The convergent CTCF loop formed by DNA extrusion is the basis of promoter enhancer interactions

Xizhe Zhang[1] and Sergio Branciamore[2]

[1]Department of Diabetes Complications & Metabolism, DMRI, City of Hope
[2]Department of Diabetes Complications & Metabolism, DMRI, City of Hope

June 28, 2021

June 28, 2021

1 abstract

The connections between HiC data and DNA extrusion have not been fully understood. In this study, we proposed a new method to reconstruct the DNA extrusion scenarios through modeling the aggregated HiC pattern. We examined several DNA extrusion scenarios which specify how promoters, enhancers, and CTCF binding sites interact with other regions. In particular, we showed that about half of all promoters tend to search enhancers in a specific direction. The underlying DNA extrusion scenario is characterized by a divergent CTCF pair and the imbalanced DNA extrusion activities which pull the promoters to a specific direction. Our findings proposed that most promoters do not contact enhancers directly. In contrast, promoters and enhancers are brought in proximity by the convergent CTCF loops.

2 Introduction

High throughput chromosome conformation capture (HiC) was developed to examine the chromatin interactions in a genome-wide and unbiased mode[1][2]. After processing, HiC data is presented as a two-dimensional heatmap in which the color intensity of each dot is proportional to the interaction strength between two corresponding genomic loci. Several important structures have been identified from the HiC data, including compartments, topological associated domains (TADs), and loops. Compartments have two types A and B, which are associated with the open or closed chromatin signatures respectively[1]. The boundaries of TADs may serve as the insulators to prevent abnormal promoter enhancer interactions (PEIs)[3][4]. Loop is recognized as a local HiC peak which represents a stable interaction between two distal genomic regions. The two anchors of loops are associated with a pair of CCCTC-binding consensus sequences (CTCF motif) whose directions are not random but face each other[5]. This observation is known as the convergent CTCF rule which can not be explained by a random diffusion model. In contrast, it is consistent with the hypothesis that the two distal loop anchors are brought in proximity by a process called DNA extrusion, which is highly dynamic and heterogeneous in a cell[2][6].

Recent studies have revealed several important details about DNA extrusion. First, the protein complexes which are responsible for reeling DNA to form loops have been identified, including condensin, cohesin, and bacterial structural maintenance of chromosomes (SMC) complex[7]. Second, it was shown that the loading of DNA extrusion complex onto the genome is dependent on NIPBL-MAUL and the releasing of DNA extrusion complex from DNA is by WAPL[8][9]. Third, as revealed by the in vitro reconstitution system, the reeling of DNA could be uni-directional and bi-directional[10][11]. Furthermore, two DNA extrusion complexes may traverse to each other and form the z-loop structure[12]. Finally, It has been shown that the HiC features, such as the domain corners, flame edges, and loops could be reproduced by modeling DNA extrusion[13].

At present, several questions about the HiC matrix remain to be answered. For example, most CTCF binding sites (CBS) are not associated with the recognized HiC features. What is the function of these sites? How promoters and enhancers find each other? Why most high-intensity interactions identified from the HiC matrix are associated with the CBSs[14]? We postulated that these questions are all relevant to DNA extrusion activities. Studying the connections between DNA extrusion and HiC data is not a trivial task.

However, it is a big challenge to rebuild the DNA extrusion process based on the HiC data directly. HiC signals could be formed not only by DNA extrusion but also by polymer interactions or random collisions. DNA extrusion may contribute to only a small fraction of HiC signals. A useful tool to overcome this obstacle is the aggregated peak analysis (APA), or pile-up analysis. APA has been widely used to enhance the HiC signals associated with certain features, like promoters and transcriptional factor binding sites[15][16][17]. We postulated that during aggregation, the signals generated by DNA extrusion could be enhanced, whereas those from the specific local interactions are averaged. It is necessary to develop a method that can reconstruct DNA extrusion from the aggregated maps.

In this study, we first presented how to reconstruct the underlying DNA extrusion scenarios from the aggregated patterns. Next, we established the DNA extrusion scenarios associated with the CBS, promoters, and enhancers. In particular, we demonstrated that about half of all promoters tend to interact with enhancers in a specific direction which is called the directional bias. The underlying scenario is characterized by a divergent CTCF pair. The observed directional bias is the result of two imbalanced DNA extrusion activities. Our analysis demonstrated that DNA extrusion does not stop or pause at enhancers or promoters. On the contrary, promoters and enhancers are brought in proximity by the DNA extrusion which forms the convergent CTCF loops.

