Identification of chromatin loops from Hi-C interaction matrices by CTCF-CTCF topology classification

Silvia Galan[1,3], François Serra[#,*] and Marc A. Marti-Renom[1,2,3,4,*]

1. CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain.

2. Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Dr. Aiguader 88, 08003 Barcelona, Spain.

3. Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain.

4. ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain.

#Current address: Computational Biology Group - Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain

*To whom correspondence should be addressed: M.A.M-R. martirenom@cnag.crg.eu, F.S. francois.serra@bsc.es

ABSTRACT

Genome-wide profiling of long-range interactions has revealed that the CCCTC-Binding factor (CTCF) often anchors chromatin loops and is enriched at boundaries of the so-called Topologically Associating Domains or TADs, which suggests that CTCF is essential in the 3D organization of chromatin. However, the systematic topological classification of pairwise CTCF-CTCF interactions has not been yet explored.

Here, we developed a computational pipeline able to classify all CTCF-CTCF pairs according to their chromatin interactions from Hi-C experiments. The interaction profiles of all CTCF-CTCF pairs were further structurally clustered using Self-Organizing Feature Maps (SOFM) and their functionality characterized by their epigenetic states. The resulting cluster were then input to a convolutional neural network aiming at the *de novo* detecting chromatin loops from Hi-C interaction matrices.

Our new method, called LOOPbit, is able to automatically detect higher number of pairwise interactions with functional significance compared to other loop callers highlighting the link between chromatin structure and function.

INTRODUCTION

The human genome consists of 2 meters of DNA and its folding in the cell nucleus is not random (Bonev and Cavalli, 2016). During the last decade, the three-dimensional (3D) organization of the genome have been associated to the regulation of multiple nuclear functions like DNA replication, repair, rearrangement and recombination, RNA processing, and transcription (Bonev and Cavalli, 2016; Dekker and Mirny, 2016; Sexton and Cavalli, 2015; Stadhouders et al., 2019). Indeed, the 3D genome is organized in a hierarchical fashion, with each layer regulating different functions. For instance, chromosomes are found in preferential areas of the nucleus, called chromosome territories (Cremer and Cremer, 2001). The molecular evidence of this observation was first brought by high-throughput Chromosome Conformation Capture (3C) experiments (Hi-C) (Lieberman-Aiden et al., 2009), which also revealed a multi-Megabase (Mb) scale structure, called chromatin compartments. In mammal genomes, there are mainly two compartments, A and B, which correlate with active, open and gene-rich and inactive, closed and gene-poor chromatin, respectively (Lieberman-Aiden et al., 2009). At a sub-Megabase level later 3C-based experiments unveiled that the genome can be further partitioned into self-interacting regions with sizes between 40 kb and 3 Mb, called Topologically Associating Domains (TADs) (Dixon et al., 2012; Nora et al., 2012). TADs modulate interactions between regulatory elements, such as promoters and enhancers, playing a role in transcriptional regulation. At their borders, TADs often harbor regions enriched in CCCTC-binding factor (CTCF) protein (Rao et al., 2014). CTCF is a transcription factor formed by 11 DNA-binding Zinc finger (ZF) that plays a major role in genome architecture (Nichols and Corces, 2015). Its functionality is dependent on the location and the relative orientation of its binding sites. Interacting CTCF pairs tend to be organized in a convergent orientation, with the binding motifs facing each other (Rao et al., 2014). The distribution and organization of CTCF leading to TAD formation has been explained by the loop extrusion model that involves CTCF but also other protein complexes like cohesin (Fudenberg et al., 2016; Sanborn et al., 2015). Briefly, the cohesin complex, which forms a ring-shaped structure upon loading onto chromatin, extrudes chromatin resulting in a growing loop of DNA. The extrusion is blocked when in both sides cohesin encounters two CTCF in a convergent orientation (Fudenberg et al., 2016; Sanborn et al., 2015). Various experiments have described the relevance of the binding motif orientation for loop formation by inverting or deleting CTCF binding sites using CRISPR/Cas9 experiments (Sanborn et al., 2015; Wutz et al., 2017). Moreover, a number of studies revealed the effect of CTCF depletion, resulting in a fainting of TADs and loop structures, while maintaining compartment organization. It suggests that CTCF and cohesin play a role in TAD and loop formation, but not necessarily in

70 compartment maintenance (Guo et al., 2015; Hyle et al., 2019; Nora et al., 2017).

Recently, several studies have applied Aggregate Peak Analysis (APA) (Rao et al., 2014), or pile-up methods (Lekschas et al., 2018) to analyze Hi-C datasets using the mean signal from selected regions of interest (Bonev et al., 2017; de Wit et al., 2013; McLaughlin et al., 2019; Pekowska et al., 2018; Ruiz et al., 2018; Schwarzer et al., 2017), such as CTCF loops. At the
75 level of single chromatin loops, their biological relevance in gene regulation, has led to the development of several loop callers. However, most of the methods present low reproducibility between biological replicates and a low correlation with different biological markers (Forcato et al., 2017). Here, we propose to use Self-Organizing Feature Maps (SOFM) (Kohonen, 1987), an artificial neural network, to classify the signal from pairs of CTCF without any prior
80 information about their topological organization. This approach allows us to obtain sub-populations of CTCF-CTCF interacting structures and to identify their epigenetic signature in an unsupervised manner. As a result, we next trained a convolutional neural network (CNN), and used the generated model in our tool, called LOOPbit, to identify chromatin loops in a robust, fast and genome-wide manner.

85

RESULTS

Classifying CTCF-CTCF interactions using Self-Organizing Feature Maps (SOFMs) and Uniform Manifold Approximation and Projection (UMAP). SOFMs are dependent on several parameters, such as the grid size (number of neurons or clusters), the learning radius, the step size, the standard deviation and the number of iterations or epochs (Kohonen, 1987). To unveil the best combination of SOFM parameters in the context of CTCF-CTCF submatrices classification, we first extracted the Hi-C submatrices of all possible *cis* pairs of CTCF peaks linearly separated between 45 kb and 1.5 Mb in the genome (**Fig. 1a**). These submatrices, here referred to as "CTCF-CTCF submatrices", spanned over 45 kb (20 kb of each side of the 5 kb bin with a given CTCF peak). Next, all extracted CTCF submatrices were input to a SOFM that resulted in a total of 1,152 SOFMs each with different combinations of parameters (**Methods**). To select the optimal set of parameters, we assessed three measures: (i) the percentage of classified submatrices where singletons (SOFM neurons with only one submatrix) were considered as non-classified; (ii) variability between neurons; and (iii) average compartment-type segregation across neurons. These three quality measures aimed at identifying the “best” classification that maximized the number of CTCF-CTCF submatrices classified, the separation between neurons (i.e., increased variability inter-neuron), and the homogenous genome compartment type within neurons. Of all varied parameters, the SOFM grid and the step sizes were the most sensitive to the final classification (**Fig. 1b** and **Supplementary Fig. 1**). The three measures mentioned above were optimal for grid size of 30, learning radius of 5, standard deviation of 0.5, step of 0.01 and, 30 epochs. Such optimal SOFM map can be represented as a grid map where each cell (neuron) is composed by a set of similar CTCF-CTCF submatrices represented by its medoid submatrix (**Fig. 1c**). The SOFM map clearly reflects a variety of CTCF-CTCF submatrices from loop forming pairs (lower-left corner) to non-interacting pairs (upper-right corner). Next, the resulting 900 SOFM neurons were further classified into CTCF-CTCF clusters by computing the Euclidean distance between the medoid submatrices and projecting it into a two-dimensional UMAP (McInnes, Healy, & Melville, 2018), a non-linear manifold dimension reduction followed by a density-based clustering algorithm (HDBSCAN) (Campello et al., 2013). The two-dimensional UMAP resulted in 10 clusters of unique CTCF-CTCF pairing patterns, from the most structured canonical cross pattern (CTCF-CTCF cluster 1 including 11 SOFM neurons with a total of 2,685 CTCF-CTCF submatrices) to a completely flat pattern with no interaction (CTCF-CTCF cluster 10 with 7 SOFM neurons and 1,565 CTCF-CTCF submatrices) (**Fig. 1d**). Our results suggest that CTCF-CTCF pairs can adopt a variety of well-defined topological signatures

