NAR Genomics and Bioinformatics, 2025, **7**, lqaf146
https://doi.org/10.1093/nargab/lqaf146
Standard Article

OXFORD

Analysis of long-range contacts across cell types outlines a core sequence determinant of 3D genome organization

Liezel Tamon [1], Zahra Fahmi[2], James Ashford[1], Rosana Collepardo-Guevara [2,3,4], Aleksandr B. Sahakyan [1,*]

[1]MRC WIMM Centre for Computational Biology, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford OX3 9DS, United Kingdom
[2]Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom
[3]Department of Physics, University of Cambridge, JJ Thomson Ave, Cambridge CB3 0HE, United Kingdom
[4]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, United Kingdom

*To whom correspondence should be addressed. Email: aleksandr.sahakyan@imm.ox.ac.uk

Abstract

The sequence-driven organizing principles of the 3D genome are crucial for interpreting the core effects of genomic variation and understanding the evolution of genome organization and function. We investigated these by isolating and analysing cell-type-persistent contacts, heavily dependent on the similarly cell-type-persistent genomic sequence. We stratified long-range contacts from a diverse set of human tissues and cell lines based on contact persistence, c_p, reflecting their presence across cell or tissue types, and present them as an atlas of contacts and cell-type invariant (CETI) hubs they form across human chromosomes. Our survey of >300 chromatin and genome features revealed their association with c_p, contrasting variable from persistent contacts in terms of co-localization with genes, 3D architectural domains, and epigenetic and sequence elements. We found persistent contacts to be predominantly comprised of adenine and thymine (AT)-rich sequences and related to heterochromatin. A key outcome is finding a link between the experimental genomic contacts and the complementarity between pairs of contacting DNA loci. This work provides evidence for a sequence determinant of genomic contacts contributing to the decoding of the relationship between sequence and structure that is crucial for functional and evolutionary studies concerning 3D genome organization.

Introduction

Within a human cell, a DNA double helix ∼2 m in length is intricately organized inside the nucleus, itself about 10 μm in diameter, while still enabling the proper functioning of molecular processes like replication and gene expression. This organization can be broadly categorised into (i) the nucleosome- and chromatin-fibre scale [1]; (ii) the intermediate (domain) scale [1], which includes loops, topologically associating domains (TADs) [2–4], lamina-associated domains (LADs), and compartments [5]; and (iii) the nuclear scale, encompassing chromosome territories, genome arrangement with respect to nuclear centre or periphery [6], and nuclear bodies [7, 8, 9].

Our collective understanding of how genome organization relates to function has progressed with the identification of factors influencing that organization. Biochemical mapping methods, imaging experiments, polymer simulations, and other computational and experimental investigations have shown that multiple factors jointly shape the 3D genome organization. These factors include the polymeric nature of the DNA [10], the complex and nonuniform information embedded in the chromatin in the form of epigenetic modifications and DNA-binding proteins [1, 11], molecular processes such as replication [12] and transcription [13], and the underlying genomic DNA sequence [14–17]. The relationship between genomic sequence and any compounding phenomenon, like 3D genome organization, remains a subject of much interest as it can crucially contribute to the interpretation of naturally occurring and *de novo* genomic variations, and to the understanding of the sequence-structural evolution of our genome.

Research on genomic 3D contacts has surged since the development of the first experimental methods to explore them [5, 18–20], establishing genome organization as dynamic and stochastic at all scales (recently reviewed [21]). The increasing amount of data from 3C and other types of methods in the context of various cell types, states, and species have outlined the multiscale nature of genome organization [2, 4, 1, 22], and the complex dynamics of contacts therein [21, 23]. Genome organization within a given species varies [24, 25] across population [26], depending on a cell type [27], state in a cell cycle [28, 29], epigenetic state [30], and throughout differentiation [31, 32]. The presence of features and patterns first described using bulk data was validated, but extensive heterogeneity across single cells of the same type and population [33] was revealed, even down to individual alleles [25, 34, 35].

In this work, we leveraged a large collection of contact data, particularly Hi-C data from a wide selection of human cell lines and primary tissues [27], to better understand the DNA

Received: May 1, 2025. **Revised:** September 17, 2025. **Accepted:** September 30, 2025

sequence basis of genomic contact formation and 3D genome organization, thereby complementing earlier studies that focused predominantly on epigenetic determinants. Various sequence features such as dinucleotide content, functional motifs and regulatory elements, repetitive regions, and DNA secondary structures like G-quadruplexes have been associated with genome organization (reviewed in [36]). We contribute to these associations by applying an approach that specifically enriches for contacts that are likely to be heavily dependent on genomic sequence. In particular, we characterized the contacts across an extensive set of human cell types to isolate and investigate the core, cell-type-persistent or invariant contacts, which are likely enriched in associations with similarly cell-type invariant (CETI) factors, i.e. DNA sequence-based features largely common in all the human cells.

Materials and methods

Computational platforms and resources

Analyses were implemented mainly in the R programming language [37]. Most computations were performed on the in-house high-performance computing facilities at MRC Weatherall Institute of Molecular Medicine, University of Oxford, using nodes with 256 GB RAM, and Intel Xeon E5-2680v3 12-core (24-thread) and Intel Xeon E7-8891v3 10-core (20-thread) processors.

Statistical analyses

Differences between two groups were mainly assessed using the Mann–Whitney–Wilcoxon (MWW) test, unless stated otherwise. For comparisons involving >2 distributions, the pairwise MWW implementation in R was used with Benjamini–Hochberg adjustment [38]. A two-sided alternative hypothesis, and a significance level of 0.05 were used by default.

Contact persistence stratification

The primary dataset of genomic contacts consisted of 21 published and newly generated Hi-C data from 14 primary tissues and 7 cell lines at 40-kb resolution (Supplementary File S1: Supplementary Table S1), which were generated, consolidated and reanalysed in Schmitt *et al.* [27] (GEO:GSE87112). Refer to Schmitt *et al.* [27] for the Hi-C summary statistics of all datasets. The processed contact matrices, containing uniquely mapped reads and normalized using HiCNorm [39], were used for the stratification. HiCNorm was used in Schmitt *et al.* [27] to adjust the observed contact frequencies accounting for variation in effective fragment length (based on the distribution of restriction fragment sites along the genome), GC content and mappability by modeling these features in a Poisson regression framework [39]. Refer to Schmitt *et al.* [27] for further details on the processing of the raw contact matrices using a custom pipeline. Intra-chromosomal, long-range contacts with a linear distance (or contact gap) ≥2 Mb were used in the analyses. The c_p of each long-range contact was calculated as the number of cell type the contact is present in, i.e. the contact had a nonzero HiCNorm c_f in the Hi-C contact matrix. The counts of long-range contacts and unique regions forming these contacts per cell type and contact persistence value are presented in Supplementary File S1: Supplementary Tables S2 and S3, respectively.

Atlas of persistent substructures in human chromosomes

Arc and network diagrams for visualizing contacts

Arc and network diagrams were built using R libraries R4RNA [40] and visNetwork [41], respectively. In the network, a node represents a region (or bin) of length equal to the Hi-C resolution. Which bins to be represented as black nodes are determined by the user-defined gap value between nodes. If that value is 50, for instance, consecutive black nodes would have 50 bins in between them, i.e. bin 1 is connected to bin 52 by a grey arrow edge, bin 52 is connected to bin 103 and so on. The length of the edge is scaled by this gap value; however, since edges behave like a spring, this length is only the value at rest, and it can change when the edge must stretch due to the positioning of the contacts. Consequently, the network representation of the persistent substructure may not be proportional to the length of the chromosome, hence the equidistant black nodes become the only distance markers. Orange nodes and edges represent the highly persistent contacts.

Identification of CETI hubs

CETI hubs are comprised of a central region highly interconnected with several other 40-kb regions. For each chromosome, hub centres were chosen based on their involvement in extremely long-range persistent contacts. All associated 2-Mb persistent contacts involving those centres were then retrieved to define their respective hubs. We have identified 38 CETI hubs, with each chromosome contributing at least one hub except for chr. 15, chr. 17, and chr. X. For these chromosomes, it was hard to identify hubs because they barely contain contacts between very distant regions compared to other chromosomes of similar length. Active and inactive X chromosomes adopt different 3D architectures with the later reported to lack compartmentalization [42]. Averaging of signal in bulk Hi-C, particularly in female samples, could have contributed to the reduced number of long-range contacts.

Feature association with contact persistence

Over 300 publicly available chromatin and genomic feature datasets from multiple cell types were used for the associations. Unless indicated otherwise, overlapping ranges means having at least 1 overlapping bp. When necessary, the coordinates of the feature regions were converted to hg19, 1-based coordinate system. To derive the isochore data, regions with GC content data from Costantini *et al.* [43] were classified into the following isochore families based on a modified version of the % GC content ranges of families reported in Jabbari and Bernardi [44]: L1 (<36.4%), L2 (37.6%–39.6%), H1 (42%–45%), H2 (46.9%–52%), and H3 (>54%).

Region-wise association

Chromatin and genomic features, which are characteristics of a region (not by a contact), were associated with the unique contact regions per c_p. The significance of feature enrichment or depletion in a set of unique contact regions was quantified through permutation testing (10 000 iterations) using custom wrapper functions based on the R library regioneR [45]. The association was quantified by calculating (i) number of contact regions overlapping with feature ranges, where an overlap of one contact region with multiple regions of a feature was counted only once, and (ii) total intersection in bp. The same

permutation test procedure was used for calculating the significance of enrichment of long genes at unique high-c_p contact regions except that the sample statistic was the mean length of genes overlapping. Enrichment and depletion of features were determined at unique 40-kb regions forming the $c_p = 21$ and $c_p \geq 19$ contacts. The background was the set of all unique long-range contact regions ($c_p \geq 1$). Further characterization of contact regions was performed through differential analysis of 7-mer composition and associations with known and *de novo* motifs using HOMER [46].

Contact-wise association

The contact-wise associations were assessed in two ways. (i) Per c_p, we determined the fraction of feature-defined contact types based on whether the two regions of a contact overlap or do not overlap with a feature. (ii) Per feature-defined contact type, we calculated the fraction of contacts with a given c_p value.

Gene-related analysis

In all analyses involving genes, those with multiple transcripts were represented either by the single longest transcript or, in the case of ties, by one of the longest transcripts, but preferring coding over non-coding ones, if applicable. The latter was the case for 962 genes out of 24 910 (∼3.86%) unique genes from the UCSC hg19 annotation table.

Expression analysis

The Genotype-Tissue Expression (GTEx) data (E-MTAB-5214, 53 tissues) (in transcripts per million or TPM) [47] was filtered to contain only genes with expression values in at least one tissue. The EMBL-EBI Expression Atlas definition of the expression levels was used in this study—not-expressed (NE): TPM/FPKM < 0.5, low-expressed (LE): TPM/FPKM = [0.5,10], medium-expressed (ME): TPM/FPKM = (10,1000], and high-expressed (HE): TPM/FPKM > 1000 [48]. The same dataset was used for the co-expression analysis. To quantify co-expression, we applied the classical method of correlating expression values of gene pairs across multiple samples or tissues. A pair of genes is part of a CETI hub if both genes have at least 1 bp overlap with any of the regions comprising a hub. Another baseline expression dataset (E-MTAB-1733) [49], containing RNA-seq data of coding genes from 27 normal tissues from 95 adult individuals, was subjected to the same pre-processing and was used for validation.

Functional term enrichment analysis with DAVID

Because DAVID can only be used to process up to 3000 inputs at a single instance, 3 sets containing 2999 genes (in some instances, DAVID maps >1 DAVID gene identifier to a gene name) were randomly sampled without replacement from 4209 unique genes co-localizing with $c_p = 21$ contact regions. The built-in medium stringency of DAVID functional annotation clustering was applied. The built-in set of *Homo sapiens* genes was used as background. Gene count is the number of $c_p = 21$ contact genes associated with each term.

Replication timing data processing and analysis

The replication timing (RT) data (192 samples) from ReplicationDomain [50] was processed using custom R scripts into a final dataset, wherein a 40-kb region has one average RT value from each of the 61 unique tissue/cell lines. The cell types represented in this collection range from embryonic stem cells and their derived endoderm/mesoderm cultures, to fully differentiated tissues, and commonly used cell lines from across human anatomy, such as blood, ovary, liver, foreskin, and neural cells (50 noncancer and 11 cancer-related). RT measurements were binned at 40-kb to match the contact data and normalized across samples by aligning each sample to a reference set by linear-model fitting. For the association with c_p, the mean and median of the average values from each group of cell lines were calculated for each 40-kb region. Only regions with data from ≥59 cell lines were considered. A region was represented by the mean of the RT measurements overlapping with it, and the consensus RT for a contact was the average of the means from the two contacting regions.

