Reconstructing the three-dimensional architecture of extrachromosomal DNA with ec3D

Biswanath Chowdhury[1†], Kaiyuan Zhu[1†], Chaohui Li[1†], Jens Luebeck[1], Owen S. Chapman[2,3], Katerina Kraft[4,5], Shu Zhang[4,6,7], Lukas Chavez[2,3], Anton G. Henssen[8,9,10], Paul S. Mischel[7,11], Howard Y. Chang[4,5], Vineet Bafna[1,12]

[1] Department of Computer Science and Engineering, University of California San Diego, San Diego, CA, USA;
[2] Department of Medicine, University of California San Diego, San Diego, CA, USA;
[3] Sanford Burnham Prebys Medical Discovery Institute, San Diego, CA, USA
[4] Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA, USA;
[5] Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA;
[6] Department of Dermatology, Stanford University School of Medicine, Stanford, CA, USA;
[7] Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA;
[8] Department of Pediatric Hematology and Oncology, Charité-Universitätsmedizin Berlin, Berlin, Germany
[9] Berlin Institute of Health, Berlin, Germany
[10] Experimental and Clinical Research Center, Max Delbrück Center for Molecular Medicine and Charité-Universitätsmedizin Berlin, Berlin, Germany
[11] Sarafan Chemistry, Engineering, and Medicine for Human Health (Sarafan ChEM-H), Stanford University, Stanford, CA, USA
[12] Halıcıoglu Data Science Institute, University of California San Diego, San Diego, CA, USA
[†] These authors contributed equally

Abstract

Extrachromosomal DNAs (ecDNAs) are large, acentric, circular DNA molecules that occur pervasively across many human cancers. EcDNA can drive tumor formation and evolution, contribute to drug resistance, and associate with poor patient survival outcomes. Beyond mediating high copy numbers, the circular topology and dynamic conformational changes of ecDNA disrupt topological domains and rewire regulatory networks, thereby conferring an important role in the transcriptional regulation of oncogenes. Here, we develop ec3D, a computational method for reconstructing the three-dimensional structures of ecDNA and analyzing significant interactions from high-throughput chromatin capture (Hi-C) data. Given a candidate ecDNA sequence and the corresponding whole-genome Hi-C as input, ec3D reconstructs the spatial structure of ecDNA by maximizing the Poisson likelihood of observed interactions. Ec3D's performance was validated using both simulated ecDNA structures with varying conformations, and Hi-C data from previously-characterized cancer cell lines. Our reconstructions reveal that ecDNAs occupy spherical configurations and mediate unique long-range interactions involved in gene regulation. Through algorithmic innovations, ec3D can

resolve complex ecDNA structures with duplicated copies of large genomic segments, identify multi-way interactions, distinguish between interactions arising from direct spatial proximity and secondary interactions resulting from alternative folding patterns or intermolecular (trans) contacts of ecDNA molecules. Our findings provide insights into how the spatial organization of ecDNA may influence gene regulation and contribute to increased oncogene expression.

Code availability: https://github.com/AmpliconSuite/ec3D

Introduction

Somatic copy number amplification of oncogenes is a major driver of cancer pathogenicity [1–3]. Recent studies[4–6] have revealed that oncogenes are often amplified by extrachromosomal DNA (ecDNA). EcDNAs are highly prevalent, occurring in approximately 15% of early stage and 30% of late-stage cancers[7], but are rarely seen in normal cells[6]. The presence of ecDNA in tumors is associated with increased pathogenicity and poor outcomes for patients[6]. While this can partially be attributed to increased oncogene expression associated with copy number amplification on ecDNA, recent results point to other contributing factors. EcDNAs have highly accessible chromatin, and their constituent genes are highly expressed, even after accounting for higher copy numbers[5,6].

In normal chromosomes, distinct compartments, known as Topologically Associating Domains (TADs), often bounded by CTCF binding sites, demarcate the regulatory elements that are accessible to a gene[8,9]. In many cancers, the integrity of TADs can be altered[10]. EcDNA formation, which often involves the joining of distal genomic segments, changes chromatin conformation, and disrupts existing topological domains, allowing for enhancer hijacking and rewiring of regulatory circuitry [11–13]. EcDNAs often cluster into hubs promoting *trans* regulatory interactions between different ecDNA molecules[14]. EcDNAs with no protein coding genes have been identified, suggesting an exclusively regulatory role in promoting oncogenesis[15]. Finally, ecDNAs are also suggested to act as roving enhancers for chromosomal genes[16].

Despite the large (10^5-10^8 bp[6,17]) size of ecDNA, their genomic compositions, including genes and regulatory elements, can be reliably identified using short and long read whole genome sequencing [15,18–20]. However, a deeper understanding of the regulatory machinery depends not only on the genomic architecture but also on the 3-dimensional conformation of circular structure. The spatial organization and the three-dimensional structure of ecDNA have, to our knowledge, not been investigated previously.

High-throughput chromosome conformation capture technique (Hi-C) is a dominant technology for characterizing the 3D genome organization[21–23], identifying TADs, and understanding long-range chromatin interactions[8]. The technique quantifies the interaction frequency between each pair of genomic loci, presented in the form of a 2-dimensional matrix. High frequency correlates with spatial proximity, which can be attributed to (a) genomic proximity, (b) structural

variation that brings distal loci together, and (c) topological constrictions in DNA structure. Computational methods have been developed to identify significant pairwise interactions suggesting spatial proximity of pairs that are distant in the reference chromosomes[24–26]. While they provide important structural and topological information, Hi-C is a 2-dimensional projection of the 3-dimensional structure and some important structural features are not immediately discernible. Therefore, these methods do not typically identify multi-way interactions or interactions induced by structural variation, with few exceptions (such as NeoLoopFinder[27]). Smaller changes in 3-dimensional configuration are not immediately apparent in the Hi-C projection. Finally, none of the existing methods account for the circular topology of ecDNA.

Many recent methods have been developed to infer the 3-dimensional structure directly from Hi-C data with increasing resolution, and they have been applied to large genomic segments including human chromosomes[28–34]. However, cancer genomes, and EcDNA in particular, present unique challenges for these methods. Most ecDNA involve complex structural variations, joining together genomic segments from different chromosomes. They may also contain multiple copies of large genomic segments, showing aggregated signals of interactions in the Hi-C matrix, which must be implicitly or explicitly de-duplicated. In this work, we present ec3D, which reconstructs the three-dimensional structure of ecDNA genome using deep Hi-C data and identifies topological constrictions and clusters of statistically significant chromatin interactions, including multi-way and crossing (non-planar) interactions. We used ec3D to reconstruct the 3-dimensional structures of ecDNAs in multiple cancer cell lines and used the structures to better characterize the unique regulatory biology of ecDNA.

Results

Overview of ec3D. Ec3D uses two types of data: (i) a local assembly of ecDNA sequence and (ii) a whole-genome Hi-C contact matrix, both aligned to the same reference genome (see **Methods** for how these data can be obtained). The input ecDNA sequence is represented by ordered and oriented genomic segments in an extended *bed* format, possibly with segments occurring multiple times. The Hi-C matrix describes the interaction frequencies for pairs of bins, each representing a genomic region of pre-specified resolution (default 5 Kbp), in either *hic* or *cool* format.

With these inputs, ec3D first extracts Hi-C submatrices corresponding to segment pairs, where both segments are chosen from the ecDNA sequence. Ec3D reassembles these submatrices into a single matrix C of dimension $N_c \times N_c$ bin pairs, representing chromatin interactions within ecDNA intervals (**Fig. 1**). Next, ec3D reconstructs the 3D structure of the input ecDNA by maximizing the joint Poisson likelihood[29,30], which models interaction frequencies C_{ij} as independent Poisson random variables with mean $\lambda = \beta d_{ij}^{\alpha}$, a decreasing function of the Euclidean distance d_{ij} between bin i and bin j, with a scaling parameter $\beta > 0$ and a power-law decay parameter $\alpha < 0$. (See **Methods** for details.)

The structure is composed of 3-dimensional coordinates for the fixed-resolution bins on ecDNA. Note that the number of bins in the reconstruction, N_e, may exceed the original dimension N_c of the Hi-C matrix C, when the input ecDNA possesses duplicated bins that represent the same genomic regions. In such cases, after determining the 3D structure of ecDNA, ec3D constructs an expanded Hi-C matrix E of dimensions $N_e \times N_e$ by redistributing the interactions to individual copies of bin pairs, proportional to their spatial distance in the reconstructed structure (see **Methods**). Next, Ec3D identifies significant interactions within this expanded matrix (**Fig. 1**). These significant interactions are subsequently clustered using the Louvain method[35]. Ec3D outputs the expanded Hi-C matrix corresponding to the ecDNA sequence, the reconstructed 3D structure coordinates as a text file (**Fig. 1**), and a dynamic structure visualization showing associated genes and clusters of significant interactions.

Ec3D reconstructs structures accurately on simulated data. Given a ground truth structure and the corresponding expanded or collapsed Hi-C matrix, ec3D can reconstruct a 3D structure with the Hi-C matrix, and its performance can be measured by comparing the ground truth and reconstructed structures. We developed an extensive suite of simulated ecDNA structures and Hi-C matrices to benchmark ec3D's performance. Very briefly, we simulated *base* structures with k ($k \in \{1, 2, 3\}$) *topological constrictions (TCs)*. (See **Methods** and **Supplementary Methods** for details.) Each TC corresponds to a pair of genomic regions that are genomically distant but spatially close. We also added multiple random local folds to the base structures (**Supplementary Fig. 1**). Structures, which share the same topological constrictions but differ in local folds, are referred to as having the same base structure. Each simulated structure is described as a $3 \times N_e$ matrix X (**Fig. 2a**), corresponding to the 3D coordinate of N_e bins. We simulated 30 random structures for each value of k, resulting in a total of 90 simulated 3D structures.

For each simulated structure, we generated 10 simulated Hi-C samples by sampling interaction frequencies from the Poisson distribution described above, with random combinations of $\alpha \in$ [-3, -0.75] and $\beta \in [1, 10]$, which cover a typical range we observed in real data. This gives 900 simulated Hi-C in total. The first 450 $N_e \times N_e$ matrices E are expanded matrices without duplicated bins (**Fig. 2b**). The other 450 $N_c \times N_c$ matrices C are collapsed matrices with duplication. To simulate Hi-C with duplicated bins, we first chose the length l (bins) of the duplicated region at random. Next, we randomly selected two ranges of bins, each of length l as being duplicated. We then generated collapsed Hi-C matrices with duplicated bins by summing the interactions for the duplicated bins from the original expanded matrix E. Thus, if the original sample with N_e bins had l bins duplicated, the dimensionality of the collapsed matrix C became $N_c \times N_c$, where $N_c = \left(N_e - l\right)$. Note that the structures were not changed when collapsed matrices were generated from expanded matrices. In a simulated 3D structure, topological

constrictions and local folds contribute to global and proximal interactions in E, which mimic the Hi-C matrix of a real ecDNA sample.

To evaluate performance, we measured the root mean square deviation (RMSD) and Pearson correlation coefficient (PCC) between the ground truth (**Fig. 2a**) and reconstructed structures (**Fig. 2c**). The median RMSD values of the 450 reconstructions without duplication was 0.058, with an RMSD interquartile range IQR=[0.032, 0.106], which was significantly lower than RMSD values computed with both two randomly selected structures with the same base structure (median RMSD 0.338, IQR=[0.268, 0.429], Wilcoxon rank-sum test, P-value $\leq$ 3.7525e-122), and two random structures with different base structures (median RMSD 0.573, IQR=[0.525,0.638], Wilcoxon rank-sum test, P-value $\leq$ 1.2276e-147) (**Fig. 2d; Supplementary Tables 1, 2**). This result suggested that ec3D can reconstruct 3D structures with high accuracy and even reconstruct smaller local folds accurately. Similar results were seen with the PCC metric - the PCC values for the reconstruction were significantly higher than those computed from random structures (**Fig. 2e; Supplementary Tables 1, 2**). Notably, samples with k = 2 and 3 topological constrictions had lower median RMSD values and higher median PCC, compared to samples with k = 1 (**Supplementary Fig. 2**). This improved performance was likely due to stronger global interactions in samples with a higher number of constrictions, resulting in more constraints on possible structures.

We next evaluated the ability of ec3D to reconstruct structures with duplicated bins. We ran ec3D on the 450 collapsed matrices and obtained median RMSD 0.122 (IQR =[0.076,0.222]), which, again was significantly better than two random structures with the same base structure (Wilcoxon rank-sum test, P-value $\leq$ 8.0688e-103) and with different base structures (Wilcoxon rank-sum test, P-value $\leq$ 5.3505e-148) (**Fig. 2d; Supplementary Tables 1, 2**). Comparisons using the PCC metric were again very similar and highly correlated with RMSD (**Fig. 2e**). Note that it is not known in advance if the duplicated regions fold into a similar local substructure. Therefore, in our simulations, we selected half of the samples to have the same local substructure in the duplicated regions, while the other half had different local substructures. The RMSD and PCC values in the two cases were very similar (**Supplementary Fig. 3**), indicating that ec3D has consistent performance regardless of the similarity of local substructures in the duplicated regions.

Because the raw RMSD/PCC values are data dependent and difficult to interpret directly, we compared the PCC (respectively, RMSD) value of ground truth versus a reconstructed structure against the PCC (RMSD) values of the ground truth versus a random structure. The vast majority (97.83%) of reconstructed structures had higher PCC than random structures (**Fig. 2f**). Similarly, 95.67% of reconstructed structures had lower RMSD than random structures (**Supplementary Fig. 4**).

Next, we tested the accuracy of ec3D estimates of the power law decay parameter, α, by measuring the correlation between the true and estimated values of α in the 900

reconstructions. The ground truth and estimated values were highly correlated (**Supplementary Fig. 5**). Defining the error as $\frac{1}{n}\sum_{i=1}^{n} |\hat{\alpha}_i - \alpha_i|/\alpha_i$, where α_i is the ground-truth, and $\hat{\alpha}_i$ the estimated value for sample i, the mean error values in estimating α, for samples without and with duplication, were 2.32% and 3.68%, respectively. The results indicated that α could be estimated accurately in most samples, regardless of whether there are duplications. The estimation accuracy was higher when the true α values were large ($\simeq$-1). To investigate this further, we reanalyzed the RMSD error of structure reconstruction of matched and duplicated groups (900 groups in total) across the different ranges of α values. We found that structure reconstruction accuracy was also better on samples with larger α values (**Fig. 2g**). Notably, α values of real data obtained from human samples tend to be close to -1, further raising confidence in the accuracy of our reconstructions on real data.

Expectedly, the negative likelihood objective decreased smoothly with iterative optimization until convergence. Broadly, the RMSD (respectively, PCC) metric also decreased (respectively, increased), but the transition was much sharper so that a relatively modest improvement in the beginning was followed by a more dramatic shift later (**Supplementary Fig. 6**). Intriguingly, the initial and final RMSD and PCC values of all runs (900×5) were positively correlated, with correlation scores 0.7037 and 0.5712, respectively (**Supplementary Fig. 7**), highlighting the importance of the initialization step in reaching optimal final structures**.**

Ec3D compute time. All samples were run on a supercomputing node equipped with two 64-core *AMD EPYC 7742* processors and 256 GB of DDR4 memory, with at most 16 threads and 2GB memory allocated for each sample. Because ec3D follows a stochastic optimization function, its running time varied from sample to sample. Running time increased with the number of bins (**Fig. 2h; Supplementary Fig. 8**). Duplications took longer time to resolve (**Fig. 2i**). Most ($\geq 90\%$) samples without duplication could be resolved within 12,000 seconds, and most samples with duplication could be resolved within 35,000 seconds.

Ec3D reveals circular structure of ecDNA linking distant segments. We applied ec3D to high coverage Hi-C data acquired from 7 cancer derived cell lines (**Supplementary Table 3**). 5 of the 7 cell lines carry ecDNA, while the remaining two, GBM39HSR and IMR-5/75, contained intrachromosomal focal amplifications that displayed as Homogeneously Staining Regions (HSRs). We used previously published reconstructions of the ecDNA and HSR sequences to obtain the genomic regions of the amplicons (**Methods, Supplementary Table 4**).

Scatter plots comparing Hi-C contact frequency and 3D distance showed a clear inverse relationship on a log-log scale, confirming the expected negative power law decay relationship between frequency and distance in 3D space suggested by the Poisson model (**Fig. 3a, Supplementary Fig. 9**). The correlation was very strong, with PCC ranging from -0.96 to -0.76. Notably, the correlation magnitude increased with increasing Hi-C contact. For medium to high contact regions, abs(PCC)≥ 0.87 PCC (**Supplementary Fig. 9**). The results indicate a more

consistent and precise prediction of spatial distances as contact frequencies increase. The observed horizontal scatter of bins for low distances was due to the regularizer term, which forced adjacent bins to have similar Euclidean distances even if their contact frequencies varied.

Previous estimates[21,30] of α range from $\alpha \simeq -3$ to $\alpha \simeq -1.5$. The optimal values of α on ecDNA structures were somewhat larger, estimated as -1.13 ± 0.22 (**Supplementary Table 3**). The significantly smaller decay of interaction strength with increasing Euclidean distance suggests that ecDNA maintain their structures despite their large size and volume.

All 5 reconstructions naturally converged to circular 3-dimensional structures in contrast with the structure of identical regions in control cell lines. For example, for the GBM39 ecDNA, a relatively simple structure was formed by a single front-to-back joining of a chr7 segment that encompasses the oncogene *EGFR* (**Fig. 3b**). High spatial proximity between the first and last bin was automatically discovered by ec3D. For comparison, we reconstructed the structure of the identical genomic region in GM12878, a cell line where *EGFR* is located on the chromosome (**Fig. 3c**). The reconstruction on GM12878 showed similarity in the smaller topological domains, but importantly, no interactions between the first and last bins.

EcDNA structures are oblate spheroidal and occupy all three dimensions. Scanning electron microscopy data on cultured cells in metaphase[5] does not reliably explain if ecDNAs occupy a sphere-like or a disk-like volume. To address this question, we first computed a minimum volume bounding cuboid that captured the overall shape of the reconstructed 3D structure (**Fig. 3d**). Had the 3D structure of ecDNA been disk-like, we would expect the smallest dimension of the cuboid to be much smaller than the largest dimension. However, the ratios between the minimum and maximum edge length of the bounding box of the 5 ecDNA structures were generally high, ranging from 0.476 (GBM39) to 0.895 (H2170) (**Supplementary Table 3**). This suggested that ecDNA structures were oblate spheroidal with a large third dimension.

We next tested if the ecDNA could be embedded in a “flatter” bounding box (i.e., with smaller edge length ratios) and still generate the observed Hi-C interactions. Specifically, we reconstructed 3D structures of the GBM39 ecDNA (amplifying *EGFR*) and RCMB56 ecDNA (amplifying *DNTTIP2*) by fixing the parameter β with optimal estimated values (β=4 for RCMB56 and β=16 for GBM39) and repeatedly halving the maximum range in the first axis without modifying the range [-1, 1] of the other two axes. By fixing the scaling factor β, we ensured that the structure was not shrinking proportionally in all axes in reconstruction. We hypothesized that for disk-like structures, decreasing the range of one axis would not impact the Poisson likelihood, as bins could still be placed on a plane orthogonal to that axis, preserving the pairwise spatial distances; however, for spherical structures, the Poisson likelihood would become worse, due to additional constraints in the 3D space disrupting expected spatial distances suggested by Hi-C interactions. For GBM39, the likelihood indeed became worse as the smallest dimension decreased from 0.25 to 0.125 (**Fig. 3e**, Wilcoxon rank-sum test, P-value

≤ 0.0045). Similarly for RCMB56, the likelihood reduced significantly as the smallest dimension decreased from 1 to 0.5, (**Supplementary Fig. 10**, P-value ≤ 0.0045), strongly suggesting a spheroidal conformation. Our results are consistent with ecDNA requiring all 3 dimensions for optimal folding, providing additional freedom for complex topological constrictions.

Ec3D reveals high structural similarities between HSR and ecDNA in isogenic lines. The cell line GBM39HSR is isogenic to GBM39EC but with an intra-chromosomal or HSR amplification of EGFR. Remarkably, the Hi-C pairwise interactions of the amplified region were highly similar (Correlation = 0.9859, **Supplementary Fig. 11a-c**). Previous findings have suggested that HSRs can be formed via reintegration of tandemly duplicated copies of ecDNA into a chromosomal locus[36], and this is supported by the similarity of the breakpoints in the isogenic cell lines.

To rebuild the structure of GBM39HSR, we duplicated the first 3 bins (**Methods**) during preparation of the collapsed matrix, and ran ec3D using this genome with duplications. The 3D reconstructions of GBM39EC and GBM39HSR were also remarkably similar (**Fig. 3b, f, Supplementary Fig. 11d**), with RMSD 0.3456 (PCC = 0.9369). By comparison, the RMSD between identical regions in GM12878 and GBM39EC was much higher at 0.3940 (PCC = 0.8504). The similarity between GBM39EC and GBM39HSR structures matched that of *random* structures with the same base structures (median RMSD=0.3380; **Fig. 2d)** suggesting that the major topological constrictions were identical, but ec3D captured fine structural differences between the ecDNA and HSR structures in a way that the Hi-C image could not (**Supplementary Fig. 11d-f**). For example, the spatial distance between chr7:54.865M-54.87M and chr7:55.08M-55.1M nearly doubled from 0.16 in GBM39EC to 0.3 in GBM39HSR.

Despite these advances, the current Hi-C data do not provide enough resolution to distinguish between different possible HSR sub-structures. Distinct structures, such as the ‘spring’ or the ‘petal’ model (**Supplementary Fig. 12**) are possible in the tandem duplication model, but resolving the fine HSR structure will likely require new technologies.

A tandem duplication model for HSR had previously been suggested for the *MYCN* amplification in the human neuroblastoma cell line IMR-5/75[12]. In the proposed architecture of this amplicon, a neo-TAD joined two genomically remote segments connecting the *ANTXR1* locus (chr2:68.9-69.2Mbp) and *LRATD1* (chr2:14.5-15.1Mbp) locus, consistent with a tandem joining of the “last” and “first” segments. Notably, the structure revealed by Helmsauer et al[12] to have two TADs was based on a collapsed matrix containing duplicated copies of Chr2:14.63M-15.1M. Ec3d automatically generated an expanded matrix resolving the duplicated region. It found that the two TADs were maintained, and that the duplicated copies of Chr2:14.63M-15.1M were part of a single TAD, with smaller substructures (**Supplementary Fig. 13**). We next asked if these duplicated regions folded into similar substructures.

Duplicated regions on ecDNA can have similar structures. One key feature of ec3D is the reconstruction of ecDNA structures with duplicated segments. Two ecDNA positive cell lines, D458 and H2170, and one HSR line, IMR-5/75, contained duplicated segments with sizes ranging from 4 bins (20 Kbp) to 163 bins (815 Kbp) (**Supplementary Table 5**). Each segment was duplicated at most two times on the two ecDNAs, including two inverted duplications in D458. We compared the significance of similarity of the local 3-dimensional structure of the duplicated regions using a permutation test (**Methods**). Of the 8 pairs of duplicated regions (6 of size at least 50 Kbp), 2 pairs in D458 and 1 pair in H2170 had significantly similar structures (**Supplementary Table 5**), including for example, duplicated bins [18, 96] and [364, 442] on H2170 (**Figure 3g**, permutation test p-value ≤ 0.016). The other 4 duplicate pairs did not have significantly similar structures, including the duplicated pair on the IMR-5/75 HSR.

We next compared identical genomic regions chr2:15.585-15.985Mbp amplified in two different cell lines. The region is amplified on ecDNA in the cell lines CHP-212, and on an HSR in IMR-5/75 (**Fig. 3h**). The RMSD value of 0.2369 was highly significant (permutation test P-value ≤ 0.0072), confirming that identical genomic sequences folded into very similar local structures despite the very different context. Together, the results suggest that the underlying DNA sequence only provides partial information for reconstructing the structure. Interactions with other factors and nuclear bodies play a role in determining structure.

Ec3D reconstruction clarifies the structure of neo-TADs in ecDNA. Recent results on Neuroblastoma cell lines revealed a class of *MYCN* amplicons that lacked key local enhancers of *MYCN*, but hijacked distal fragments containing previously discovered super-enhancers known to mediate Neuroblastoma progression[12]. Hi-C data from the cell line CHP-212 showed the formation of a neo-TAD connecting *MYCN* to distal super-enhancers. Our 3D reconstruction (**Fig. 4a**) provides a clear delineation of the TAD comprising 3 distal regions on Chr2, containing both *MYCN* and super-enhancers, but not some local *MYCN* enhancers.