120 observed in a Hi-C matrix. Next, we set ourselves to functionally characterize each of the detected CTCF-CTCF clusters.

Functional characterization of the CTCF-CTCF clusters. As described before, segregation of A/B compartment types between neurons was used as a metric to select the optimal set of 125 parameters for the SOFM classifier. However, such measure was agnostic to the medoid submatrix topology representing each neuron. Interestingly, clusters with clear loop topology (that is, CTCF-CTCF clusters 1 to 5) are enriched in A compartment at the anchor points of the loops. Conversely, clusters with non-interacting CTCF-CTCF (that is, clusters 6 to 10) are enriched in B compartment (**Fig. 2a**). This separation between clusters 1 to 5 and 6 to 10 is 130 also observed when revealing enrichments in CTCF pairs with different directionalities**Error! Bookmark not defined.** (Rao et al., 2014), CTCF-CTCF clusters with clear interaction patterns are enriched in convergent-oriented CTCF-CTCF (clusters 1 to 4, **Fig. 2b**), while non-interacting patterns are enriched in divergent-oriented CTCF-CTCF (clusters 7 to 10, **Fig. 2c)**. Parallel-oriented CTCF-CTCF are enriched in mid-interacting signal clusters (clusters 5 and 135 6, **Fig. 2d**). CTCF-CTCF clusters with loop topology (that is, clusters 1 to 3) have a mean genomic distance between CTCF peaks of between ~634 kb and ~680 kb (**Fig. 2e**). CTCF-CTCF clusters 7 to 10, which correspond to enrichment of B compartment and divergent orientation, present shorter mean genomic distances spanning from ~510 kb to ~629 kb (**Fig. 2e**). To assess whether chromatin state correlate with the CTCF-CTCF clusters, we next 140 assessed the chromatin state enrichment in the anchors of CTCF sites for each cluster (**Methods**). Interestingly, pairs of CTCFs that form loop-like structures are enriched with enhancer-promoter interactions (with one anchor points labelled as “enhancer” chromatin state and the other as “promoter”, **Methods**) (**Fig. 2f**). Anchors falling in “heterochromatin” state have an almost opposite distribution among clusters. CTCF-CTCF clusters forming loops 145 (clusters 1 to 5) are depleted of heterochromatin types while non-loop clusters (cluster 6 to 9) are enriched in heterochromatic marks (**Fig. 2g**). Interestingly, CTCF-CTCF cluster number 10, which corresponds to the most B compartment cluster, has no heterochromatic anchors but is enriched in polycomb-promoter pairs (**Fig. 2h**). CTCF-CTCF pairs with polycomb in both anchors are enriched in mid-interacting clusters (that is, clusters 4 to 6) and depleted in non-150 interacting pairs (that is, clusters 7 to 10) (**Fig. 2i**). Finally, enhancer-enhancer pairs, as expected, are more present in interacting clusters (clusters 1 to 4) and less abundant in non-interacting clusters (clusters 7 to 10) (**Fig. 2j**). Our analysis indicates the existence of clusters or types of CTCF-CTCF pairs that gradually expand from enhancer-promoter, convergent,

mid-range interacting pairs in A compartment, to a polycomb-heterochromatin, divergent, short-range, non-interacting pairs in B compartment. Taken together, our results based solely in the interaction pattern of CTCF-CTCF proteins reveal two major types of CTCF-CTCF pairing that can further characterize well defined functional states.

Loop calling using LOOPbit, a CNN trained with looping and non-looping CTCF-CTCF pairs. Next, we randomly subset a total of 6,000 CTCF-CTCF submatrices from clusters 1 to 5 as loop forming pairs and another 6,000 CTCF-CTCF submatrices from clusters 7 to 10 as non-loop forming pairs. The two datasets, loop and non-loop, where next used to train a convolutional neural network (CNN), that we called LOOPbit, that aims at automatically assign a probability of a CTCF-CTCF pair forming a loop in the central bin of a 9x9 submatrix extracted from genome-wide Hi-C experiments (**Fig. 3a** and **Methods**). LOOPbit, which was trained by a 20% leave-out of the used data, was then assess for accuracy and compared to other loop-calling methods using a recently published benchmark (Forcato et al., 2017). LOOPbit, similarly to other published methods including HICCUPS (Rao et al., 2014), GOTHiC (Mifsud et al., 2017), HOMER (Heinz et al., 2010), diffHiC (Lun and Smyth, 2015), HIPPIE (Hwang et al., 2015), and Fit-Hi-C (Ay et al., 2014), detects more interacting loops with an increase of the number of valid interactions in the Hi-C experiment (**Fig. 3b**), which indicates the dependency of such loop-detection methods on the depth of the Hi-C experiment. A solution to minimize this effect is to reduce data resolution (*i.e.*, 5 kb to 40 kb) (**Supplementary Fig. 2a**). Loops detected by LOOPbit were of similar average size as of HICCUPS, diffHiC and HIPPIE (~200 kb) and larger than those by GOTHiC (~80 kb) and HOMER (~100 kb) but shorter than those by Fit-Hi-C (~10 Mb) (**Fig. 3c**). This loop size measure is again dependent on the resolution of the data as loops called at 40 kb resolution increased to ~1 Mb of size for all methods (**Supplementary Fig. 2b**). Next, we assessed the ability of LOOPbit to reproduce loop detection by measuring the Jaccard Index (JI) using replicates of previously published Hi-C experiments (**Methods** and **Supplementary Table 2**). Despite that LOOPbit yields continuous probability clouds instead of pinpointing single cells in the matrix, we used the exact same benchmark measures previously published (Forcato et al., 2017) to calculate JI between replicates (that is to consider two loops identical if both their anchoring bins are the same). With such benchmark, LOOPbit results are slightly more reproducible in average than compared to HOMER and Fit-Hi-C, similar to GOTHiC, diffHiC and HIPPIE and lower reproducibility than HICCUPS (**Fig. 3d**). Next, we compared the accuracy of the loop callers in terms of the biological relevance of the predicted loops. LOOPbit was able to detect higher

percentage of enhancer-promoter loops than any other caller (**Fig. 3e**, top panel). Concomitantly, LOOPbit also detects less loops between heterochromatic anchors (**Fig. 3e**, middle panel) and has similar levels of non-expected loops (that is, between enhancer and heterochromatin, **Fig. 3e**, bottom panel). Interestingly, and likely particular to LOOPbit as it was trained using CTCF-CTCF selected loops, ~56% of all detected loops in the training had an annotated CTCF site in both anchor points (**Methods**). Altogether, our results indicate that LOOPbit has similar tolerance to sparse input Hi-C data as other loop callers, a characteristic related to the intrinsic difficulty of detecting loops (Forcato et al., 2017). Nonetheless, chromatin loops detected by LOOPbit show an enrichment in functional signatures when compared to the other methods. Therefore, we next focused on assessing LOOPbit applicability in a real case biological scenario.

