

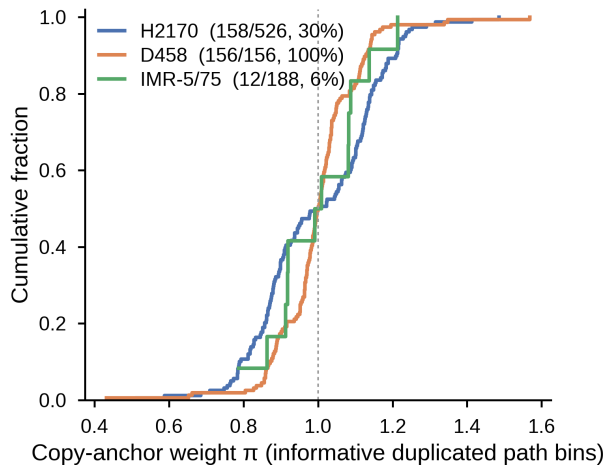
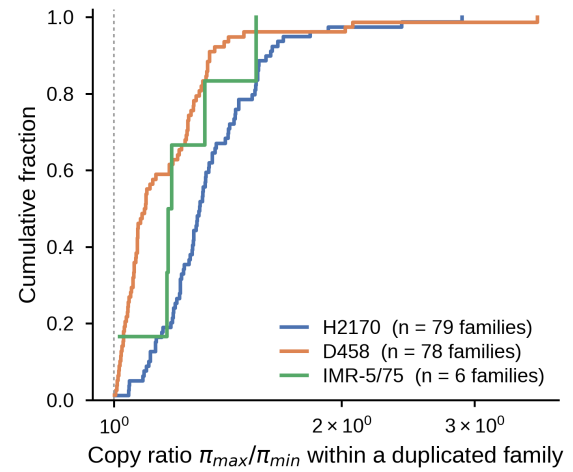
Supplementary Files

Copy aware deconvolution of Hi-C interactions in extrachromosomal DNA using CADET