Single-cell RNA-seq data of CHP-212 had previously revealed that of the 6 genes present on ecDNA, 4 were overexpressed (*LPIN1, TRIB2, DDX1,* and *MYCN*), but the expression of two genes *GREB1, NTSR2* remained at basal level, with median expression at least 5X lower than the minimum median expression of the overexpressed genes (**Supplementary Fig. 14**[37]). The 3D structure clarifies this observation by revealing a topological domain containing *LPIN1, TRIB2, DDX1,* and *MYCN* along with the super-enhancers, while excluding the other two genes. We also observed that the expression of ncRNA *GACAT3* and the pseudogenes *RNU5E-7P, RPLP1P5* was not impacted despite being in the same topological domain. Intriguingly, *LPIN1* is split in the ecDNA, and the upstream region of *LPIN1* is connected to a novel enhancer region (**Fig. 4a**). Moreover, the circularization removes the region immediately upstream of *GREB1*. Thus, ecDNA can alter the regulation of genes through a combination of structural variation and 3D conformational change.

We also investigated a TAD on the Medulloblastoma cell line D458, which is a 2.5 Mbp molecule amplifying the oncogenes *MYC* (chr8) and *OTX2* (chr14) on an ecDNA. Earlier results had suggested that a DNase-hypersensitive region (DHS1[38]) containing a putative enhancer located 80 Kbp from the *OTX2* gene on chr14 was essential for proliferation of the cell line[17], and it was speculated that DHS 1 might be hijacked by *MYC* to drive proliferation. However, DHS1 was found to not influence *MYC* activity on D458[17]; instead, it enhanced *OTX2* expression in other Medulloblastoma cell lines[38]. Ec3D analysis suggests a neo-TAD that includes DHS1, *OTX2,* and the lncRNA *OTX2-AS1*, but not *MYC*, providing more clarity for the observed experimental data (**Supplementary Fig. 15**). We also noted that an inversion of the *OTX2* region brought *OTX2-AS1* closer to the enhancer on the ecDNA, in contrast to their positioning on the reference genome.

Ec3D reconstructions enable identification and clustering of significant Hi-C interactions. We used ec3D to identify the mechanisms of significant interactions (SIs) between pairs of bins in an expanded matrix. We used 3 methodologies (**Methods, Supplementary Table 6**), each capturing a subset of the possible interactions. Briefly, ref-SI captured SIs relative to expectation on the reference genomes. It was the most general method for capturing significant interactions. The next method, circ-SI, captured SIs after conditioning on the ecDNA sequence, thereby removing interactions due to the joining of distal segments (structural variations) leading to ecDNA formation. The third measure, spatial-SI, captured interactions that could be directly attributed to higher spatial proximity in the ec3D reconstruction, relative to their genomic distance on the ecDNA. Thus, circ-SI and spatial-SI captured decreasing subsets of the interactions predicted by ref-SI.

The number of ref-SI interactions in the ecDNA of GBM39 and RCMB56 was significantly larger relative to the identical region in controls GM12878 and IMR90, and the difference was most pronounced at larger genomic distances due to circularization (two-sample Kolmogorov-Smirnov test P-values: GBM39-GM12878 = 0.024, GBM39-IMR90 = 0.00016, RCMB56-GM12878 = 1.2e-36, RCMB56-IMR90 = 1.1e-71; **Fig. 4b, Supplementary Fig. 16**). The results did not change even after rescaling the control matrices to correct for the higher copy number of ecDNA (**Methods**). In fact, the number of significant interactions reached a local maximum in most cases without the need for rescaling); Furthermore, the proportion of distance-dependent significant interactions remained consistent despite variation in the total number of interactions.

Ec3D reconstruction captures "crossing" interactions. We next investigated if ecDNA could have significant interactions with a non-planar topology, unlike interactions on TADs that are represented as diagonal blocks consistent with a planar topology. One discernible feature of a complex or non-planar 3D fold is the presence of "crossing" significant interactions, which can be described by 4 remote loci, or two interacting pairs (x, z) and (y, w), such that $x < y < z < w$, corresponding to a topological constriction with k=2 (**Supplementary Fig. 17**). While the smaller ecDNA structures (e.g. GBM39, **Fig. 3b**) encompassed only a single topological constriction, other, larger ecDNAs contained multiple topological constrictions with

crossing interactions (**Supplementary Table 3**). For example, we identified 15 crossing interactions from the D458 ecDNA[17]. Among these interactions, one between 500 Kbp distal sites on the *MYC-PVT1* locus crossed another interaction that connected the region upstream of *TMEM260* in chromosome 14 with a region upstream of *CASC8* on chromosome 8 (**Fig. 4c**). The results suggest that ecDNA can promote novel interactions utilizing not only structural variation but also complex topological constrictions.

Ec3D reconstructions identify multi-way interactions. We used Louvain clustering (**Methods)** to obtain clusters of ref-SI interactions for each of the cell lines, suggestive of complex regulatory networks. In D458, we identified 6 ref-SI clusters (**Supplementary Table 7)**. One of these was a clique-like interaction among multiple loci on chr8 and chr14: chr8:127.95-128.02Mb*,* the *PVT1* locus; chr8:128.44-128.58Mb; chr8:128.70-128.74Mb; and chr14:56.80Mb-56.88Mb, the *OTX2* locus (**Supplementary Fig. 18a**). A second cluster (cluster 4) from the same cell line showed a star-like connectivity where a central region containing *MYC*, *PVT1* interacted with multiple distal loci situated ~430 Kbp 5' upstream, and ~800 kb, ~1.01 Mb, and ~1.06 Mb 3' downstream of the *MYC/PVT1* region (**Supplementary Fig. 18b**). The regions upstream and downstream of MYC/PVT1 are devoid of coding genes but contain ncRNA including *CASC8* and *CASC21*. These findings support earlier studies that show the co-amplification and ecDNA formation of these two distinct regions in multiple acute myeloid leukemia samples[37]. It is also notable that the ncRNA *PVT1* appears (partially) with 4 copies in the ecDNA with SV driven proximity to *OTX2*, *TMEM260* (natively on Chr14) and *CASC8*, *MYC* (Chr8) consistent with its role in mediating gene fusions[41].

Ec3D reveals 'differential' Hi-C interactions. As described earlier, we used circ-SI to identify significant interactions that cannot be attributed to structural variations. We next used spatial-SI to identify significant interactions that were specifically due to spatial proximity. Indeed, in the simpler structures such as GBM39, interactions in circ-SI were also identified using spatial-SI (**Supplementary Fig. 19**). Surprisingly, in the ecDNA of RCMB56 **(Supplementary Fig. 20)** and D458, we observed many *differential* interactions–interactions in circ-SI that were not identified by spatial-SI. These interactions could not be attributed either to sequence proximity created by structural variation in ecDNA or to spatial proximity corresponding to topological constrictions. We also did not find evidence of other structural variations that could indicate heterogeneity of ecDNA in the sample.

Many mechanistic reasons could explain these differential interactions. They could, for example, occur due to heterogeneity of ecDNA structure. Another intriguing hypothesis is that these differential interactions are *trans-interactions,* where regulatory elements in one ecDNA are utilized by a different ecDNA in the same hub, as has been suggested previously[14,42].

We explored the occurrence of known regulatory sites in regions with differential interactions. In RCMB56, the region from chr1:86.905M-86.935M interacted with multiple distal regions, including chr1:93.845M-93.885M, containing the oncogene *DNTTIP2*, and

chr1:94.410M-94.430M, containing *ABCD3* (**Fig. 4d**). An H3K27Ac peak, reflective of an active enhancer, was prominent at chr1:86.915M in multiple tissue types[43] **(Supplementary Fig. 20)**. Similarly, we observed a multi chromosomal trans-interaction between chr8:127.73M-chr8:127.745M and chr14:56.645M-56.675M in D458 (**Fig. 4e**), where the chr8 region contained the oncogene *MYC,* and the chr14 region contained an active enhancer mark (**Supplementary Fig. 21**).

Discussion

EcDNAs are circular acentric molecules that are exclusively and ubiquitously found in cancer cells, where they are responsible for oncogene amplification and increased pathogenicity. Their unusual shape and highly accessible chromatin allow for enhancer hijacking and regulatory rewiring. Here, we add another layer of understanding of ecDNA, by presenting the first algorithm to reconstruct its 3-dimensional structure using chromatin capture data.

While these are large molecules and not expected to have a rigid structure, our results on extensive simulation experiments and on real data suggest that most proximities are accurately captured by ec3D. Larger α values imply relatively stronger interactions even between spatially distant regions, adding more information for our structure reconstruction. Because ecDNA are formed by joining multiple distinct genomic segments and are circular, Hi-C interaction is exactly strong between a pair of bins when they are brought proximal either because of the structural variation or because of a topological constriction. In all experiments, ec3D reconstructions consistently showed strong inverse correlations between the spatial distance of bin pairs and the strength of their Hi-C interactions.

Compared with *Multidimensional Scaling* (MDS) based methods[24,28,31,44–46], which attempt to minimize a stress function that measures a discrepancy between the “wish distances” and the 3D distances of the structure, the Poisson model[30] allows more flexible handling of duplicated segments, as one either has to compute wish distances between each copy of a duplicated bin and other bins by splitting the interactions; or introduce a stress function to measure the discrepancy between the expected and observed interactions in duplicated regions.

The challenge of 3D reconstruction of DNA structures can potentially be addressed by other complementary methodologies. Multiplexed imaging of hundreds of genomic loci by sequential hybridization has the potential to elucidate 3-dimensional structures of entire chromosomes at single-cell levels, albeit with a tradeoff between throughput and resolution[47,48]. Here, we focused on chromatin capture data due to higher resolution and ease of data acquisition. The complex multichromosomal configuration of ecDNAs also makes Hi-C a more appropriate technology. As imaging technologies improve, they may help with resolving native ecDNA structures at a single-cell level. Newer exciting developments, such as the optical reconstruction of chromatin architecture (ORCA) promise genomic resolution of 2 Kbp[49], and our future work will explore synergies between these different methodologies.

The Hi-C data are derived from a population of cells, and each cell carries many copies of the same (or very similar) ecDNA species. It is possible that the 3D sequence of ecDNA varies at a single molecule level. Here, we focused on providing a single, representative structure that is practical for generating hypotheses, designing experiments, and integration with other data types (e.g., ChIP-seq) which are often gathered at the bulk level. Furthermore, consensus structures have proven useful in revealing the components of regulatory machinery in a region, including topologically associating domains (TADs) and chromatin loops[8]. In the samples that we analyzed, a single ecDNA structure dominated and its sequence could be unambiguously obtained. Moreover, there was a strong correlation between the number of interactions and the predicted distance across all molecules tested, providing confidence in the predicted structure.

Because ecDNAs are large molecules, with flexibility of DNA conformations, we hypothesized that their 3-dimensional structure was not entirely intrinsic, but was impacted by interactions with proteins, including proteins involved in gene regulation. It had been shown previously that ecDNAs generate new topologically associated domains and rewire the regulatory circuitry with previously inaccessible enhancer regions hijacked by oncogenes. To test this phenomenon more, we first looked at the volume occupied by ecDNA. Our results suggest that ecDNAs fully occupy a 3-dimensional volume, making their shape less disk-like and more oblate spheroidal. The 3-dimensional shape allows for more complex patterns of interaction, and possibly rewires the regulatory circuitry in ways that could be quite different from the chromosome. Indeed, our analysis of significant interactions revealed many interesting cases; we found crossing interactions which would not be possible in a planar structure; examples of clique-like and star-like interactions implying proximity of multiple regions (multiple enhancer elements regulating a gene); and also possible evidence of trans-interactions between different ecDNA molecules. These early findings provide new hypotheses that can be tested in future work, for example, through changes in differential interactions upon dissociation of ecDNA hubs.

We used ec3D to investigate the structure of amplified regions on isogenic lines which were mostly identical except for the location of focally amplified region, which is either extrachromosomal or intrachromosomal. Remarkably, the amplicon had very similar structures suggestive of similar regulatory patterns. Indeed, in addition to neo-TADs, chromosomal TADs have also been observed on ecDNA. These reconstruction data also shed light on the possible 3-dimensional structure of HSRs. As the resolution of Hi-C data improves, we can use our methods to better distinguish between different HSR configurations.

The ec3D algorithm can work even when the ecDNA contains duplicated segments whose interactions are all collapsed in the input Hi-C data. We investigated the fine structure of duplicated regions and found that while some duplicated regions have very similar structures, others do not, consistent with the idea that the 3D structure of ecDNA is not intrinsic to its sequence but is mediated by interacting proteins. We even found a significantly similar structure

of the same region amplified in two different cell lines, suggesting common patterns of regulatory wiring across different samples.

There are many future avenues for improving the basic methodology. Clearly, the technology requires a complete and correct ecDNA primary sequence. This is often challenging with short read based reconstruction which could be ambiguous, and miss many critical breakpoints. Here, we selected ecDNA structures that were tractable and showed minimal cell-to-cell heterogeneity. This method should not be deployed straight out of the box into unvalidated structures, or patient samples where there is often greater heterogeneity than in cancer cell lines. However, with long-read technologies, the prediction of the ecDNA sequence is possible even for complex ecDNA[18].

Newer methods for single-cell Hi-C[50,51] will allow for measurements of cell to cell variability of ecDNA structures, and also help elucidate the structures of multiple ecDNAs in the same sample. Methods are also being developed that disrupt the tethering of ecDNA to chromosomes, or to other ecDNA[14]. Future work aimed at studying the change in structure due to disruption of tethering could help identify the DNA elements involved in tethering, resolving an important biological problem.

In summary, ec3D provides a new tool for the exploration of the regulatory biology of extrachromosomal DNA and other focal amplifications.

Methods

Modeling genomic duplications in Hi-C. The input ecDNA genome often contains duplicated segments of a reference genome. Standard Hi-C mapping and binning methods are unable to separate the interactions on (and between) each distinct copy of a duplicated segment; instead, we observe the sum of interactions given by all copies of that segment. Formally, we refer to the **collapsed** matrix as the Hi-C matrix where each duplicated segment occurs only one time; and the **expanded** matrix as the Hi-C matrix representing the structure of ecDNA where all duplicated segments occur as many times as they are duplicated. Note that only the collapsed matrix is observed. The expanded matrix, which must be inferred, determines the structure of ecDNA and the significant interactions on ecDNA.

To differentiate collapsed matrices and expanded matrices, we use the following notations throughout the method description:

- N_e: total number of fixed resolution bins in the expanded matrix. We typically use 5K or 10K resolution. The size of the expanded Hi-C matrix is $N_e * N_e$.
- N_c: total number of bins involved in ecDNA in the collapsed Hi-C matrix, which is of size $N_c * N_c$.

- L_i: denotes the genomic coordinates (at 5K resolution) corresponding to bin i. Note that in an expanded matrix, different bins may have the same genomic location if they come from duplicated segments on ecDNA.
- $\boldsymbol{\mathcal{R}}_i$: For each bin i in the collapsed matrix, $\boldsymbol{\mathcal{R}}_i = \left\{a \in \{1, \cdots, N_e\} \mid L_i = L_a\right\}$ denotes the set of indices in the expanded Hi-C matrix that have the same genomic location as bin i. The bin i is denoted as *unique*, if $|\boldsymbol{\mathcal{R}}_i| = 1$, and *duplicated* otherwise.
- C_{ij}: #interactions between bins i, j in the collapsed Hi-C matrix.
- E_{ab}: #interactions between bins a, b in the expanded Hi-C matrix.

We make the assumption that the observed number of interactions C_{ij} between a pair of bins i, j is given by:

$$C_{ij} = \sum_{a \in \boldsymbol{\mathcal{R}}_i} \sum_{b \in \boldsymbol{\mathcal{R}}_j} E_{ab}. \quad (1)$$

The following methods are developed based on this principle.

Preparing ecDNA Hi-C matrices. Ec3D's three-dimensional reconstruction only depends on interactions within the ecDNA intervals. Therefore, we first create an ecDNA Hi-C matrix by extracting, reassembling and reorienting the submatrices corresponding to interactions between pairs of segments composing the ecDNA. The input ecDNA sequence is given as a list $S = [(s_1, o_1), (s_2, o_2), (s_3, o_3), \cdots]$ of ordered and oriented genomic segments, where each s_i denotes a genomic interval and $o_i \in \{' + ', ' - '\}$ indicates the orientation of s_i. The Hi-C data is provided as a matrix of interactions between genomic bins from the whole genome. As a first step, we map each segment to a collection of bins, allowing for duplications, to obtain the N_e bins that are amplified by the ecDNA. For each pair of segments $\left(s_i, s_j\right)$, we extract the corresponding submatrix of binned Hi-C interactions, and reassemble these submatrices into a single matrix E of size $N_e * N_e$ bins according to their order in S, with inverted segments ($o_i = ' - '$) reoriented (Fig. 1). Next, we iteratively remove all rows (and columns) a' if there exists a column $a < a'$ with $L_{a'} = L_a$ in E. This results in a collapsed matrix C, to be used subsequently. Additionally, we keep the mapping of indices from the expanded matrix E to the collapsed matrix C to query the indices in each $\boldsymbol{\mathcal{R}}_i$.

Normalizing ecDNA Hi-C matrices. The Hi-C data is typically normalized, for example, using ICE normalization[52], to correct for bin-to-bin variation by ensuring that for each bin i in the normalized matrix C^{ICE}, $\sum_j C_{ij}^{ICE} = 1$.

Within the ecDNA Hi-C matrix, we also ignore the copy numbers contributed by the normal chromosomes as they are much smaller than the ecDNA copy numbers, and copy numbers are uniform across the ecDNA. However, normalization must account for duplications of genomic regions within the ecDNA. With an expanded matrix E, we could enforce $\sum_a E_{ab}^{ICE} = 1$. Instead, we work directly with the collapsed matrix, and aim to compute $C_{ij}^{ICE} = \sum_{a \in \boldsymbol{\mathcal{R}}_i} \sum_{b \in \boldsymbol{\mathcal{R}}_j} E_{ab}^{ICE}$. But since E^{ICE} is not known, we approximated C_{ij}^{ICE} through a generalized version of ICE normalization such that in the normalized matrix C^{ICE} (of the reassembled matrix C), $\sum_i C_{ij}^{ICE} = |\boldsymbol{\mathcal{R}}_i|$, where $|\boldsymbol{\mathcal{R}}_i|$ is the multiplicity of genomic bin i on ecDNA. Finally, to keep the original scale of interactions, we multiply a constant $r = (\sum_i \sum_j C_{ij})/N_c$ to the normalized matrix C^{ICE} and work on the scaled matrix $r \cdot C^{ICE}$ in the following steps. We implemented the normalization procedure above using the iced package[53].

Reconstructing the 3D structure of ecDNA. Given a normalized Hi-C matrix for ecDNA C^{ICE} (or $r \cdot C^{ICE}$), we compute a single **consensus** (of multiple copies of ecDNA in a mixture of cells) 3D structure of the ecDNA. Formally, we compute a vector $X \in \mathbb{R}^{N_e \times 3}$ of dimension $N_e \times 3$ - where $X_a = (x_{a1}, x_{a2}, x_{a3})$ represents the coordinate of bin a ($a \in 1, \cdots, N_e$). Define

$$d_{ab} = ||X_a - X_b||_2 = \sqrt{(x_{a1} - x_{b1})^2 + (x_{a2} - x_{b2})^2 + (x_{a3} - x_{b3})^2}$$

as the Euclidean distance between bin a and bin b given the coordinates of X_a and X_b.

The normalized interaction frequency C_{ij}^{ICE} is modeled as a Poisson random variable, relating to d_{ij}. Specifically, for a pair of unique bins i, j, the expected number of interactions is given by $\lambda_{ij} = \mathbb{E}\left[C_{ij}^{ICE}\right] = \beta d_{ij}^{\alpha}$, for parameters $\alpha < 0, \beta > 0$, which are estimated separately for each dataset. The parameter α describes the rate of power law decay of Hi-C interactions due to spatial distances, and β can be treated as a scaling factor. Moreover, the likelihood of observing C_{ij}^{ICE} interactions between a pair of bins i, j is given by a Poisson(-like) distribution

$$\mathcal{L}(C_{ij}^{ICE}, X) = \frac{(\lambda_{ij})^{C_{ij}^{ICE}} \exp(-\lambda_{ij})}{\Gamma(C_{ij}^{ICE} + 1)}. \qquad (2)$$

When the bin pairs are not unique, we define $\lambda_{ij} = \mathbb{E}\left[C_{ij}^{ICE}\right] = \sum_{a \in \boldsymbol{\mathcal{R}}_i} \sum_{b \in \boldsymbol{\mathcal{R}}_j} \beta d_{ab}^{\alpha}$, and the likelihood is computed based on the new expectations.

We aim to maximize the log likelihood of the overall collapsed matrix C^{ICE} or minimize $-\ln\left(\prod_{i,j} \mathcal{L}(C_{ij}^{ICE}, X)\right)$. Additionally, we control the variance between consecutive bins a and $a+1$ using a regularization term proportional to

$$Reg(X) = Var(d_{a,a+1}) = \frac{1}{N_e - 1}\left(\sum_{a=1}^{N_e-1} d_{a,a+1}^2 - \left(\sum_{a=1}^{N_e-1} d_{a,a+1}\right)^2\right). \quad (3)$$

This is based on the assumption that every pair of consecutive bins should be spaced approximately equally in Euclidean space. The overall optimization problem is given as follows

$$\min -\ln\left(\mathcal{L}(C^{ICE}, X)\right) + \gamma \cdot Reg(X) \quad (4)$$

$$\sim \min \sum_i \sum_j \left(\lambda_{ij} - C_{ij}^{ICE} \ln\left(\lambda_{ij}\right)\right) + \frac{\gamma}{N_e - 1}\left(\sum_{a=1}^{N_e-1} d_{a,a+1}^2 - \left(\sum_{a=1}^{N_e-1} d_{a,a+1}\right)^2\right) \quad (5)$$

where a constant term is ignored for the minimization. The weight γ of the regularization term is provided as a user input, and by default we set γ to $0.05 \cdot N_e$.

Implementation details. The optimization is done iteratively, for X and α (and β), with l-BFGS[54] algorithm implemented in SciPy:

1. Start with an initial estimation of X;
2. Minimize the negative log likelihood with respect to α and β by fixing X;
3. Minimize the negative log likelihood over X after fixing α and β;
4. Iterate steps 2 and 3 until convergence or reaching an upper bound of rounds (by default we set the maximum round to 1000).

To determine convergence we look at the value of objective function in the last 10 rounds and set the convergence criteria to $|obj_i - obj_{i-10}| / \max(obj_i, \cdots, obj_{i-10}) < \epsilon$, where obj_i and obj_{i-10} are objective values at the current round and 10 rounds before, respectively. To avoid local minimums due to non-convexity, we run the initialization and iterative optimization 5 times, with random initialization of X for running MDS (see below for the initialization of X), and keep the final X which leads to the best objective value. In the optimization process, we require that the three dimensions are bounded by [-1, 1], but do not enforce any limit on β to allow flexible scaling of the structures.

Initialization of X. We found that initialization plays an important role in deriving the optimal coordinates X (**Results, Supplementary Fig. 7**), and therefore we try to initialize X sufficiently close to the final solution by initializing X with running a procedure similar to multidimensional scaling (MDS)[30].