Somatic cancer single nucleotide variant data processing and analysis

The dataset ($N = 38\ 428\ 969$) was downloaded from the ICGC Data Portal (Release 28, 27 March 2019) [51] from all human autosomes from 2320 samples. The single nucleotide variants (SNVs) were categorized based on their location relative to transcript components according to a hierarchical assignment of SNVs in the following decreasing order of priority, exon > intron > intergenic. To quantify the vulnerability of each contact to SNVs, we first calculated, for each region, (i) the number of mutated sites with at least one mutation (Nmutsite), and (ii) the total number of mutations (Nmut). Both metrics were normalized to the number of base pairs that can be mutated depending on the SNV type and location, denoted herein as $\text{Nmutsite}_{\text{norm}}$ and $\text{Nmut}_{\text{norm}}$, respectively. The consensus value for a contact was then equal to the mean of the metric values from the two contacting regions.

Contact sequence complementarity analyses

Contact sequence complementarity calculation

Sequence complementarity of contacts ($c_{||}$) was estimated using three measures. (i) Calculating $c_{||}$ *via* global sequence alignment using Levenshtein distance (i.e. minimum number of single-character edits to transform one sequence to another) ($c_{||}^{\text{align}}$) implemented using the open-source C/C++ library, edlib [52]. Substitutions, insertions, or deletions were penalized by 1 regardless of the base identity. (ii) Calculating $c_{||}$ *via* matching of 7-mer counts ($c_{||}^{\text{k-mer}}$) involved counting the occurrence of all possible 7-mers in both strands of a region and normalizing these counts to the length of the region. The $c_{||}^{\text{k-mer}}$ of a contact was then calculated as the sum of the absolute differences between the 7-mer normalized counts of a pair of regions in contact. (iii) Calculating $c_{||}$ *via* crude estimation of hybridization energy of regions in contacts ($c_{||}^{G}$) was possible with published free energy parameters for unique, perfectly matched DNA triplets [53]. Free energy parameters for 7-mers were derived from the triplet energy parameters by sliding a 3-bp window along the 7-mers and averaging the parametric values of triplets present. The $c_{||}^{G}$ of a contact is then given by the sum of the products of the 7-mer parametric values with the matching 7-mer counts of regions in contact. All complementarity values were normalized to the length of contact regions. The $c_{||}^{\text{k-mer}}$ and $c_{||}^{\text{align}}$ were negated so they positively correlate with the degree of complementarity.

Shuffling of contact regions

Shuffling, performed per chromosome and per c_p, was done using our general-purpose optimization library, ROptimus [54], in order to maximize the number of fake or shuffled long-range contacts that would not be present in the real or original set. Duplicates were also not allowed.

Contact map generation based on c_p, c_f, and $c_{||}$

The Hi-C_p maps include all long-range contacts from all cell types. The other Hi-C_f data not part of the main contact dataset were downloaded as .hic from the 4DN portal [55] and the sparse contact matrices were retrieved using the R library strawr [56]. For all contact maps with gradient colouring, the upper and lower limits of the colour scale are the upper and lower whisker values (Q1 − 1.5 × IQR, interquartile range, and Q3 + 1.5 × IQR). Values outside these limits are coloured using the corresponding most extreme colour.

Repeat element association with contact persistence

Repeat subfamily sites and copy number ranking were derived from the UCSC hg19 RepeatMasker annotation table. For calculating sequence complementarity using repeat-masked genome, only $c_{||}^{k\text{-mer}}$ was calculated as detailed above. Contacts formed by regions with >50% of their sequence masked were excluded. In addition, those involving regions with at least one missing bp in the unmasked genome were excluded to be consistent with the analysis using the unmasked genome.

Results

Contact persistence to focus on core genomic contacts

We quantified the persistence of genomic contacts across human cell types (Supplementary File S1: Supplementary Table S1) by integrating long-range contacts (40 kb resolution) between regions separated by ≥2 Mb. The latter threshold was set to be greater than the size of most TADs [57, 58] (Supplementary File S1: Supplementary Fig. S1). This was done considering our focus on core, sequence-driven components of genome organization, which would benefit from minimizing the effect of specific mechanisms, such as the reported role in TAD formation of the CCCTC-binding factor (CTCF)/cohesin-mediated loop extrusion predominantly demonstrated at the submegabase scale [22, 59, 60], and the pronounced driver role of cell-type specific transcription at shorter-scale organization [13]. The assembly of such contacts was possible by using 21 high-quality, reanalysed Hi-C datasets from various primary tissues and cell lines generously made available by Ren and co-authors [27] (Fig. 1A). Only intra-chromosomal contact data were available, so our analyses were restricted to those. We chose this data as the primary resource because maximizing cell type coverage was more important for our objectives than using fewer datasets with higher resolution and/or inter-chromosomal contacts. This limitation, however, does not preclude the relevance of our findings to inter-chromosomal and shorter-range contacts.

With the integration, the original contact matrices, denoted here as Hi-C_f (based on the conventional contact frequency measure, c_f), were converted into a single map termed Hi-C_p (Fig. 1B). In a Hi-C_p map, each contact is represented by a persistence score (c_p) from 1 to 21, equal to the number of human cell types it is present in Fig. 1B and C. The value of c_p is independent of the exact c_f value of a given contact across cell types. A contact in a given cell type contributes to c_p as long as c_f (HiCNorm-normalized [39]) >0 for uniquely mapped contacts. We did not apply statistical tests to enrich for interactions that have higher c_f than the expected value based on linear distance or gap between contacting regions because our goal was not to prioritize the potential for the most functionally relevant contacts. It should also be emphasized that contacts not flagged as "significant" by statistical enrichment tests are not guaranteed to be non-contacts or noise [61]. By design, we wanted to work with all contacts with firm evidence of happening, i.e. a robust chimeric read representing that contact. For this reason, we also ensured that only uniquely mapped reads were included, i.e. excluding chimeric reads that could not be confidently assigned to a single pair of contacting regions. In addition, working with longer-range contacts (>2 Mb, greater than the average TAD size determined using various methods [57]) was a way to mitigate inclusion of significant noise from the method, e.g. getting chimeric reads for contacts that are just extremely close to each other linearly. Furthermore, if a given contact was merely an artefact that happened to pass all our checks, we then had the voting of all the considered cell types to define the persistence of the contact. Therefore, artefacts are unlikely to have high contact persistence (c_p) and thus would have minimal influence on the conclusions (see Supplementary File S1: Supplementary Section 1 for extended explanation). The contact persistence scores were also found to be robust against any outlying datasets based on sequencing depth and similarity of cell types and tissues (Supplementary File S1: Supplementary Fig. S2).

Out of the 112 861 349 long-range contacts examined, the proportion of contacts decreases with increasing persistence, with the least and most persistent ones comprising 17.448% and 0.012%, respectively (Supplementary File S1: Supplementary Table S4 and Supplementary Fig. S3). The c_p stratification is independent of the exact c_f, but we did find that persistent contacts tend to have relatively high c_f, even when accounting for contact gap variation across c_p (Supplementary File S1: Supplementary Figs S4 and S5). Persistent contacts tend to be shorter in range, with median contact gap lengths between 2.4 and 2.7 Mb for $c_p \geq 19$ contacts (Supplementary File S1: Supplementary Fig. S6). Interestingly, in Yang *et al.* [62], authors reported that short-range contacts, with contact gaps of around 2.5 Mb, primarily have high relative contact frequencies conserved across the lymphoblastoid cells of humans and three other primates—chimpanzee, bonobo and gorilla. By directly associating their data with our c_p data, we did find that our persistent contacts mostly correspond to the contacts they found to have conserved high-c_f pattern (Supplementary File S1: Supplementary Figs S7 and S8).

With this curated data of long-range contacts, we investigated the foundations of contact persistence to facilitate the study of the core, invariant determinants (as well as its implications) of higher-order genome organization, primarily by contrasting variable from persistent contacts (Fig. 1D). In Fig. 2, we visualize these persistent contacts as arcs in an arc diagram (Fig. 2A and C) and as edges in a network diagram (Fig. 2B and D), particularly highlighting outlying persistent contacts that are extremely long-range. For instance, out of its 10 253 $c_p \geq 19$ contacts, chr. 1 has 303, 90, and 5

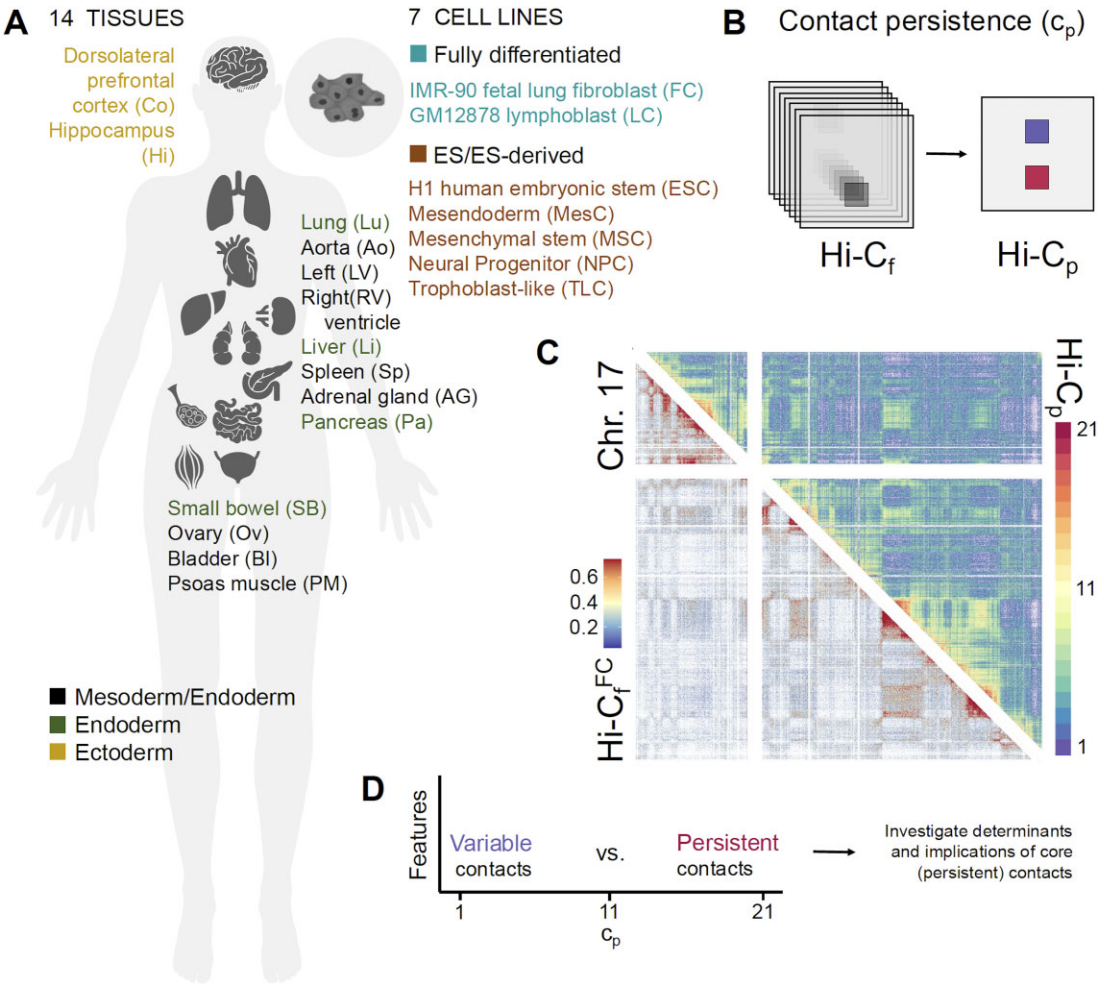

Figure 1. Contact persistence to isolate and investigate the core genomic contacts, their determinants and implications. (**A**) Human tissue and cell line sources of the Hi-C (or Hi-C_f) datasets [27] used in this study. The icons indicate the sources, and the colours denote tissue origin, cell line source and differentiation state (Supplementary File S1: Supplementary Table S1). (**B**) Diagram representation of c_p derivation. Black squares in Hi-C_f maps represent contacts with HiCNorm $c_f > 0$ (uniquely mapped reads), i.e. considered present in a dataset. (**C**) Chr. 17 FC (IMR-90 foetal lung fibroblast cells) Hi-C_f compared with Hi-C_p of long-range contacts from all cell types. (**D**) The strategy of this study to reveal core determinants of genomic contacts—investigating the core, persistent contacts by contrasting with variable ones to reveal their implications and determinants, which are likely to be similarly persistent across cell types, i.e. sequence determinants.

highly persistent contacts joining intervals linearly separated by ≥8, 26, and 200 Mb, respectively. In addition, the network diagrams effectively show how highly persistent contacts can bring together far-off regions and form clusters of highly interconnected regions that are present amongst most cell types and that we call in this text as CETI hubs (Fig. 2B and D).