3 Results

3.1 The aggregated pattern could be reproduced by a collection of compatible DNA extrusion scenarios

For the convenience of description, all aggregated maps in this study are divided into three sub-domains (Fig1a). The two smaller triangle areas which are symmetrical to the center are named subdomain A (SDA, left to the center) and subdomain B (SDB, right to the center). The square area locating between subdomain A and B is named the insulated domain (ISD).

The CTCF aggregated pattern (CAP, Fig1b) shows the following features. First, the signals within SDA and SDB are higher than those within the ISD which form two

boundaries. Second, there are signals enriched within the boundaries, which forms two strip-like structures (referred to as "strip" in this study)[18][19]. But it is important to note that the strips and boundaries are independent of each other. Third, the whole pattern is approximately symmetrical to the center. Of note, these features can not be recognized from most component maps (randomly selected from a total of 25894 maps, FigS1a), which is true for all the aggregated patterns in this study.

To explain this observation, we must find out what signals are enriched by aggregation. We divided the HiC signals into two classes. The class I signals are contributed by DNA extrusion activities, and the class II signals are formed by the random collision between DNA fragments. The class II signals are specific to the local regions and are much more abundant than class I signals. Aggregation should not enrich the class II signals.

However, it is not straightforward to confirm that aggregation will enrich the class I signals. The class I signals from the local HiC data are very limited which makes it difficult to model the corresponding DNA extrusion scenarios. Alternatively, it is possible to model the compatible DNA extrusion scenarios which demonstrate the shared and special characteristics associated with the component regions. For example, as the component genomic regions of CAP all contain the CBS at the center, the compatible DNA extrusion scenarios should have the center insulators as the shared characteristics. On the other hand, the special characteristics refer to the local DNA extrusion behaviors, such as the positions of DNA extrusion starting points (DSPs) and the velocity of anchor moving (See methods for details).

We generated 2000 images based on the fixed and variable parameters described above. Aggregating these images has reproduced the major features of CAP, including the strips, the boundaries, and the overall symmetrical structure (Fig1c). We noticed that most component images will form part of the strip which will be aggregated into the whole strip finally. Furthermore, the boundaries are formed because the center insulators prevent the DNA extrusion activities from entering the ISD. (FigS1b). The characterized DNA extrusion movement shared by most component regions is that one anchor is stopped at the center insulator, while the other anchor continues moving.

Altogether, we concluded that the aggregated HiC pattern reflects a collection of DNA extrusion scenarios that are compatible with the component regions. Piling up the images generated by these DNA extrusion scenarios can reproduce the major features of the aggregated HiC pattern. In another word, the aggregated pattern can be used to rebuild the compatible DNA extrusion scenarios.

3.2 The directional bias of CBSP and CBSM can be explained by DNA extrusion

The convergent CTCF rule indicates that the CBS with plus motif (CBSP) or minus motif (CBSM) interact with other genomic regions with strong directional bias[5][20][14]. However, most CBSs are not associated with the recognized HiC features and the relevant DNA extrusion scenarios have not been elaborated. We then generated the aggregated maps centered on the genome-wide CBSP or CBSM. The corresponding patterns are named the CTCF plus aggregated pattern (CPAP, Fig2a) and CTCF minus aggregated pattern respectively (CMAP, FigS2a). The most striking difference between CPAP and CMAP is that the strip locates between the ISD and SDA in CMAP but between the ISD and SDB in CPAP.

In the following, we used CPAP as the example to illustrate the underlying DNA

extrusion scenario. As discussed in the last section, the CAP (Fig1b) requires that the center CBS is an insulator and the DNA extrusion starting points (DSP) are randomly selected from regions upstream or downstream of the center. As CPAP shows only one strip, does it mean that the DNA extrusion events are initiated only downstream of the center? The pattern generated by the corresponding scenario indeed shows one strip locating at the expected position, but it loses all the signals within the SDA at the same time(Fig2b).