Note that the naive MDS requires the expanded matrix to work with. To obtain the expanded matrix for MDS, we redistribute the normalized interactions C_{ij}^{ICE} to E_{ab} for all $a \in \boldsymbol{\mathcal{R}}_i$, $b \in \boldsymbol{\mathcal{R}}_j$ in proportional to d_{ab}^{-3} (i.e., with the assumption that $\alpha = -3$; β, the scaling factor, can be canceled out here). Thus,

$$E_{ab} = \frac{d_{ab}^{-3}}{\sum_{a' \in \boldsymbol{\mathcal{R}}_i} \sum_{b' \in \boldsymbol{\mathcal{R}}_j} d_{a'b'}^{-3}} \cdot C_{ij}^{ICE} \text{ if } i \neq j. \quad (6)$$

When $a, b \in \boldsymbol{\mathcal{R}}_i$, we set

$$E_{ab} = Avg(C^{ICE}_{i'j':\boldsymbol{\mathcal{R}}_{i'}=\{a'\},\boldsymbol{\mathcal{R}}_{j'}=\{b'\},\, d_{a'b'}=d_{ab}}) \quad (7)$$

(i.e., the average of all unique bin pairs a' and b' with the same distance as bin pair a, b; and we use genomic distance as defined below) when redistributing the diagonal elements. Since the Euclidean distance is not known, we use a *circular* genomic distance on ecDNA as a proxy: $d_{ab} = g_{ab} = \min(|a - b|, N_e - |a - b|)$ - the shortest distance between bin a and b on the circular ecDNA structure. To better compute X we allow some flexibility in redistributing interactions by treating E_{ab} as variables in the optimization process and adding a stress function to penalize the discrepancies between $\sum_{a \in \boldsymbol{\mathcal{R}}_i} \sum_{b \in \boldsymbol{\mathcal{R}}_j} E_{ab}$ and C_{ij}.

Specifically, the objective of MDS can be written as

$$\min \sum_{a=1}^{N_e} \sum_{b=1}^{N_e} \left(d_{ab} - \delta_{ab}\right)^2 / \delta_{ab}^2 + \sum_{i=1}^{N_c} \sum_{j=1}^{N_c} \left(\sum_{a \in \boldsymbol{\mathcal{R}}_i} \sum_{b \in \boldsymbol{\mathcal{R}}_j} E_{ab} - C_{ij}^{ICE} \right)^2 / C_{ij}^{ICE} \quad (8)$$

where $\delta_{ab} = (E_{ab}/\beta)^{-1/3}$ is the wish distance. Again we set $\alpha = -3$ regardless of the true/optimal values as MDS is run just for initialization purposes.

Resolving HSRs created through reintegration of ecDNA. We preprocessed data to reconstruct the structure of HSRs formed by head-to-tail recombination of the ecDNA sequence and subsequent chromosomal re-integration. We ran CoRAL[18] to obtain a single copy composing this underlying tandem-duplication like HSR genome (see section **ecDNA genome reconstruction from WGS data** below) and duplicated the first 3 bins, representing 15 Kbp of sequence during preparation of the collapsed matrix. The predicted CoRAL sequence along

with the 15 Kbp duplication was provided as input to ec3D for structure reconstruction. Ec3D automatically normalized the collapsed matrix and reconstructed HSR structures.

Identifying significant interactions. Increased Hi-C interactions can be attributed to three main factors: (a) reference genome proximity, which leads to spatial proximity, (b) spatial proximity induced by structural variants (SVs)[55–58], and (c) spatial proximity introduced by a conformational change. Furthermore, due to the higher copy number of ecDNA and potential formation[14,42], significant interactions may also reveal *trans* interactions between two ecDNA molecules. As per previous methods[24–26], we define significant interactions as pairs of bins (a, b) $(a, b = 1, \cdots, N_e)$ with interaction frequencies E_{ab} much more than expected at a given genomic distance. We first introduce a unified method in ec3D that computes significant interactions for an abstract definition of genomic distance here. In the next subsection, we describe different choices of genomic distance function that allow us to distinguish interactions due to SVs from interactions due to conformational change.

Specifically, we always model the interactions at each genomic distance g using a Negative Binomial distribution with mean μ_g and variance σ_g $(\sigma_g > \mu_g)$. The statistical significance (P-value) of E_{ab} is computed as the probability of observing at least E_{ab} interactions with the underlying distribution: $P_{ab} = \mathbb{P}(e \geq E_{ab})$, $e \sim NB(\mu_g, \sigma_g)$, $\forall a, b$ satisfying $g_{ab} = g$. Then we correct all resulting P-values for multiple testing using the Benjamini-Hochberg procedure to compute an adjusted P-value (i.e., q value) for each bin pair (a, b). By default, pairs of bins with q value < 0.05 are denoted as significant interactions. We noticed that significant interactions often occurred clumped with their neighboring bin pairs in the Hi-C matrix, at high resolutions such as 5K. Therefore, we implemented an option to only output the locally maximal significant interactions, i.e., those with interaction frequencies greater than their top, bottom, left and right neighbors.

The mean (μ_g) and variance (σ_g) of the number of interactions at each genomic distance g are estimated by computing the empirical mean and variance interactions E_{ab} for all a, b satisfying $g_{ab} = g$, after detecting and removing outliers using the IQR method[59].

The computation of significant interactions also requires the expanded matrix E_{ab}. To compute the expanded matrix, we first redistribute raw interactions C_{ij} to E_{ab} for all $a \in \boldsymbol{\mathcal{R}}_i$, $b \in \boldsymbol{\mathcal{R}}_j$ similar to equations (6) and (7) described in **Initialization of** X, but using the optimal values of spatial distance d_{ab} and α:

$$E_{ab} = \frac{d_{ab}^{\alpha}}{\sum_{a' \in \boldsymbol{\mathcal{R}}_i} \sum_{b' \in \boldsymbol{\mathcal{R}}_j} d_{a'b'}^{\alpha}} \cdot C_{ij} \text{ if } i \neq j;$$

and

$$E_{ab} = Avg(C_{i'j':\mathcal{R}_{i'}=\{a'\},\mathcal{R}_{j'}=\{b'\},\, d_{a'b'}=d_{ab}})$$

otherwise. The resulting expanded matrix E_{ab} is then renormalized with ICE normalization.

Finally, we exclude potential false positive calls due to an artifact of ICE normalization. For example, in RCMB56 matrix, row (or column) 27 has only a few non-zero entries, potentially due to a mapping/binning artifact of HiC-pro, but ICE normalization forces this row to have the same sum of interactions as other rows. As such the interaction counts were boosted by ICE normalization, making them returned as significant in the P-value calculation. We postprocess significant interactions by removing all rows with much less non-zero entries than average again with the IQR method. Ec3d implements a user option to remove interactions in certain rows/columns.

Choosing genomic distance between two bins. The circular genomic distance $g^c_{ab} = \min(|a-b|, N_e - |a-b|)$ defined in section **Initialization of** X implicitly removes the effect of SV breakpoints joining remote genomic segments (in S) on ecDNA. In contrast, to capture both SV-driven and conformation-driven significant interactions, we define the genomic distance between bin a and b as their genomic distance on the reference genome:

$$g^r_{ab} = \min(|L_a - L_b|, g_{max}) \qquad (9)$$

when L_a and L_b are located on the same chromosome, where g_{max} is a sufficiently large genomic distance with no (or few) interactions on expectation at this distance; and $g_{ab} = g_{max}$ when L_a and L_b are located on different chromosomes. Notice that genomic distance is not continuous in Hi-C - two adjacent values differ by a fixed resolution, e.g., 5 Kbp. We set g_{max} to the size of ecDNA to avoid large gaps between two genomic distances. We refer to **ref-SI** as bin-pairs (a, b) where the number of Hi-C interactions between loci in bins a and b was significantly higher than expected for the reference genomic distance g^r_{ab} between the bins; **circ-SI** as pairs (a, b) where the number of Hi-C interactions was significantly higher than expected for the *circular* distance g^c_{ab}.

The number of bin pairs on ecDNA usually decreases (linearly) with increasing reference genomic distance. Bin pairs that do not share the same genomic distance with any other pairs are unlikely to be identified as significant, due to the way P-values are calculated. Therefore, we sort all bin pairs according to their genomic distances, and partition them into groups, each with at least $\frac{N_e}{2}$ bin pairs, by greedily merging bin pairs that are similar in genomic distance. The mean and variance of the number of interactions are estimated separately for each group and used to compute nominal p-values. If circular genomic distance is used, this partition is not

needed as the number of bin pairs at each genomic distance remains the same (i.e., N_e), except for a maximum distance with $\frac{N_e}{2}$ bin pairs when N_e is even.

Identifying candidate *trans*-interactions. Earlier research has revealed that ecDNA forms hubs with regulatory interactions between different ecDNA molecules[14,42]. Therefore, it is possible that ecDNA Hi-C data includes interactions between distinct copies of ecDNA molecules. To identify *cis* interactions within ecDNA, we can optionally compute significant interactions with respect to the ratio g^c_{ab}/d_{ab} between circular genomic distances and spatial distances. Specifically, for each distance g, and all bin pairs (a, b) such that $g^c_{ab} = g$, we fit a Negative Binomial distribution for the ratio g/d_{ab}. Pairs of bins with significantly high ratio after FDR corrections corresponded to significant interactions relative to their spatial proximity. We refer to this third measure of significant interactions as **spatial-SI**. Significant interactions computed from Hi-C using circ-SI that are not found using spatial-SI are suggestive of "secondary" interactions. These interactions can result from alternative 3D conformations, secondary SVs (not participating in the ecDNA sequence), or *trans* interactions between ecDNAs.

Identifying significant interactions from rescaled matrices. To show that significant interactions on ecDNA were not due to their higher copy numbers, we rescaled the case (i.e., ecDNA) and control (i.e., extracted from the same intervals from non-amplified cell lines) Hi-C matrices by a factor ranging from 0.25 to 4, and then identified significant interactions with the same procedure but from the rescaled matrices. The results suggested that the number of significant interactions is not monotonically increasing with the total number of interactions, and the pattern of significant interactions as a function of increased genomic distance remains the same. In fact, the number of significant interactions reached a local maximum in most cases without rescaling. We note that the variance of interactions at each genomic distance decreased quadratically to the downscale factor, breaking the negative binomial property when the rescaling factor becomes too small. As such we only tested rescaling factors that preserve larger variance than mean interactions at 90% of all distinct genomic distances.

Clustering significant interactions. EcDNA often exhibits complex conformations that form multi-way interactions among different regions within its structure to amplify the oncogene and other associated gene expression. The connectivity of these multi-way interactions (e.g. star-like shape or clique-like) indicates different types of interacting pathways. To identify multi-way interactions, we build an interaction network $G = (\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$ from all significant interactions where the node set $\boldsymbol{\mathcal{V}}$ include all bins involved in a significant interaction and the edge set $\boldsymbol{\mathcal{E}}$ indicates the actual interactions. We detect *communities* in the interaction network G by using Louvain clustering. Louvain clustering[35] partitions nodes into clusters while maximizing the modularity score (density of links within clusters compared to links between clusters).

Simulations. We simulated ecDNA 3D structures and their corresponding Hi-C data to assess the effectiveness of ec3D in 3D structure reconstruction. At the highest level, we introduced the notion of *topological constrictions* to simulate the effect of major conformational changes on ecDNA structures. Topological constrictions generalize chromatin loops - which typically connect a pair of bins (x, y) that are genomically far - by specifying two broader intervals of bins $[x, x + \Delta x]$, $[y, y + \Delta y]$ where the neighboring bins around x and y are generally genomically distant but spatially close, resulting in strong off-diagonal Hi-C interactions. Increased number of topological constrictions usually indicates more complex 3D structures.

Each simulated structure was obtained by sampling evenly spaced points from a circular 3D curve. We generated a diverse set of *base structures* by varying three key parameters, which determine the shape of the underlying 3D curve and the number of points to be sampled. First, we incorporated $k \in \{1, 2, 3\}$ topological constrictions. Second, we varied the spatial distance between the two intervals that participate in a topological constriction. Third, we simulated structures of different sizes by varying the total number of points $N_e \in \{250, 500, 750\}$. In addition, we introduced *local folds* on each base structure by randomly disturbing the positions of small collections of continuous points. See Supplementary Fig. 1 and Supplementary Methods for details. In these simulated structures, each point can be treated as the spatial placement of a genomic bin at a fixed Hi-C resolution.

We next generated an expanded Hi-C matrix E of size $N_e \times N_e$ from each simulated circular structure with N_e bins as follows. For each pair of bins $a, b \in \{1, 2, \cdots, N_e\}$, we sampled the interaction counts E_{ab} from a Poisson distribution with mean β/d_{ab}^{α} [30], where d_{ab} represents the Euclidean distance between bins a and b. The parameters α and β were randomly chosen from [-3, -0.75] and $[1, 10]$, respectively. Next, we simulated duplications by designating contiguous ranges of bins as duplicated regions in each Hi-C matrix E, and summing up the interaction frequencies of duplicated bins in E to obtain the collapsed matrix C of size $N_c \times N_c$. To evaluate ec3D's ability to reconstruct structures where duplicated regions fold into different conformations, we finally designed our simulations in a way that half of the samples had the same local substructures for the duplicated regions, while the other half had different local substructures.

Using the procedure described above, we randomly generated 10 structures for each combination of $k \in \{1, 2, 3\}$ and $N_e \in \{250, 500, 750\}$, which led to 90 structures in total. And for each simulated structure, we generated 5 expanded matrices E without duplication and 5 collapsed matrices C with duplication by varying α and β, giving 450 expanded matrices and 450 collapsed matrices.

Performance metrics (RMSD, PCC). Similar to other 3D reconstruction methods, we measure the (dis)similarity between two 3D structures X and X' of the same size N through root mean squared distance (RMSD) and Pearson correlation coefficient (PCC):

$$RMSD(X,\ X') = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - X_i')^2}\ ,$$

$$PCC(X,\ X') = \frac{\sum_{i,j}(d_{ij}-\bar{d})(d_{ij}'-\bar{d}')}{\sqrt{\sum_{i,j}(d_{ij}-\bar{d})^2\sum_{i,j}(d_{ij}'-\bar{d}')^2}}\ ,$$

where d_{ij} is the Euclidean distance between bin i and bin j, and $\bar{d} = \frac{1}{N^2}\sum_{i,j} d_{ij}$.

However, due to the flexibility of coordinates with respect to rigid transformation in reconstruction, we first aligned X and X' by translation, scaling, and rigid body rotation using the Kabsch-Umeyama algorithm[60]. For a brief summary, the algorithm works in three steps. (i) Move the centroid of both structures to the origin by subtracting X and X' with their centroid $\overline{X}$ and $\overline{X'}$. (ii) Rescale the two structures by their maximum diameters. (iii) Rotate X by singular value decomposition (SVD) to align it with X' in the optimal orientation. Namely, we computed SVD of $X \cdot X'^{T} = VSW^{T}$ and rotated X by $(W \cdot V^{T}) \cdot X$.

Similarity of 3D structure in duplicated regions. We used a permutation test to measure the significance of similarity between the local structures of duplicated regions D_1 and D_2 on the same ecDNA, or correspond to the local 3D reconstruction of the same genomic interval but from different samples (e.g., ecDNA and HSR amplicon of *MYCN*). Specifically, to compare duplicated regions D_1 and D_2 on the same ecDNA, we randomly sampled 5,000 regions S_i ($i = 1, 2, \cdots, 5000$) of the same size from the same molecule, and computed the fraction of times, a random pair (D_1, S_i) had smaller RMSD or larger PCC compared to the duplicated pair (D_1, D_2) as the empirical P-value. To compare 3D reconstruction of the same genomic interval from different samples, we sampled S_i from the larger genome.

Minimum bounding box analysis. A minimum volume bounding box can be used to describe the overall 3D shape of an ecDNA structure. We implemented both the "rotating calipers" method[61] and PCA to compute the bounding box. Rotating calipers method takes $O(N_e^3)$ time and computes an exact solution; PCA takes $O(N_e)$ time for the $N_e \times 3$ structure matrix X and gives a good practical solution, though without approximation guarantee[62]**.** In fact, we mainly focused on the ratio between the largest dimension and the smallest dimension of the bounding box, which can separate disk-like structures from spherical structures. Both methods suggested extremely similar ratios (reported in Results) - even if the optimum bounding boxes computed by rotating calipers turn out smaller than PCA bounding boxes.

We additionally tested if the reconstructed ecDNA structures could be placed into a "flatter" bounding box (i.e., with smaller edge length ratios) and still generate the observed Hi-C

interactions. Specifically, we reconstructed 3D structures of the ecDNA by optimizing the objective function described in equation (5) with the scaling parameter β fixed, but with the maximum range of the first axis repeatedly halving from [-1, 1] to [$-\frac{1}{32}$, $\frac{1}{32}$]. The other two axes remain in the range [-1, 1]. By fixing β, we ensured that the structure was not shrinking proportionally in all axes in reconstruction. Decreasing the range of one axis would not impact the Poisson likelihood of a disk-like structure, as bins could still be placed on a plane orthogonal to that axis, preserving the pairwise spatial distances; while for spherical structures the Poisson likelihood would become worse, due to additional constraints in the 3D space disrupting expected spatial distances suggested by Hi-C interactions.

Hi-C data preparation. We downloaded the raw Hi-C data of CHP-212 and IMR-5/75 from Helmsauer et al.[12], D458 and RCMB56 from Chapman et al.[17]. We downloaded high coverage GM12878 and IMR90 Hi-C (as control samples without ecDNA amplification) from 4D nucleome (https://data.4dnucleome.org/). The HiC library for H2170 was prepared using the Arima-HiC kit. Hi-C libraries for GBM39EC and GBM39HSR were prepared following a standard protocol to investigate chromatin interactions[23]. Samples were sequenced using Illumina NovaSeq in 150 bp paired-end reads, with 3 replications for both GBM39EC and GBM39HSR. We combined these replications into a single matrix in our structural reconstruction.

We processed the raw Hi-C reads with HiC-Pro version 3.1.0[53]. This process included aligning the reads to the human reference genome (hg38), removing duplicate reads, assigning reads to restriction fragments, filtering for valid interactions, and generating binned contact matrices. For Arima Hi-C we set the restriction enzyme to ^GATC and G^ANTC, and trimmed 5 bases from the 5' end of both read 1 and read 2 before alignment as per their user guide. Otherwise, we set the restriction enzyme to Dpnii, and did not trim the reads. We generated contact matrices at resolutions ranging from 2 Kbp to 1 Mbp, but focused on 5 Kbp resolution mostly in our analysis, allowing for detailed description of chromatin interactions. The HiC-Pro output was converted into *cooler* format (*.*cool* or *.*mcool*)[63] required by ec3D. Note that ec3D also supports *.*hic* format as input compatible for visualization and analysis with Juicebox tools[64], and internally converts *.*hic* input to *.*cool* format using hic2cool (https://github.com/4dn-dcic/hic2cool).

ecDNA genome reconstruction from WGS data. Ec3D requires an ecDNA sequence as input. For GBM39, GBM39HSR, CHP-212, IMR-5/75 (HSR), and H2170, we assembled the ecDNA genomes from Oxford Nanopore WGS by running CoRAL[18]. We ran CoRAL with a non-default command line argument '--min_bp_support 10.0' (i.e., with minimum coverage cutoff 10 times the diploid coverage for breakpoints) to eliminate redundant breakpoints which could result from non-ecDNA structural variations or heterogeneous ecDNA sequences. We extracted the cycle with the largest predicted copy number from CoRAL's output as the (primary) ecDNA sequence. For RCMB56 we ran AmpliconArchitect[19] with default parameters on paired end short reads and again selected the cycle with the largest CN as its ecDNA sequence. The resulting ecDNA sequence of RCMB56 also agreed with optical genome mapping (OGM) contigs[17]. For D458, we reused the ecDNA sequence from Chapman et al.[17] computed by AmpliconReconstructor[65] from

WGS and OGM contigs. Compared with AmpliconArchitect output consisting of multiple small cycles (as part of the ecDNA sequence), OGM provided a single consensus ecDNA sequence of D458 that was supported by all informative contigs[17].

References

1. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
2. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
3. Steele, C. D. *et al.* Signatures of copy number alterations in human cancer. *Nature* **606**, 984–991 (2022).
4. Turner, K. M. *et al.* Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* **543**, 122–125 (2017).
5. Wu, S. *et al.* Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature* **575**, 699–703 (2019).
6. Kim, H. *et al.* Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat. Genet.* **52**, 891–897 (2020).
7. Luebeck, J. *et al.* Extrachromosomal DNA in the cancerous transformation of Barrett's oesophagus. *Nature* **616**, 798–805 (2023).
8. Dixon, J. R. *et al.* Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. *Nature* **485**, 376–380 (2012).
9. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
10. Akdemir, K. C. *et al.* Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat. Genet.* **52**, 294–305 (2020).
11. Morton, A. R. *et al.* Functional Enhancers Shape Extrachromosomal Oncogene Amplifications. *Cell* **179**, 1330-1341.e13 (2019).
12. Helmsauer, K. *et al.* Enhancer hijacking determines extrachromosomal circular MYCN amplicon architecture in neuroblastoma. *Nat. Commun.* **11**, 5823 (2020).
13. RP Koche, E. R.-F., K. Helmsauer, M. Burkert, IC MacArthur, J. Maag, R. Chamorro, N. Munoz-Perez, M. Puiggros, GH Dorado. Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. *Nat Genet* **52**, 29–34 (2020).
14. Hung, K. L. *et al.* ecDNA hubs drive cooperative intermolecular oncogene expression. *Nature* **600**, 731–736 (2021).
15. Hung, K. L. *et al.* Targeted profiling of human extrachromosomal DNA by CRISPR-CATCH. *Nat. Genet.* **54**, 1746–1754 (2022).
16. Yi, E., Chamorro González, R., Henssen, A. G. & Verhaak, R. G. W. Extrachromosomal DNA amplifications in cancer. *Nat. Rev. Genet.* **23**, 760–771 (2022).
17. Chapman, O. S. *et al.* Circular extrachromosomal DNA promotes tumor heterogeneity in high-risk medulloblastoma. *Nat. Genet.* **55**, 2189–2199 (2023).
18. Zhu, K. *et al.* CoRAL accurately resolves extrachromosomal DNA genome structures with

long-read sequencing. Preprint at https://doi.org/10.1101/2024.02.15.580594 (2024).
19. Deshpande, V. *et al.* Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat. Commun.* **10**, 392 (2019).
20. Giurgiu, M. *et al.* Decoil: Reconstructing Extrachromosomal DNA Structural Heterogeneity from Long-Read Sequencing Data. in *Research in Computational Molecular Biology* (ed. Ma, J.) vol. 14758 406–411 (Springer Nature Switzerland, Cham, 2024).
21. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
22. Jin, F. *et al.* A high-resolution map of three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
23. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).
24. Ay, F., Bailey, T. L. & Noble, W. S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.* **24**, 999–1011 (2014).
25. Carty, M. *et al.* An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. *Nat. Commun.* **8**, 15454 (2017).
26. Wolff, J., Backofen, R. & Grüning, B. Loop detection using Hi-C data with HiCExplorer. *GigaScience* **11**, giac061 (2022).
27. Wang, X. *et al.* Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nat. Methods* **18**, 661–668 (2021).
28. Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465**, 363–367 (2010).
29. Hu, M. *et al.* Bayesian inference of spatial organizations of chromosomes. *PLoS Comput. Biol.* **9**, e1002893 (2013).
30. Varoquaux, N., Ay, F., Noble, W. S. & Vert, J.-P. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics* **30**, i26–i33 (2014).
31. Lesne, A., Riposo, J., Roger, P., Cournac, A. & Mozziconacci, J. 3D genome reconstruction from chromosomal contacts. *Nat. Methods* **11**, 1141–1143 (2014).
32. Szałaj, P. *et al.* An integrated 3-Dimensional Genome Modeling Engine for data-driven simulation of spatial genome organization. *Genome Res.* **26**, 1697–1709 (2016).
33. Paulsen, J. *et al.* Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. *Genome Biol.* **18**, 21 (2017).
34. Wang, H., Yang, J., Zhang, Y., Qian, J. & Wang, J. Reconstruct high-resolution 3D genome structures for diverse cell-types using FLAMINGO. *Nat. Commun.* **13**, 2645 (2022).
35. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
36. Bailey, C., Shoura, M. J., Mischel, P. S. & Swanton, C. Extrachromosomal DNA—relieving heredity constraints, accelerating tumour evolution. *Ann. Oncol.* **31**, 884–893 (2020).
37. Stöber, M. C. *et al.* Intercellular extrachromosomal DNA copy-number heterogeneity drives neuroblastoma cell state diversity. *Cell Rep.* **43**, 114711 (2024).
38. Wortham, M. *et al.* Chromatin Accessibility Mapping Identifies Mediators of Basal Transcription and Retinoid-Induced Repression of OTX2 in Medulloblastoma. *PLoS ONE* **9**,

e107156 (2014).
39. L′Abbate, A. *et al.* MYC-containing amplicons in acute myeloid leukemia: genomic structures, evolution, and transcriptional consequences. *Leukemia* **32**, 2152–2166 (2018).
40. Bagchi, A. Methods and compositions for treating myc-driven cancers. (2021).
41. Tolomeo, D., Agostini, A., Visci, G., Traversa, D. & Storlazzi, C. T. *PVT1*: A long non-coding RNA recurrently involved in neoplasia-associated fusion transcripts. *Gene* **779**, 145497 (2021).
42. Hung, K. L. *et al.* Coordinated inheritance of extrachromosomal DNAs in cancer cells. *Nature* **635**, 201–209 (2024).
43. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
44. Ben-Elazar, S., Yakhini, Z. & Yanai, I. Spatial localization of co-regulated genes exceeds genomic gene clustering in the Saccharomyces cerevisiae genome. *Nucleic Acids Res.* **41**, 2191–2201 (2013).
45. Rieber, L. & Mahony, S. miniMDS: 3D structural inference from high-resolution Hi-C data. *Bioinformatics* **33**, i261–i266 (2017).
46. Zhang, Y., Liu, W., Lin, Y., Ng, Y. K. & Li, S. Large-scale 3D chromatin reconstruction from chromosomal contacts. *BMC Genomics* **20**, 186 (2019).
47. Jia, B. B., Jussila, A., Kern, C., Zhu, Q. & Ren, B. A spatial genome aligner for resolving chromatin architectures from multiplexed DNA FISH. *Nat. Biotechnol.* **41**, 1004–1017 (2023).
48. Su, J.-H., Zheng, P., Kinrot, S. S., Bintu, B. & Zhuang, X. Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. *Cell* **182**, 1641-1659.e26 (2020).
49. Mateo, L. J. *et al.* Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature* **568**, 49–54 (2019).
50. Zhang, R., Zhou, T. & Ma, J. Multiscale and integrative single-cell Hi-C analysis with Higashi. *Nat. Biotechnol.* **40**, 254–261 (2022).
51. Chang, L. *et al.* Droplet Hi-C enables scalable, single-cell profiling of chromatin architecture in heterogeneous tissues. *Nat. Biotechnol.* (2024) doi:10.1038/s41587-024-02447-1.
52. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
53. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
54. Liu, D. C. & Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program.* **45**, 503–528 (1989).
55. Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
56. Wang, S. *et al.* HiNT: a computational method for detecting copy number variations and translocations from Hi-C data. *Genome Biol.* **21**, 73 (2020).
57. Wang, X. *et al.* Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nat. Methods* **18**, 661–668 (2021).
58. Dubois, F., Sidiropoulos, N., Weischenfeldt, J. & Beroukhim, R. Structural variations in

cancer and the 3D genome. *Nat. Rev. Cancer* **22**, 533–546 (2022).
59. Tukey, J. W. *Exploratory Data Analysis*. (Addison-Wesley Pub. Co, Reading, Mass, 1977).
60. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. Sect. A* **32**, 922–923 (1976).
61. O'Rourke, J. Finding minimal enclosing boxes. *Int. J. Comput. Inf. Sci.* **14**, 183–199 (1985).
62. Dimitrov, D., Knauer, C., Kriegel, K. & Rote, G. Bounds on the quality of the PCA bounding boxes. *Comput. Geom.* **42**, 772–789 (2009).
63. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311–316 (2020).
64. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* **3**, 99–101 (2016).
65. Luebeck, J. *et al.* AmpliconReconstructor integrates NGS and optical mapping to resolve the complex structures of focal amplifications. *Nat. Commun.* **11**, 4374 (2020).