Feature associations of persistent contacts

To examine the identity, causes and implications of a contact persistent in many cell identities, we looked for significant associations of c_p with a set of ~300 features comprised of CETI, sequence features (e.g. CpG island (CGI) motifs, *de novo* motifs, non-B DNA motifs, isochores) and cell-type-specific ones (e.g. A/B compartments and subcompartments, histone modifications, and transcription factor binding sites) (Supplementary File S1: Supplementary Table S5 for list of features and sources). The lengths of most features are below the 40-kb contact resolution even when combining consecutive or overlapping intervals. Exceptions include genomic regulatory blocks (GRBs), A/B compartments and subcompartments, as well as forests and prairies—regions of high and low CGI density, respectively [63]. GRBs are regions containing syntenic clusters of highly conserved non-coding elements (CNEs) shown to coincide with TADs [64]. LAD marker LMNB1 and isochore regions have lengths both below and above the contact resolution. Isochores are large DNA segments, typically a few hundred kilobases in size, characterized by relatively uniform GC content [43]. They are classified into GC-poor L and GC-rich H isochore families, which are further divided into subfamilies based on increasing GC content: L1, L2, H1, H2, and H3 [44].

We found, at high-c_p regions (region-wise), a significant enrichment of heterochromatin-related domains and features, i.e. B-compartments and subcompartments, LAD marker LMNB1 [65] and repressive chromatin mark H3K9me3 (Fig. 3A, left, Supplementary File S1: Supplementary Fig. S9, and Supplementary File S2). This could, in part, be attributed to a prior observation that about 25% of the human genome is consistently classified as B-compartment across multiple cell types and tissues [27]. At the sequence level, AT-rich

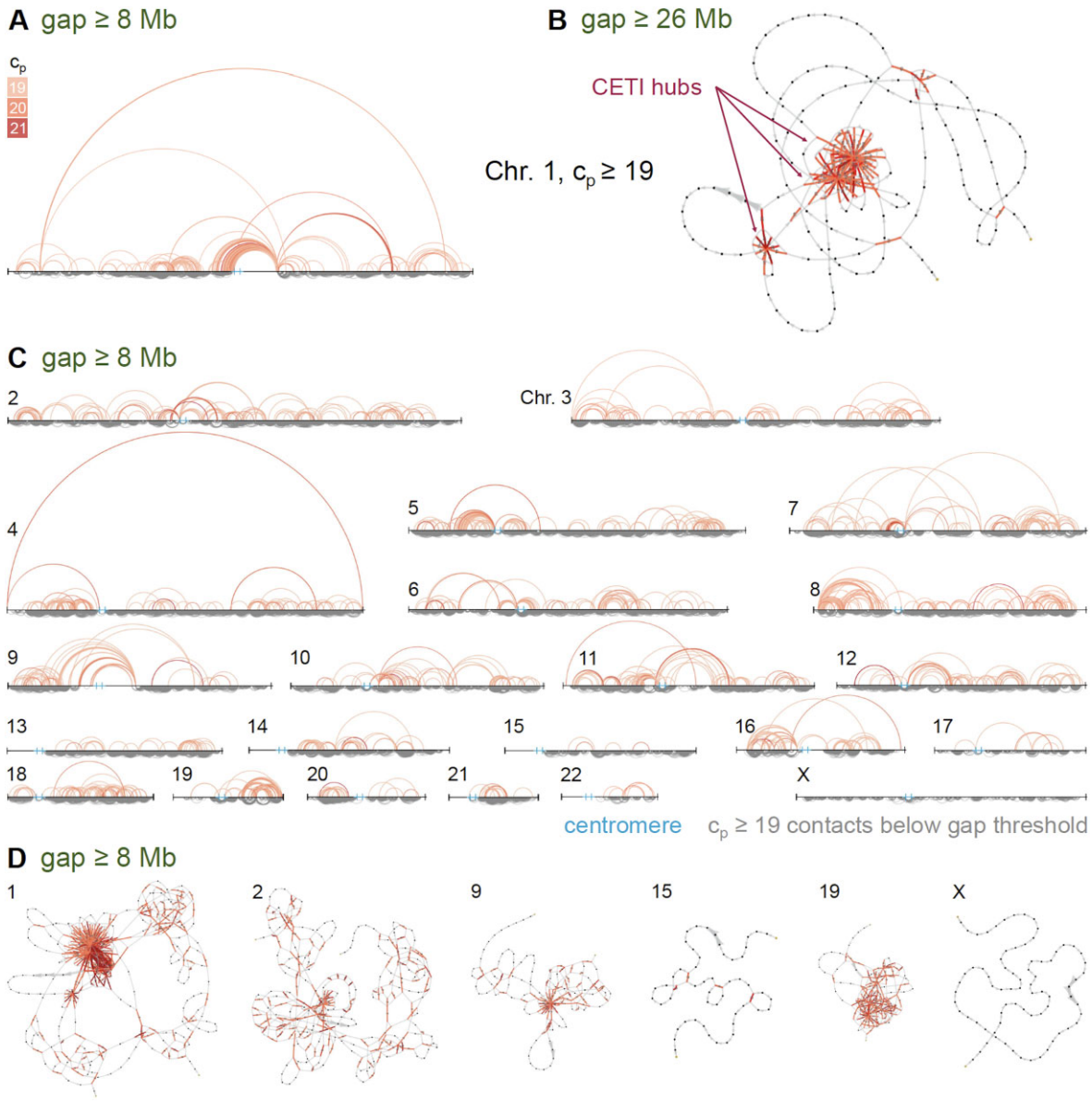

Figure 2. Visualization of persistent contacts and the chromosome organization they mediate. Contacts displayed have $c_p \geq 19$ (c_p denoted by arc or edge colour). Centromeric regions are marked by cyan lines and by thick edges on arc and network diagrams, respectively. (**A**) Arc diagram for chr. 1. The upper side has contacts with gap ≥200 40-kb bins (or 8 Mb), while the bottom side shows the rest of $c_p \geq 19$ contacts (≥2 Mb). (**B**) Network diagram for chr. 1 with gap ≥650 40-kb bins (or 26 Mb). CETI hubs of interconnected regions are indicated. (**C**) Arc diagrams corresponding to panel (A), but for the rest of human autosomes and chr. X. (**D**) Representative network diagrams, analogous to panel (B), for some chromosomes generated using $c_p \geq 19$ contacts with gap ≥200 40-kb bins (or 8 Mb).

features are consistently enriched, particularly A-phased repeats, L1 and L2 isochores, and CpG-depleted prairie sequences [63] (Fig. 3A, left). Analysis of enriched 7-mers at persistent contact regions show no strong motif, but they tend to contain more A/T over G/C bases (Supplementary File S1: Supplementary Table S6 and Supplementary Figs S10 and S11). At 40-kb resolution, however, the differences in the AT content across c_p is not drastic (Supplementary File S1: Supplementary Fig. S12), with values across c_p being close to the recently calculated genome-wide average of 40.9% GC [66]). Motif analysis using longer motifs highlighted 8-bp *de novo* motifs as the most significantly enriched hits based on the HOMER recommended criteria [46] (Supplementary File S1: Supplementary Figs S13 and S14). The proportion of contacts overlapping with such enriched features increases with c_p; however, not all persistent contacting regions contain them as shown in Supplementary File S1: Supplementary Fig. S15. This is expected, as different contacts may be influenced by different factors, potentially reflecting their functional roles, such as transcription regulation or chromatin compaction. Further analysis showed that the enrichment of the aforementioned features is accompanied by the depletion of euchromatin-related domains and GC-rich features namely A-compartments and subcompartments, CpG island, putative G-quadruplex sequences, H1, H2, and H3 isochores, and CpG-dense forest sequences [63] (Fig. 3A, right, and Supplementary File S1: Supplementary Fig. S9). Findings translate contact-wise, with the proportion of contacts associ-