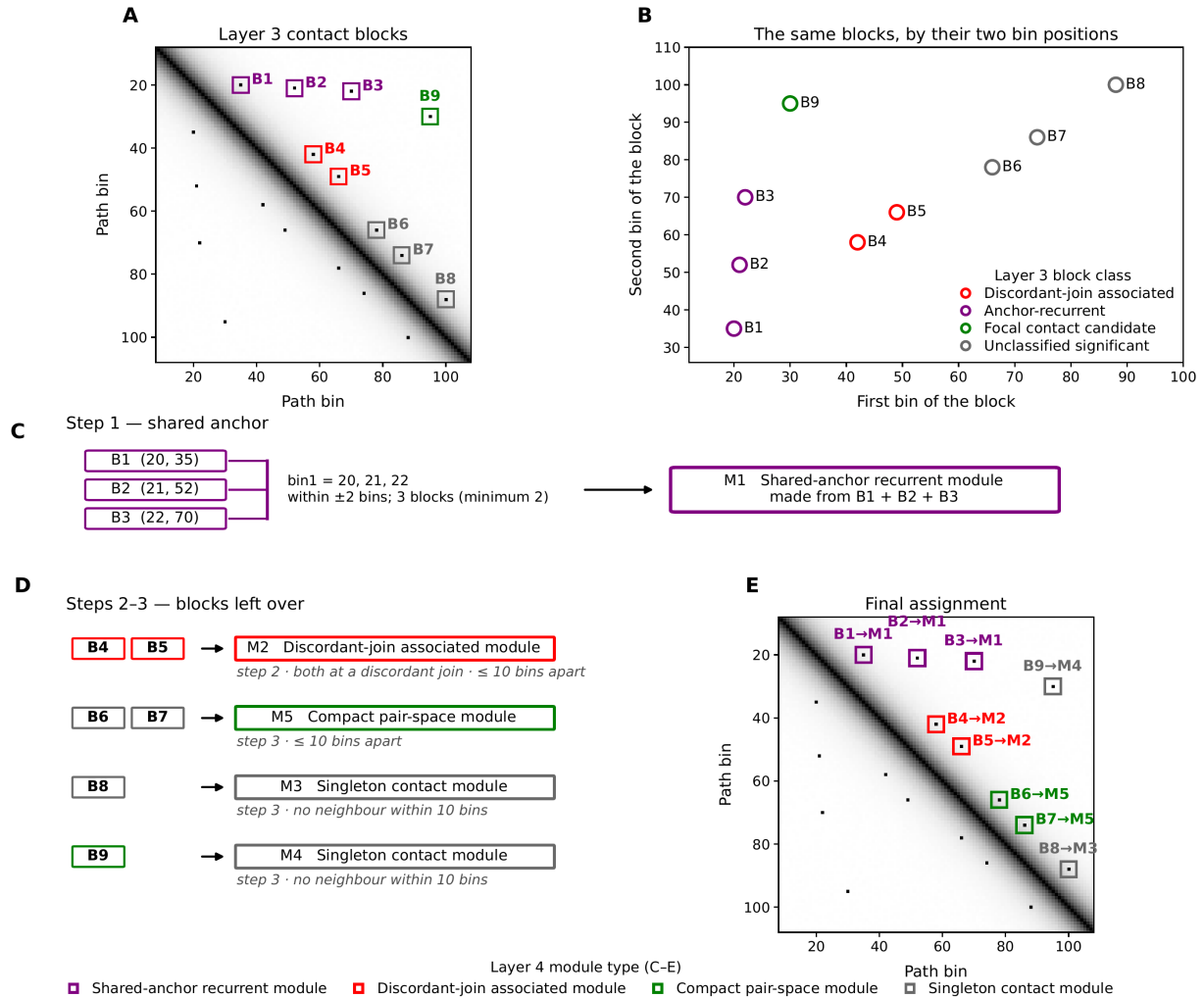
Biswanath Chowdhury

Contents

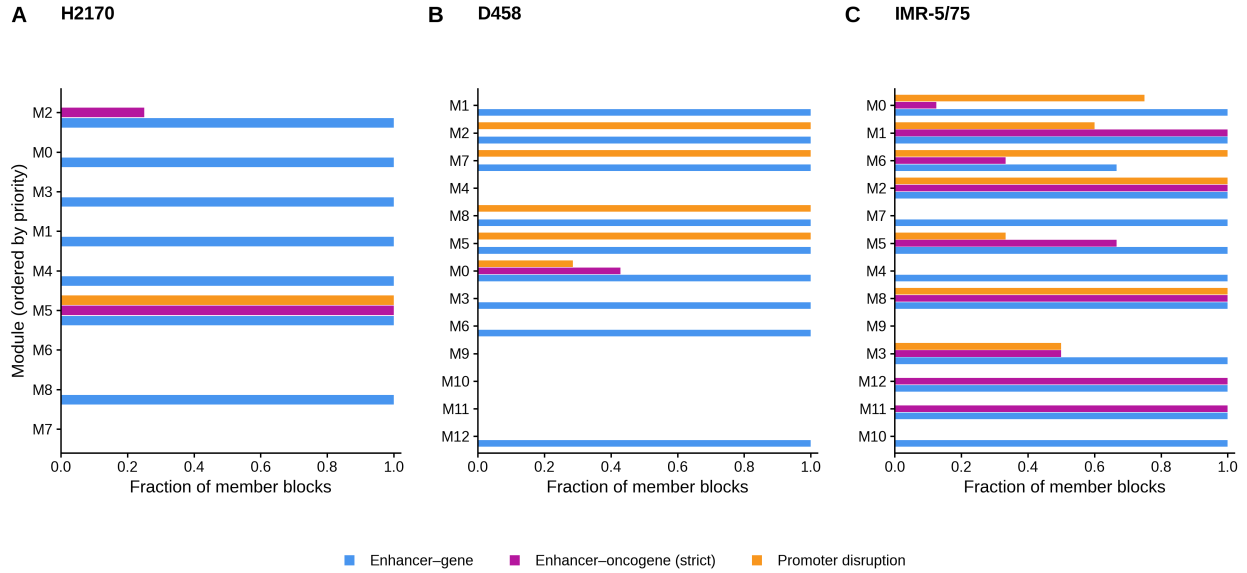
- Supplementary Figure S1
- Supplementary Figure S2
- Supplementary Figure S3
- Supplementary Figure S4
- Supplementary Figure S5
- Supplementary Figure S6
- Supplementary Figure S7
- Supplementary Figure S8
- Supplementary Table S1 (separate file)
Supplementary_Table_S1_family_enrichment_alpha.xlsx
- Supplementary Table S2 (separate file)
Supplementary_Table_S2_block_regulatory_annotation.xlsx
- Supplementary Table S3 (separate file)
Supplementary_Table_S3_CADET_ec3D_relevant_results.xlsx
- Supplementary Table S4 (separate file)
Supplementary_Table_S4_layer8_target_recurrence.xlsx
- Supplementary Note 1
- Supplementary Note 2