Acknowledgments

We thank Dr. Mahidhar Tatineni and Nicole Wolter from SDSC for providing detailed guidance on running ec3D on the SDSC expanse clusters. We thank members from the Bafna, Chang, and Mischel laboratories as well as all members from team eDyNAmiC for helpful discussion and feedback. A.G.H. is supported by the *Deutsche Krebshilfe* (German Cancer Aid) Mildred Scheel Professorship program – 70114107.

Figure legends

Fig. 1 Overview of the ec3D workflow. Ec3D takes the amplicon sequence coordinates and chromatin capture (Hi-C) data as input, and outputs the 3-dimensional coordinates of each fixed-resolution bin. It also resolves the structure of duplicated regions within ecDNA. Finally, it computes and reports significant interactions between pairs of bins.

Fig. 2 Performance of ec3D on simulated data. a, A simulated 3D circular structure (ground truth) with 250 bins. **b**, A simulated Hi-C matrix generated from the structure in (**a**). **c**, The reconstructed structure computed by running ec3D on the dataset in (**b**). **d,** Distribution of RMSD values in 4 different groups: No duplication - ground-truth versus reconstructed structures without duplication; Duplication - ground-truth versus reconstructed structures with duplication; Same base - random pairs of structures with the same base structure; Different base - random pairs of structures with different base structures. Each group has n=450 samples. P-values were calculated by Wilcoxon rank-sum test for two samples (*P-value $\leq$ 8.0688e-103, **P-value $\leq$ 3.7525e-122, ***P-value $\leq$ 1.2276e-147, ****P-value $\leq$ 5.3505e-148). **e,** Distribution of PCC values in 4 different groups. Each group has n=450 samples. P-values were calculated by Wilcoxon rank-sum test for two samples (*P-value $\leq$ 5.7944e-112, **P-value $\leq$ 2.7649e-140, ***P-value $\leq$ 1.2626e-148, ****P-value $\leq$ 5.9450e-149). **f**, PCC values of ground truth versus reconstructed structures (PCC-reconstructed) compared to PCC values of

ground truth versus random structures (PCC-random). A data point at the bottom right of the red dashed line indicates that the reconstructed structure is more similar to the ground truth than a random structure. **g**, Distribution of RMSD values (ground truth versus reconstructed structures) over α values. **h, i,** Running time of ec3D for reconstructing structures without duplication (**h**) and with duplication (**i**). The y-axis represents the proportion of simulated samples (each plot has 450 in total) whose reconstruction was completed by a specific time point.

Fig. 3. Structural properties of ecDNA. a, Correlation between Hi-C interaction frequencies and spatial (Euclidean) distances from GBM39. The Spearman and Pearson correlation coefficients suggested a negative power law decay of interaction frequencies on spatial distances. Color gradient representing the genomic distance was overlaid on each scatter plot point, with warmer colors indicating shorter genomic distances and cooler colors indicating longer distances. **b,** 3D structure of GBM39 ecDNA with oncogenes amplified on the ecDNA. Genes are highlighted, and red crosses represent structural variations leading to ecDNA formation**. c,** 3D structure of the same chromosomal segment (Chr7:54.7M-56.1M) on a control cell line, GM12878, reconstructed by ec3D. **d,** Size of the minimum volume bounding box enclosing GBM39 ecDNA, suggesting an oblate spheroidal structure. **e,** Optimal values of ec3D's objective function (y-axis) after fixing the scaling parameter β and limiting the maximum range of the first axis to force a flatter structure (x-axis). A box plot describing the 5 final objective values with random initialized X was made for each length limit ranging from 0.03125 - 2. **f,** 3D structure of GBM39HSR reconstructed by ec3D. **g,** 3D structure of H2170 ecDNA. The duplicated segment Chr8:128.4M-128.9M showed significant similarity (RMSD = 0.2759, permutation test P-value = 0.01). **h**, a segment from chromosome 2:15.58M-15.98M amplified on two different cyclic (ecDNA/HSR) structures with significant similarity (RMSD = 0.2369, permutation test P-value = 0.007).

Fig. 4 Significant interactions on ecDNA. a, 3D structure of CHP-212 ecDNA. Dashed box encloses the *MYCN* Neo-TAD in CHP-212 connecting 3 distal segments from Chr2, and the dashed line divides a region with overexpressed genes (bottom) from the other region with base-level gene expression (top). **b,** Distribution of number of significant interactions (ref-SI) as a function of reference genomic distances in RCMB56 ecDNA, compared to control cell lines GM12878 and IMR90. **c,** Representative crossing interactions on D458 structure shown with green dotted lines. **d,** Differential circ-SI interactions that are spatially distant, suggesting trans-interactions in RCMB56. Both interactions connect an active enhancer at *SELENOF* locus to remote oncogenes (*DNTTIP2* and *ABCD3*). **e,** Differential interactions in D458 connecting *MYC* and a remote enhancer on Chr14.

Supplementary Fig. 1. Simulating samples with topological constrictions. a, A simulated structure with k=1 topological constriction and its corresponding expanded Hi-C data and reconstructed structure. **b,** k=2. **c,** k=3.

Supplementary Fig. 2. Simulation test results grouped by the number of topological constrictions in simulated structures. The RMSD and PCC values are calculated by comparing ground truth and reconstructed structures in simulated data.

Supplementary Fig. 3. Results of simulation tests with duplication that have the same or different local substructures.

Supplementary Fig. 4. RMSD values of ground truth versus reconstructed structures (RMSD1) and RMSD values ground truth versus random structures (RMSD2) in simulation tests. The data points at the top left of the red dashed line indicate that the reconstructed structures have higher correlation with the ground truth than random structures.

Supplementary Fig. 5. Estimated α values versus ground truth α values in simulation tests.

Supplementary Fig. 6. Updates on objective values, RMSD, and PPC during Poisson optimization on simulated samples.

Supplementary Fig. 7. Final RMSD (and PCC) values versus initial RMSD (and PCC) values in simulation tests. A majority of RMSD and PCC values were improved through optimization, showing smaller RMSD and larger PCC compared with initial values.

Supplementary Fig. 8. Running time as a function of sequence length. The sequence length is measured as the number of bins, where each bin corresponds to a 5 Kbp region. n=150 samples were simulated for each category.

Supplementary Fig. 9. Correlation between Hi-C interaction frequencies (x-axis) and spatial (Euclidean) distances (y-axis), from CHP-212 (a), D458 (b), GBM39 (c), GBM39 HSR (d), H2170 (e), IMR-5/75 HSR (f), and RCMB56 (g). The Spearman and Pearson correlation coefficients are computed separately for high interaction frequencies (the largest quantile, marked by 'x'), and low interaction frequencies (the first 3 quantiles, marked by 'o'). Color gradient representing the genomic distance was overlaid on each scatter plot point, with warmer colors indicating shorter genomic distances and cooler colors indicating longer distances.

Supplementary Fig. 10. Optimal values of ec3D's objective function (y-axis) for RCMB56 when fixing the scaling parameter β and limiting with the maximum range of the first axis (x-axis). A box plot describing the 5 final objective values with random initialized X was made for each length limit ranging from 0.03125 - 2. The other two axes remain in the default range [-1, 1].

Supplementary Fig. 11. Similarities of ecDNA and HSR structures from isogenic cell lines GBM39 and GBM39HSR. a, b, c, normalized Hi-C interaction frequencies and their

correlations. x axis: interaction frequency of GBM39EC; y axis: the corresponding interaction frequency of GBM39HSR. **d,** Euclidean distances between each pair of 5K bins in ec3D reconstructions. x axis: Euclidean distances on GBM39EC structure; y axis: the corresponding distances on GBM39HSR structure. Bin pairs showed significantly larger or smaller distances between the two structures are marked by green. Black arrow points to the example bin pair in panels **e** and **f**. **e, f,** ec3D reconstructions of GBM39 ecDNA and GBM39 HSR, with an example of notable differences indicated by the green lines.

Supplementary Fig. 12. EcDNA and potential HSR models. a, A stacked model of ecDNA with the collapsed matrix. **b,** A 'spring-like' HSR model with the collapsed matrix has a signal very similar to the stacked ecDNA. **c,** A 'petal-like' HSR model with the collapsed matrix also provides end-to-end contacts but has fewer additional interactions. Each color in the structures represents one copy of duplication.

Supplementary Fig. 13. Collapsed matrix (a) and expanded matrix (b) of IMR-5/75 (HSR). Both matrices display a two TAD structure, where the second TAD incorporates a joining of the last and the first segment of the cyclic structure, supporting a tandem duplication model of HSR. Ec3D reconstruction and the expanded matrix clarify the first TAD.

Supplementary Fig. 14. Single cell expression level of genes (left) and lncRNAs (right) amplified on CHP-212 ecDNA.

Supplementary Fig. 15. Ec3D reconstruction of D458 (a) better clarifies the (sub)structure of a neo-TAD (b). The neo-TAD involves an inversion of a segment 56.8-57Mb from Chr14, which brings together a distal enhancer DHS 1 and *OTX2*, but not *MYC* as suggested by ec3D.

Supplementary Fig. 16. Cumulative distribution of significant interactions (ref-SI, y-axis) as a function of increasing genomic distance (x-axis). Both GBM39 and RCMB56 showed more long range interactions (a, b) than controls GM12878 and IMR90, corresponding to slower increases in the cumulative distributions; and similar trends of significant interactions (c, d) with upscaled and downscaled matrices. Note that in RCMB56, downscaling to 0.5x breaks the Negative Binomial property in most of the genomic distances, and as such we omitted that curve.

Supplementary Fig. 17. Cartoon illustration of "crossing" interactions in a 3D structure comprising two topological constrictions (TCs).

Supplementary Fig. 18. The clique-like interactions (a) involving Chr8 and Chr14; and the star-like interactions (b) centered at *MYC/PVT1* locus, in D458.

Supplementary Fig. 19. Differential interactions in GBM39. There are no obvious differences (c) between circ-SI (a) and spatial-SI (b), except a few loops (i.e., TAD boundaries).

Supplementary Fig. 20. Differential interactions in RCMB56. The most remarkable differential interactions (c) that occur in circ-SI (a) but not spatial-SI (b) all involve an active enhancer at *SELENOF* locus. These interactions suggest potential trans interactions, due to the lack of spatial proximity in our 3D structure reconstruction.

Supplementary Fig. 21. Differential interactions in D458, excluding interactions from all duplicated segments. The most remarkable differential interactions (c) that occur in circ-SI (a) but not spatial-SI (b) are between the MYC locus on Chr8 and some enhancer region on Chr14. These interactions suggest potential trans interactions, due to the lack of spatial proximity in our 3D structure reconstruction.

Supplementary Methods

Base structures. We simulated three circular base structures with k ($k \in \{1, 2, 3\}$) topological constrictions. Each base structure can be represented by a parametric function with a variable $\theta \in [0, 2\pi]$ and hyperparameter $p \in [0.90, 0.99]$ that controls the interaction frequency of the constrictions. If p is high, the constrictions contribute stronger interactions, and vice versa.

- $k = 1$
  - $x = \cos(\theta)$
  - $y = \sin(\theta) - p\sin(\theta)^4$ if $\sin(\theta) > 0$ else $\sin(\theta) + p\sin(\theta)^4$
  - $z = \cos(\theta)^2$
- $k = 2$
  - $x = \cos(\theta) - p\cos(\theta)^4$ if $\cos(\theta) > 0$ else $\cos(\theta) + p\cos(\theta)^4$
  - $y = \sin(\theta) - p\sin(\theta)^4$ if $\sin(\theta) > 0$ else $\sin(\theta) + p\sin(\theta)^4$
  - $z = \sin(\theta)^2$
- $k = 3$
  - $x = \cos(\theta) \cdot \frac{1}{4}(1 + p\cos(3\theta))$
  - $y = \sin(\theta) \cdot \frac{1}{4}(1 + p\cos(3\theta))$
  - $z = \frac{1}{4}\sin(3\theta)$

Given a set of coordinate functions, we generated 10,000 3D points with $\theta_i = \frac{i}{10000} \cdot 2\pi$, where $i = 0, 1, \ldots, 9999$, and we calculated the arc length between each pair of adjacent points by $l_i = \int_{\theta_i}^{\theta_{i+1}} \sqrt{f'(\theta)^2 + g'(\theta)^2 + h'(\theta)^2}\, d\theta$. Then we got 10,000 data points (l_i, θ_i) and used SciPy to interpolate the function F with respect to $\sum_{t=0}^{i} l_i$ and θ_i. As a result, we can generate n points

that are evenly spaced on the 3D structure by calculating $\theta_i = F(\frac{i}{n}L)$ for $i = 0, 1, ..., n$, where L is the total arc length, i.e., $L = \int_0^{2\pi} \sqrt{f'(\theta)^2 + g'(\theta)^2 + h'(\theta)^2}\, d\theta$.

Local folds. Given a pair of adjacent points X_i and X_{i+1} on a base structure, we added local folds between the two points by random walks. We denote difference between the two points as a vector $\vec{d} = X_{i+1} - X_i$. If we want to generate a local fold with m bins, where m is an even number, then we have the average distance vector $\vec{\Delta} = \frac{2 \cdot \vec{d}}{m}$. Given the step length $l = L/n$, we can generate a step vector $\vec{s} = (l \cdot \sin\theta \cdot \cos\phi, l \cdot \sin\theta \cdot \sin\phi, l \cdot \cos\phi)$ by randomly sampling θ and ϕ from the range $[0, 2\pi]$. To guarantee that the local fold starts from X_i and ends at X_{i+1}, we need to generate a complementary step vector $\vec{s'} = \vec{\Delta} - \vec{s}$. As a result, we can generate a set of m setp vectors $S = \{\vec{s}_1, \vec{s}_2, ..., \vec{s}_m\}$, where $\vec{s}_j + \vec{s}_{m/2+j} = \vec{\Delta}, \forall j \in \{1, 2, ..., m/2\}$. To generate a local fold with m points, we can simply add the accumulated sum of step vectors to the starting point X_i, i.e., $P = \{X_i + \sum_{k=1}^{j} s_k \mid j = 1, 2, ..., m\}$. Note that $X_i + \sum_{s \in S} s = X_{i+1}$ since $\vec{d} = \sum_{s \in S} s$. In this way, we can generate multiple local folds with m bins, where $m \in \{16, 18, 20, 22\}$, and the adjacent local folds have the space in terms of the number of bins d, which is either 4 or $22 + 4 \cdot (k - 1)$. The smaller value makes adjacent local folds form as a TAD; the greater value makes two local folds far apart.

Hi-C simulation. Given a 3D structure, we simulated pairwise Hi-C interaction frequencies by randomly sampling integers from the Poisson distribution with mean βd_{ij}^{α}, where d_{ij} represents the Euclidean distance between bin i and bin j. Furthermore, we simulated duplication by designating n ($n \in \{1, 2, ..., ceil(N_e/10)\}$) pairs of local folds as duplicated regions. Then we summed up the Hi-C interactions of all duplicated bins, assigned the results to the copy with the least index, and remove the other copies in the Hi-C matrix.

Figure 1

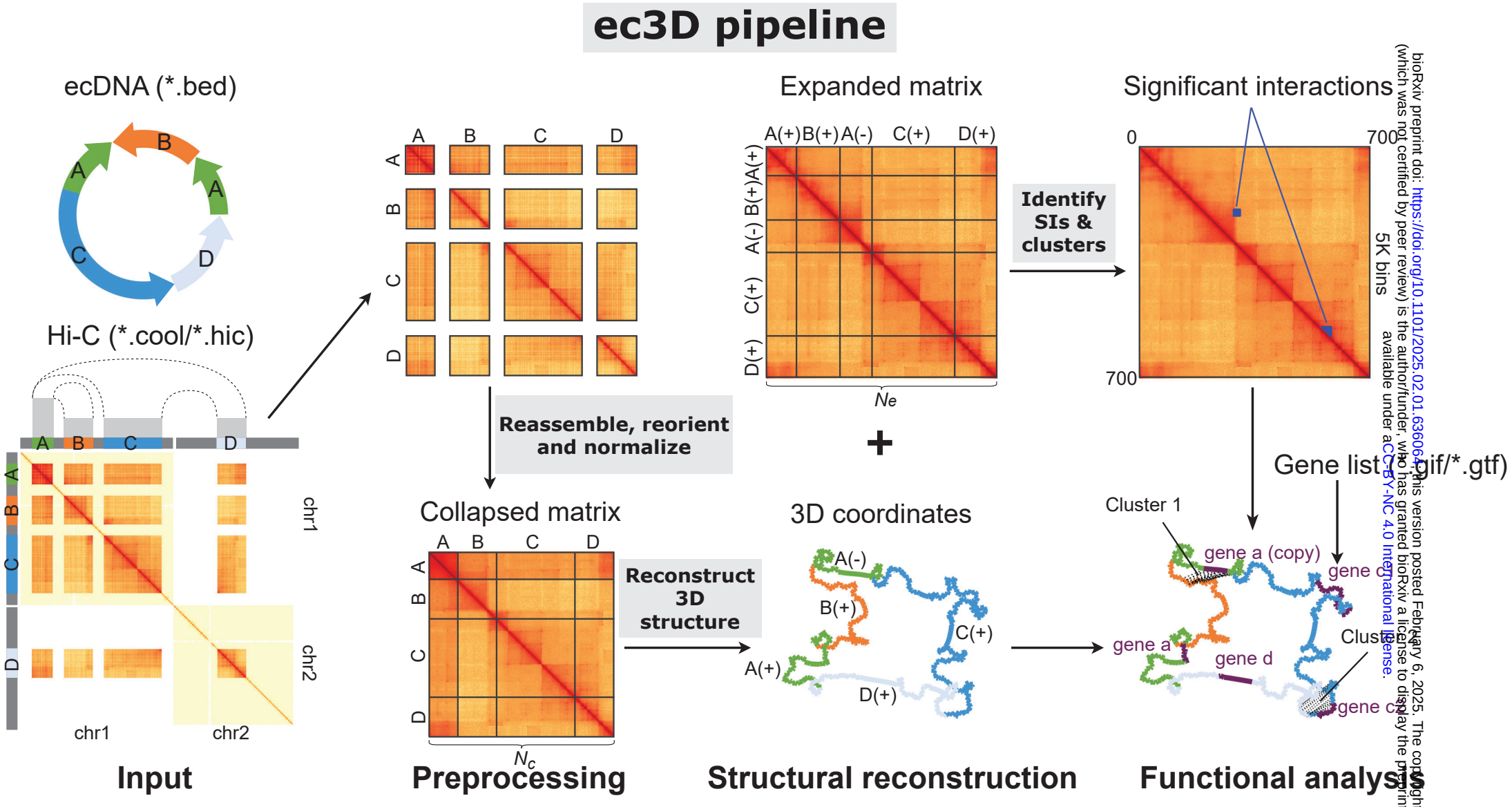


Figure 2

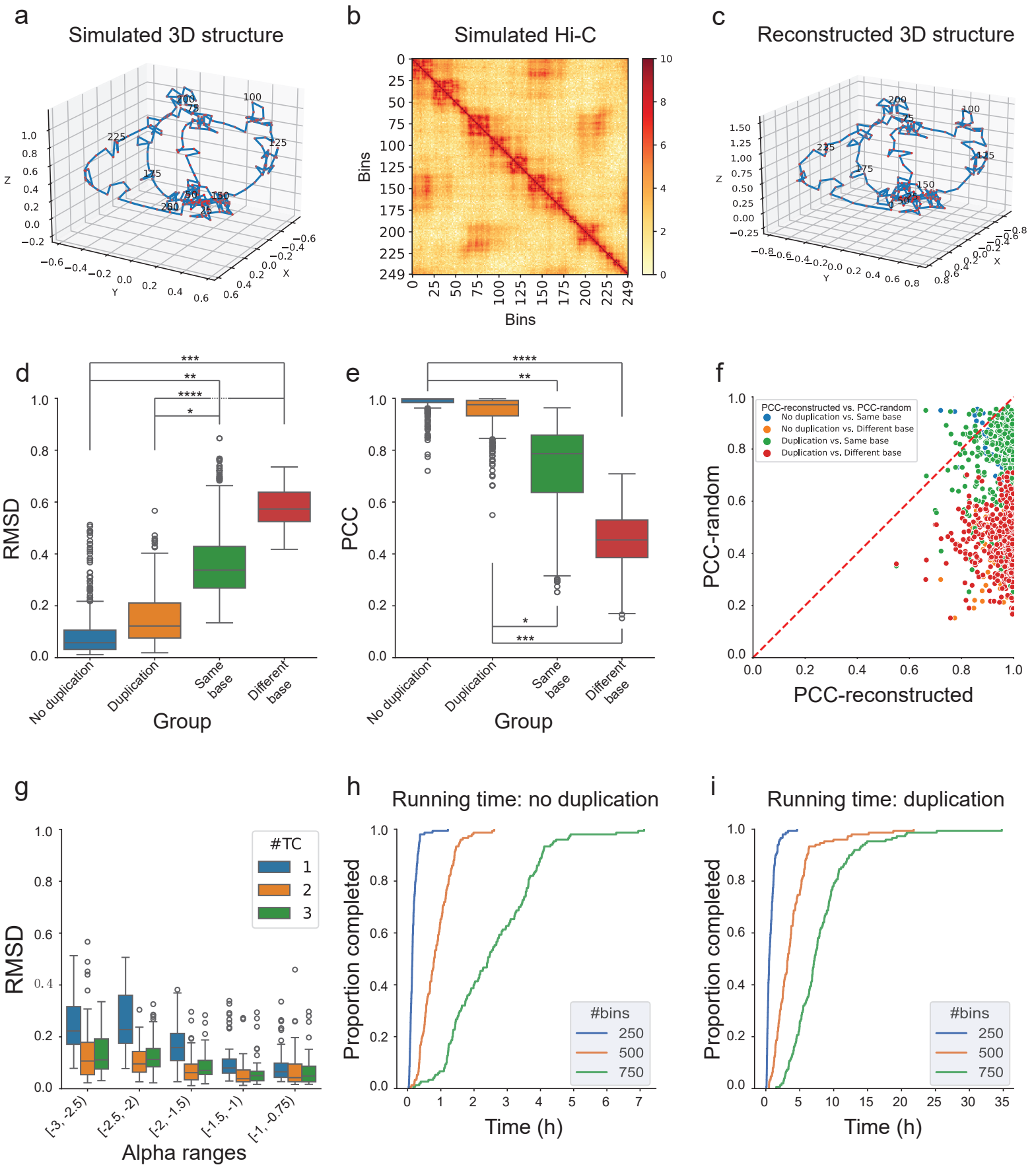

a
Simulated 3D structure
b
Simulated Hi-C
Bins
Bins
c
Reconstructed 3D structure
d
RMSD
No duplication
Duplication
Same base
Different base
Group
e
PCC
No duplication
Duplication
Same base
Different base
Group
f
PCC-reconstructed vs. PCC-random
No duplication vs. Same base
No duplication vs. Different base
Duplication vs. Same base
Duplication vs. Different base
PCC-random
PCC-reconstructed
g
RMSD
#TC
1
2
3
[-3, -2.5)
[-2.5, -2)
[-2, -1.5)
[-1.5, -1)
[-1, -0.75)
Alpha ranges
h
Running time: no duplication
Proportion completed
#bins
250
500
750
Time (h)
i
Running time: duplication
Proportion completed
#bins
250
500
750
Time (h)