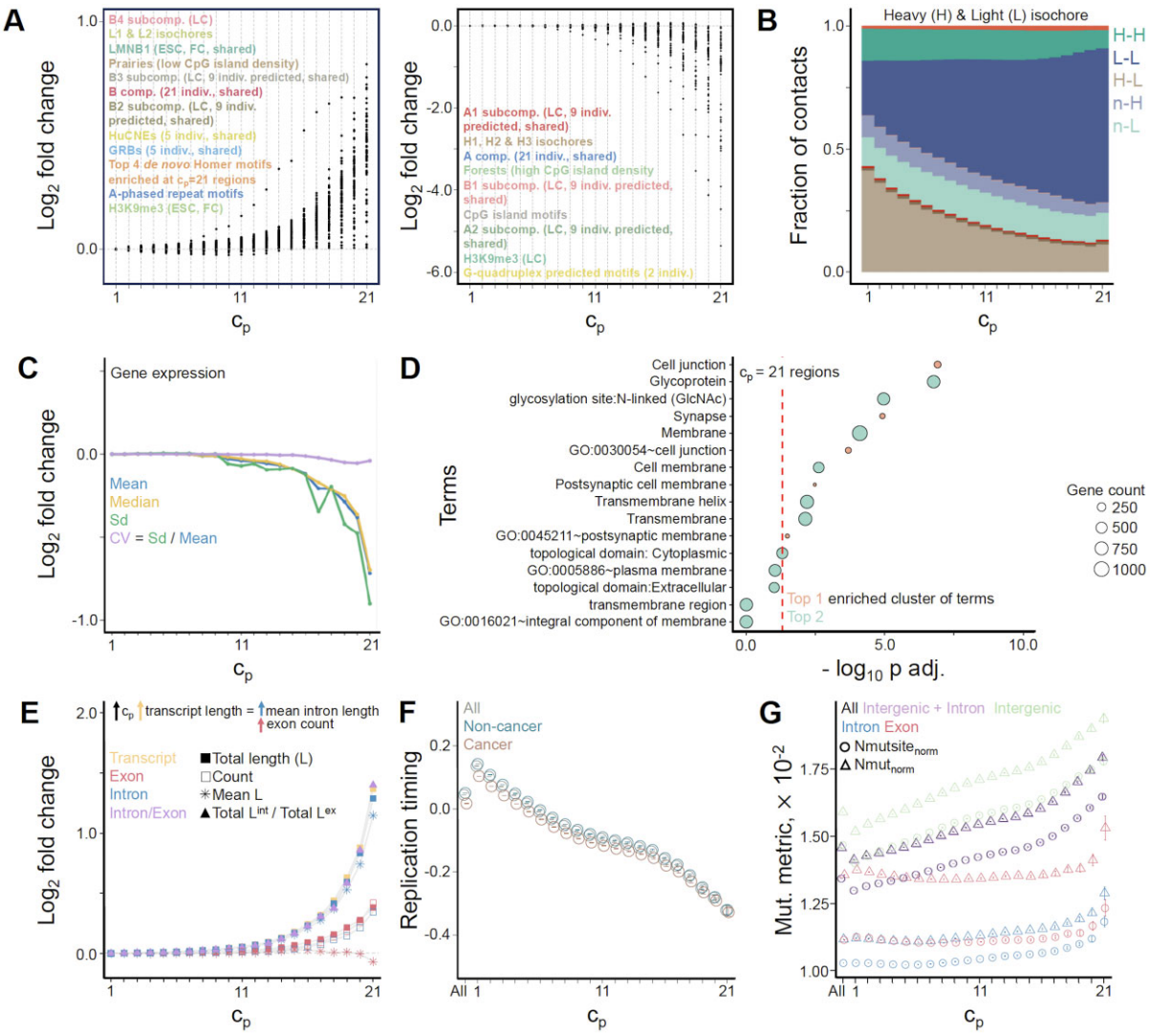

Figure 3. Persistent contacts enriched for contacts with features associated with heterochromatin and preferential AT sequence composition. The $\log_2$ fold changes are relative to the value at $c_p = 1$. (**A**) Log_2 fold change of the proportion of unique contact regions across c_p overlapping with significantly enriched (left) and depleted (right) features at unique $c_p = 21$ contact regions (permutation testing, Supplementary File S1: Supplementary Fig. S9 heatmap for complete result of permutation tests and Supplementary File S2 for data behind the heatmap). Feature names are written in descending order, based on the median of absolute $\log_2$ fold changes at $c_p = 19{:}21$ of datasets within each feature type. Cell-type specific features can have multiple datasets (indicated in parentheses). "Indiv." refers to individual data, "shared" refers to regions shared by or common to all individual data for that feature, and "predicted" refers to subcompartment regions predicted by SNIPER [67]. (**B**) Fraction of contacts overlapping isochore families across c_p (Supplementary File S1: Supplementary Figs S15 and S16 for similar plots of other enriched features). Only dominant contact types are shown in legend; "n" means no overlap. (**C**) Log_2 fold change of the mean of various cross-tissue expression metrics across c_p (Supplementary File S1: Supplementary Fig. S17). The $c_p \leq 3$ and $c_p \geq 19$ distributions are significantly different except for CV. Only genes with data in ≥70% of the tissues were considered. (**D**) Top 2 most significant DAVID clusters of enriched functional terms at $c_p = 21$ (Supplementary File S1: Supplementary Fig. S19). Dashed line at $\log_{10}(0.05) \sim 1.301$. (**E**) Log_2 fold change of the mean of derivative lengths and counts of various genic elements across c_p (Supplementary File S1: Supplementary Fig. S20). Mean length of $c_p = 21$ and $c_p \geq 19$ genes significantly higher than $c_p \geq 1$ genes (permutation testing). (**F**) Mean cross-tissue RT ($\log_2$ ratio of early and late signals) across c_p using data from all ($N = 61$), noncancer only ($N = 50$) and cancer-related only ($N = 11$) cell lines; error bars at 95% confidence intervals (Supplementary File S1: Supplementary Fig. S23). All pairwise comparisons of the distributions are significantly different. (**G**) Mean somatic cancer SNV frequency metrics across c_p (Supplementary File S1: Supplementary Figs S24–S26); error bars at 95% confidence intervals. $\text{Nmutsite}_{\text{norm}}$ and $\text{Nmut}_{\text{norm}}$ are the number of mutated sites with at least one mutation, and the total number of mutations, respectively, normalized to the total bp that can be mutated depending on the SNV type and location. The c_p = "All" refers to all contacts with a c_p value.

ated with the enriched features increasing with c_p as demonstrated here for H/L isochores (Fig. 3B, and Supplementary File S1: Supplementary Figs S15 and S16). Consistent with enrichment of heterochromatin-related features, we also observed that genes at persistent contacts have a relatively lower expression across tissues, evident in the lower distribution of cross-tissue mean and median values of expression (Fig. 3C and Supplementary File S1: Supplementary Fig. S17). In alignment, the fraction of tissues with low expression for a gene increases with c_p, accompanied by a decrease in fraction of tissues with medium and high expression (Supplementary File S1: Supplementary Fig. S17). Note that we can not necessarily expect the contacts persistent across cell types to overlap with constitutively expressed genes, particularly housekeeping genes, because chromatin contacts serve variable functions, such as facilitating expression or compaction [1]. We did find that housekeeping genes overlap significantly less with persis-

tent contacts than with variable contacts (Supplementary File S1: Supplementary Fig. S18).

Genes co-localizing with persistent contacts are functionally different, showing enrichment for terms related to synapse, glycoproteins, and (trans)membrane (Fig. 3D and Supplementary File S1: Supplementary Fig. S19). Analysis of CETI hub genes also show different enriched terms across hubs (Supplementary File S3). With the most significant cluster of terms in Fig. 3D showing enrichment of neuronal genes that are known to be long [68], we also investigated gene length variation and found that $c_p = 21$ contact genes are also significantly longer, driven by higher mean intron length and greater count of exons as c_p increases (Fig. 3E and Supplementary File S1: Supplementary Fig. S20). This functional diversity could translate to some degree of variation in expression, for instance, due to some subsets of persistent genes being more tightly coregulated than others, some subsets not tightly repressed as the others given that a few persistent contacts overlap with nonheterochromatin-related features. This could explain why the normalized variation is not significantly different between variable and persistent gene sets (Fig. 3C). With co-expression analysis across tissues, we did find that high-c_p gene pairs within a hub have higher correlations than those high-c_p gene pairs not within a hub (Supplementary File S1: Supplementary Figs S21 and S22). This supports the functional relevance of CETI hubs based on harbouring functional and coordinately expressed units formed by extremely long-range contacts present across multiple cell types.

To further probe the implications of the persistent genome organization, we associated contact persistence with two molecular phenotypes, RT (using aggregated data from different cell types) and somatic mutation frequency (using cancer SNV data). Consistent with the enrichment of heterochromatin-related features [69], we found that persistent contacts occur between regions that are late-replicating (Fig. 3F and Supplementary File S1: Supplementary Fig. S23), with this pattern persisting across both diseased (i.e. cancer) and nondiseased contexts. Persistent contacts also generally have higher SNV frequencies regardless of mutation type (Fig. 3G and Supplementary File S1: Supplementary Figs S24–S26). We observed significantly increased frequencies in the context of the whole contact region and even when considering only introns or intergenic portions, which could potentially be attributed to the impaired access of repair machinery to the persistent contact sites [70, 71]. Exons, however, have an overall lower mutational burden attributed to more efficient mismatch repair activity (compared to introns) [72], which could explain why the observed increasing vulnerability with persistence is not pronounced at exons (Supplementary File S1: Supplementary Fig. S26). To further characterize the mutational mechanisms at persistent contacts, the analyses could be repeated grouping somatic mutations per mutational signature. Degasperi *et al.* defined 38 SNV-based mutational signatures, consisting of signatures similar to COSMIC mutational signatures and some unexplained ones [73].

Higher sequence complementarity between persistent contacts

Sequence-specific interactions involving the recognition and association of complementary or homolog nucleic acid sequences are commonplace occurrences in various molecular processes (Fig. 4A, left). The protein-independent, preferential interaction of identical DNA duplexes (DNA self-assembly) with the help of biological cations has also been demonstrated *in vitro* [74–78], even for DNA occurring in nucleosomes (nucleosome self-assembly) [79] (see Fig. 4A, right, and Supplementary File S1: Supplementary Section 3 for description of each paper). The sequence identity awareness or recognition involved in these processes, whether happening directly and/or indirectly, prompted us to hypothesize that the degree of sequence similarity or complementarity between regions is associated or could contribute to the tendency of regions to come in contact, regardless of cell type or state. Associating the similarity of sequences with contact persistence, we have taken a general approach, defining the similarity to be independent of any specific chromatin feature or genomic pattern, by using different measures of complementarity between sequences, $c_{||}$. These measures were derived from (i) matching of short-span 7-mer counts between sequences in contact ($c_{||}^{\text{k-mer}}$) (Fig. 4B), (ii) long-span global sequence alignment using edit distance ($c_{||}^{\text{align}}$) (Fig. 4C), and (iii) approximate estimation of hybridization free energy ($c_{||}^{G}$, decreasing trend means increasing $c_{||}$). The global alignment ($c_{||}^{\text{align}}$) metric measures the complementarity of full 40-kb sequences. But in the plausible scenario, where the whole 40-kb regions and their full complementarity are not essential for the contacts, k-mer matching ($c_{||}^{\text{k-mer}}$) is more appropriate, as it only matches the counts of short-span k-mers within the contacting *i* and *j* regions. We used 7-mers due to the finding that it is the minimum span defined by context-dependent spontaneous mutation rates, which suggests that the genome is optimized by those rates at this sequence scale [68]. This $c_{||}^{\text{k-mer}}$ metric accounts for the sequence complementarity of shorter sequences or k-mers within regions, and is not heavily influenced by the exact arrangement of k-mers within the 40-kb sequence. The third method, $c_{||}^{G}$, crudely estimates the hybridization energy using published free energy parameters from short oligonucleotides [46]. This method is still an inadequate estimator of the real hybridization free energies because it ignores multiple conformations possible, along with potential mismatches, internal loops, and dangling ends. Therefore, the use of this metric in this work was mainly for observing the general trend across c_p. Its major advantage over the other metrics is that it could somehow differentiate the hybridization abilities of the short spans depending on the base composition.

Remarkably, for the alignment and k-mer-based metrics, we observe a stepwise increase in the complementarity of sequences with rising persistence (Fig. 4B and C). Consistently, the hybridization of contacting sequences is more favourable with increasing c_p as reflected by the decreasing trend of the hybridization-free-energy-based $c_{||}^{G}$ score (Supplementary File S1: Supplementary Fig. S27C). The observation holds valid even when limiting the drastic contact gap variation across c_p by observing $c_{||}$ trends at narrower ranges of contact gaps (Fig. 4D), and when completely removing the effect of gap or distance (Supplementary File S1: Supplementary Fig. S28), which influences both contact formation and sequence similarity. When shuffling the contacting DNA regions within a c_p category to form fake contacts not actually paired in that c_p category (Supplementary File S1: Supplementary Figs S29 and S30, and Supplementary Table S7), fake contacts exhibit a similar trend of increasing $c_{||}$ with increasing c_p (Fig. 4B and C), which could be explained by persistent contacts be-