A π distribution**B Imbalance between the two copies of a family**

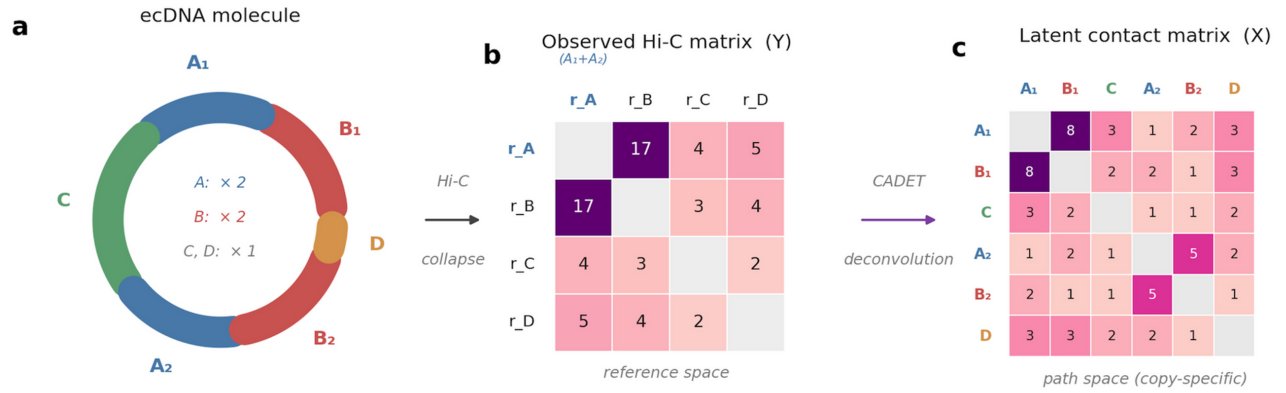
Supplementary Figure S1. Copy-anchor π across amplified genomes. **A**, ECDF of π over informative duplicated path bins where a path bin whose π deviates from the neutral value of 1. H2170: 158 of 526 path bins, D458: 156 of 156, IMR-5/75: 12 of 188. In the remaining duplicated families, the local evidence did not prefer either copy and both members kept $\pi = 1$. **B**, the same families expressed once each as the copy ratio $\pi_{\max}/\pi_{\min}$ within a duplicated family, on a log axis. Since π is normalized to mean 1 within a family and every family here has exactly two copies, the two members always sum to 2. Panel A therefore plots each family twice, symmetric around 1, whereas panel B plots it once.