Usability of LOOPbit to capture chromatin loops in a CTCF-targeted degradation system. Next, we assessed whether LOOPbit was able to detect differential loops between three previously published Hi-C datasets where CTCF was targeted for degradation (Nora et al., 2017). The three datasets included non-treated cells, auxin treated cells with degraded CTCF, and auxin wash-off cells where CTCF recovered non-treated levels. These three experiments, which had around 450 million valid Hi-C pairs each, allowed to test whether LOOPbit was sensitive to the depletion of CTCF for detecting loops. First, we assessed the JI value between the three experiments (**Fig. 4a**). As expected, the JI between non-treated and wash-off experiments is higher (~0.1 JI and **Fig. 4a**). Importantly, LOOPbit results confirmed that the auxin-treated loops detected by LOOPbit were different than those for non-treated or wash-off replicas (0.0 JI and **Fig. 4a**). Next, we assessed whether the detected loops have enrichment of CTCF ChIP-Seq signal around the anchor points. Indeed, the majority of the loops detected by LOOPbit are enriched in CTCF at anchoring bins. This enrichment drops for the auxin-treated cells and is recovered for the wash-off cells (**Fig. 4b**). Then we identified that the distance between loop anchors in auxin-treated cells was significantly higher than in non-treated and wash-off datasets (**Fig. 4c**). Finally, we show an example from the region chr3:47,600,000-49,000,000 where LOOPbit detects two loops for non-treated and wash-off, that vanish for auxin-treated cells (**Fig. 4d**). In summary, LOOPbit detects chromatin loops from Hi-C interaction matrices that are enriched for functional interactions and, as expected, for those mediated by the CTCF transcription factor.

DISCUSSION AND CONCLUSION

In this work we introduce the use of the structure signal deconvolution in the context of colocalizing DNA-binding proteins. This methodology aims at identifying clusters of different structural patterns. Until now, most methods detecting structural patterns associated to DNA-binding proteins use aggregate peak analysis (APA) to show an average interaction pattern for different targets of interest. Unfortunately, APA is blind to small subsets with specific interaction patterns, different from the average structure that could potentially be related to specific functions.

Here, we deconvolved the genomic average CTCF-CTCF interaction pattern, and, based solely on structural features, we were able to classify each CTCF-CTCF interaction into distinct subpopulations within the range of distances between CTCF pairs of 45 kb to 1.5 Mb. According to their structural patterns, ten CTCF clusters were obtained (**Fig 1d**). Each CTCF-CTCF cluster has a specific pattern in terms of genome compartment location and epigenetic state. The first observation is that the genomic distance between the CTCF pairs as well as their orientation are relevant features to classify CTCF-CTCF interactions between those that structurally form and do not form loops (Bonev and Cavalli, 2016; Rao et al., 2014). Also expected, but less evident is that loop forming clusters (that is, those from cluster 1 to cluster 5) are enriched in A compartment and drive principally enhancer-enhancer and enhancer-promoter interactions. Then CTCF-CTCF clusters with sparse or blur signal of interaction are often appearing in heterochromatin and polycomb chromatin states spanning shorter genomic distance, which may indicate their participation in silent chromatin (Ogiyama et al., 2018) or simply highlight the overlapping of the signal of polycomb-polycomb driven interactions with CTCF-CTCF driven interactions (Narendra et al., 2015; Van Bortle and Corces, 2012). Interestingly, we could capture the polycomb interacting network, which has been observed to be essential for cell differentiation and identity. Thus, cluster number 10, which is mostly associated to compartment B, no loop structure, few interactions, divergent CTCF binding sites and short genomic distances between anchor points, contained a surprising enrichment of promoter-polycomb states which could be explained by polycomb driving interactions here instead of CTCF. While clusters 4 to 6, presented a loop pattern and were mostly associated to A compartment, were enriched in polycomb at both loop anchors, suggesting the formation of chromatin loops for a proper cell identity regulation. Together these findings define subcategories of CTCF-CTCF interactions within which few (CTCF clusters 1 to 5 representing ~30% of the CTCF-CTCF interactions studied) are consistent with the most accepted model

of CTCF loops bringing together promoters and enhancers to initiate transcription (Guo et al., 2015).

The deconvolution of the average signal between CTCF-CTCF pairs allowed us thus to identify two major types of interactions, one forming a canonical loop (clusters 1 to 5) and a second not forming the canonical loop (clusters 7 to 10). This allowed us to generate a bona fide set of CTCF-CTCF submatrices that generate loops versus those that do not. Next, we used those two sets to train a CNN to develop the loop-caller LOOPbit, which was technically and biologically tested against several previously benchmarked methods (Forcato et al., 2017). LOOPbit, which was applied to 33 Hi-C experiments at 5 kb resolution and 5 experiments at 40 kb resolution, resulted in a number and length of loops similar to all other benchmarked methods and suffered from the same limitations. Indeed, loop-callers cannot easily replicate findings when comparing replicates of low sequencing depths or binned at very high resolutions. Fortunately, when high sequencing depth data is available, LOOPbit increases its reproducibility.

Importantly, loops called by LOOPbit were found to be particularly relevant in terms of biological function. We found a clear enrichment of promoter-enhancer loops and depletion of loops between anchor points in heterochromatin state (Guo et al., 2015; Rao et al., 2014). When compared to other callers, LOOPbit detected almost twice as many promoter-enhancer loops. This feature is likely to be a direct consequence of the SOFM classification and its ability to capture the functional information embedded in CTCF-pairings structures. LOOPbit was also able to identify differentially called loops between experiments where the CTCF protein was degraded by auxin-induced degron system (Nora et al., 2017). Interestingly, loops detected by LOOPbit in auxin-treated samples were significantly enriched in long-range distances, this is in agreement with the loss of local organization of TADs and the preserved regulation of long-range enhancer-promoter interactions by the remaining ~20% of CTCF (Hyle et al., 2019; Elphege P Nora et al., 2017; Splinter et al., 2006).

In summary, we have shown that signal deconvolution is able to define sub-classes of CTCF-driven chromatin loops with specific structural features, which allowed us to train with an increased specificity an Artificial Intelligence algorithm to finally call functionally relevant loops. Beyond the scope of this study, our methodology opens the possibility to train our CNN on loops driven by other DNA binding proteins.