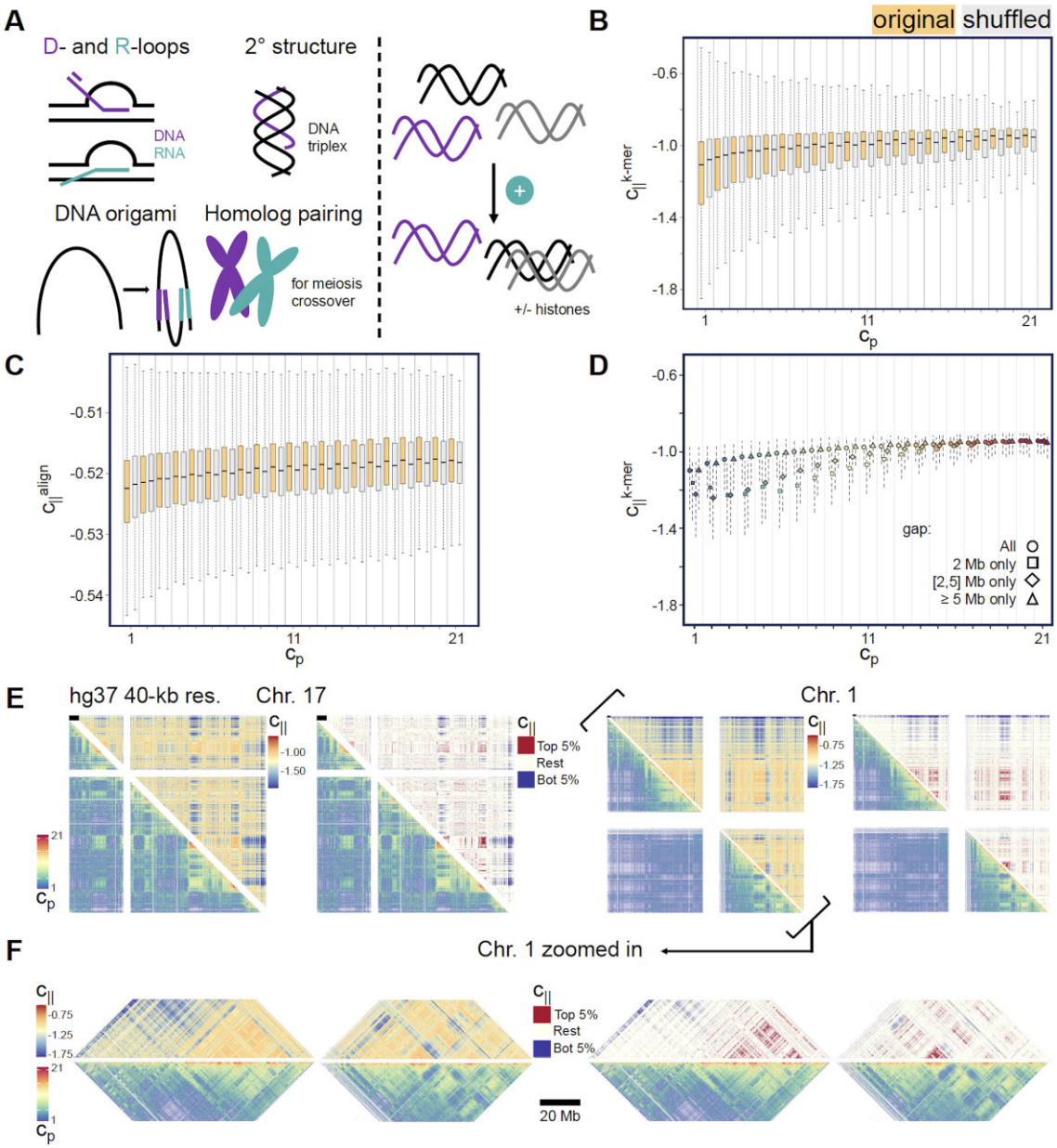

Figure 4. Higher sequence complementarity between persistent contacts. (**A**) Studies on the preferential association of identical DNA duplexes *in vitro* [74–79] (right) motivate the implication of sequence complementarity, $c_{||}$, in contact formation, reinforced by many known sequence-dependent processes (left), involving single- and double-stranded nucleic acids (e.g. D- [80] and R-loop [81, 82] formation, triplex formation [83, 84], DNA origami nanotechnology [85], and homolog pairing [86]). Sequence complementarity of original (orig.) contacts calculated based on (**B**) short-span k-mer matching ($c_{||}^{k\text{-mer}}$) and (**C**) long-span global alignment using edit distance ($c_{||}^{align}$); $c_{||}$ of shuffled (shuff.) contacts. Pairwise comparisons of orig. distributions between any two neighbouring c_p values show significant differences except for some comparisons between the highly persistent contacts ($c_p \geq 17$). Each orig. versus shuff. distribution pair is also significantly different (Supplementary File S1: Supplementary Fig. S31) (**D**) $c_{||}^{k\text{-mer}}$ across c_p at specific contact gap ranges (Supplementary File S1: Supplementary Figs S32A and B). Shown are medians of distributions and the dashed segments extend to the 25th and 75th percentiles. Pairwise comparisons of the five most persistent and five most variable distributions show significant differences. (**E**) Hi-C_p versus Hi-$C_{||}^{k\text{-mer}}$ using actual and binned values of $c_{||}$ (see Supplementary File S5 for other chromosomes). Brackets highlight the zoomed in area shown in panel (F). Scale bars at the top left corner are 4 Mb long. (**F**) Zoom in on chr. 1 highlighted regions in panel (E).

ing more similar due to the enrichment of specific feature types as discussed previously. However, fake contacts show significantly lower $c_{||}$ distribution, as compared to that of real ones, more pronounced for high-c_p values (Fig. 4B and C, and Supplementary File S1: Supplementary Fig. S31). This suggested that the degree of complementarity is specific to real pairs of sequences in contact and not solely dependent on a single sequence motif or pattern present in all persistent regions.

We thus found a link between the Hi-C-inferred contact persistence and a simple metric, $c_{||}$, calculable between any two sequences, without any training or assumption borrowed from experimental data. Consistent with the observed positive correlation between $c_{||}$ and c_p, the complementarity-based map, Hi-$C_{||}$ (generated by directly plotting the calculated $c_{||}$ values and binned at a matching resolution), shows areas of both high and low signal like those in the Hi-C_p (Fig. 4E and F, highlighted with brackets). Similar areas of matching

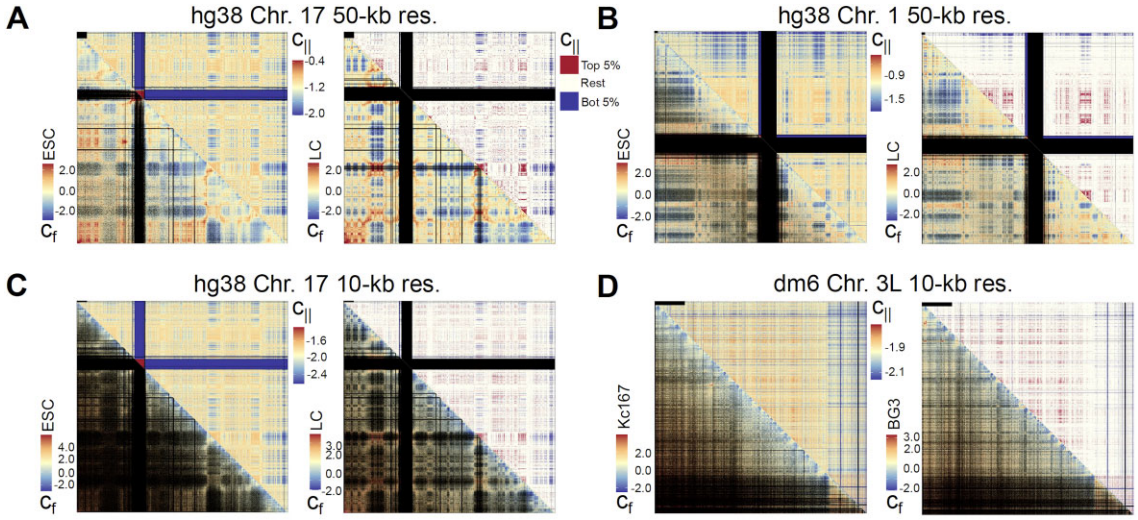

Figure 5. Genome-wide $c_{||}$ values recapitulate some Hi-C features in additional human datasets of different resolutions and in another species, *Drosophila*. Hi-C_f versus Hi-$C_{||}^{k\text{-mer}}$ using $\log_2$ observed over expected c_f, and actual and binned values of $c_{||}$. Areas with no contact signal are shown in black to enhance visibility of contact-derived structures. Scale bars at the top left corner are 4 Mb long. Human (hg38) cell types: ESC (H1-hESC) embryonic and LC (GM12878) lymphoblastoid cells at (**A**, **B**) 50-kb and (**C**) 10-kb resolution (res.); (**D**) *Drosophila* (dm6) cell types: Kc167 embryonic and BG3 neuronal cells at 10-kb resolution (Supplementary File S4 for other chromosomes).

high and low signals can also be seen when comparing $c_{||}$ with c_f data from human embryonic and lymphoblastoid cell lines (Fig. 5A–C), and *Drosophila* embryonic and differentiated (neuronal) cell types (Fig. 5D) at different resolutions (Supplementary File S4 for complete set of contacts maps across chromosomes using data from different human and *Drosophila* cell lines). These similarities are emphasized in the versions of the hybrid contact maps, where $c_{||}$ is binned to distinguish the top and bottom 5% contacts based on value. Extending the insights obtained from the human genome to the *Drosophila* genome reinforces the relevance of sequence complementarity in genome organization. When interpreting these maps, however, it should be noted that the Hi-$C_{||}$ maps do not account for the polymer properties of a DNA and, as such, do not display the expected decay of contact probability with increasing distance [87], as observed in the experimental Hi-$C_{||}$ and Hi-C_p maps. This is reflected in the high-signal areas at Hi-$C_{||}$ maps between regions extremely far from each other in the linear genome (Fig. 3E). Regardless of the high complementarity, these highly off-diagonal interactions would likely diminish due to the long stretch of DNA between their regions. Methods like polymer simulations would be necessary to generate contact maps more accurately recapitulating the genome architecture profile dictated by sequence complementarity.

Repeats and the observed sequence complementarity of contacts

Based on the DNA/nucleosome self-assembly phenomenon [74–79], repetitive elements, representing similar sequences, have been proposed to mediate a sequence-dependent phase separation of the genome [88]. Consistently, a combination of computational [89–91] and experimental [92] findings across eukaryotic species show that the self-clustering of repeat families is indeed correlated with genome organization. Given these, we determined whether transposons, which have brought similar sequences to different parts of the genome, could be reinforcers of the observed higher complementarity at persistent contacts by using metrics to measure site distribution of a subfamily between two contacting regions. Likely, the most influential transposon remnants would be from subfamilies with higher copy numbers that could cover a high proportion of contacts. But since it is not guaranteed that their sites would be distributed across regions in contact, we used metrics to measure site distribution between two DNA regions in contact. We calculated the shared number of sites of a transposon in the two sites per contact. For example, given i_1 and j_1 regions that are in contact and have 5 and 1 mammalian-wide interspersed repeat (MIR) sites, respectively (i.e. site total of 6), the contact has 1 shared number of MIR site. This metric, however, could not differentiate contacts with equal shared numbers but different numbers of sites in each of their regions. To address this, we calculated site skew, adopting the concept of the GC skew. Site skew is equal to $|N_i - N_j| / (N_i + N_j)$, where N_i and N_j are the number of transposon sites in the i and j regions forming a contact. A site skew of 0 means that the site total in a contact is split equally between its two regions. A site skew of 1 means that the sites in a contact are all located in one of its regions. Given another contact formed by i_2 and j_2 regions having 3 and 1 MIR sites, respectively, we could not differentiate site distribution between the two contacts using the shared number metric because both contacts have 1 shared site. In contrast, although they both contribute one pair of sites, site skew would suggest that sites tend to be more distributed between the two regions of the second contact with the lower value of 0.50 (versus 0.67).

In Fig. 6, we highlight four candidate subfamilies, MIRb, MIR, L2a, and L2c, likely to be most influential to the observed higher complementarity of persistent contacts (green text in Fig. 6). These subfamilies, among the ones with the highest copy numbers, have at least 1 shared number of sites in >50% of persistent contacts ($c_p \geq 19$) (Fig. 6A), and lower median site skew at persistent compared with variable contacts ($c_p \leq 3$) (Fig. 6B). Based on the site distribution metrics, the sites of these subfamilies tend to be more distributed between persistent contact regions even relative to other high-copy-number subfamilies like AluJb and AluSx (~2000 sites