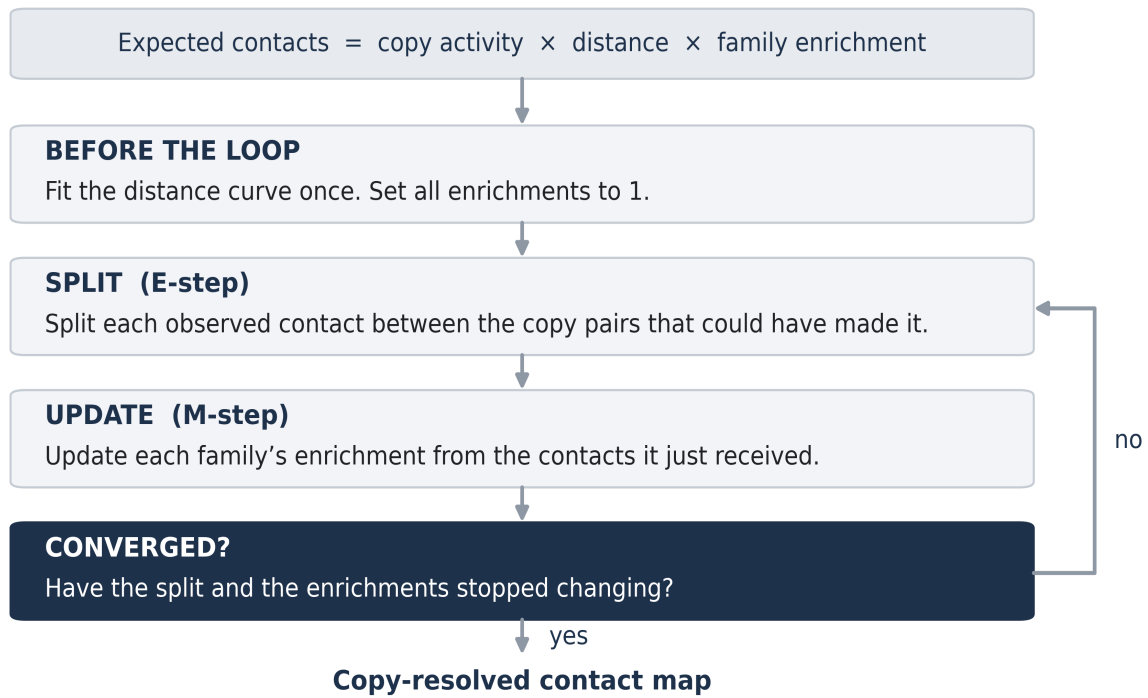
Supplementary Figure S2. How Layer 3 contact blocks become Layer 4 modules. The grouping shown is what the Layer 4 routine `build_modules` returns for these nine blocks, using the same settings as all three datasets: blocks count as sharing an anchor if their anchor bins fall within 2 bins of each other, at least 2 blocks are needed to form a shared anchor, and two blocks count as neighbors if both of their bins are within 10 bins. **A**, nine contact blocks built by Layer 3 on a contact map, colored by the class Layer 3; each box is 5 bins across, the window Layer 3 uses when it merges neighboring focal pixels or candidates into a block. **B**, the same nine blocks placed by their bin positions. **C**, step 1: B1, B2 and B3 all have one end at bin positions 20, 21 or 22, close to count as one event, and three blocks share it; they are not merged into a single region but stay three separate blocks that share same module name. **D**, steps 2 and 3, applied to the rest which step 1 did not take: blocks beside a discordant join are grouped first if they are close to each other, then the others which gives compact pair-space modules where a neighbor exists within the tolerance and singleton modules where none does. The order is important: B4 to B9 are only considered because B1 to B3 have already been assigned. **E**, the result, colored by module type. A block in Layer 3 class and its Layer 4 module type are separate things: B9 is a focal contact candidate that ends up as a singleton module.



Supplementary Figure S3. Layer 6 regulatory logic per chromatin module. **A–C**, for H2170, D458 and IMR-5/75, the fraction of member blocks in each module satisfying the enhancer-gene, enhancer-oncogene and promoter-disruption criteria. Modules are ordered by structural priority, the same order as Figure 4 **G–I**. The criteria are evaluated independently and are not mutually exclusive, so fractions within a module need not sum to one. The enhancer-oncogene category counts only blocks meeting the strict criterion; H2170 modules 0 and 1 meet the permissive criterion but not the strict one, which is why they carry no enhancer-oncogene bar although the text describes them as permissive enhancer-oncogene context modules.

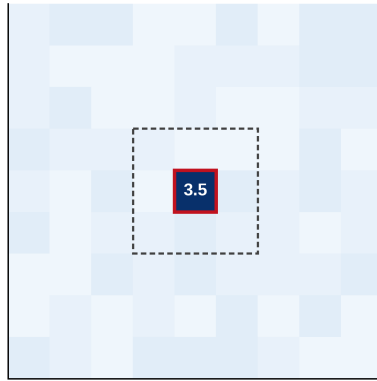


Supplementary Figure S4. Copy ambiguity in an amplified genome. An ecDNA carrying two copies of segments A and B is sequenced by Hi-C. Because both copies of a duplicated bin map to the same reference position, the observed reference-space matrix Y is a collapsed sum over every compatible copy-specific pair. The latent path-space matrix X that separates those contributions.