METHODS

CTCF ChIP-Seq and Hi-C data.

CTCF ChIP-Seq experiments of Human B-lymphocyte cell line (GM12878) were downloaded
290 from the ENCODE database (https://www.encodeproject.org/; dataset ids: ENCSR000DRZ, ENCSR000AKB and ENCSR000DZN) (Davis et al., 2018). After processing the peaks following the ENCODE pipelines available at https://github.com/ENCODE-DCC, only common peaks from all three experiments were kept resulting in a total of 52,844 CTCF peaks. Next, the orientation of the CTCF binding motifs was assessed by means of the MEME and FIMO
295 motif-based sequence analysis tools (Grant et al., 2011; Ma et al., 2014). Only peaks with a statistically significant CTCF motif were kept (p-value < 0.05), which resulted in a total of 41,816 *ChIP-Seq+motif* peaks. This filter discarded ~21% of the original detected CTCF ChIP-Seq peaks.

In situ Hi-C datasets for GM12878 cell line (Rao et al., 2014) were downloaded from the GEO
300 database (**Supplementary Table 1**) and its replicates merged and parsed using TADbit as previously described (Serra et al., 2017). Two resolution matrices (that is, at 100 kb and at 5 kb) were obtained and normalized using OneD (Vidal et al., 2018) with default parameters. The 100 kb Hi-C matrices were next used to calculate chromosome compartmentalization (Imakaev et al., 2012; Lieberman-Aiden et al., 2009) using TADbit (Serra et al., 2017). The
305 5 kb resolution matrices were further parsed to subtract 45 kb squared submatrices centered in each axis of the matrices to any two of the 15,597 isolated CTCF *ChIP-Seq+motif* peaks (that is, ~38% of all selected peaks above). Isolated peaks were defined as those *ChIP-Seq+motif* peaks with no other peak within a 45 kb window span from the center of the peak. This additional filter ensured that the observed signal in a Hi-C submatrix was due to a
310 particular pair of CTCF-CTCF peak and not multiple pairs. Finally, we subtracted a total of 130,655 submatrices between any pair of isolated CTCF-CTCF peaks spanning any distance between 45 kb and 1.5 Mb, which ensured to select most pairs within the size of a typical TAD in the human genome (~900 kb).

315 **Submatrix analysis, deconvolution, classification and clustering.**

All 130,655 submatrices between any pair of isolated CTCF-CTCF peaks were next analyzed with our Python based package called Meta-Waffle and available in GitHub (https://github.com/3DGenomes/metawaffle). Meta-Waffle takes as input a list of coordinates pairs corresponding to the selected pairs of selected peaks. Meta-Waffle then will extract,

320 analyze, deconvolve and classify the submatrices using a Self-Organizing Feature Map (SOFM) approach (Kohonen, 1987) available in http://neupy.com/pages/home.html. Briefly, Meta-Waffle first extracts the submatrices from an input Hi-C map, which in this application were 45 kb per 45 kb (9x9 bins) of size. Second, the submatrices values are next re-scaled between 0 and 1 using a sigmoid function. Third, the re-scaled submatrices are input to the
325 SOFM based on a series of empirically optimized parameters including the SOFM grid size (*i.e.,* 10x10, 20x20, 30x30, 40x40, 45x45, and 50x50), learning radius (*i.e.,* 1,2, and 5), standard deviation (*i.e.,* 0.5, 0.1, and 0.01), steps (*i.e.,* 1, 0.5, 0.1, and 0.01), and Epochs (*i.e.,* 30, 90, 130, and 200). Combination of the tested parameters resulted in a total of 1,152 generated SOFMs. Assessing which set of parameters results in the best sub-matrices
330 classification is not trivial as there is no a "standard of gold" for classifying CTCF-CTCF Hi-C sub-matrices. Therefore, we devised three measures to be maximized by the selected SOFM optimal parameters. First, a percentage of classified sub-matrices, which allowed to select the optimal parameters that result in minimal singleton cells (that is, cells with only one sub-matrix) after the SOFM classification. Second the variability between neurons. And third, a
335 compartment segregation score that maximizes the segregation of SOFM cells in the two main genome compartments. Using the mean compartment signal per SOFM cell, the segregation compartment score was calculated as the subtraction between A and B compartment frequencies in each SOFM map. The SOFM optimal parameters that maximized both of the devised scores were: Grid size: 30 x 30; Learning radius: 5; Standard deviation: 0.5; Step:
340 0.01; and Epochs: 30. Finally, the optimal SOFM cells were used to generate a hierarchical clustering by means of the Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018), which can be used as an effective pre-processing step to enhance the performance of density-based clustering. The UMAP was computed using a local neighboring size of 6 for the manifold approximation, an effective minimum distance between embedded
345 points of 0.3 and 100 epochs to increase its accuracy. The final 10 clusters of Hi-C CTCF-CTCF cells were obtained by using HDBSCAN (Campello and D.M.J., 2013) with default parameters, considering a minimum clusters size of 6 and a minimum of 6 samples in a neighborhood for a point to be considered a core point.

350 **Chromatin states integration.**

The chromatin 15-state model for the GM12878 cell line was downloaded from the Roadmap Epigenomics Consortium (Roadmap Epigenomics et al., 2015). Following a similar previously published protocol (Forcato et al., 2017), all 15 states were merged into 4 major classes:

promoter (Active TSS, Flanking Active TSS, Bivalent/poised TSS Flanking bivalent TSS), enhancer (Enhancers, Genic enhancers, Bivalent enhancers), repressed polycomb (Repressed Polycomb, Weak repressed Polycomb) and heterochromatin (Heterochromatin, Quiescent/Low). Next, the genome was segmented into 5 kb bins, and each bin was classified based on the overlap (> 50 bp) with any of the four major chromatin states. A bin could be assigned to one or more categories. Next, interactions between bins were classified as "Not expected" (interacting promoter/enhancer bin with heterochromatin bin), "Promoter-Enhancer" (interacting promoter bin with an enhancer bin), and "Heterochromatin-Heterochromatin", (interacting heterochromatin bins).

The LOOPbit CNN.

LOOPbit is a trained Convolutional Neural Network (CNN) to predict the localization of loops from Hi-C interaction matrices. LOOPbit was trained using the TensorFlow platform (https://www.tensorflow.org/) using a common pattern: input matrix flattening - dense layer (ReLu) - dropout - dense layer (Softmax). As an input, the CNN takes tensors of shape (9, 9) that will be flattened in the first layer. The dense layers were used to perform the classification of input Hi-C cells into two different classes: loop and no-loop. The dropout layer is needed to avoid overfitting of the model (**Fig 3a**). The LOOPbit CNN was trained by a 20% leave-out of the data used as test and 80% of the data for training. The training dataset obtained by sub-sampling a total of 6,000 randomly selected CTCF-CTCF submatrices from the SOFM clusters with clear signal of looping (loops) and another 6,000 randomly selected CTCF-CTCF submatrices from the SOFM clusters with no signal of looping (no-loop) (**Fig 3a**). In particular, loop submatrices were extracted from clusters 1, 2, 3, 4 and 5, and no-loop submatrices from clusters 7, 8, 9 and 10 (**Fig 1d**). The trained model used a 0.2 drop of the data and 1,024 neurons resulting in a classification accuracy of 85.9% and 89.6% for the loop and no-loop class, respectively. As LOOPbit retrieves a loop probability per matrix cell, the chromatin loops are defined according to the probability values similarity within the neighbors in a delimited area, which was set to 3 bins for the benchmark and 5 bins for the CTCF depleted experiment. Thus, multiple neighboring matrix cells with high loop probability would be considered to be the same chromatin loop.