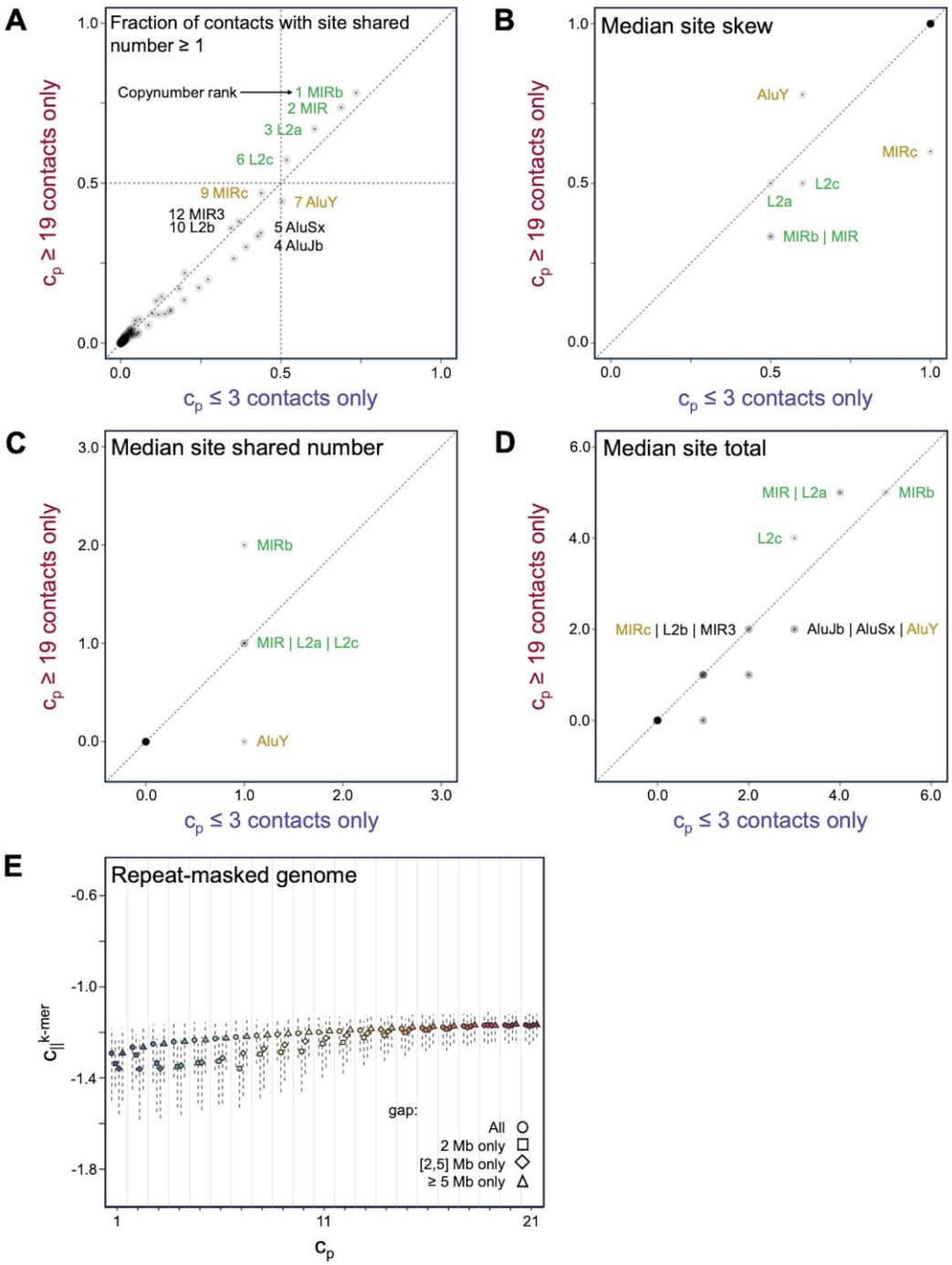

Figure 6. Repeat contribution to the observed sequence complementarity of contacts. Transposon subfamily site distribution at persistent ($c_p \geq 19$) versus variable contacts ($c_p \leq 3$). Given *i* and *j* forming a contact, and N_i and N_j being the number of subfamily sites in *i* and *j*, (**A**) shows the fraction of contacts with at least 1 shared number of site, i.e. $\min(N_i, N_j) = 1$. In green text are subfamilies with ≥1 shared number at >50% of contacts. The number beside the subfamily name denotes the copy number ranking. (**B**–**D**) show medians of the distributions at persistent and variable contacts of the following metrics: (**B**) site skew equal to $|N_i - N_j| / (N_i + N_j)$, (**C**) shared number of sites equal to $\min(N_i, N_j)$, and (**D**) total number of sites of a contact equal to $N_i + N_j$. In yellow text, are additional subfamilies with median site skew not equal to 1 for both distributions. Only contacts with at least one site can have a valid site skew value. Each data point, representing a subfamily, is coloured with a transparency level of <1, whereby darker areas indicate the overlap of many points. For all named subfamilies, the variable and persistent distributions of the metrics are significant. (**E**) $c_{||}^{k\text{-mer}}$ across c_p at specific contact gap ranges using repeat-masked genome (Supplementary File S1: Supplementary Fig. S32C and D).

more than L2c). MIRb was the only subfamily that had higher median shared number at persistent contacts (Fig. 6C). MIRb does have the highest copy number among the subfamilies, but it is not the only one that could have 2 shared sites because the other three subfamilies had median site total of ≥4 sites (Fig. 6D) at persistent contacts. MIR, L2a, and L2c have mean shared numbers at persistent contacts significantly higher than the variable counterparts. As for the rest of the subfamilies, since only contacts with at least 1 site could be considered in the site skew calculation, having a median site skew of 1 means that most of the long-range, intra-chromosomal contacts these subfamilies have inserted on have no shared site. With recent studies linking them to genome organization [93, 94], we also applied this type of analysis to satellite DNA subfamilies. For all the subfamilies, the median number of shared sites and site skew are 0 and 1, respectively, for both persistent and variable contacts. Comparing the distributions of the two metrics, certain subfamilies, namely ALR/Alph, MSR1 and HSAT5, tend to be more distributed between persistent contacting regions (Supplementary File S1: Supplementary Fig. S33). This motivates further investigations of the role of satellite DNAs in 3D genome organization, especially with the availability of the telomere-to-telomere (T2T) human genome assembly [94].

Given that repeats represent a significant amount of sequence, about half of the human genome, it is not surprising that repeats are key contributors to the observed complementarity phenomenon. Finding that some high-copy-number subfamilies, particularly the ancient MIRb and MIR transposons as well as L2a and L2c, tend to have more shared sites that are more distributed at persistent contacts support this and is consistent with earlier aforementioned studies. It should be noted, however, that the site distribution metrics do not directly measure the individual and collective contribution of subfamilies to the total degree of complementarity of contacts based on the length and sequence of the remnants. Interestingly, when we used the repeat-masked genome to analyse a smaller subset of long-range contacts, whose regions were mostly devoid of annotated repeats, even the unmasked portions of persistent contacts in that subset were found to be more similar than that of variable contacts, suggesting that this higher complementarity at the given resolution is a characteristic of persistent contacts that the annotated simple and interspersed repeat sites could not solely account for (Fig. 6E, and Supplementary File S1: Supplementary Fig. S32C and D).

Discussion

With this work, we aimed to further understand the sequence basis of 3D genome organization, which is an important foundation for evolutionary and functional investigations such as the prediction of the effects of genomic variations, and characterization of the role of non-coding sequences that make up the majority of eukaryotic genomes. We used an extensive collection of human contact data from multiple cell types and tissues to enrich our analysis for contacts that are persistent across different cell contexts and thus likely to be heavily dependent on cell-context-invariant features such as the DNA sequence.

Persistent contacts are predominantly associated with AT-rich sequence features and B-compartment/heterochromatin-related features across different cell types, which generally are less dictated by the cell identity, particularly the case for constitutive heterochromatin regions in contrast to A-compartment active regions. This observation is linked to the proposed model that B compartmentalization, which was not linked to any sequence pattern except for preferring AT-rich regions, could be the "default" state of sequences [95], unless sequences become A-compartment material, such as by having active TSS sites [95], which are prone to cell-type specificity depending on the required gene expression profile. Also, with AT-rich duplexes found to interact more favourably than GC-rich ones *in vitro* [96], AT-rich features enriched at contacts, could contribute to persistence across cell states, demonstrating how a nonspecific or core sequence effect can have influence on which regions preferentially interact. Also, this tight clustering of AT-rich sequences, typically characterizing heterochromatin regions, could help explain why certain chromosomes, like chromosomes 1 and 9, known to have a large portion of constitutive heterochromatin, contain prominent CETI hubs, i.e. large clusters of persistent long-range contacts (Fig. 2).

The mechanisms involved in the preferential interaction of similar regions, shown here to be correlated with genome organization, are yet to be elucidated in detail but have been described in terms of direct and indirect mechanisms. The direct recognition of sequence identity is supported by *in vitro* [75–77, 79, 96] and *in vivo* [97] studies, particularly in the fungal model *Neurospora crassa*, where it was observed that genomic regions align without the involvement of known complexes, crucial for homology recognition. In addition, similar DNA sequences could have similar proteins and RNAs co-localized, which could be the ones driving the preferential interaction or phase separation [92, 98]. Both mechanisms could potentially contribute, and the combination and identity of factors may vary for certain subsets of contacts depending on whichever sequence or epigenetic features are present. Our analyses do not thoroughly characterize specific combinations of mechanisms at play, but the association studies have identified a general characteristic, sequence complementarity between regions, that is most pronounced for contacts persistent across different cell contexts. Hence, it is a potential core sequence determinant of genome organization, which contributes to the "default" tendency of interactions between regions in any cell context, much like the observed favourable association of AT-rich duplexes [96]. The implication of sequence complementarity in genome organization along with findings from the characterization of the persistent contacts also corroborate with associations previously brought up elsewhere, providing insight on the relationship between genome organization and function. For instance, as summarized in Mazur *et al.* [99], pairing of homologous DNA duplexes have been proposed to have a role in initiating heterochromatin formation and transcription silencing [100], in mediating cytosine methylation [101], which produces mutation hotspots across the genome, and in generating supercoiling, which could affect topoisomerase-dependent long genes [102].

Findings presented here prompt further computational and experimental investigations. The contribution of sequence to genome organization could be quantitatively assessed, across multiple scales of organization, by comparing models of genome organization based on the complementarity of sequences with experimental data. Other sequence-derived features such as k-mer contents calculated in this study and the recently reported quantum mechanical properties of genomic sequences [103] could also be incorporated. These models can be generated through molecular simulations, as demon-

Name	Source	Access
Datasets		
Homo sapiens GRCh37/hg19	Ensembl Release 73 (September 2013) [107]	GCA_000001405.13 https://ftp.ensembl.org/pub/release-73/fasta/homo_sapiens/dna
Homo sapiens GRCh38/hg38	Ensembl Release 108 (October 2022) [108]	GCA_000001405.28 https://ftp.ensembl.org/pub/release-108/fasta/homo_sapiens/dna
Drosophila melanogaster dm6	Adams *et al.* [109], dos Santos *et al.* [110]Downloaded from NCBI assembly GCA_000001215.4.	https://www.ncbi.nlm.nih.gov/assembly/GCF_000001215.4
Hi-C contact data and TAD boundaries from 21 *H. sapiens* cell lines and tissues (hg19)	Schmitt *et al.* [27]	GEO:GSE87112
TAD boundaries from 2 *D. melanogaster* cell lines (dm6)	Chathoth *et al.* [111]	GEO:GSE122603
Phylo-HMRF evolutionary states	Yang *et al.* [62]	GEO:GSE128800
Bed files of ∼300 features (e.g. chromatin domains, transcription factors, histone modifications, sequence motifs)	Mainly ENCODE [112, 113] and NIH Roadmap Epigenomics Project [114]. Refer to Supplementary File S1: Supplementary Table S5 for other sources.	Supplementary File S1: Supplementary Table S5
Genes/gene predictions and repeat annotations (Feb. 2009 GRCh37/hg19)	Downloaded from UCSC Table Browser [115].	https://genome.ucsc.edu/cgi-bin/hgTables
Two baseline expression datasets	GTExConsortium [47]Fagerberg *et al.* [49]Downloaded from EMBL-EBI Expression Atlas [48].	E-MTAB-5214 E-MTAB-1733 https://www.ebi.ac.uk/gxa/home
Replication timing data (hg19)	ReplicationDomain [50]	https://www.replicationdomain.org
Somatic cancer SNV data (hg19)	ICGC Data Portal (Release 28, 27 March 2019) [51]	https://dcc.icgc.org
Triplet-based free energy parameters	Tulpan *et al.* [53]	https://doi.org/10.1186/1471-2105-11-105
H1-hESC Hi-C contact data (.hic, hg38)	4DN portal [55]Akgol Oksuz *et al.* [116] 03/08/2025 17:53:00	https://data.4dnucleome.org/files-processed/4DNFIIMZB6Y9
GM12878 Hi-C contact data (.hic, hg38)	4DN portal [55]Krietenstein *et al.* [117]	https://data.4dnucleome.org/files-processed/4DNFI1UEG1HD
GM12878 Micro-C contact data (.hic, hg38)	4DN portal [55]Krietenstein *et al.* [117]	https://data.4dnucleome.org/files-processed/4DNFI2TK7L2F/
Hi-C contact data (.hic) from 2 *D. melanogaster* cell lines (dm6)	Chathoth *et al.* [111]	Email correspondence with authors
Software		
All original computer code	This paper	https://github.com/SahakyanLab/GenomicContactDynamics
R4RNA (1.18.0, R package)	Tsybulskyi *et al.* [40]	https://www.bioconductor.org/packages/release/bioc/html/R4RNA.html
visNetwork (2.0.9, R package)	Almende *et al.* [41]	https://cran.r-project.org/web/packages/visNetwork/index.html
regioneR (1.22.0, R package)	Gel *et al.* [45]	https://github.com/bernatgel/regioneR
HOMER (4.10)	Heinz *et al.* [46]	https://homer.ucsd.edu/homer
DAVID (6.8)	Huang *et al.* [118, 119]	https://david.ncifcrf.gov
edlib (1.2.6)	Šošić and Šikić [52]	https://github.com/Martinsos/edlib
ROptimus (3.0.0, R package)	Johnson *et al.* [54]	https://cran.r-project.org/web/packages/ROptimus/index.html
strawr (0.0.9, R package)	Cherniavsky Durand *et al.* [56]	https://cran.rstudio.com/web/packages/strawr/index.html

strated by our preliminary analysis (Supplementary File S1: Supplementary Figs S34 and S35), and evaluated against experimental data not only from different cell types, but also from other layers, e.g. across cell cycle and species. Quantifying the similarity of regions using the complementarity measures, which goes beyond defining similarity based on a particular feature like repeat site content, could enable dissection of how the degree and pattern of similarity or homology between regions could influence genome organization, which are relevant based on *in silico* [104], *in vitro* [105], and *in vivo* [97] mechanistic studies of DNA duplex association [99], and phase separation mechanisms [106]. Finally, experimental investigations are crucial, particularly, to validate and disentangle the potential direct and/or indirect contributions of genomic sequence to the 3D organization. *In vitro* studies have already shown the preferential association of identical DNA duplexes and similar experimental setups could be leveraged to determine whether the degree of complementarity between sequences influences these associations using oligonucleotide designs that vary in terms of total complementarity, patterns of complementarity (e.g. periodically spaced complementary tracts of varying lengths), and GC content, given the differing propensities of GC-rich and AT-rich duplexes to associate. Bound proteins as well as complementary RNAs, already shown to be important drivers of associations between genomic regions, could then be added to the system to test how they alter the patterns and degree of associations.