Figure 3

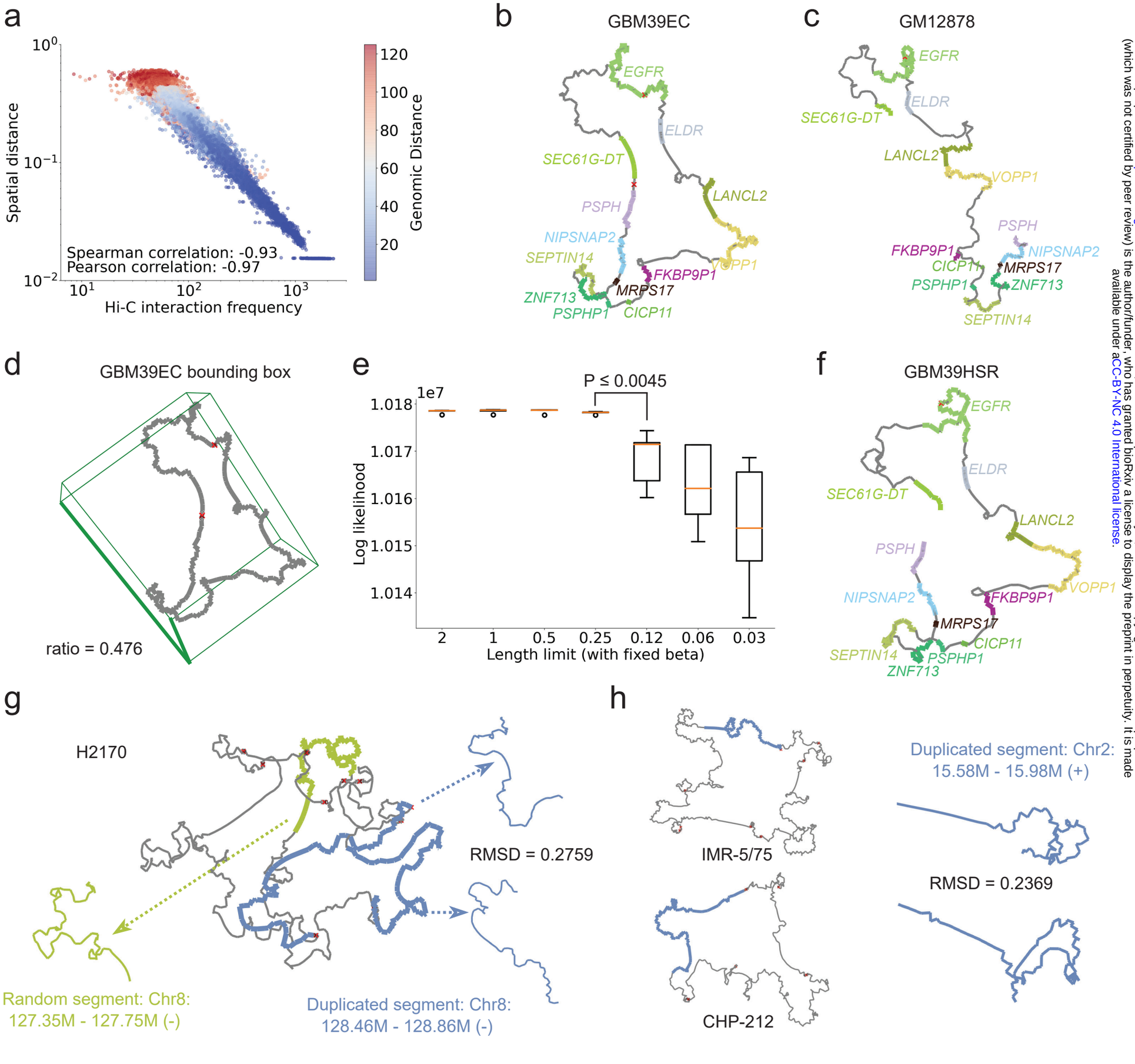
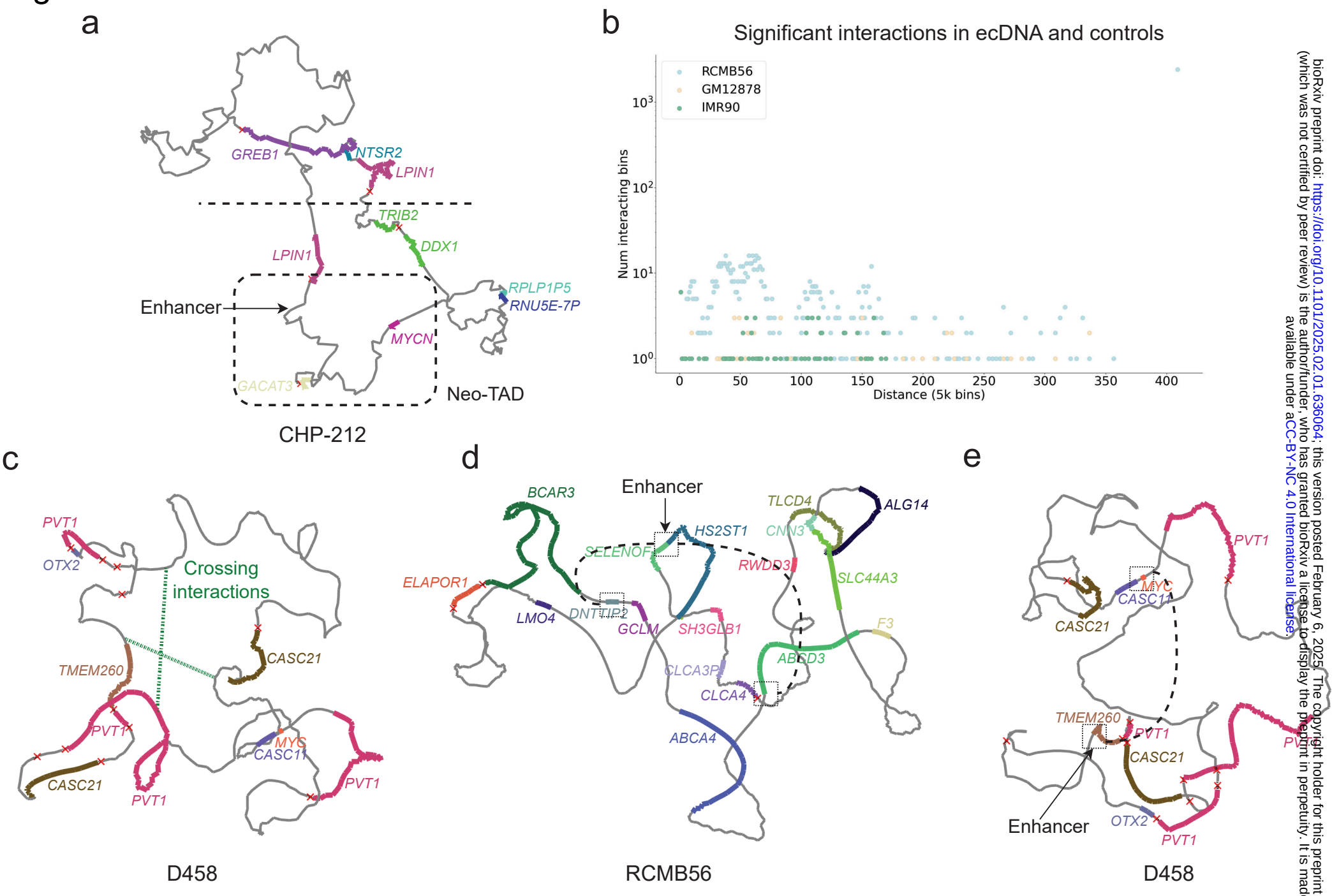
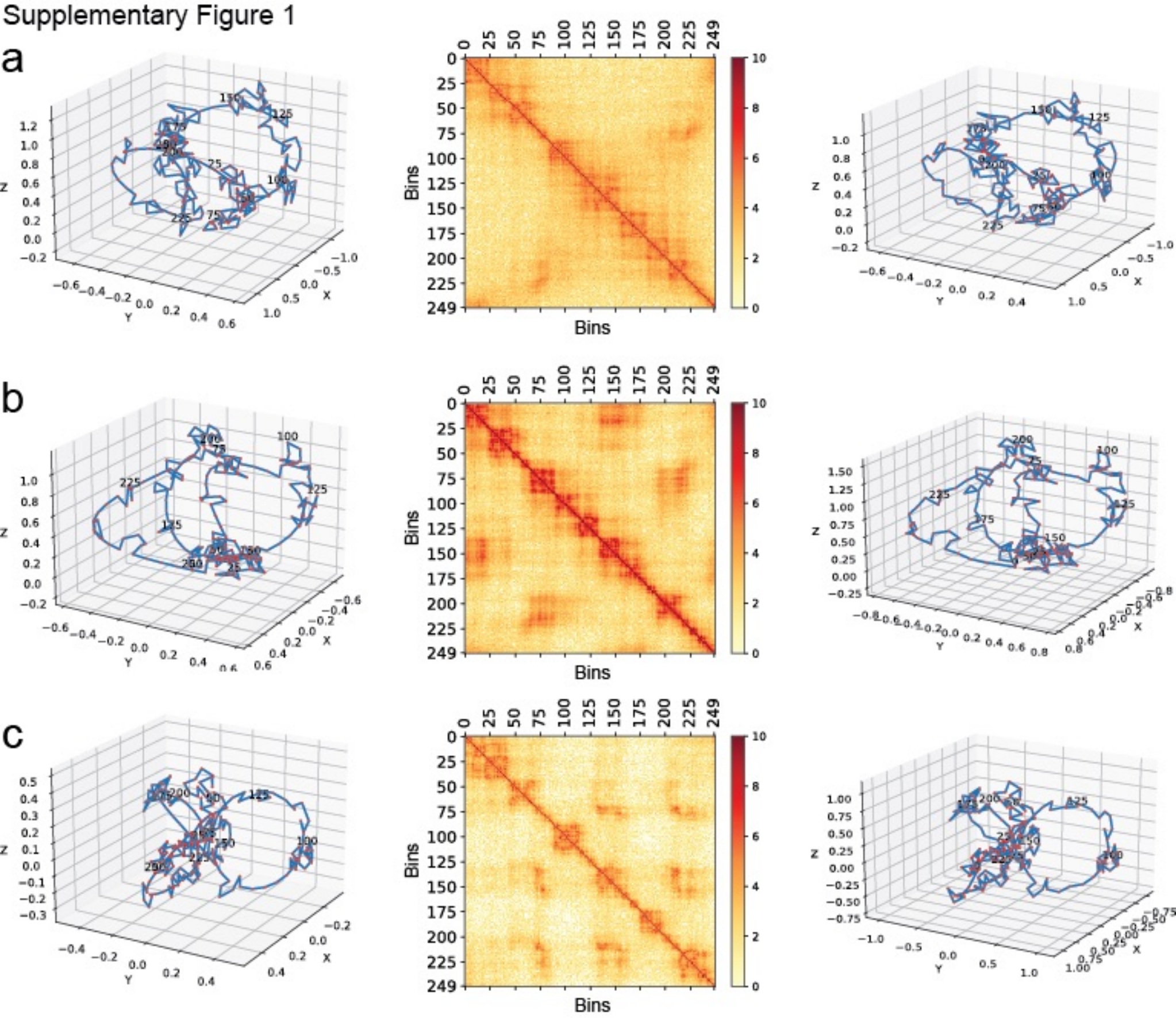