LOOPbit benchmark.

The trained model was then use to predict loop localization in a different Hi-C matrix with different goals:

1. *Large-scale benchmark for loop detection:* To benchmark the accuracy of LOOPbit detecting loops from a Hi-C input matrix, we used 38 previously published Hi-C datasets (**Supplementary Table 2**). To avoid biases derived from the data processing pipelines, all 38 datasets were all processed using the same protocol as the training datasets from GM12878 cell line (see above). Then, the Hi-C experiments were analyzed with LOOPbit, scanning chromosome-wise using a windows of 9x9 bins and a step of 1 bin, to predict loops at two different resolutions (5 kb and 40 kb). The predicted localization of loops was next assessed using several measures. First, a Jaccard Index (JI) was computed to assess reproducibility between replicates of the same experiment. Two predicted loops were considered to be identical when they shared exactly the same anchoring bins in both replicates. We also computed the reproducibility allowing the 70% of overlapping between the chromatin loops (**Supplementary Fig. 3**). Second, to characterize the possible biological relevance of the predictions, the enrichment of diverse chromatin marks at loop anchors was calculated. For the benchmarking, the 15-states chromatin models for GM12878, IMR90 and h1-ESC cell lines were downloaded from the Roadmap Epigenomics Consortium (Roadmap Epigenomics et al., 2015) and analyzed as described above. For the fly late embryos dataset, the 16-chromatin states model was download from modENCODE (Ho et al., 2014). As previously, the states were also merged into 4 major classes: promoter (Promoter), enhancer (Enhancer 1, Enhancer 2), repressed polycomb (PC repressed 1, PC repressed 2), and heterochromatin/Low (Heterochromatin1, Heterochromatin 2, Low signal 1, Low signal 2, Low signal 3). Similarly, the enrichment analysis of the chromatin states at the loop anchors was done described above. Additionally, the assessment presence of CTCF sites and their orientation in the base of the predicted loops was performed only for *cis* interactions identified in the Hi-C maps at 5 kb resolution. Briefly, CTCF ChIP-seq experiments were downloaded (**Supplementary Table 3**) and peaks processed using HOMER motif analysis with default parameters. The *cis* interactions conserved in at least 2 replicates within each dataset (with exception of Jin H1-hESC with one replicate) were intersected with the list of motifs of the CTCF peaks. Then, an interaction was considered convergent if the interacting bin closer to the p-terminus of the chromosome contained one CTCF motif on the forward strand (+ orientation), and the

420 interacting bin closer to the q-terminus of the chromosome contained one CTCF motif on the reverse strand (– orientation) (Forcato et al., 2017).

2. *Control of a CTCF depleted experiment*. To assess whether LOOPbit still detects loops in a CTCF depleted cell, we downloaded the Hi-C and CTCF ChIP-Sequencing data from the GSE98671 GEO entry (Nora et al., 2017). The dataset include three samples:
425 (i) untreated cells, (ii) cells treated for 2 hours with a degron system targeting CTCF with about 20% remaining CTCF after treatment, and (iii) wash-off cells with recovered CTCF levels. The three biological replicates were merged and parsed as indicated above. As in the other benchmark, we assess the accuracy of LOOPbit detecting loops by means of the Jaccard Index between replicates as well as chromatin marks
430 enrichment at loop anchors (see above).

CODE AVAILABILITY

Meta-Waffle as well as LOOPbit are available on GitHub: (https://github.com/3DGenomes/metawaffle and https://github.com/3DGenomes/loopbit, respectively).

ACKNOWLEDGEMENTS

We thank the authors of (Forcato et al., 2017) for providing the dataset used to make possible the comparison of LOOPbit with other loop callers. We thank the authors of (E. P. Nora et al., 2017) for helping with the interpretation of CTCF targeted-degraded Hi-C datasets. We thank David Castillo, Luciano Di Croce, Sandra Peiró and François Le Dily for their continuous discussions and support to the development of Meta-Waffle and LOOPbit.

FUNDING

This work was partially supported by the European Research Council under the 7th Framework Program FP7/2007-2013 (ERC grant agreement 609989), the European Union's Horizon 2020 research and innovation programme (grant agreement 676556) and the Spanish Ministerio de Ciencia e Innovación (BFU2017-85926-P). We also knowledge support from "Centro de Excelencia Severo Ochoa 2013-2017", SEV-2012-0208, the Spanish ministry of Science and Innovation to the EMBL partnership and the CERCA Programme/Generalitat de Catalunya to the CRG. Finally, we acknowledge support of the Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III, the Generalitat de Catalunya through Departament de Salut and Departament d'Empresa i Coneixement and the Co-financing by the Spanish Ministry of Science and Innovation with funds from the European Regional Development Fund (ERDF) corresponding to the 2014-2020 Smart Growth Operating Program to CNAG.

CONFLICT OF INTEREST

None declared.

460 **REFERENCES**

Ay, F., Bailey, T.L., and Noble, W.S. (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res *24*, 999-1011.

Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. Nat Rev Genet *17*, 661-678.

465 Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.P., Tanay, A.*, et al.* (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell *171*, 557-572 e524.

Campello, R.J.G.B., and D.M.J., S. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. Advances in Knowledge Discovery and Data Mining *7819*, 160-172.

470 Campello, R.J.G.B., Moulavi, D., and Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates (Berlin, Heidelberg: Springer Berlin Heidelberg).

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet *2*, 292-301.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain,
475 K., Baymuradov, U.K., Narayanan, A.K.*, et al.* (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res *46*, D794-D801.

de Wit, E., Bouwman, B.A., Zhu, Y., Klous, P., Splinter, E., Verstegen, M.J., Krijger, P.H., Festuccia, N., Nora, E.P., Welling, M.*, et al.* (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature *501*, 227-231.

480 Dekker, J., and Mirny, L. (2016). The 3D Genome as Moderator of Chromosomal Communication. Cell *164*, 1110-1121.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376-380.

485 Forcato, M., Nicoletti, C., Pal, K., Livi, C.M., Ferrari, F., and Bicciato, S. (2017). Comparison of computational methods for Hi-C data analysis. Nat Methods *14*, 679-685.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of Chromosomal Domains by Loop Extrusion. Cell reports *15*, 2038-2049.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given
490 motif. Bioinformatics *27*, 1017-1018.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y.*, et al.* (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900-910.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C.,
495 Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576-589.