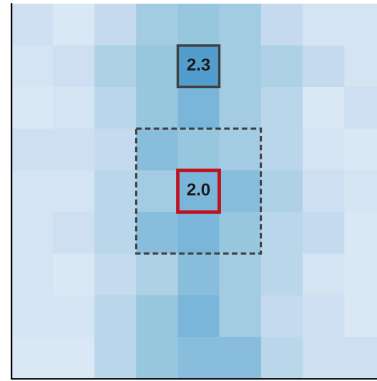
Supplementary Figure S5. The Layer 1 expectation-maximization loop. The expected contacts between two path bins are $\hat{\lambda}_{ij} = \pi_i \pi_j \alpha_{ij} f(d_{ij})$, where f is the distance backbone fitted by PAVA and then fixed, π is the copy-anchor weight and α is the enrichment shared by every pair in the same reference bin family. Each E-step divides every observed reference-space count among its compatible path-bin pairs in proportion to the current rates; each M-step re-estimates α and rescales it to geometric mean 1. The loop stops when the latent matrix and α both correlate above 0.999 with the previous iteration and the log-likelihood change falls below the tolerance.

A Focal contact: kept



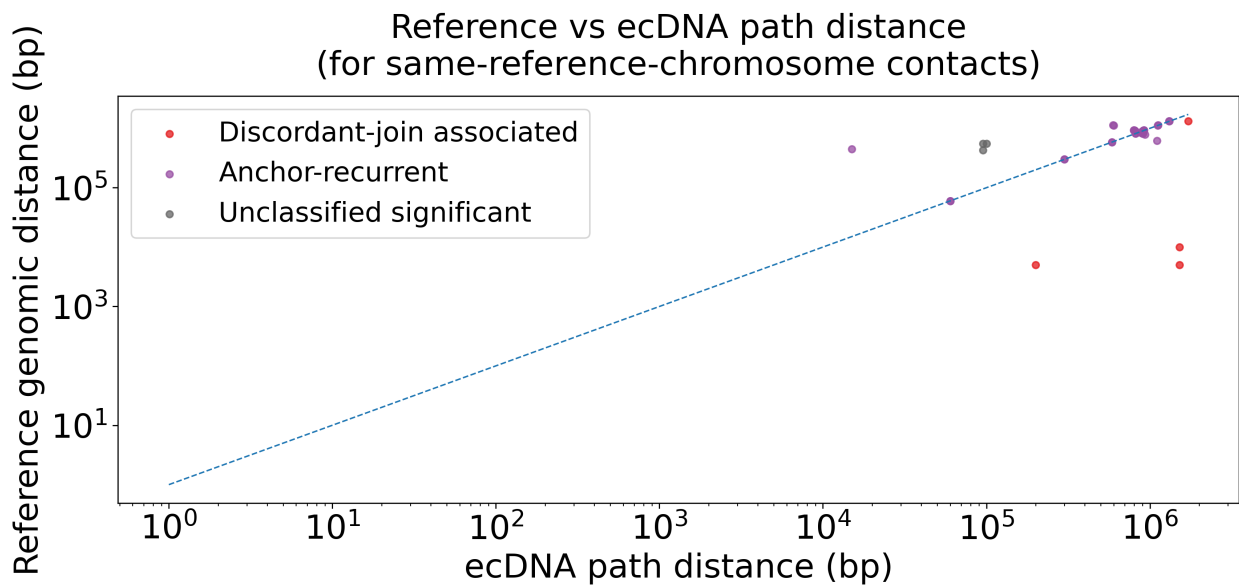
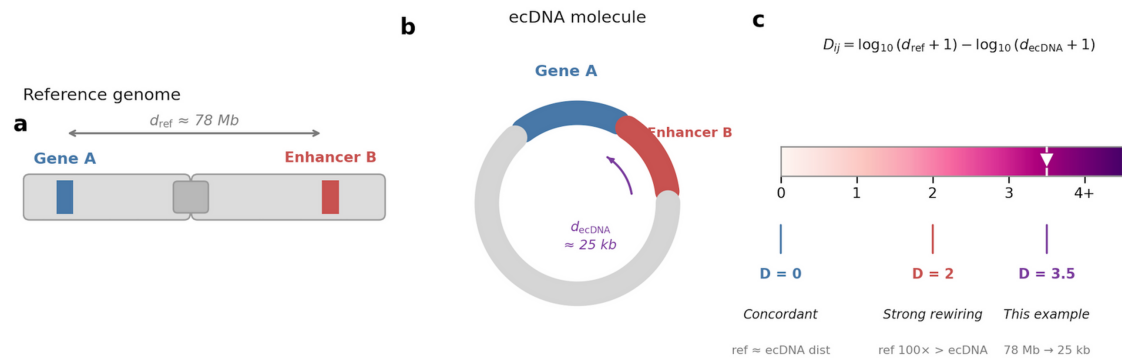
met fold change $3.9\times \geq 1.2\times$
met local peak $3.5 \geq 1.0$