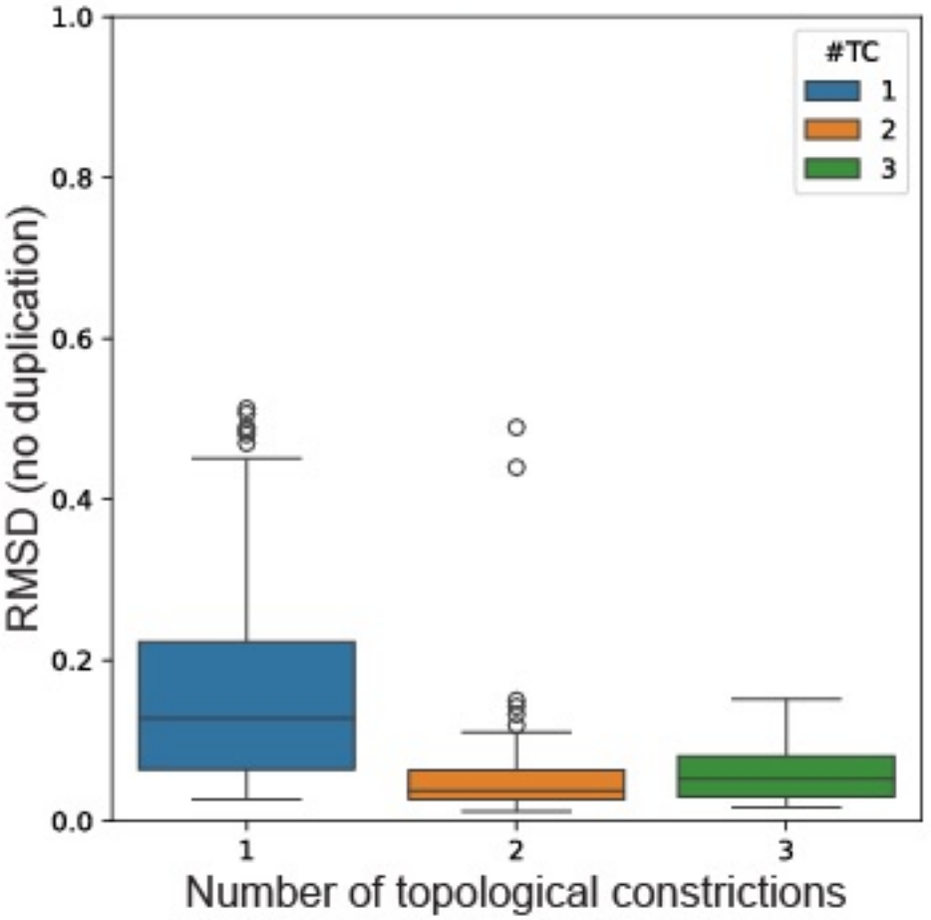
Figure 4

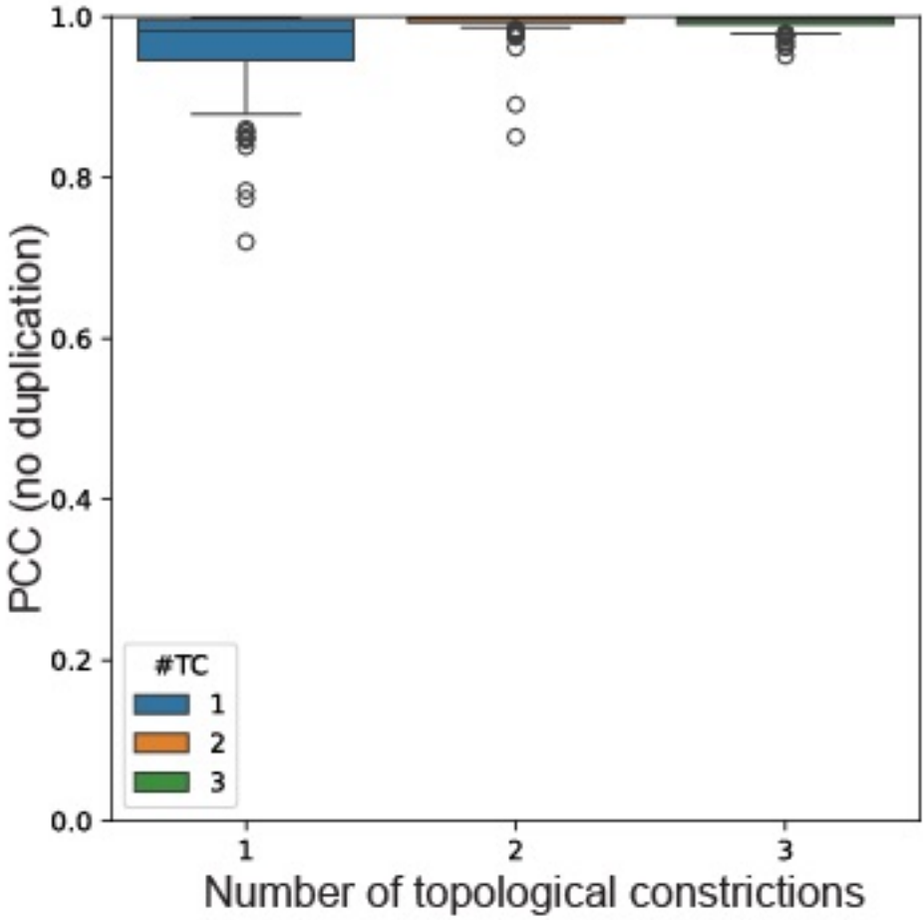


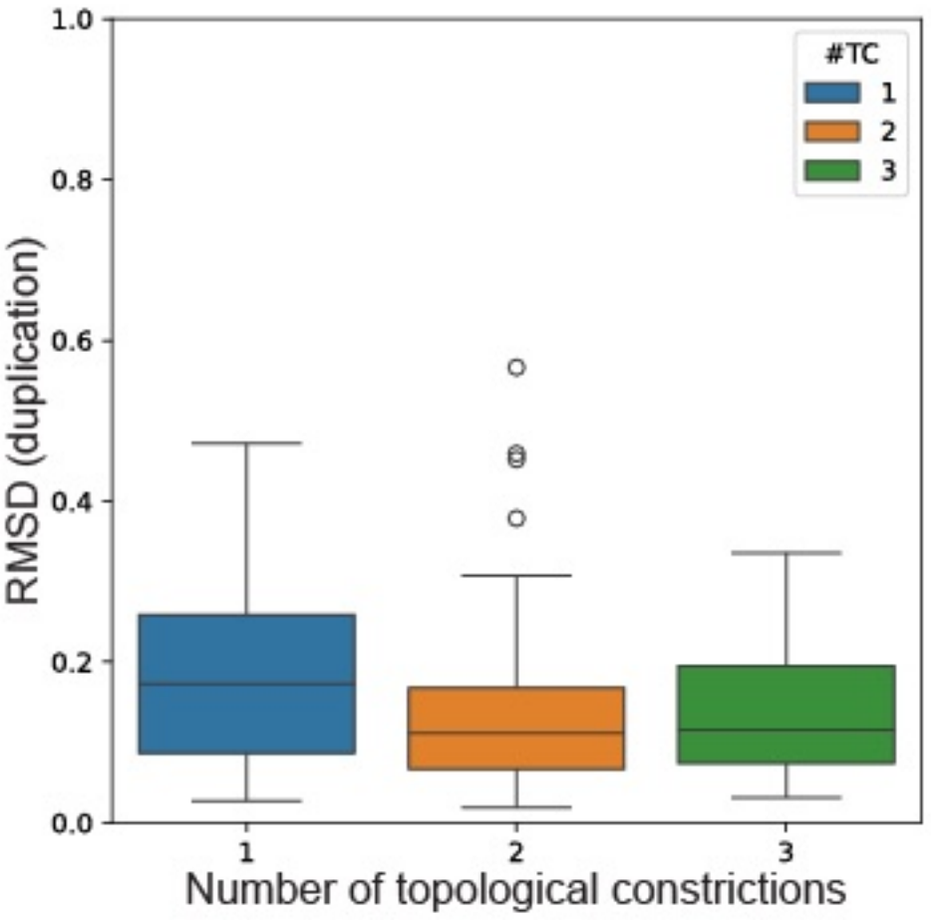
Supplementary Figure 1

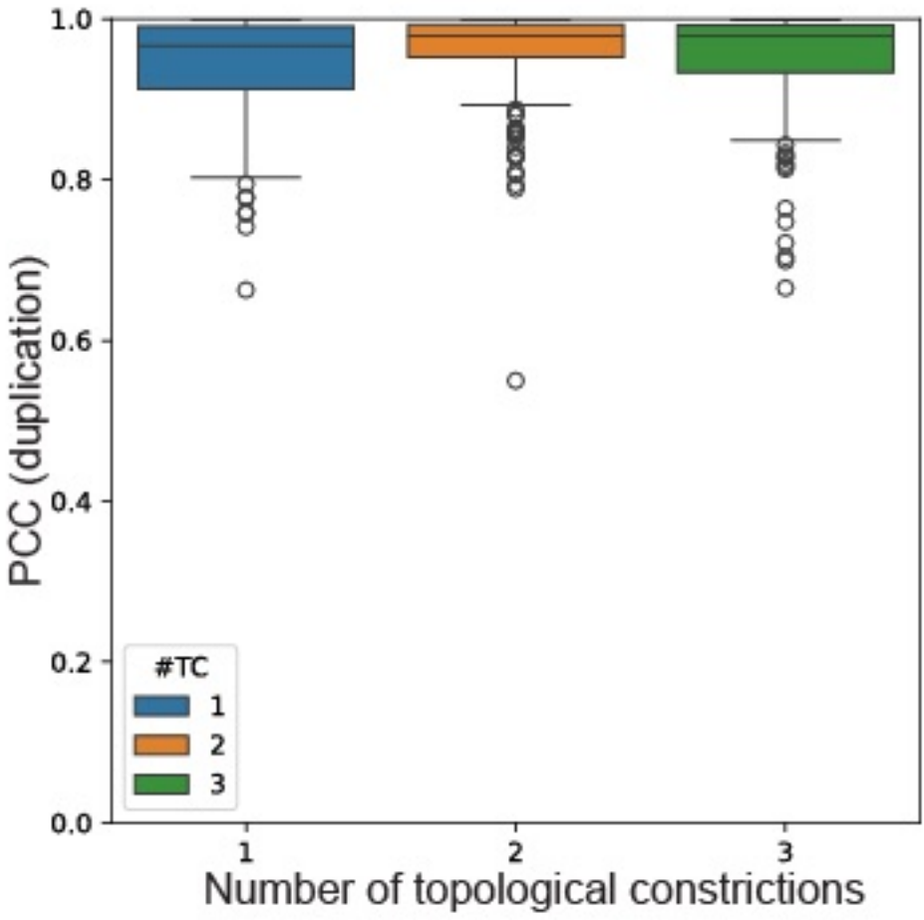


Supplementary Figure 2

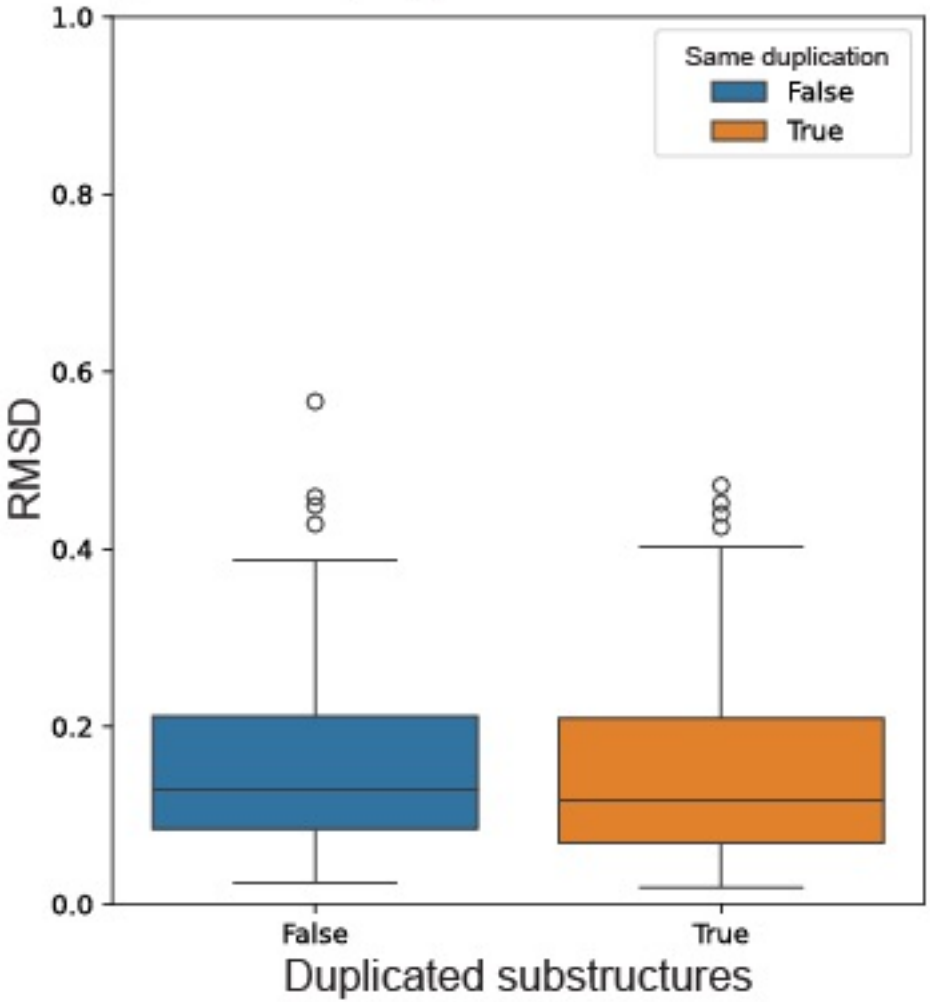


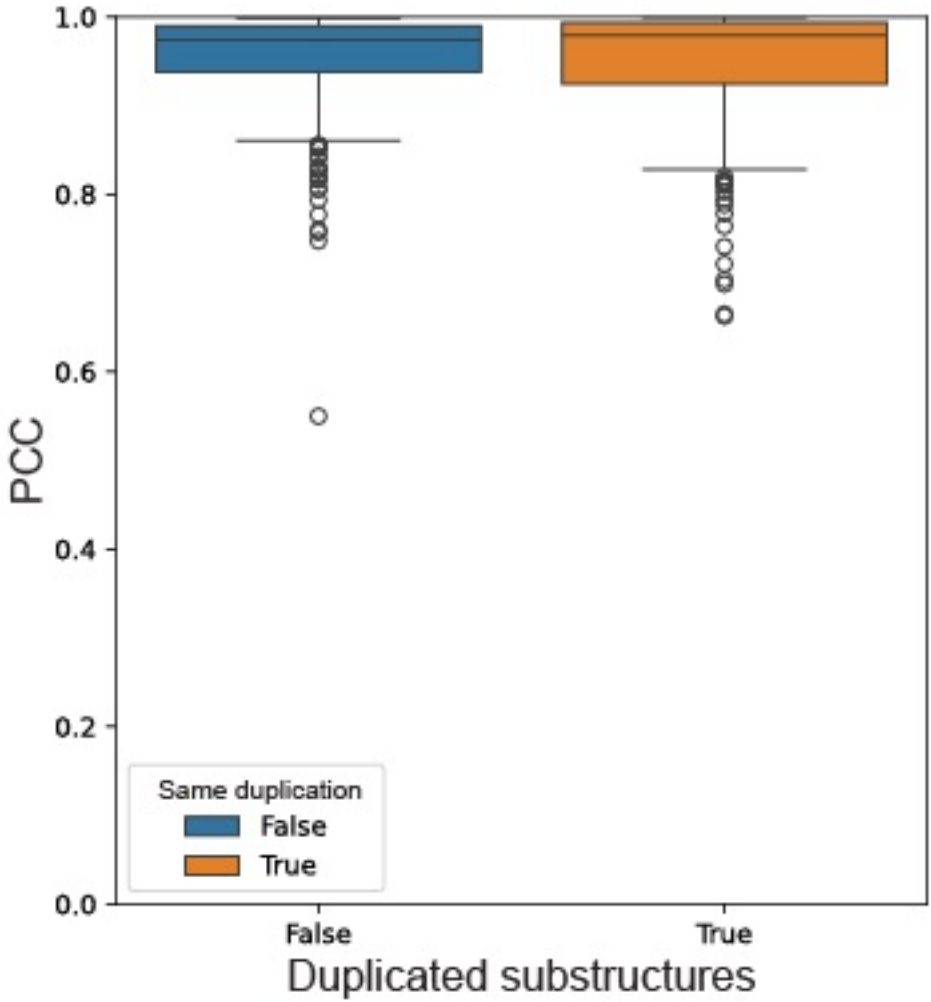




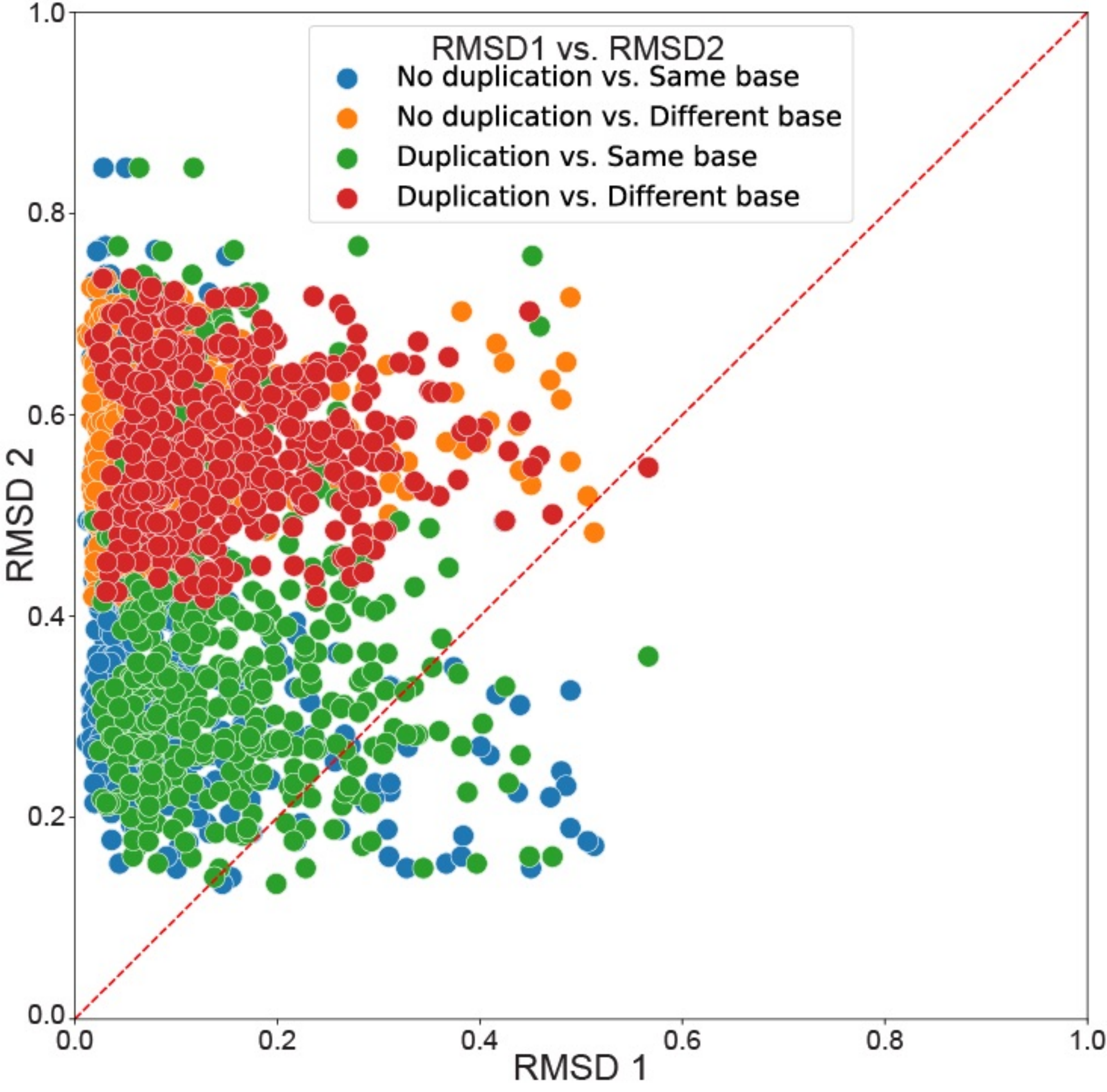


Supplementary Figure 3

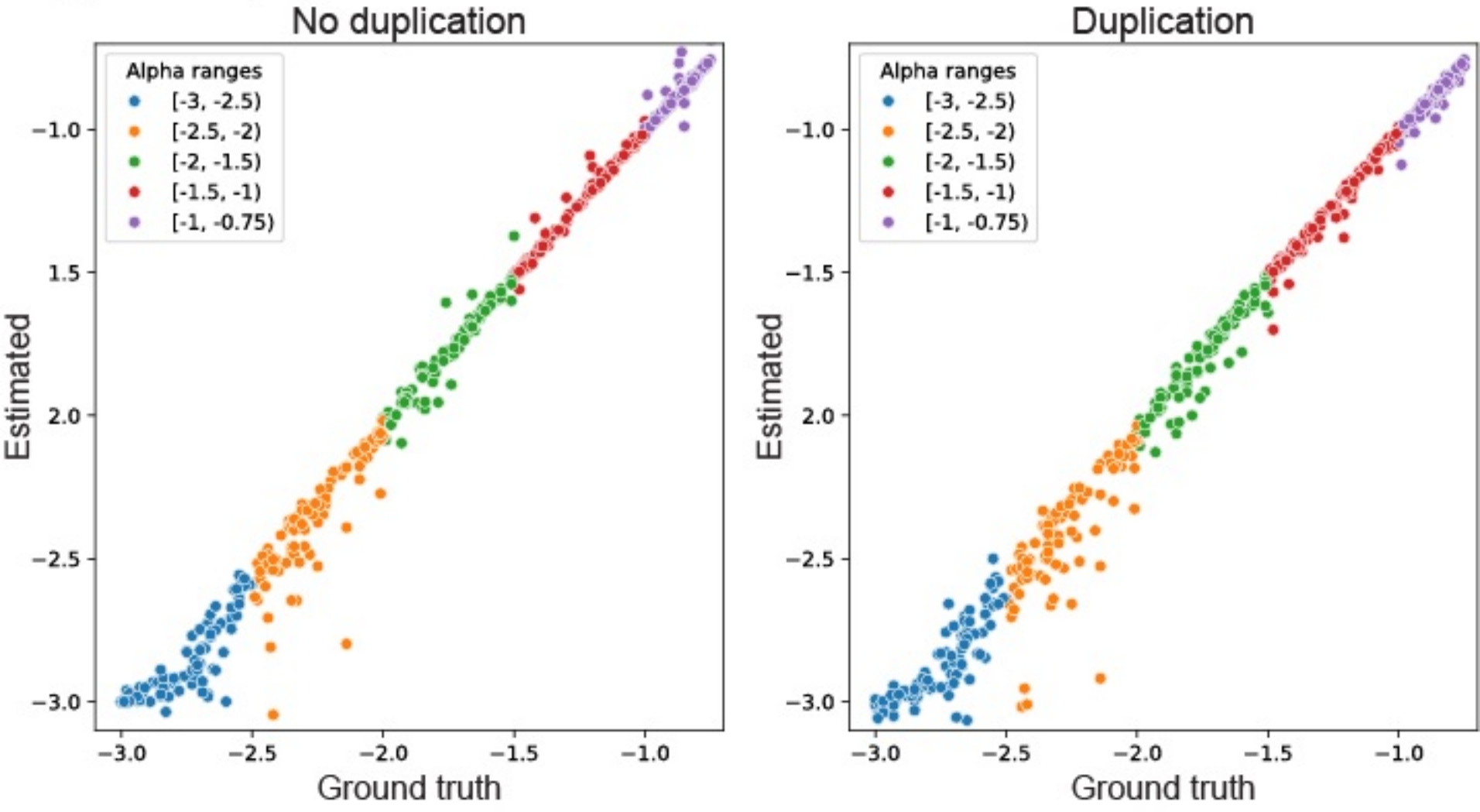




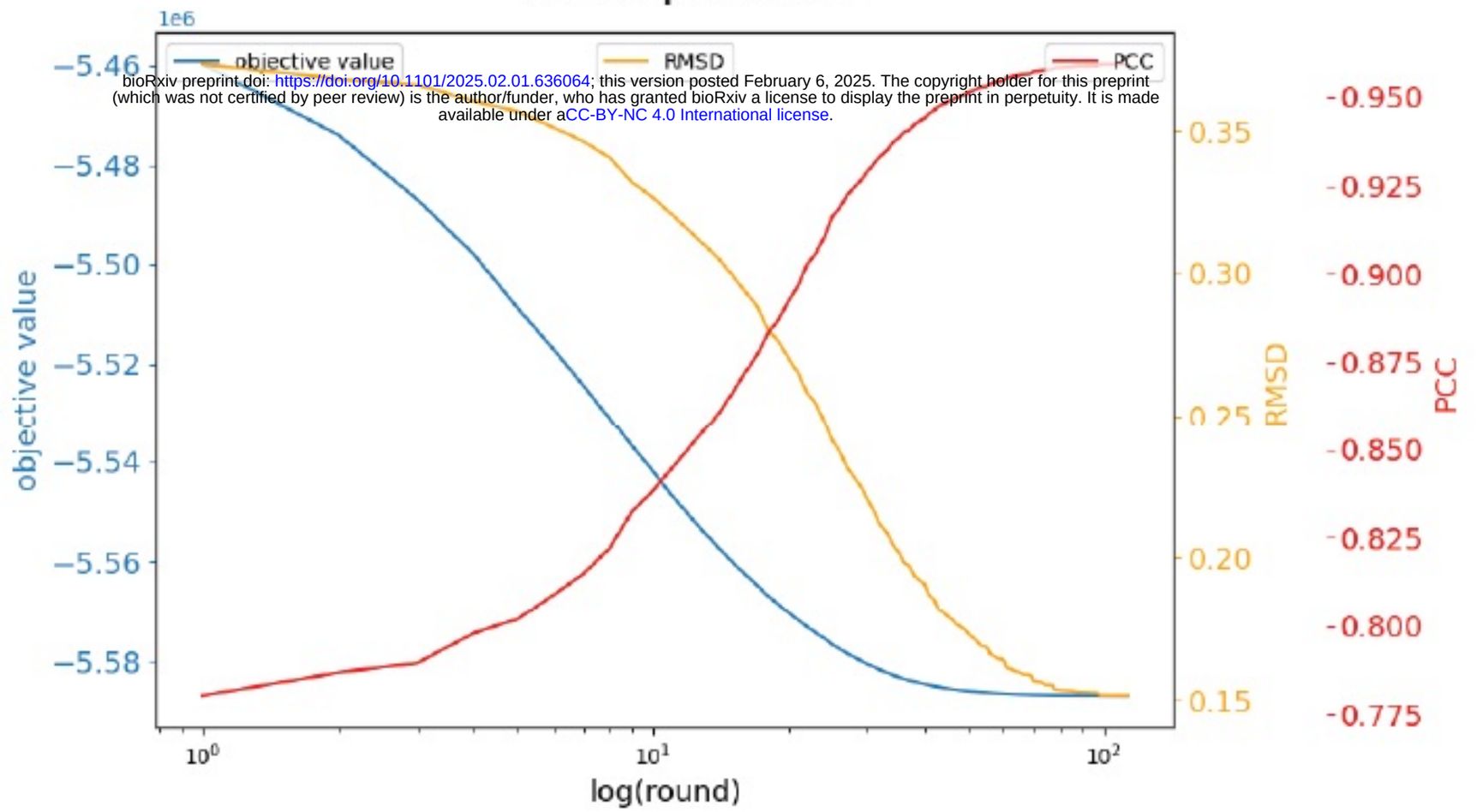
Supplementary Figure 4

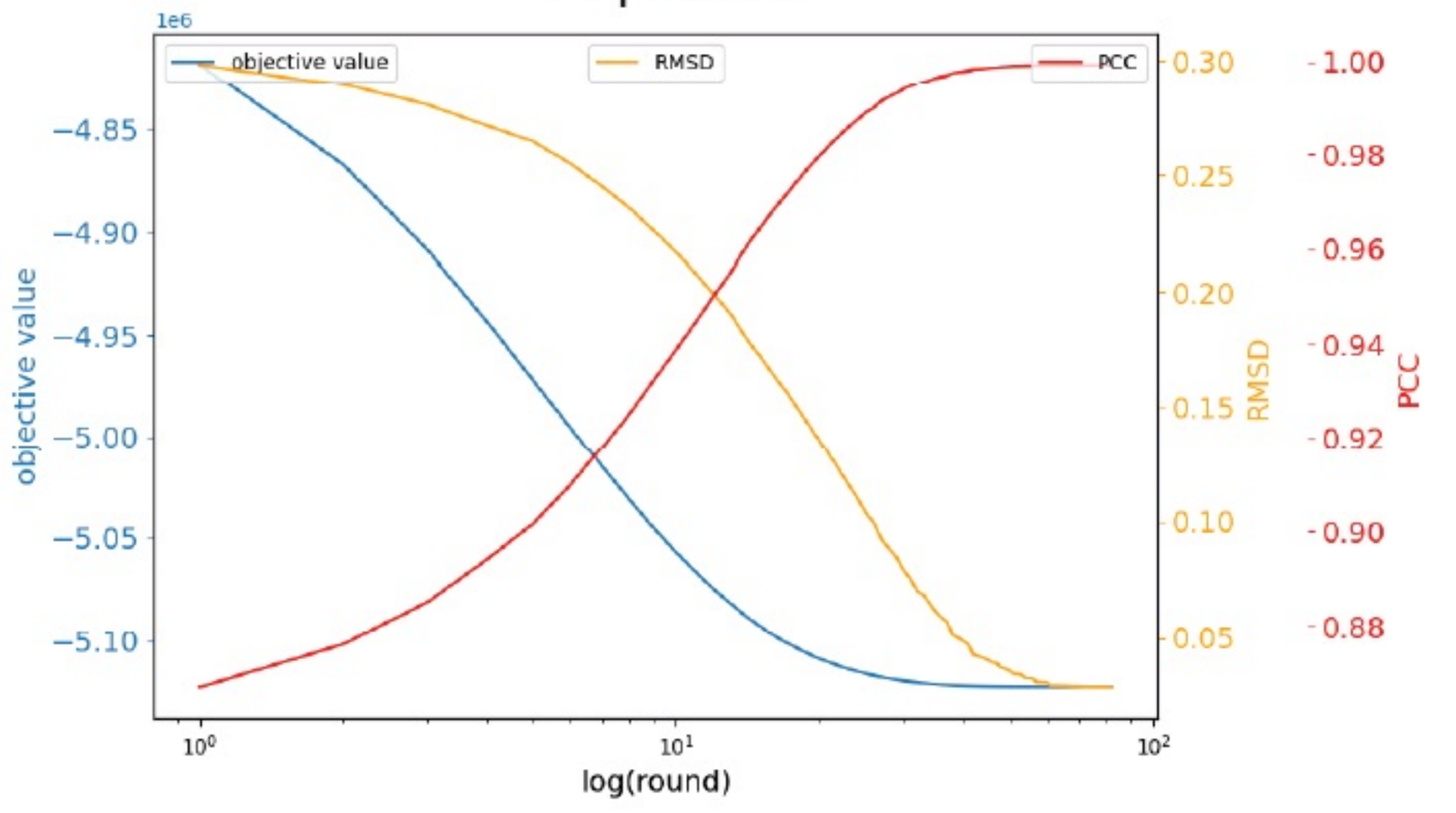