Ho, J.W., Jung, Y.L., Liu, T., Alver, B.H., Lee, S., Ikegami, K., Sohn, K.A., Minoda, A., Tolstorukov, M.Y., Appert, A.*, et al.* (2014). Comparative analysis of metazoan
500 chromatin organization. Nature *512*, 449-452.

Hwang, Y.C., Lin, C.F., Valladares, O., Malamon, J., Kuksa, P.P., Zheng, Q., Gregory, B.D., and Wang, L.S. (2015). HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements. Bioinformatics *31*, 1290-1292.

Hyle, J., Zhang, Y., Wright, S., Xu, B., Shao, Y., Easton, J., Tian, L., Feng, R., Xu, P., and Li,
505 C. (2019). Acute depletion of CTCF directly affects MYC regulation through loss of enhancer-promoter looping. Nucleic Acids Res *47*, 6699-6713.

Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods *9*, 999-1003.

510 Kohonen, T. (1987). Adaptive, associative, and self-organizing functions in neural computing. . Appl Opt, *26(23)*, 4910-4918.

Lekschas, F., Bach, B., Kerpedjiev, P., Gehlenborg, N., and Pfister, H. (2018). HiPiler: Visual Exploration of Large Genome Interaction Matrices with Interactive Small Multiples. IEEE transactions on visualization and computer graphics *24*, 522-531.

515 Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O.*, et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289-293.

Lun, A.T., and Smyth, G.K. (2015). diffHic: a Bioconductor package to detect differential
520 genomic interactions in Hi-C data. BMC Bioinformatics *16*, 258.

Ma, W., Noble, W.S., and Bailey, T.L. (2014). Motif-based analysis of large nucleotide data sets using MEME-ChIP. Nat Protoc *9*, 1428-1450.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv *1802.03426*.

525 McLaughlin, K., Flyamer, I.M., Thomson, J.P., Mjoseng, H.K., Shukla, R., Williamson, I., Grimes, G.R., Illingworth, R.S., Adams, I.R., Pennings, S.*, et al.* (2019). DNA Methylation Directs Polycomb-Dependent 3D Genome Re-organization in Naive Pluripotency. Cell reports *29*, 1974-1985 e1976.

Mifsud, B., Martincorena, I., Darbo, E., Sugar, R., Schoenfelder, S., Fraser, P., and Luscombe,
530 N.M. (2017). GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. PLoS One *12*, e0174744.

Narendra, V., Rocha, P.P., An, D., Raviram, R., Skok, J.A., Mazzoni, E.O., and Reinberg, D. (2015). CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. Science *347*, 1017-1021.

535 Nichols, M.H., and Corces, V.G. (2015). A CTCF Code for 3D Genome Architecture. Cell *162*, 703-705.

Nora, E.P., Goloborodko, A., Valton, A.L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell
540 *169*, 930-944 e922.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J.*, et al.* (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381-385.

Ogiyama, Y., Schuettengruber, B., Papadopoulos, G.L., Chang, J.M., and Cavalli, G. (2018).
545 Polycomb-Dependent Chromatin Looping Contributes to Gene Silencing during Drosophila Development. Mol Cell *71*, 73-88 e75.

Pekowska, A., Klaus, B., Xiang, W., Severino, J., Daigle, N., Klein, F.A., Oles, M., Casellas, R., Ellenberg, J., Steinmetz, L.M.*, et al.* (2018). Gain of CTCF-Anchored Chromatin Loops Marks the Exit from Naive Pluripotency. Cell Syst *7*, 482-495 e410.

550 Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S.*, et al.* (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665-1680.

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-
555 Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J.*, et al.* (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317-330.

Ruiz, J.L., Tena, J.J., Bancells, C., Cortes, A., Gomez-Skarmeta, J.L., and Gomez-Diaz, E. (2018). Characterization of the accessible genome in the human malaria parasite Plasmodium falciparum. Nucleic Acids Res *46*, 9414-9431.

560 Sanborn, A.L., Rao, S.S., Huang, S.C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J.*, et al.* (2015). Chromatin extrusion explains key

features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci U S A *112*, E6456-6465.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., C, H.H., Mirny, L.*, et al.* (2017). Two independent modes of chromatin organization revealed by cohesin removal. Nature *551*, 51-56.

Serra, F., Bau, D., Goodstadt, M., Castillo, D., Filion, G.J., and Marti-Renom, M.A. (2017). Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput Biol *13*, e1005665.

Sexton, T., and Cavalli, G. (2015). The role of chromosome domains in shaping the functional genome. Cell *160*, 1049-1059.

Stadhouders, R., Filion, G.J., and Graf, T. (2019). Transcription factors and 3D genome conformation in cell-fate decisions. Nature *569*, 345-354.

Van Bortle, K., and Corces, V.G. (2012). Nuclear organization and genome function. Annu Rev Cell Dev Biol *28*, 163-187.

Vidal, E., le Dily, F., Quilez, J., Stadhouders, R., Cuartero, Y., Graf, T., Marti-Renom, M.A., Beato, M., and Filion, G.J. (2018). OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. Nucleic Acids Res *46*, e49.

Wutz, G., Varnai, C., Nagasaka, K., Cisneros, D.A., Stocsits, R.R., Tang, W., Schoenfelder, S., Jessberger, G., Muhar, M., Hossain, M.J.*, et al.* (2017). Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. EMBO J *36*, 3573-3599.

Yang, T., Zhang, F., Yardimci, G.G., Song, F., Hardison, R.C., Noble, W.S., Yue, F., and Li, Q. (2017). HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome Res *27*, 1939-1949.

