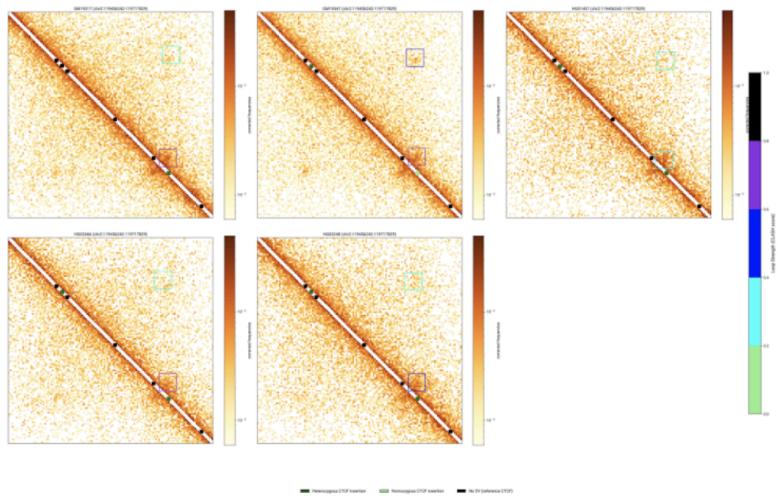
Supplementary Figure 37. Representative locus (chr4:85373292–85784871) illustrating the expected relationship between m^5C methylation and loop strength across samples. In HG02666, the upstream loop-associated CTCF site is significantly more methylated (darker green) than the other samples. This corresponds to both the lowest CTCF occupancy at that site and the weakest loop formation across samples. For clarity, only relevant CTCF sites are shown.



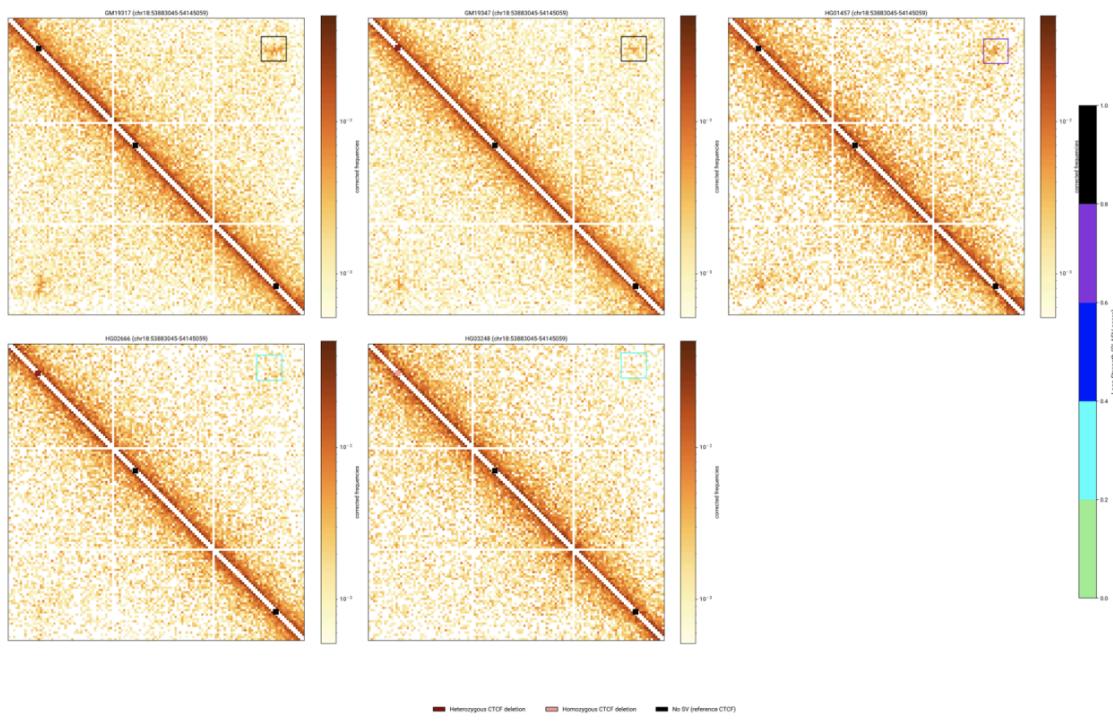
Supplementary Figure 38. Enrichment of eQTLs among loop-altering SNPs (putative iQTLs; n = 668 total SNP-containing CTCF loci) at Δ loop-strength thresholds of 0.2 to 0.6.
Although this analysis lacks statistical power ($p > 0.05$ for all thresholds) due to small sample size (n = 11 iQTLs at CLASH ≥ 0.4 ; n = 2 overlapping eQTLs), the observed trend is consistent with expectations.



Supplementary Figure 39. Distribution of loop-strength differences for loci containing structural-variant insertions ($n = 70$, left) or deletions ($n = 65$, right), compared with samples lacking the variant at the same locus, indicating a minimal global effect of structural variation on loop strength.



Supplementary Figure 40. Example locus (chr2:119456242-119717829) showing a structural-variant insertion that introduces a homozygous CTCF site (light green) in GM19347, in contrast to the other samples with heterozygous insertions (dark green), leading to increased loop strength in GM19347. Stable CTCF sites are shown in black.



Supplementary Figure 41. Example locus (chr18:53883045-54145059) showing a structural-variant deletion that removes a CTCF site, present as heterozygous deletions (dark red) in GM19347 and HG02666 and as a homozygous deletion (light red) in HG03248. Stable CTCF sites are shown in black. Samples without any deletions exhibit strong

loop formation, samples with heterozygous deletions show variable loop strength (as expected given non-haplotype-resolved Hi-C maps), and homozygous deletions exhibit minimal loop formation.