B Broad stripe: dropped



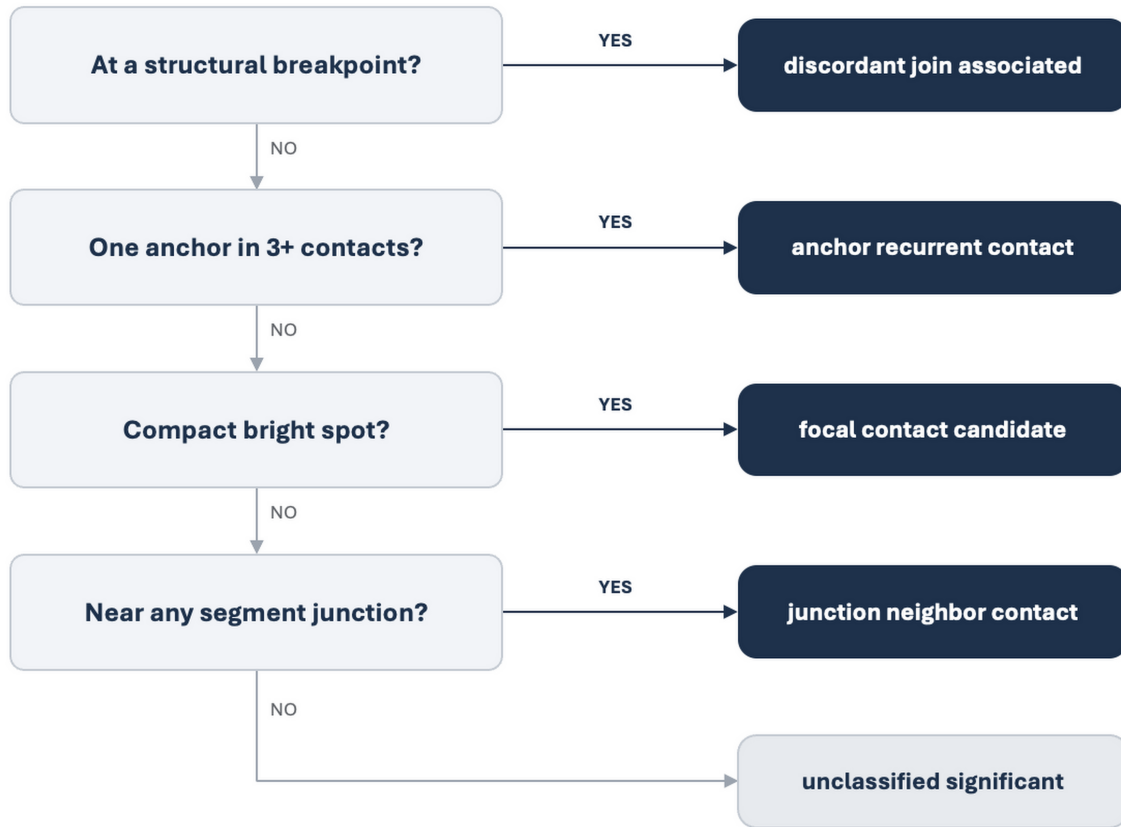
met fold change $1.4\times \geq 1.2\times$
not met local peak $2.0 < 2.3$

Supplementary Figure S6. The local focality test in Layer 2. A candidate pixel is kept only if it passes two tests. The fold-change: its observed/expected value must be at least 1.2 times the mean of the surrounding background. The local-peak: it must be no smaller than any focal pixel in that background. The background is a 9×9 window (outer radius 4) with the central 3×3 block (inner radius 1, dashed) removed. The candidate pixel is outlined in red. **A**, a focal contact clears both tests and is kept. **B**, clears the fold-change test at 1.4 times the background mean, but fails in the local-peak test. A pixel of 2.3, outlined in gray, resides in the background against the 2.0.