In summary, this study contributes to the understanding of the relationship between genome sequence and structure by implicating a single parameter, sequence complementarity, as

a core factor contributing to the formation of genomic contacts. Along with other works that aimed to understand the encoding of the structure into the sequence (reviewed in [14–17]), this study shows that the complementarity between different parts of the genome may play a role in this encoding, and suggests that organizing mechanisms, such as phase separation, are not agnostic to the underlying DNA sequence. Consequently, the DNA could be involved in both direct and indirect manner, and we encourage experimental validation that will help delineate the contribution of sequence to genome organization, in conjunction with earlier-characterized protein and epigenetic determinants.

Acknowledgements

L.T. is grateful to the Jardine Foundation for supporting her DPhil studies. J.A. is thankful to the MRC for funding his DPhil through a WIMM Studentship. The Sahakyan laboratory including this project has been supported by the UK Medical Research Council (MRC) for the MRC Strategic Alliance Funding (MC_UU_12025). UCSF Chimera, used here for 3D genome model visualization, was developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311. The cell-MDCK icon by DBCLS (greyscaled in this paper) in Fig. 1A of this study is licensed under CC-BY 4.0 Unported and was retrieved from bioicons (https://bioicons.com).

Author contributions: Liezel Tamon (Conceptualization [equal], Formal analysis [equal], Methodology [equal], Software [equal], Validation [equal], Visualization [equal], Writing–original draft [equal]), Zahra Fahmi (Formal analysis [supporting], Methodology [supporting], Writing–review & editing [supporting]), James Ashford (Software [supporting], Writing–review & editing [equal]), Rosana Collepardo-Guevara (Funding acquisition [supporting], Methodology [supporting], Resources [supporting], Supervision [supporting]), and Aleksandr B. Sahakyan (Conceptualization [equal], Funding acquisition [equal], Methodology [equal], Resources [equal], Writing-original draft [equal], Supervision [equal]).

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

This research has been supported by the UK Medical Research Council (MRC) through the MRC Strategic Alliance Funding (MC_UU_12025). Funding to pay the Open Access publication charges for this article was provided by Oxford University.

Data availability

All code used in this study is publicly available *via* the GitHub repository: https://github.com/SahakyanLab/GenomicContactDynamics, and is also archived on Zenodo: https://zenodo.org/records/15068445. This work is purely computational and relied on publicly available datasets and software, as detailed below.

References

1. Misteli T. The self-organizing genome: principles of genome architecture and function. *Cell* 2020;183:28–45. https://doi.org/10.1016/j.cell.2020.09.014
2. Dixon JR, Selvaraj S, Yue F *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;485:376–80. https://doi.org/10.1038/nature11082
3. Nora EP, Lajoie BR, Schulz EG *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 2012;485:381–5. https://doi.org/10.1038/nature11049
4. Sexton T, Yaffe E, Kenigsberg E *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* 2012;148:458–72. https://doi.org/10.1016/j.cell.2012.01.010
5. Lieberman-Aiden E, Berkum NL, Williams L *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the Human genome. *Science* 2009;326:289–93. https://doi.org/10.1126/science.1181369
6. Crosetto N ,Bienko M. Radial organization in the mammalian nucleus. *Front Genet* 2020;11:33. https://doi.org/10.3389/fgene.2020.00033
7. Wang Q Sawyer IA, Sung MH, Cajal bodies are linked to genome conformation. *Nat Commun* 2016;7:10966. https://doi.org/10.1038/ncomms10966
8. Chen H Zhang Y, Wang Y, Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *J Cell Biol* 2018;217:4025–4048. https://doi.org/10.1083/jcb.201807108
9. Quinodoz SA Ollikainen N, Tabak B, Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* 2018;174:744–757. https://doi.org/10.1016/j.cell.2018.05.024
10. Rosa A, Zimmer C. Chapter nine - computational models of large-scale genome architecture. *Int Rev Cell Mol Biol* 2014;307:275–349. https://doi.org/10.1016/B978-0-12-800046-5.00009-6
11. Sefer E, Kingsford C. Semi-nonparametric modeling of topological domain formation from epigenetic data. *Algorithms Mol Biol* 2019;14:4. https://doi.org/10.1186/s13015-019-0142-y
12. Sima J, Chakraborty A, Dileep V *et al.* Identifying cis elements for spatiotemporal control of mammalian DNA replication. *Cell* 2019;176:816–830.e18. https://doi.org/10.1016/j.cell.2018.11.036
13. Hsieh THS, Cattoglio C, Slobodyanyuk E *et al.* Resolving the 3D landscape of transcription-linked mammalian chromatin folding. *Mol Cell* 2020;78:539–553.e8. https://doi.org/10.1016/j.molcel.2020.03.002
14. Quan H, Yang Y, Liu S *et al.* Chromatin structure changes during various processes from a DNA sequence view. *Curr Opin Struct Biol* 2020;62:1–8. https://doi.org/10.1016/j.sbi.2019.10.010
15. Mondal M, Yang L, Cai Z *et al.* A perspective on the molecular simulation of DNA from structural and functional aspects. *Chem Sci* 2021;12:5390–409. https://doi.org/10.1039/D0SC05329E
16. Bernardi G. The "genomic code": DNA pervasively moulds chromatin structures leaving no room for "junk". *Life* 2021;11:342. https://doi.org/10.3390/life11040342
17. King JT, Shakya A. Phase separation of DNA: from past to present. *Biophys J* 2021;120:1139–49. https://doi.org/10.1016/j.bpj.2021.01.033
18. van Steensel B, Dekker J. Genomics tools for unraveling chromosome architecture. *Nat Biotechnol* 2010;28:1089–95. https://doi.org/10.1038/nbt.1680

19. Kempfer R, Pombo A. Methods for mapping 3D chromosome architecture. *Nat Rev Genet* 2020;21:207–26. https://doi.org/10.1038/s41576-019-0195-2
20. Jerkovic´ I, Cavalli G. Understanding 3D genome organization by multidisciplinary methods. *Nat Rev Mol Cell Biol* 2021;22:511–28. https://doi.org/10.1038/s41580-021-00362-w
21. Sood V, Misteli T. The stochastic nature of genome organization and function. *Curr Opin Genet Dev* 2022;72:45–52. https://doi.org/10.1016/j.gde.2021.10.004
22. Schwarzer W, Abdennur N, Goloborodko A *et al.* Two independent modes of chromatin organization revealed by cohesin removal. *Nature* 2017;551:51–6. https://doi.org/10.1038/nature24281
23. Finn EH, Misteli T. Molecular basis and biological function of variability in spatial genome organization. *Science* 2019; 365: eaaw9498.https://doi.org/10.1126/science.aaw9498
24. Wachsmuth M, Knoch TA, Rippe K. Dynamic properties of independent chromatin domains measured by correlation spectroscopy in living cells. *Epigenetics Chromatin* 2016;9:57. https://doi.org/10.1186/s13072-016-0093-1
25. Finn EH, Pegoraro G, Brandão HB *et al.* Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell* 2019;176:1502–1515.e10. https://doi.org/10.1016/j.cell.2019.01.020
26. Waszak SM, Delaneau O, Gschwind AR *et al.* Population variation and genetic control of modular chromatin architecture in humans. *Cell* 2015;162:1039–50. https://doi.org/10.1016/j.cell.2015.08.001
27. Schmitt AD, Hu M, Jung I *et al.* A compendium of chromatin contact maps reveals spatially active regions in the Human genome. *Cell Rep* 2016;17:2042–59. https://doi.org/10.1016/j.celrep.2016.10.061
28. Naumova N, Imakaev M, Fudenberg G *et al.* Organization of the mitotic chromosome. *Science* 2013;342:948–53. https://doi.org/10.1126/science.1236083
29. Nagano T, Lubling Y, Várnai C *et al.* Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* 2017;547:61–7. https://doi.org/10.1038/nature23001
30. Boettiger AN, Bintu B, Moffitt JR *et al.* Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature* 2016;529:418–22. https://doi.org/10.1038/nature16496
31. Torre-Ubieta L, Stein JL, Won H *et al.* The dynamic landscape of open chromatin during Human cortical neurogenesis. *Cell* 2018;172:289–304.e18. https://doi.org/10.1016/j.cell.2017.12.014
32. Paulsen J, Liyakat Ali TM, Nekrasov M *et al.* Long-range interactions between topologically associating domains shape the four-dimensional genome during differentiation. *Nat Genet* 2019;51:835–43. https://doi.org/10.1038/s41588-019-0392-0
33. Nagano T, Lubling Y, Stevens TJ *et al.* Single-cell hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 2013;502:59–64. https://doi.org/10.1038/nature12593
34. Bintu B, Mateo LJ, Su JH *et al.* Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* 2018;362: eaau1783. https://doi.org/10.1126/science.aau1783
35. Su JH, Zheng P, Kinrot SS *et al.* Genome-scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell* 2020;182:1641–1659. https://doi.org/10.1016/j.cell.2020.07.032
36. Tamon L, Ashford J, Nicholls M *et al.* The emerging sequence grammar of 3D genome organisation. https://doi.org/10.5281/zenodo.16737235
37. R Core Team. R: A language and environment for statistical computing. 2021. https://www.R-project.org (7 August 2025, date last accessed).
38. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 1995;57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
39. Hu M, Deng K, Selvaraj S *et al.* HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* 2012;28:3131–3. https://doi.org/10.1093/bioinformatics/bts570
40. Tsybulskyi V, Mounir M, Meyer IM. R-chie: a web server and R package for visualizing cis and trans RNA–RNA, RNA–DNA and DNA–DNA interactions. *Nucleic Acids Res* 2020;48:e105. https://doi.org/10.1093/nar/gkaa708
41. Almende BV, Thieurmel B, Robert T. visNetwork: Network Visualization Using "Vis.Js" Library. 2019. https://CRAN.R-project.org/package=visNetwork (12 July 2022, date last accessed).
42. Bauer M, Vidal E, Zorita E *et al.* Chromosome compartments on the inactive X guide TAD formation independently of transcription during X-reactivation. *Nat Commun* 2021;12:3499. https://doi.org/10.1038/s41467-021-23610-1
43. Costantini M, Clay O, Auletta F *et al.* An isochore map of human chromosomes. *Genome Res* 2006;16:536–41. https://doi.org/10.1101/gr.4910606
44. Jabbari K, Bernardi G. An isochore framework underlies chromatin architecture. *PLoS One* 2017;12:e0168023. https://doi.org/10.1371/journal.pone.0168023
45. Gel B, Díez-Villanueva A, Serra E *et al.* RegioneR: an R/bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* 2016:32:289. https://doi.org/10.1093/bioinformatics/btv562
46. Heinz S, Benner C, Spann N *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010;38:576–89. https://doi.org/10.1016/j.molcel.2010.05.004
47. The GTEX consortium, Ardlie KG, Deluca DS *et al.* The genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015;348:648–60. https://doi.org/10.1126/science.1262110
48. Papatheodorou I, Moreno P, Manning J *et al.* Expression Atlas update: from tissues to single cells. *Nucleic Acids Res* 2019;48:gkz947. https://doi.org/10.1093/nar/gkz947
49. Fagerberg L, Hallström BM, Oksvold P *et al.* Analysis of the Human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* 2014;13:397–406. https://doi.org/10.1074/mcp.M113.035600
50. Weddington N, Stuy A, Hiratani I *et al.* ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data. *BMC Bioinformatics* 2008;9:530. https://doi.org/10.1186/1471-2105-9-530
51. Zhang J, Bajari R, Andric D *et al.* The International Cancer Genome Consortium Data Portal. *Nat Biotechnol* 2019;37:367–9. https://doi.org/10.1038/s41587-019-0055-9
52. Šošić M, Šikić M. Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. *Bioinformatics* 2017;33:1394–5. https://doi.org/10.1093/bioinformatics/btw753
53. Tulpan D, Andronescu M, Leger S. Free energy estimation of short DNA duplex hybridizations. *BMC Bioinformatics* 2010;11:105. https://doi.org/10.1186/1471-2105-11-105
54. Johnson NAG, Tamon L, Liu X *et al.* ROptimus: a parallel general-purpose adaptive optimization engine. *Bioinformatics* 2023;39:btad292. https://doi.org/10.1093/bioinformatics/btad292
55. Reiff SB, Schroeder AJ, Kırlı K *et al.* The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. *Nat Commun* 2022;13:2365. https://doi.org/10.1038/s41467-022-29697-4
56. Cherniavsky Durand N, Saad Shamim M. strawr: Fast Implementation of Reading/Dump for .Hic Files. 2024. https://CRAN.R-project.org/package=strawr (1 July 2023, date last accessed).
57. Zufferey M, Tavernari D, Oricchio E *et al.* Comparison of computational methods for the identification of topologically