FIGURES

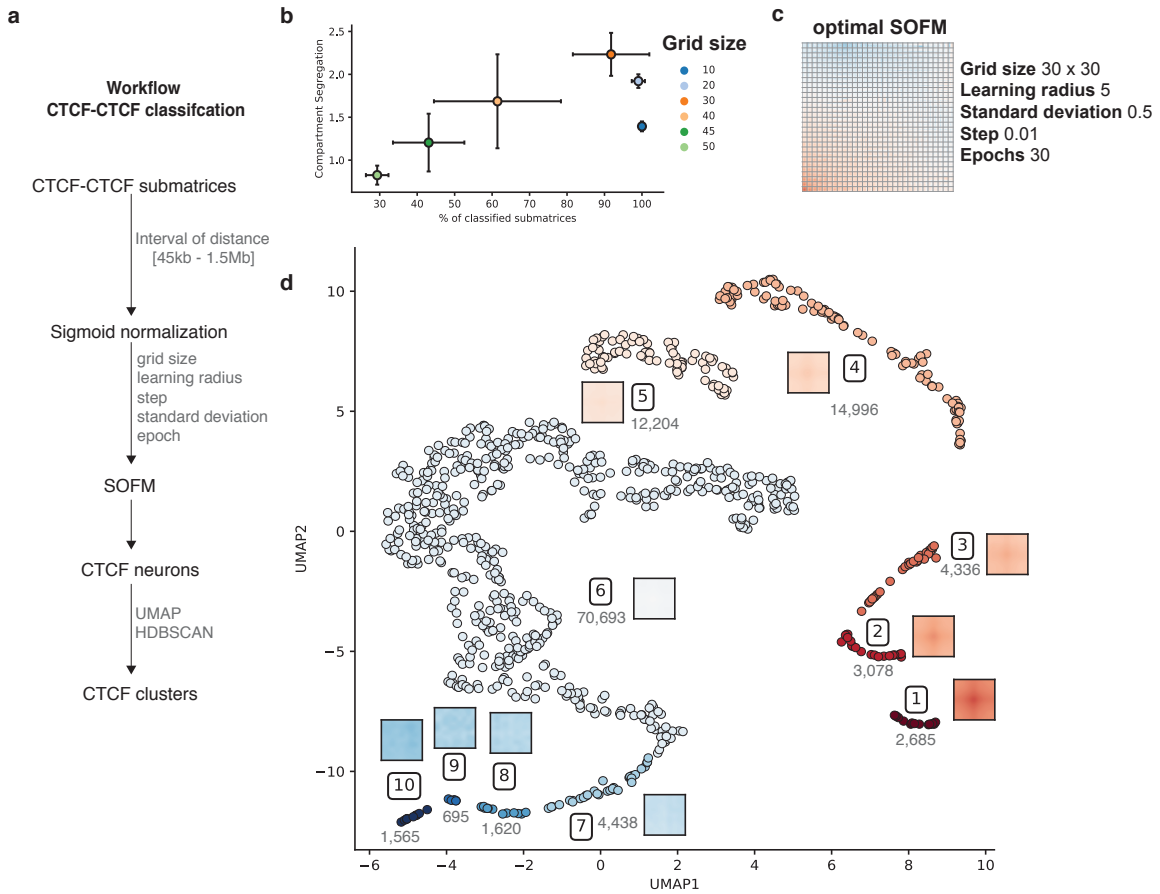
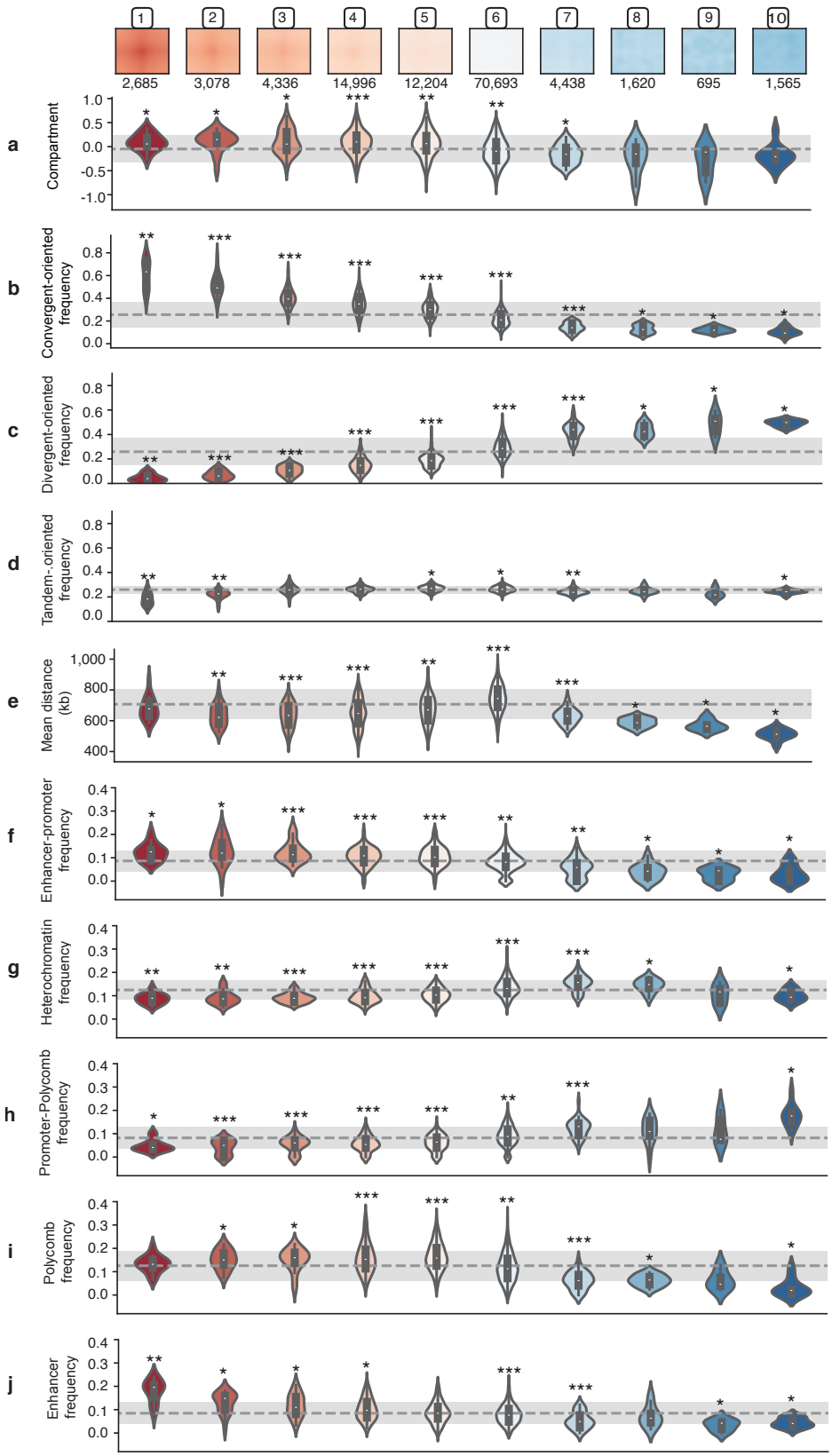