Supplementary Figure S7. Distance discordance. Upper panel: a–c, for a contact between two loci, D compares how far apart the loci are in the reference genome with how far apart they are along the reconstructed amplicon. D score near 0 means the rearrangement has not changed their separation; large positive D marks loci brought close together by the rearrangement. **Lower panel**: the same quantity measured in real data: reference genomic distance against ecDNA path distance for every H2170 contact whose two anchors lie on the same reference chromosome (25 blocks), on logarithmic axes, with the dashed line marking equal distances. Points are colored by Layer 3 contact class using the palette of **Figure 4A–C**. Contacts above the diagonal have positive D score, contacts below it has negative D score.

Layer 3: Classification Decision Tree



Supplementary Figure S8. Layer 3 structural classification. Each focal contact candidate is assigned to one primary class by a fixed order: association with a discordant segment join, then recurrence of one anchor across three or more contacts, then a compact focal appearance, then proximity to any segment junction. Contacts satisfying none of these are retained as unclassified significant contacts.

Supplementary Tables

The Four supplementary tables are provided as separate Excel files.

Supplementary Table S1. Family level enrichment α in the three amplicons. Per dataset summary of α across reference bin families, and the family level α .

File: Supplementary_Table_S1_family_enrichment_alpha.xlsx

Supplementary Table S2. Per block regulatory annotation for the CADET interaction blocks in the three amplicons. The table reports module assignment and Layer 3 block class; hg38 anchor coordinates; overlapping and nearest OncoKB genes; enhancer-like ENCODE cCRE counts and densities within the ± 25 -kb annotation window; enhancer-burden tier relative to the amplicon background; and the block-level functional evidence.

File: Supplementary_Table_S2_block_regulatory_annotation.xlsx

Supplementary Table S3. Positional comparison between CADET D458 discovery contacts and ec3D circ-SI or spatial-SI interactions.

File: Supplementary_Table_S3_CADET_ec3D_relevant_results.xlsx

Supplementary Table S4. Layer 8 candidate-target recurrence across chromatin modules. For each dataset and candidate target, the table reports the number and identity of nominating modules, represented evidence classes, block-level regulatory support, summed module evidence scores, and the module-aware and global-shuffle empirical null summaries, including empirical P values and FDR values.

File: Supplementary_Table_S4_layer8_target_recurrence.xlsx

Supplementary Note 1

Layer 8 summarizes candidate-target support across modules

Layer 8 summarized the functional annotations of the final modules at the target level. They document how many times each candidate target was assigned in the different modules and the combined evidence scores for these modules. Because of the limited number of final modules and the dependence of target recurrence on the module construction and functional annotations in the previous layers, I used this analysis descriptively rather than as an additional statistical test.

Six candidate targets were nominated in the nine final modules in H2170. MYC was assigned most frequently in three modules, followed by IKZF3 in two modules. RARA, MIR1207, and two other gene loci were assigned in a single module each. MYC had the highest combined module evidence score (1.67) versus IKZF3 (1.22).

I additionally explored the recurrence pattern in H2170 using two empirical null models. The module-aware null permuted target identities within modules of the same evidence class while the global null permuted targets across all modules. Neither permutation analysis provided significant recurrence of MYC or of IKZF3. For MYC, the empirical P-values were 1.0 under the module-aware null and 0.48 under the global null, and for IKZF3 they were 0.50 and 0.36, respectively. No target survived multiple testing correction. I therefore only use the recurrence of MYC and IKZF3 to describe the support across multiple modules without claiming additional statistical support for regulatory activity.