Supplementary Figure 5

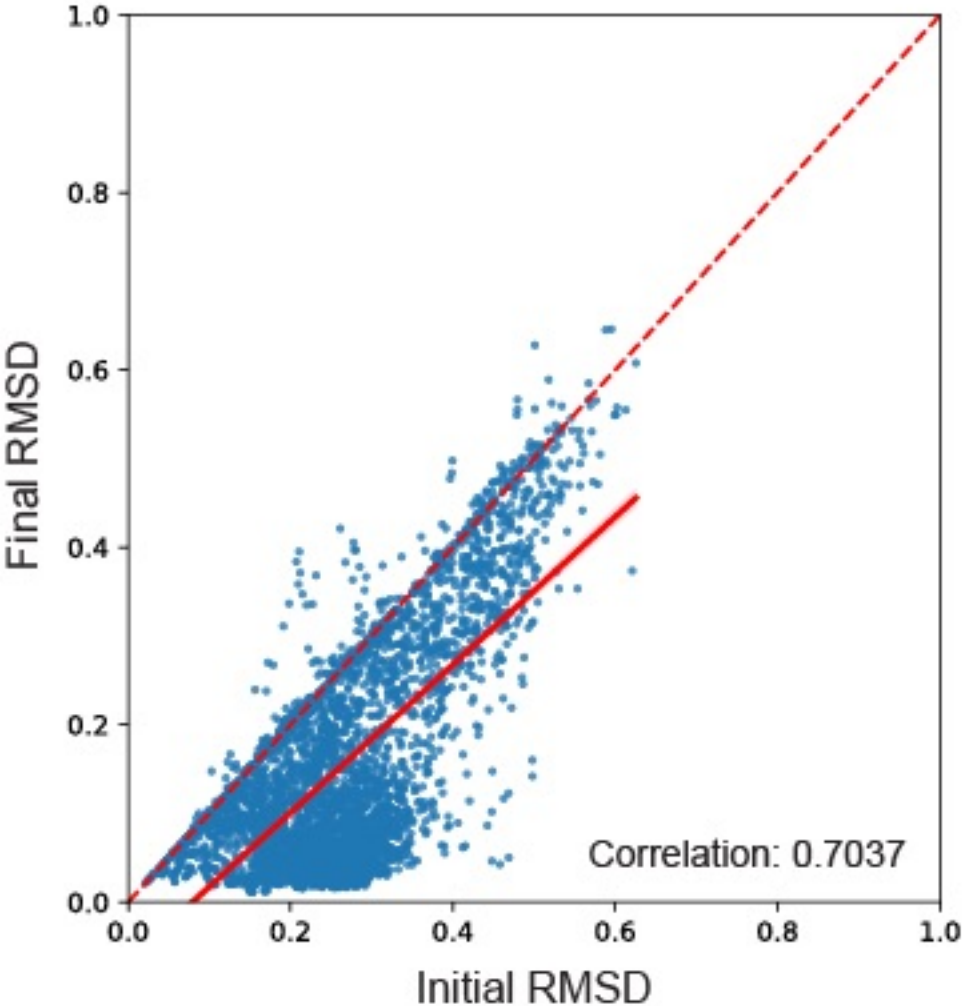


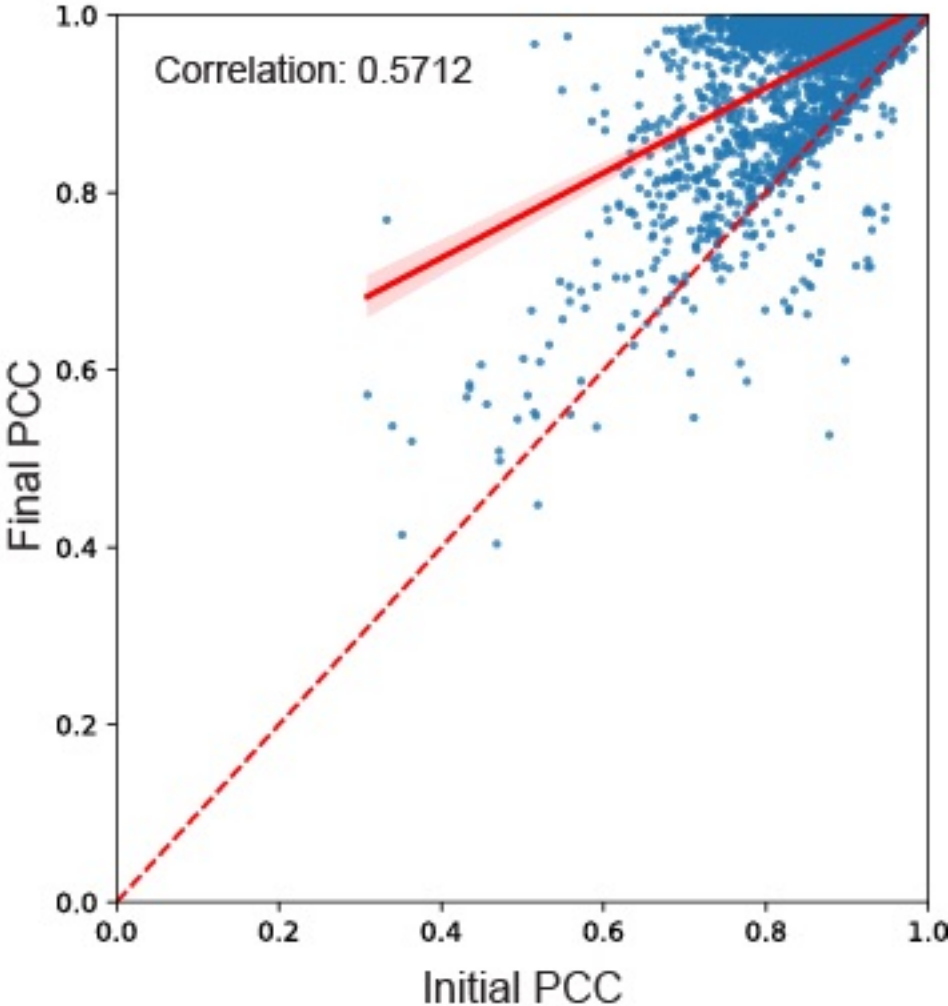
Supplementary Figure 6



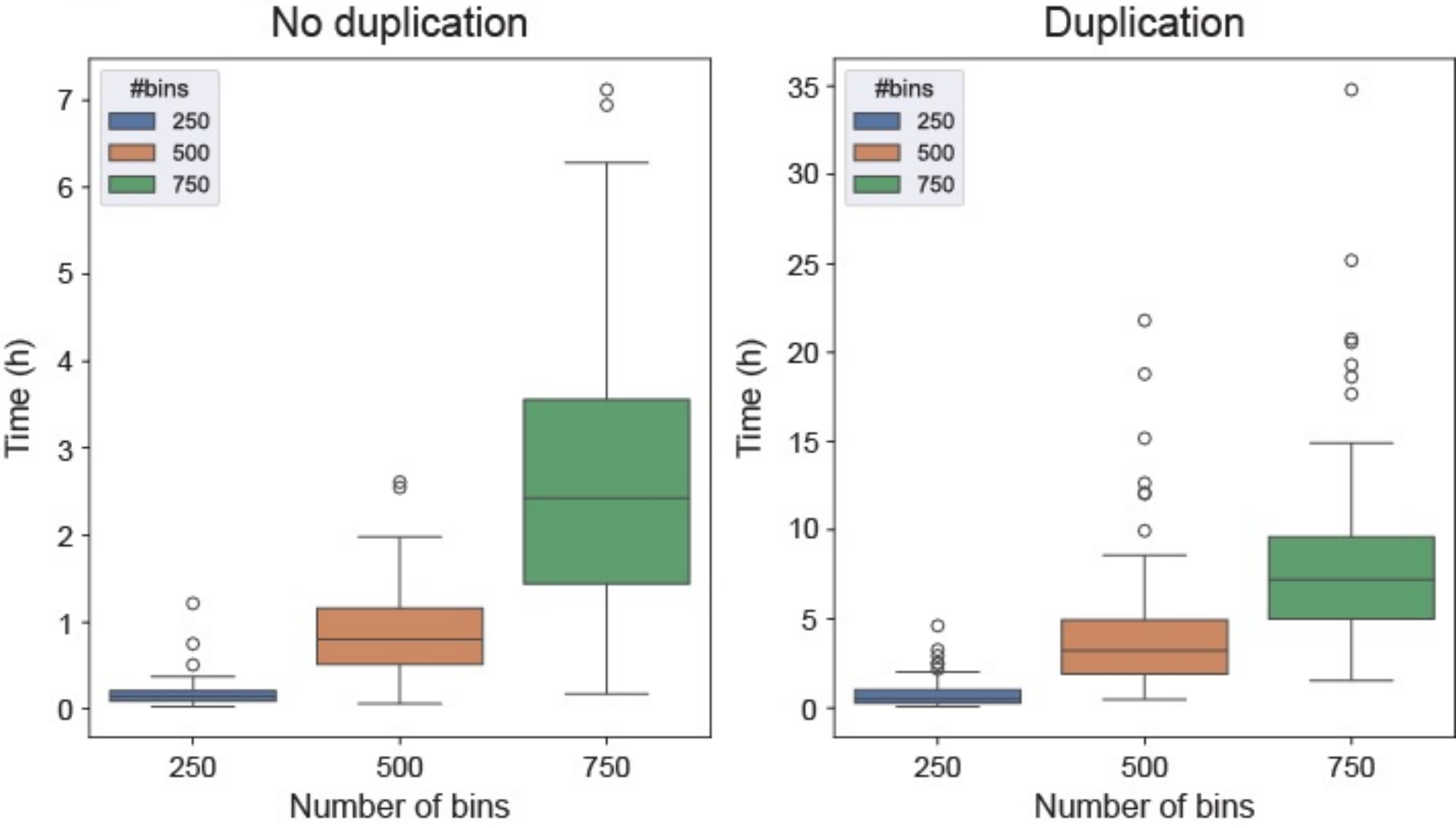


Supplementary Figure 7

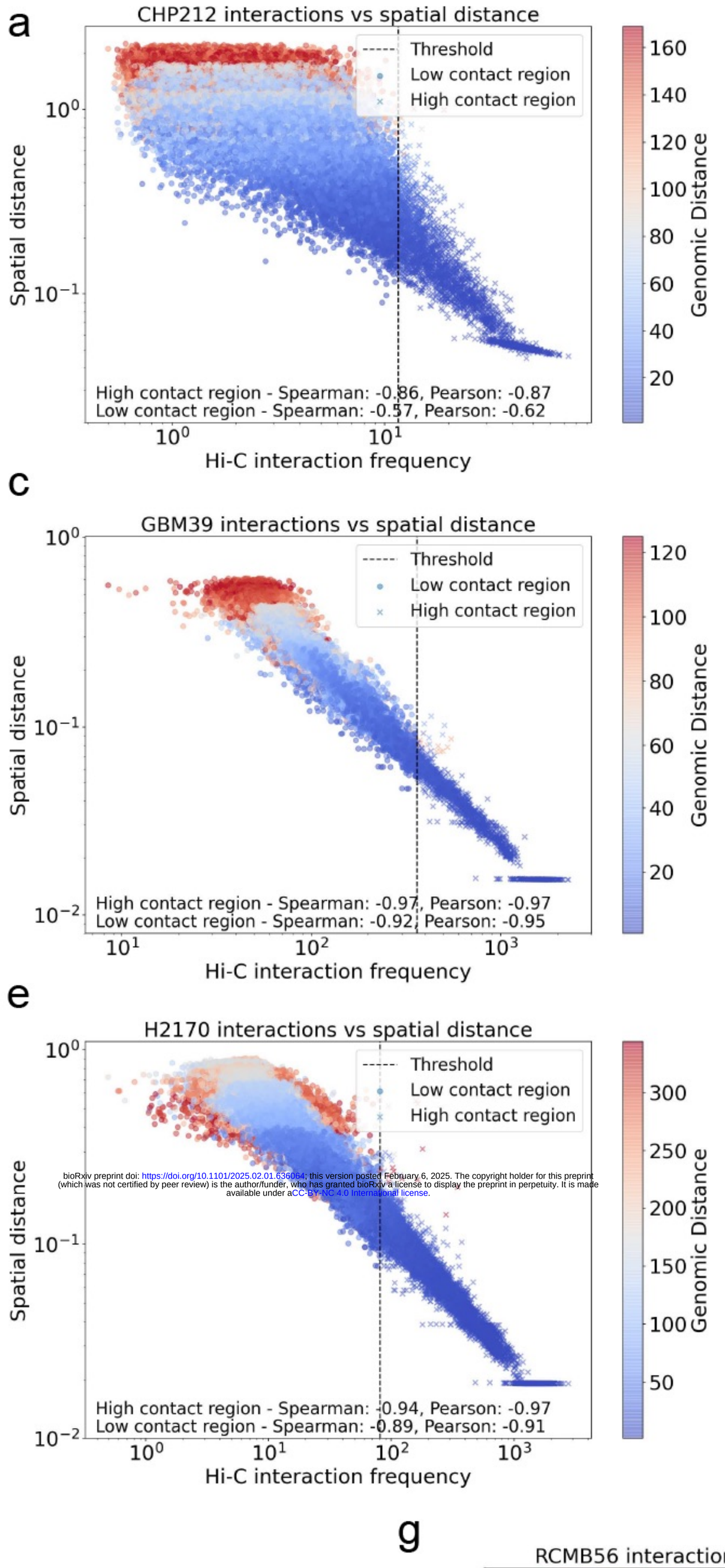




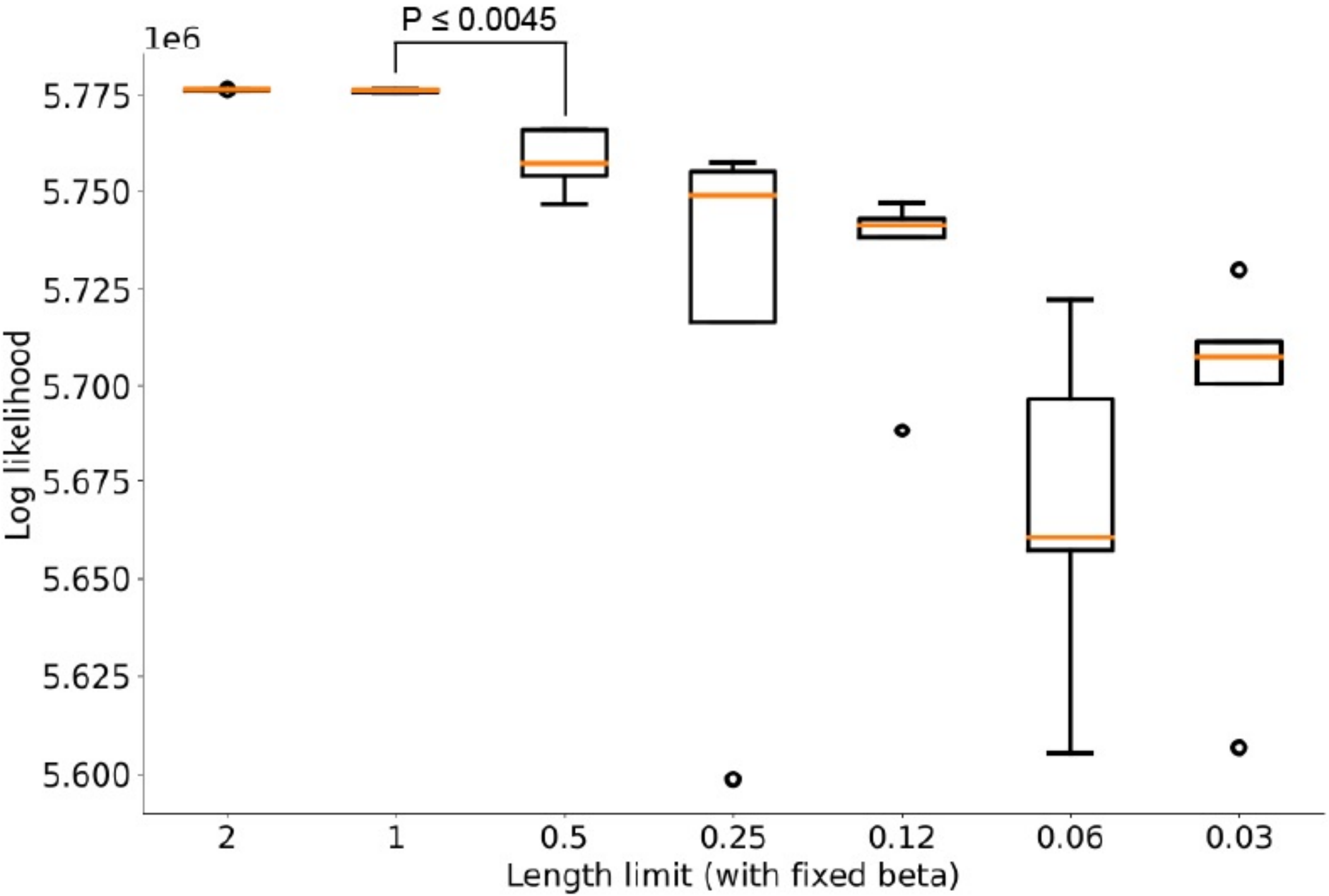
Supplementary Figure 8



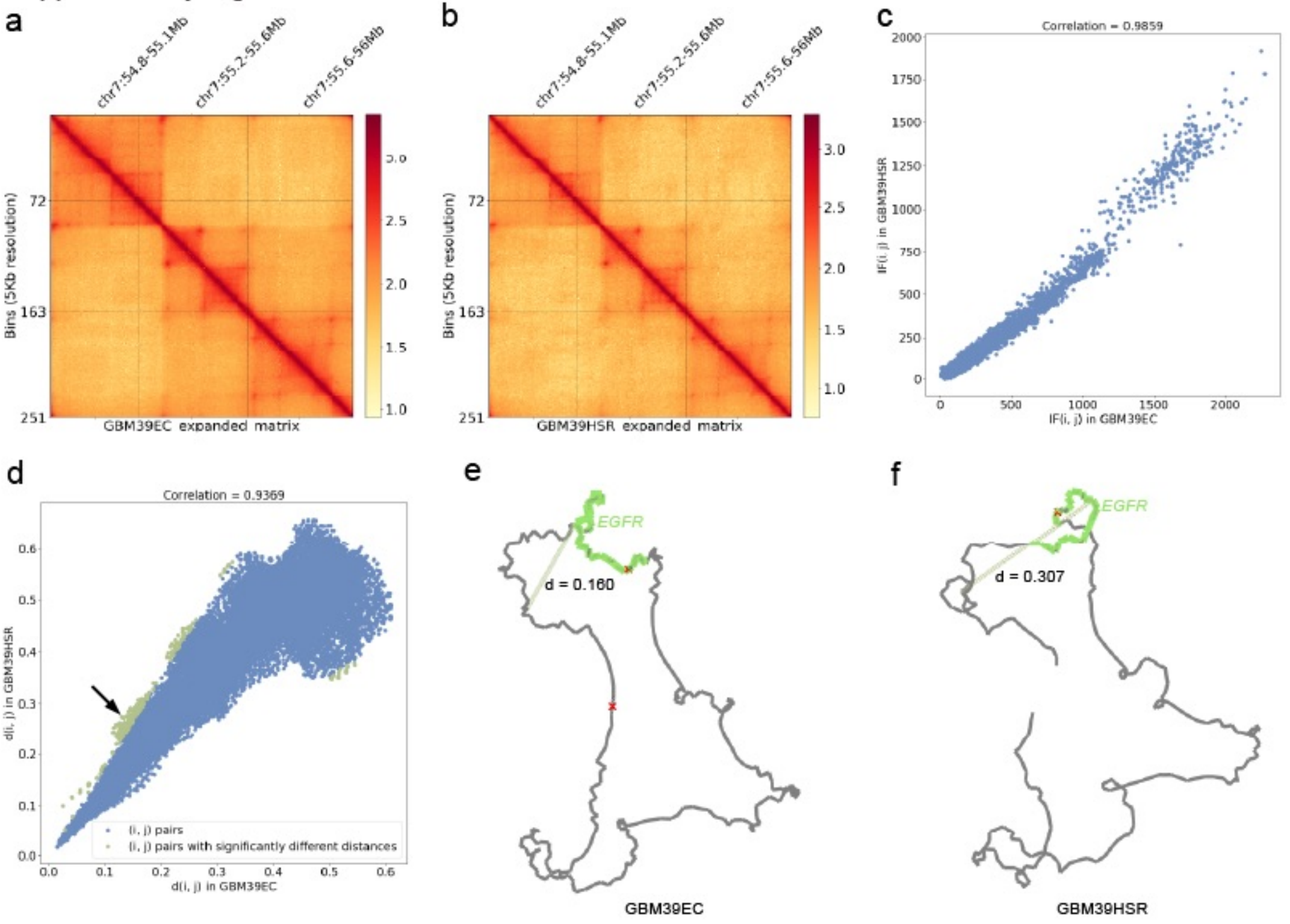
Supplementary Figure 9



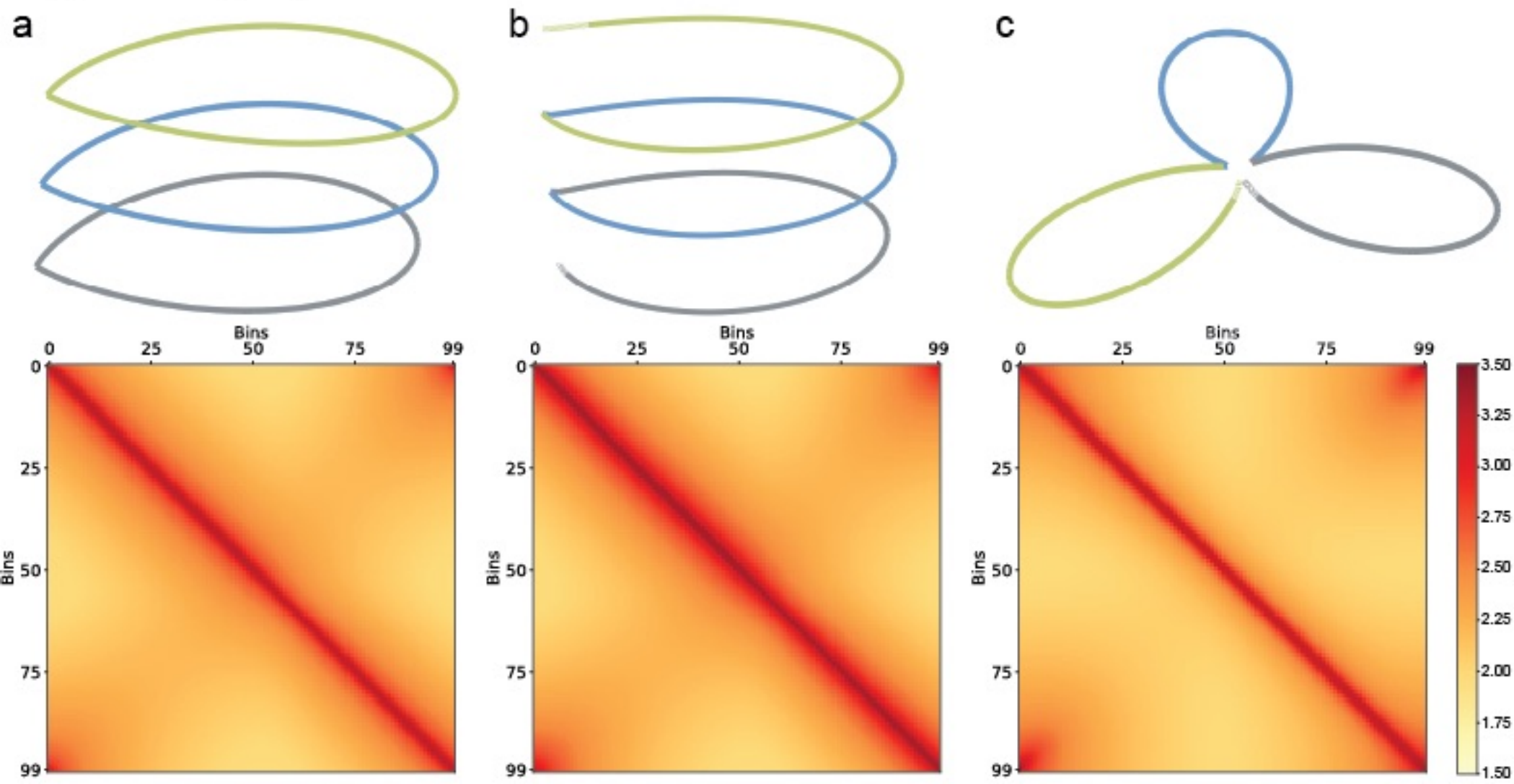
Supplementary Figure 10



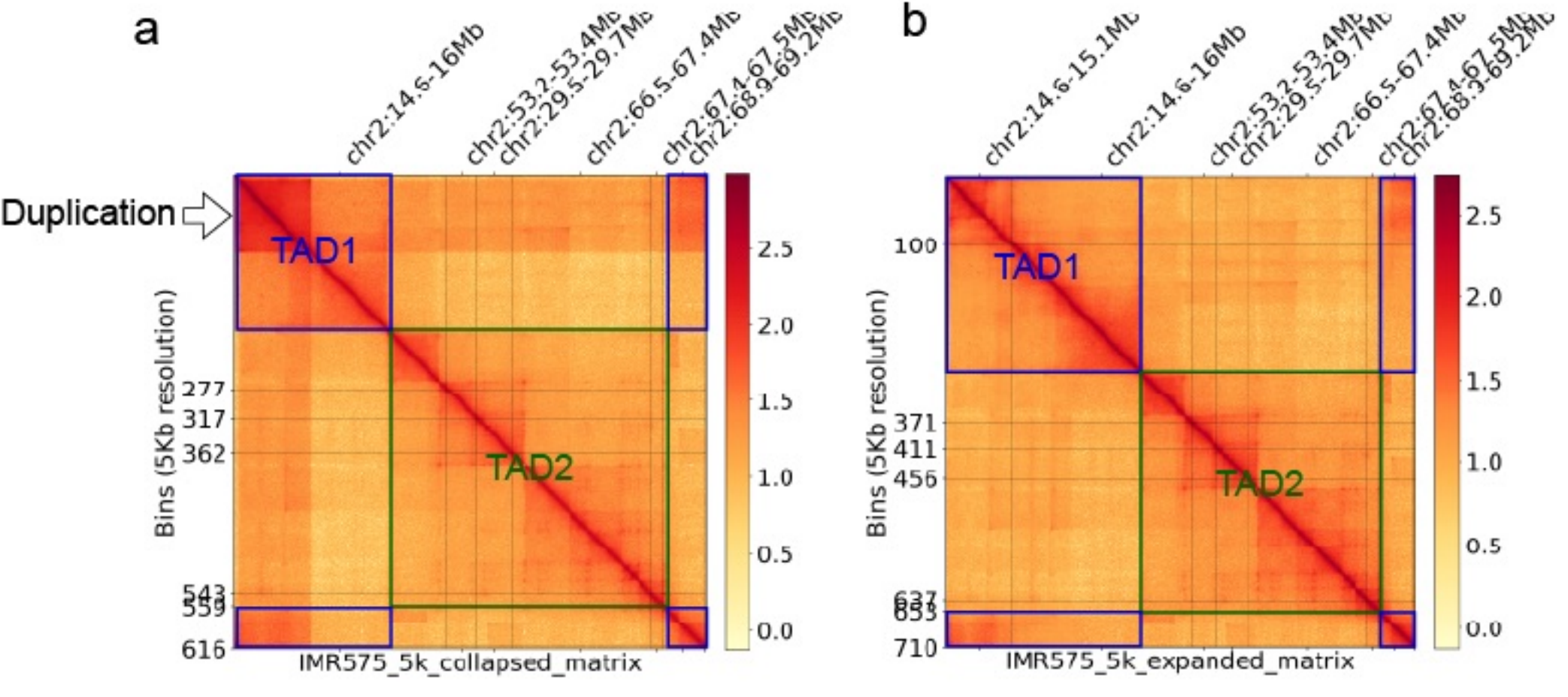
Supplementary Figure 11