associating domains. *Genome Biol* 2018;19:217. https://doi.org/10.1186/s13059-018-1596-9
58. Sefer E. A comparison of topologically associating domain callers over mammals at high resolution. *BMC Bioinformatics* 2022;23:127. https://doi.org/10.1186/s12859-022-04674-2
59. Fudenberg G, Imakaev M, Lu C *et al.* Formation of chromosomal domains by loop extrusion. *Cell Rep* 2016;15:2038–49. https://doi.org/10.1016/j.celrep.2016.04.085
60. Ganji M, Shaltiel IA, Bisht S *et al.* Real-time imaging of DNA loop extrusion by condensin. *Science* 2018;360:102–5. https://doi.org/10.1126/science.aar7831
61. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* 2015;72:65–75. https://doi.org/10.1016/j.ymeth.2014.10.031
62. Yang Y, Zhang Y, Ren B *et al.* Comparing 3D genome organization in multiple species using phylo-HMRF. *Cell Syst* 2019;8:494–505.e14. https://doi.org/10.1016/j.cels.2019.05.011
63. Liu S, Zhang L, Quan H *et al.* From 1D sequence to 3D chromatin dynamics and cellular functions: a phase separation perspective. *Nucleic Acids Res* 2018;46:9367–83. https://doi.org/10.1093/nar/gky633
64. Nash AJ, Lenhard B. A novel measure of non-coding genome conservation identifies genomic regulatory blocks within primates. *Bioinformatics* 2019;35:2354–61. https://doi.org/10.1093/bioinformatics/bty1014
65. Guelen L, Pagie L, Brasset E *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 2008;453:948–51. https://doi.org/10.1038/nature06947
66. Piovesan A, Pelleri MC, Antonaros F *et al.* On the length, weight and GC content of the human genome. *BMC Res Notes* 2019;12:106. https://doi.org/10.1186/s13104-019-4137-z
67. Xiong K, Ma J. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nat Commun* 2019;10:5069. https://doi.org/10.1038/s41467-019-12954-4
68. Sahakyan AB, Balasubramanian S. Long genes and genes with multiple splice variants are enriched in pathways linked to cancer and other multigenic diseases. *BMC Genomics* 2016;17:225. https://doi.org/10.1186/s12864-016-2582-9
69. Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* 2012;488:504–7. https://doi.org/10.1038/nature11273
70. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* 2015;521:81–4. https://doi.org/10.1038/nature14173
71. Sabarinathan R, Mularoni L, Deu-Pons J *et al.* Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* 2016;532:264–7. https://doi.org/10.1038/nature17661
72. Frigola J, Sabarinathan R, Mularoni L *et al.* Reduced mutation rate in exons due to differential mismatch repair. *Nat Genet* 2017;49:1684–92. https://doi.org/10.1038/ng.3991
73. Degasperi A, Amarante TD, Czarnecki J *et al.* A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies. *Nat Cancer* 2020;1:249–63. https://doi.org/10.1038/s43018-020-0027-5
74. Kornyshev AA, Leikin S. Sequence recognition in the pairing of DNA duplexes. *Phys Rev Lett* 2001;86:3666–9. https://doi.org/10.1103/PhysRevLett.86.3666
75. Inoue S, Sugiyama S, Travers AA *et al.* Self-assembly of double-stranded DNA molecules at nanomolar concentrations. *Biochemistry* 2007;46:164–71. https://doi.org/10.1021/bi061539y
76. Baldwin GS, Brooks NJ, Robson RE *et al.* DNA double helices recognize mutual sequence homology in a protein free environment. *J Phys Chem B* 2008;112:1060–4. https://doi.org/10.1021/jp7112297
77. Danilowicz C, Lee C, Kim K *et al.* Single molecule detection of direct, homologous, DNA/DNA pairing. *Proc Nat Acad Sci USA* 2009;106:19824–9. https://doi.org/10.1073/pnas.0911214106
78. Ohyama T. New aspects of magnesium function: a key regulator in nucleosome self-assembly, chromatin folding and phase separation. *Int J Mol Sci* 2019;20:4232. https://doi.org/10.3390/ijms20174232
79. Nishikawa J, Ohyama T. Selective association between nucleosomes with identical DNA sequences. *Nucleic Acids Res* 2013;41:1544–54. https://doi.org/10.1093/nar/gks1269
80. Kasamatsu H, Robberson DL, Vinograd J. A novel closed-circular mitochondrial DNA with properties of a replicating intermediate. *Proc Natl Acad Sci USA* 1971;68:2252–7. https://doi.org/10.1073/pnas.68.9.2252
81. Thomas M, White RL, Davis RW. Hybridization of RNA to double-stranded DNA: formation of R-loops. *Proc Natl Acad Sci USA* 1976;73:2294–8. https://doi.org/10.1073/pnas.73.7.2294
82. Kim A, Wang GG. R-loop and its functions at the regulatory interfaces between transcription and (epi)genome. *Biochim Biophys Acta* 2021;1864:194750. https://doi.org/10.1016/j.bbagrm.2021.194750
83. Felsenfeld G, Rich A. Studies on the formation of two- and three-stranded polyribonucleotides. *Biochim Biophys Acta* 1957;26:457–68. https://doi.org/10.1016/0006-3002(57)90091-4
84. Buske FA, Mattick JS, Bailey TL. Potential *in vivo* roles of nucleic acid triple-helices. *RNA Biol* 2011;8:427–39. https://doi.org/10.4161/rna.8.3.14999
85. Rothemund PWK. Folding DNA to create nanoscale shapes and patterns. *Nature* 2006;440:297–302. https://doi.org/10.1038/nature04586
86. Barzel A, Kupiec M. Finding a match: how do homologous sequences get together for recombination? *Nat Rev Genet* 2008;9:27–37. https://doi.org/10.1038/nrg2224
87. Fudenberg G, Mirny LA. Higher-order chromatin structure: bridging physics and biology. *Curr Opin Genet Dev* 2012;22:115–24. https://doi.org/10.1016/j.gde.2012.01.006
88. Tang SJ. Potential role of phase separation of repetitive DNA in chromosomal organization. *Genes* 2017;8:279. https://doi.org/10.3390/genes8100279
89. Cournac A, Koszul R, Mozziconacci J. The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. *Nucleic Acids Res* 2016;44:245–55. https://doi.org/10.1093/nar/gkv1292
90. Nikumbh S, Pfeifer N. Genetic sequence-based prediction of long-range chromatin interactions suggests a potential role of short tandem repeat sequences in genome organization. *BMC Bioinformatics* 2017;18:218. https://doi.org/10.1186/s12859-017-1624-x
91. Winter DJ, Ganley ARD, Young CA *et al.* Repeat elements organise 3D genome structure and mediate transcription in the filamentous fungus Epichloë festucae. *PLoS Genet* 2018;14:e1007467. https://doi.org/10.1371/journal.pgen.1007467
92. Lu JY, Chang L, Li T *et al.* Homotypic clustering of L1 and B1/Alu repeats compartmentalizes the 3D genome. *Cell Res* 2021;31:613–30. https://doi.org/10.1038/s41422-020-00466-6
93. Yang M, Ma J. UNADON: transformer-based model to predict genome-wide chromosome spatial position. *Bioinformatics* 2023;39:i553–62. https://doi.org/10.1093/bioinformatics/btad246
94. Haws SA, Simandi Z, Barnett RJ. Phillips-Cremins JE. 3D genome, on repeat: higher-order folding principles of the heterochromatinized repetitive genome. *Cell* 2022;185:2690–707. https://doi.org/10.1016/j.cell.2022.06.052
95. Zhou J. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat Genet* 2022;54:725–34. https://doi.org/10.1038/s41588-022-01065-4

96. Yoo J, Kim H, Aksimentiev A *et al.* Direct evidence for sequence-dependent attraction between double-stranded DNA controlled by methylation. *Nat Commun* 2016;7:11045. https://doi.org/10.1038/ncomms11045
97. Gladyshev E, Kleckner N. Direct recognition of homology between double helices of DNA in Neurospora crassa. *Nat Commun* 2014;5:3509. https://doi.org/10.1038/ncomms4509
98. Ding DQ, Okamasa K, Katou Y *et al.* Chromosome-associated RNA–protein complexes promote pairing of homologous chromosomes during meiosis in schizosaccharomyces pombe. *Nat Commun* 2019;10:5598. https://doi.org/10.1038/s41467-019-13609-0
99. Mazur AK, Nguyen TS, Gladyshev E. Direct homologous dsDNA–dsDNA pairing: how, where, and why? *J Mol Biol* 2020;432:737–44. https://doi.org/10.1016/j.jmb.2019.11.005
100. Gladyshev E, Kleckner N. DNA sequence homology induces cytosine-to-thymine mutation by a heterochromatin-related pathway in Neurospora. *Nat Genet* 2017;49:887–94. https://doi.org/10.1038/ng.3857
101. Bender J. Cytosine methylation of repeated sequences in eukaryotes: the role of DNA pairing. *Trends Biochem Sci* 1998;23:252–6. https://doi.org/10.1016/S0968-0004(98)01225-0
102. King IF, Yandava CN, Mabb AM *et al.* Topoisomerases facilitate transcription of long genes linked to autism. *Nature* 2013;501:58–62. https://doi.org/10.1038/nature12504
103. Masuda K, Abdullah AA, Sahakyan AB. Quantum mechanical electronic and geometric parameters for DNA k-mers as features for machine learning. *Sci Data* 2023;11:911. https://doi.org/10.1101/2023.01.25.525597
104. Mazur AK. Homologous pairing between long DNA double helices. *Phys Rev Lett* 2016;116:158101. https://doi.org/10.1103/PhysRevLett.116.158101
105. Wang X, Zhang X, Mao C *et al.* Double-stranded DNA homology produces a physical signature. *Proc Natl Acad Sci USA* 2010;107:12547–52. https://doi.org/10.1073/pnas.1000105107
106. Erdel F, Rippe K. Formation of chromatin subcompartments by phase separation. *Biophys J* 2018;114:2262–70. https://doi.org/10.1016/j.bpj.2018.03.011
107. Flicek P, Amode MR, Barrell D *et al.* Ensembl 2014. *Nucleic Acids Res* 2014;42:D749–D755. https://doi.org/10.1093/nar/gkt1196
108. Cunningham F, Allen JE, Allen J *et al.* Ensembl 2022. *Nucleic Acids Res* 2022;50:D988–D995. https://doi.org/10.1093/nar/gkab1049
109. Adams MD, Celniker SE, Holt RA *et al.* The genome sequence of *Drosophila melanogaster*. *Science* 2000;287:2185–95. https://doi.org/10.1126/science.287.5461.2185
110. dos Santos G, Schroeder AJ, Goodman JL *et al.* FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res* 2015;43:D690–D697. https://doi.org/10.1093/nar/gku1099
111. Chathoth KT, Zabet NR. Chromatin architecture reorganization during neuronal cell differentiation in Drosophila genome. *Genome Res* 2019;29:613–25. https://doi.org/10.1101/gr.246710.118
112. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57–74. https://doi.org/10.1038/nature11247
113. Davis CA, Hitz BC, Sloan CA *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* 2018;46:D794–D801. https://doi.org/10.1093/nar/gkx1081
114. Kundaje A, Meuleman W, Ernst J *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518:317–30. https://doi.org/10.1038/nature14248
115. Karolchik D, Hinrichs AS, Furey TS *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 2004;32:D493–D496. https://doi.org/10.1093/nar/gkh103
116. Akgol Oksuz B, Yang L, Abraham S *et al.* Systematic evaluation of chromosome conformation capture assays. *Nat Methods* 2021;18:1046–55. https://doi.org/10.1038/s41592-021-01248-7
117. Krietenstein N, Abraham S, Venev SV *et al.* Ultrastructural details of mammalian chromosome architecture. *Mol Cell* 2020;78:554–565. https://doi.org/10.1016/j.molcel.2020.03.003
118. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;4:44–57. https://doi.org/10.1038/nprot.2008.211
119. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009;37:1–13. https://doi.org/10.1093/nar/gkn923
120. Ou HD Phan S, Deerinck TJ, ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science* 2017;357:eaag0025. https://doi.org/10.1126/science.aag0025

Received: May 1, 2025. **Revised:** September 17, 2025. **Accepted:** September 30, 2025