In D458, MYC was assigned in five of 13 modules, more frequently than any other readable gene, and the other targets in one to three modules. However, the evidence scores for the five MYC modules were not equal. Only one module had strict enhancer-oncogene evidence and the other four were supported primarily by more general local regulatory context. The seven-block recurrent module provided the strongest functional support for MYC with three strict enhancer-oncogene pairings, seven permissive enhancer-oncogene pairings and two contacts with promoter disruption evidence. The recurrence of MYC in five modules documents the presence of multiple types of functional support rather than five independent enhancer-MYC interactions.

In IMR-5/75, ALK was assigned in five of 13 interpreted modules, MYCN in three and ETAA1 in two, and the other candidate targets in single modules. Consequently, ALK had the highest summed module evidence score, followed by MYCN and ETAA1. The recurrence counts document how many times each target was supported by individual modules. I did not use the Layer 8 target label shuffling as a formal test for whether ALK or MYCN was assigned more frequently than expected because it maintains the observed number of target labels. Instead, the analysis describes the amount and composition of evidence for module-level support of each candidate target.

In summary, Layer 8 offers an optional summary of the module and functional annotation analyses at the target level.

Supplementary Note 2

Identifiability and the role of the family-level enrichment α

CADET models the Poisson rate on each pair of path bins (i, j) :

$$\lambda_{ij} = \pi_i \pi_j \alpha_{ij} f(d_{ij})$$

where f is the monotone distance backbone, π the copy-anchor weights and α the residual enrichment. Let $\mu_{ij} = \pi_i \pi_j f(d_{ij})$ ($f(d_{ij})$ is B_{ij} in the manuscript) for the part of the rate that is fixed before the EM begins, and let $F(u, v)$ be the family of path-bin pairs compatible with the observed reference-bin pair (u, v) .

One identifiable enrichment parameter per family

A family is observed through a single number, the aggregated count Y_{uv} . One observation cannot identify more than one free enrichment parameter, so α is necessarily a family-level quantity. This is a property of the data.

Consequence for the EM

In the E step the allocation weight is

$$w_{ij} = \mu_{ij} \alpha_{ij} / \sum_{(k,l) \in F(u,v)} \mu_{kl} \alpha_{kl}$$

If α is constant within the family, it cancels from numerator and denominator, leaving $w_{ij} = \mu_{ij} / M_{uv}$ with $M_{uv} = \sum_F \mu_{kl}$. Substituting $X_{ij} = Y_{uv} w_{ij}$ into the M step gives

$$\hat{\alpha}_{ij} = (X_{ij} + \tau \mu_{ij}) / ((\tau + 1) \mu_{ij}) = ((Y_{uv} / M_{uv}) + \tau) / (\tau + 1),$$

with shrinkage strength $\tau = 1$, so that the update is the midpoint between the family's observed-to-expected ratio and 1. Values are clipped to $[0.1, 10]$ and the matrix is then rescaled to geometric mean 1, so reported α can exceed 10 by the reciprocal of that scale factor. The upper bound was reached by 7 reference-bin families in H2170 and 8 in D458 and by none in IMR-5/75. The lower bound was never reached.

Since α is initialized to 1, family constancy holds at initialization and is preserved at every iteration. The allocation weights therefore never move from their initial values, and the EM reaches its fixed point after one complete pass. In the three datasets the stopping rule terminated after two (D458) or three (H2170, IMR-5/75) iterations, and α is constant within each reference-bin family to machine precision.

The EM derivation is what establishes that the allocation weights are the posterior probabilities of each latent assignment. Copy-specific allocation is determined by π and $f(d)$, both estimated before the loop, and not by iterative refinement. The copy-discriminating information enters through π , which is computed from contacts with uniquely mapping neighbours in a ± 50 -bin window and is therefore based on data the allocation step does not otherwise use.

α is not redundant. It is constant within a family but varies widely between families, and it is that between-family variation that distinguishes a contact that exceeds its copy-aware distance

expectation from one that does not. Across families α ranged from 0.555 to 11.107 in H2170, 0.504 to 10.076 in D458 and 0.513 to 4.540 in IMR-5/75, with 4537, 4450 and 10042 families respectively exceeding 1.5. Per family values for every discovery contact are given in Supplementary Table S1.