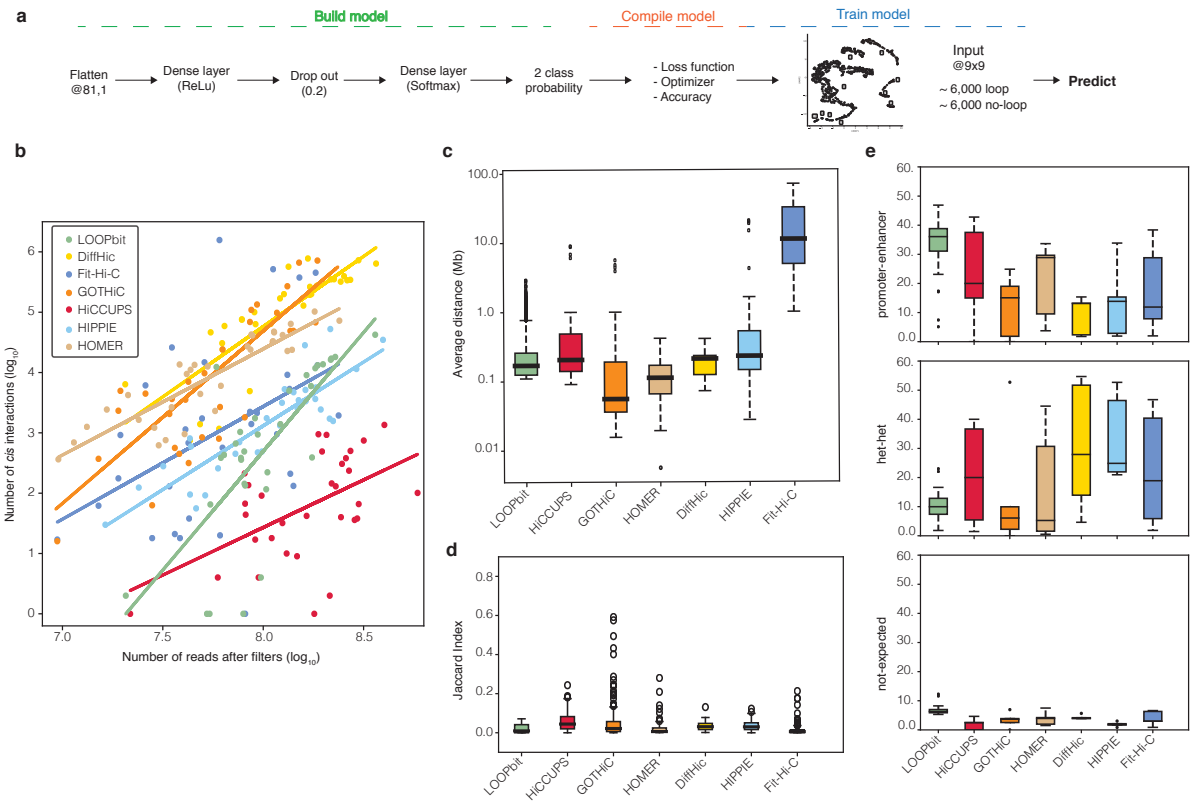
Figure 1. General overview of signal structure deconvolution. a. Schematic workflow to obtain CTCF clusters. **b.** SOFM parameters selected based on the percentage of classified matrices and the compartment segregation value (see also **Supplementary Fig. 1**). **c.** SOFM map showcasing the medoid of each neuron for optimal SOFM with the highest compartment segregation as well as highest percentage of classified CTCF-CTCF submatrices. **d.** Low dimensional representation of the SOFM neurons using a UMAP algorithm followed by a clustering method using HBDSCAN. A total of 10 clusters were obtained. Each point corresponds to a neuron in the SOFM map in panel c. Clusters are represented by the medoid signal and the number of CTCF-CTCF submatrices per cluster.



600

Figure 2. Distribution of multiple genomic features throughout CTCF clusters. In the first row the representative medoid of each CTCF cluster is represented. Below the different

genomic features per cluster. In the first row the compartment type distribution, then the binding CTCF motif orientation, the distance between the two CTCFs and finally the chromatin
605 state enrichment of the CTCF pairs. The dashed gray lines mark the mean, whereas the standard deviation is marked with lighter grey. Statistical significance against the mean is calculated using Wilcoxon test, $p < 0.05$ (*), $p < 0.001$ (**), $p < 0.0001$ (***).



610 **Figure 3. LOOPbit technical and biological benchmarking. a.** CNN workflow, model building with the multiple layers to transform the input data. Model compilation to assess the accuracy and optimization of the model. Finally, training of the model using the data coming from the CTCF-CTCF SOFM deconvolution. **b.** Representation of the number of reads after filters and the number of identified *cis*-interactions by LOOPbit in all experiments at a 5 kb

615 resolution (n=32). **c.** Average distance between the identified loop-anchors of all the Hi-C experiments at 5 kb resolution (n=32). **d.** Boxplot representing the Jaccard Index, in here the overlapping between exact loop-anchors were considered to be the same loop between the same replicates (n=39). **e.** Proportion of the identified cis interactions based on the chromatin states at their anchoring points of the datasets at 5 kb resolution (n=32).

620

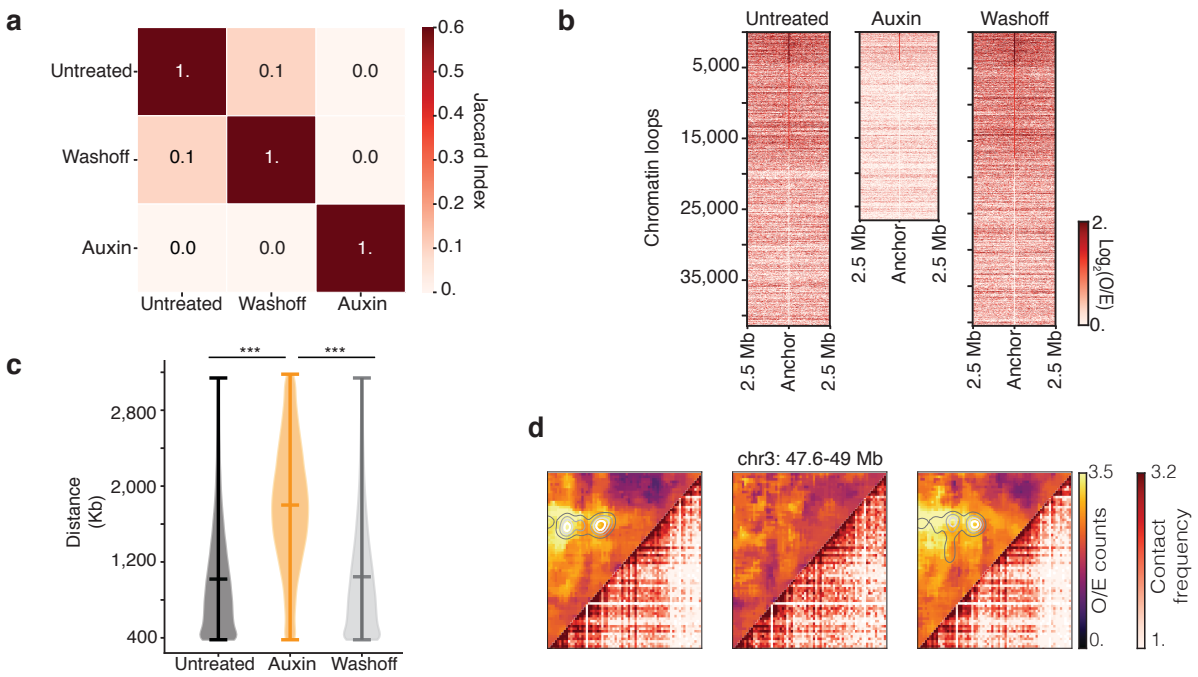


Figure 4. Applicability of LOOPbit to AID-CTCF dataset (E. P. Nora et al., 2017). a. Jaccard Index (JI) between the three samples. **b.** CTCF enrichment in the loop anchors in the untreated, CTCF-degraded and wash-off datasets. **c**. Distance between loop-anchors in the three conditions. Statistical significance against the mean is calculated using Wilcoxon test, $p < 0.05$ (*), $p < 0.001$ (**), $p < 0.0001$ (***). **d.** Example of an output of LOOPbit from a genomic region (chr3:47,600,000-49,000,000), showing half of the normalized Hi-C matrix, and half with the contact frequencies. The density of probabilities provided by LOOPbit define the position of the chromatin loops (white and grey lines).