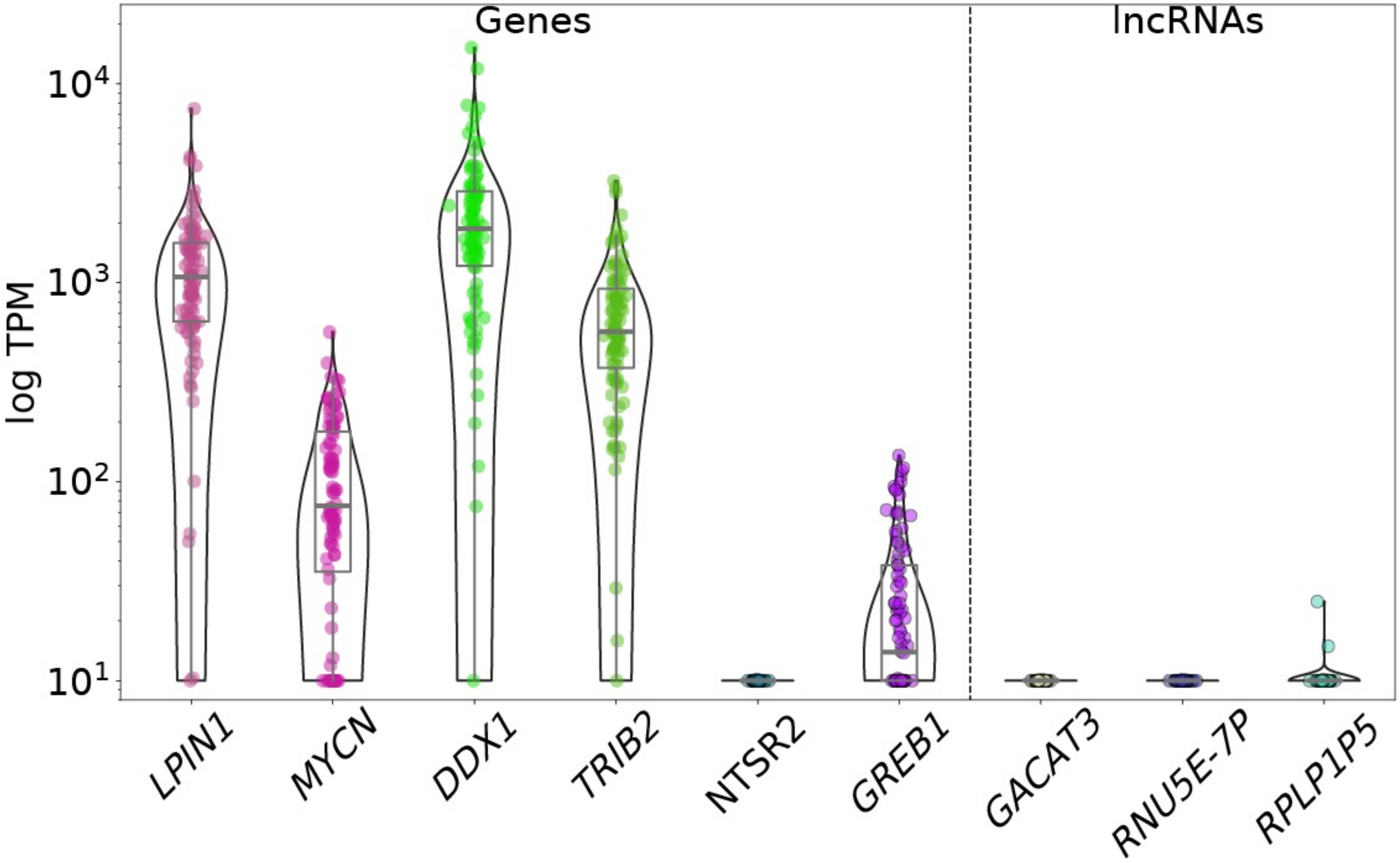
Supplementary Figure 12



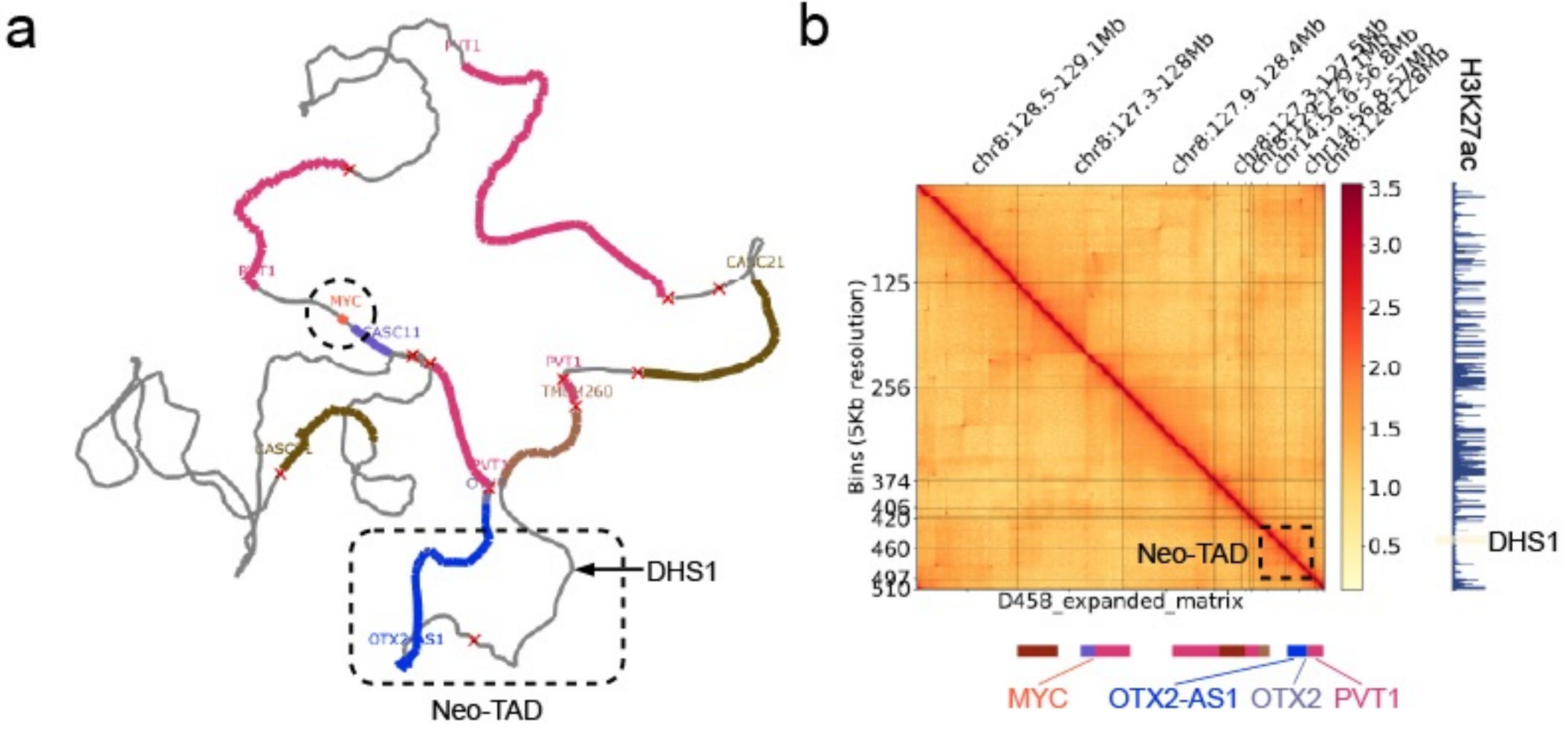
Supplementary Figure 13



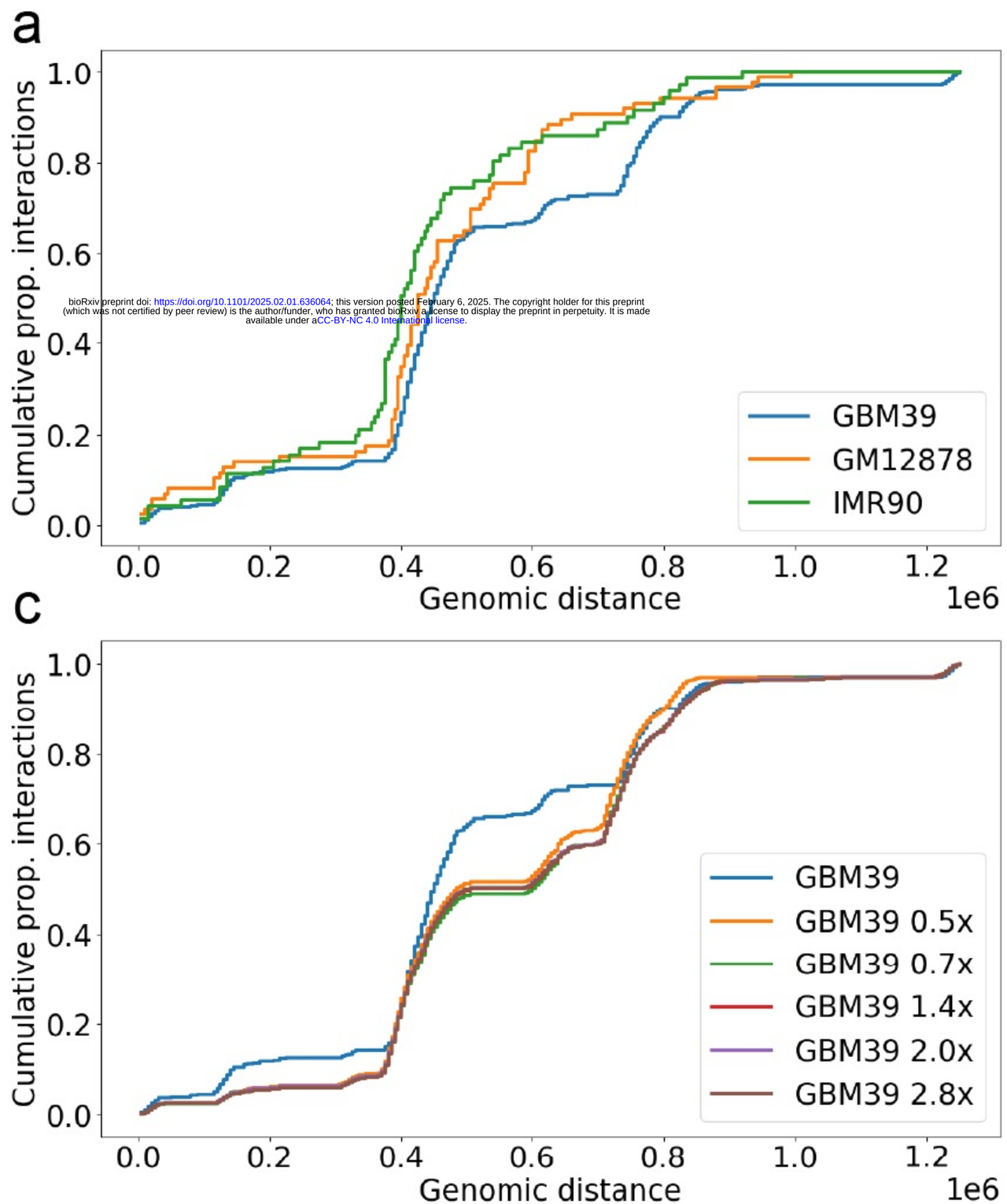
Supplementary Figure 14



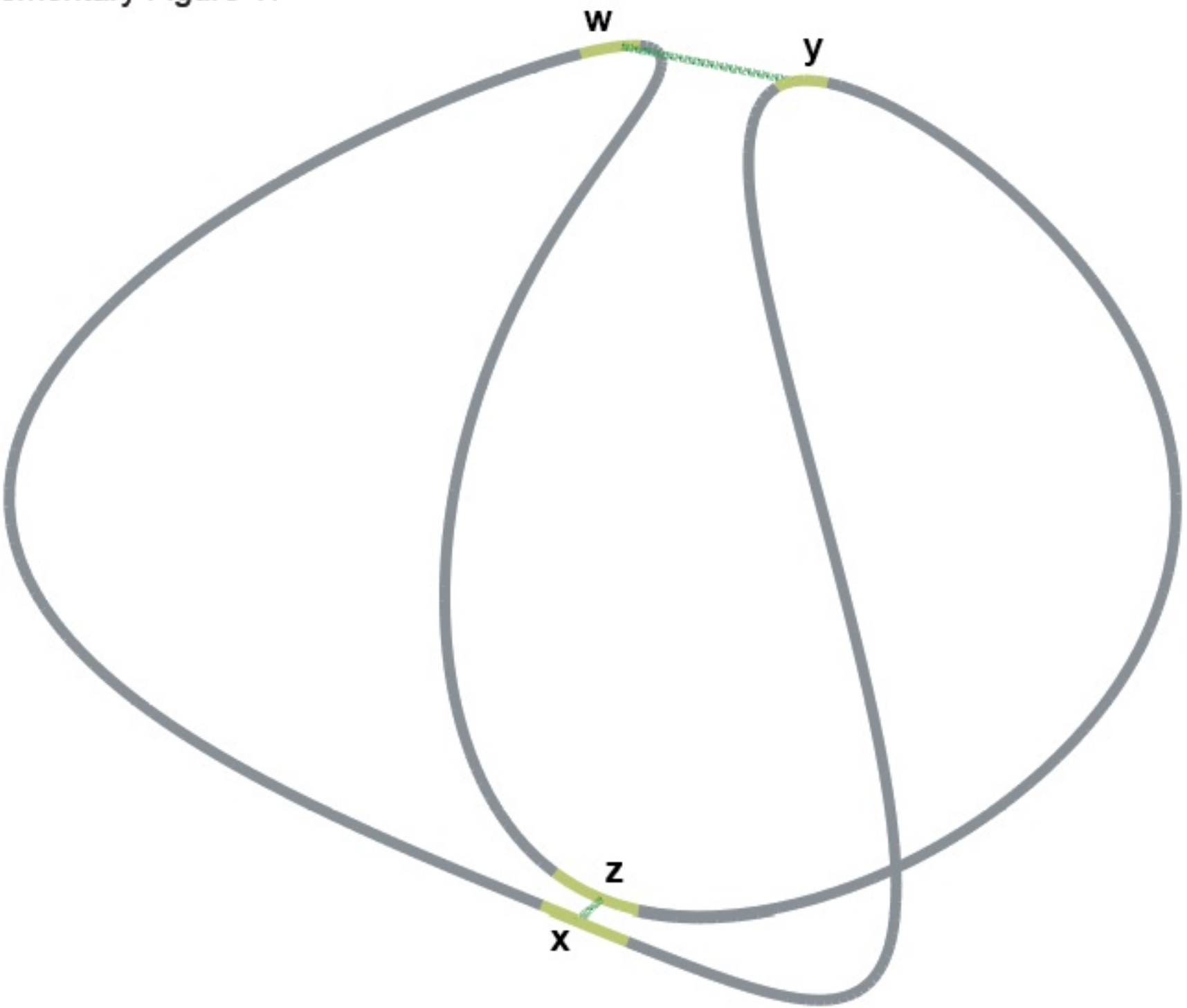
Supplementary Figure 15



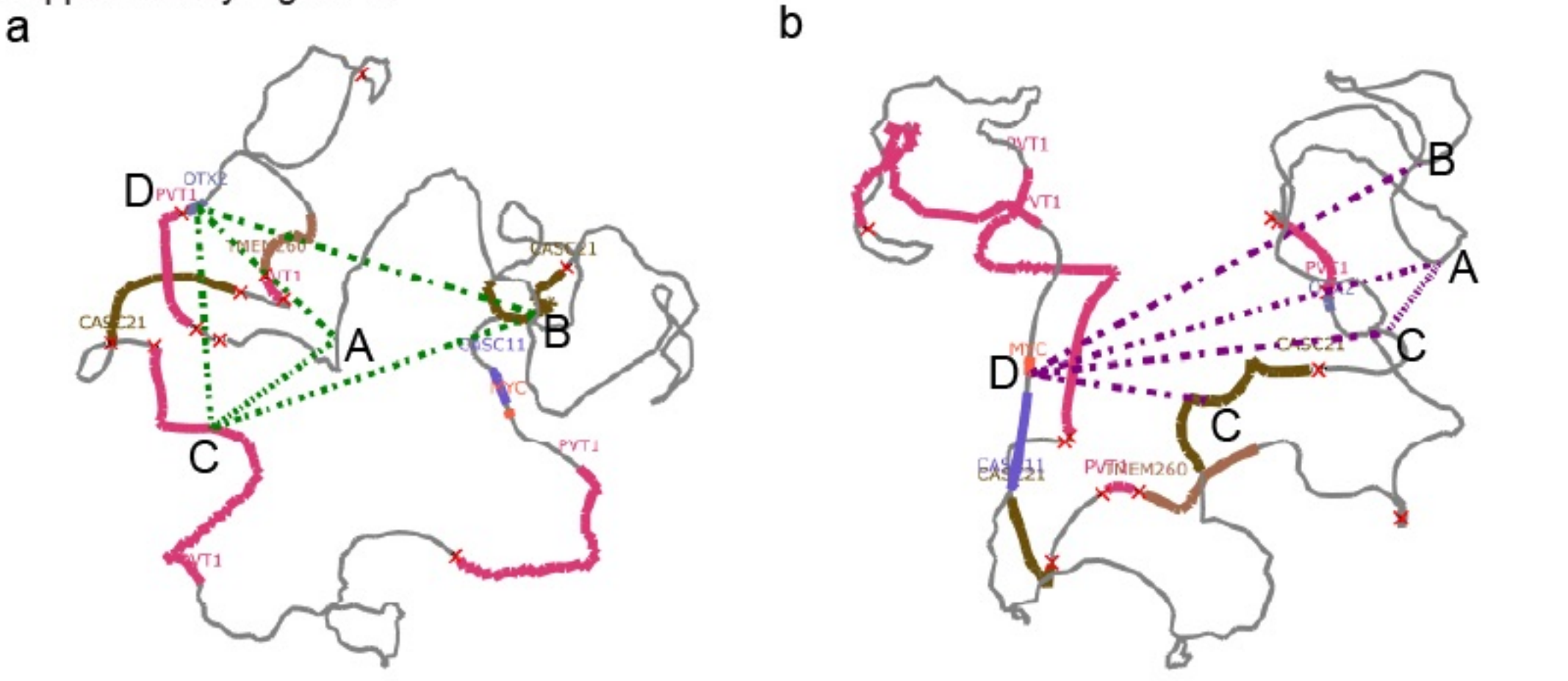
Supplementary Figure 16



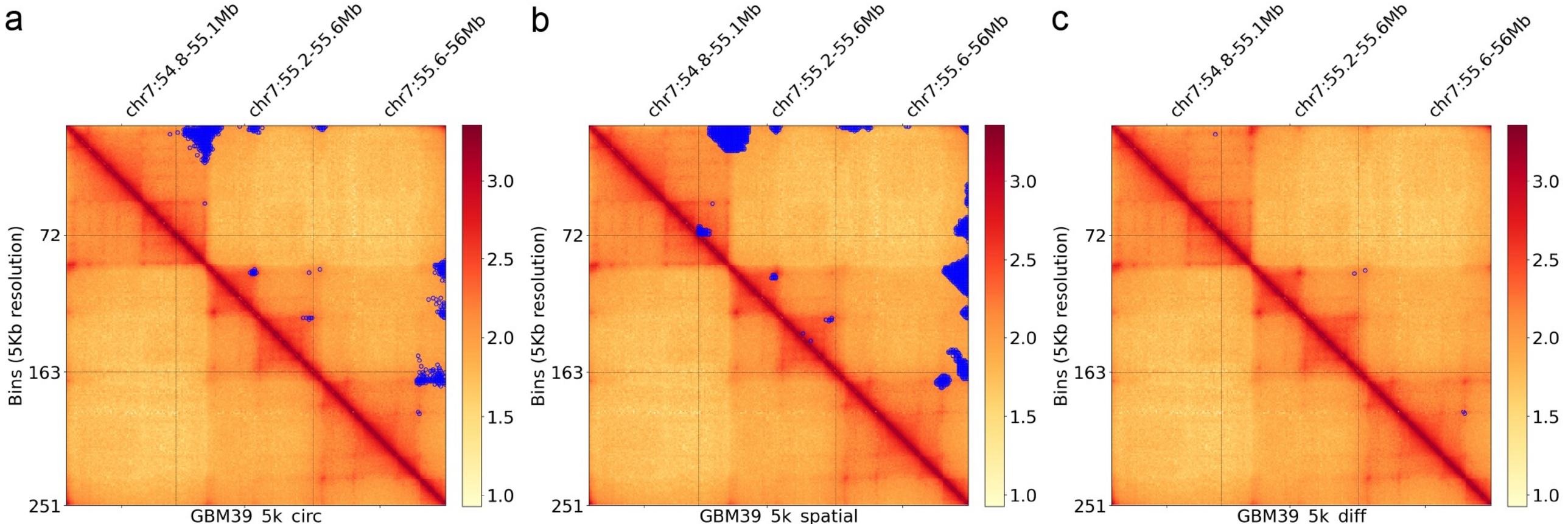
Supplementary Figure 17



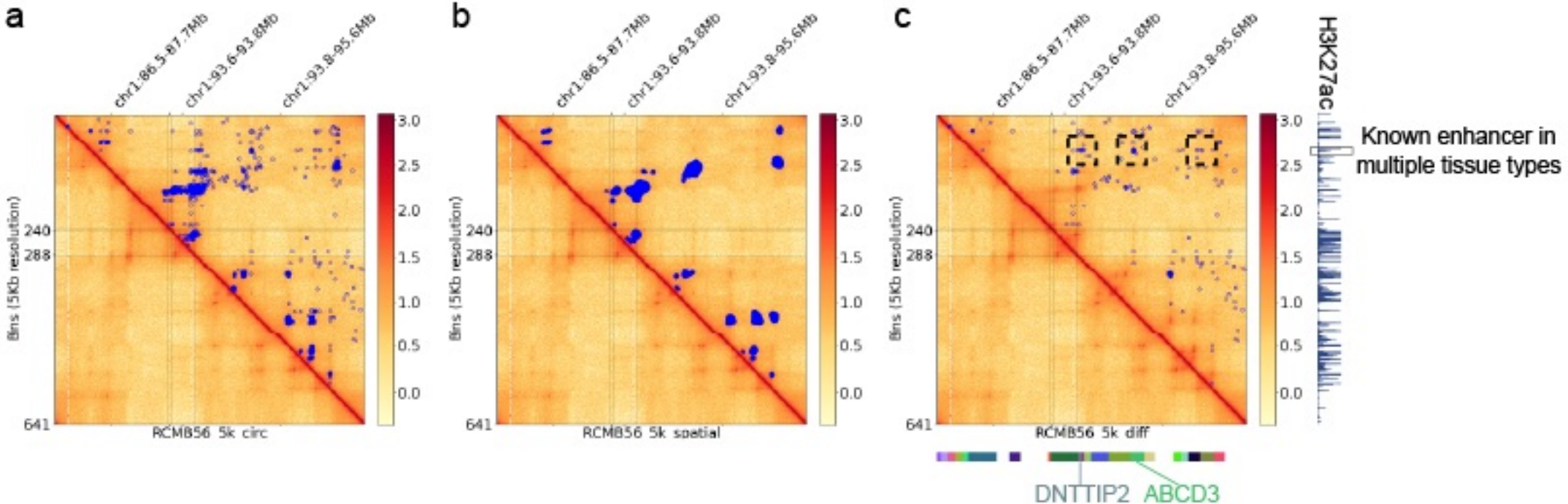
Supplementary Figure 18



Supplementary Figure 19



Supplementary Figure 20



Supplementary Figure 21