Another possible scenario is that DNA extrusion events could be initiated both upstream and downstream but those starting upstream can traverse the center insulator. Loop traversing is not consistent with the DNA extrusion behavior proposed by previous studies[21][22][18][19]. However, a recent study demonstrated that loop traversing is possible for the condensin mediated DNA extrusion reconstituted in vitro[12]. The scenario in which the upstream loop extrusion can traverse the center insulator generates a pattern showing one strip and signals within the SDA and ISD. Different from CPAP, this pattern loses the boundary between the ISD and SDA(Fig2c).

Finally, we reproduced the whole CPAP by the following DNA extrusion scenario(Fig2d). First, DNA extrusion events are initiated both upstream and downstream. Second, only the downstream ones can get access to the center. Third, DNA extrusion events initiated from the upstream are blocked by the CBSM upstream which can not affect the center CBSP. This scenario indicates that one anchor of the DNA extrusion initiated downstream is blocked by the center CBSP, while the other anchor moves further downstream, which is manifested as the directional bias of the CBSP. The modeling on CMAP follows the same way(See FigS2b to FigS2d).

Of note, the upstream CBSM and the center CBSP form the divergent CTCF pair which is the essential structure to generate CPAP. Moreover, it also implicates that two DNA extrusion colliding at the CBS is not a frequent event because this will generate two strips in CPAP (or CMAP). For the genomic regions containing the tandem CBS, our findings suggest that these CBSs are associated with one dominant DNA extrusion event. Otherwise, loop collision at the CBS should be observed frequently (See the discussion below).

3.3 Half of the promoters show strong directional bias to interact with enhancers

The directional bias associated with CBSP and CBSM can be explained by DNA extrusion. We next try to reconstruct the DNA extrusion scenarios associated with the promoters, which may reveal how PEIs are formed. We first generated the promoter aggregated patterns in several different cellular context(Fig3a, FigS3a, and FigS3b). The promoter aggregated patterns are similar to the results reported previously [15][17], which does not show directional bias.

However, the directional bias could be associated with a subset of promoters. To represent a promoter's directional bias, we computed a value (Directional Bias index, DBI) which is equal to the difference between the number of upstream PEIs (PEIs) and the downstream PEIs (See methods for the definition of PEIs). The promoters that only contact one enhancer are excluded from analysis since their directional bias can not be judged by this method. Based on this value, promoters could be categorized into three groups, named G1 (17150 counts), G2(17163 counts), and G3 (35833 counts). G1 promoters have the positive DBI and G2 promoters have the negative DBI. G3 is all other promoters that are not G1 or G2.

Consistent with the DBI, the aggregated PEI signals centered on G1 and G2 promoters show strong directional bias (Fig3b). Of note, the directional bias is not due to the imbalanced distribution of enhancers relative to the promoters. For details, we observed 5.4 times more signal on the stronger side than the weaker side for G1 promoters(Fig3b, left panel), whereas there are only 1.4 fold more enhancers located upstream than that located downstream for the G1 promoters (Fig3c, left panel, $p <$ 1e-16, binomial distribution). A similar tendency is easily recognized for G2 promoters (Right panel in Fig3b and Fig3c). In general, G1 and G2 promoters have much higher expression levels than the G3 promoters indicating that the directional bias could be important for the promoter's activity (FigS3c).

3.4 The directional bias of G1/G2 promoters is dependent on the neighboring Ctcf binding sites

The aggregated pattern above is about the PEI signals only. We next generated the aggregated HiC patterns centered on G1 promoters (G1AP, Fig4a) and G2 promoters (G2AP, FigS4a). We found that The overall structure between G1AP and CMAP is quite similar, indicating that these two patterns have strong connections. Similar relationships can be observed between G2AP and CPAP. Moreover, the G1AP and G2AP show a more diffusive strip than that in CMAP and CPAP, suggesting the underlying DNA extrusion scenario is not equivalent.

The center element of the aggregated pattern is critical to reconstructing the associated DNA extrusion scenario. We then first examined the colocalization between G1 promoters (the center of G1AP) and CBSM (the center of CMAP). About 10% of G1 promoters locate at the same genomic bin (10kb) with a CBSM. This number increases to 60% if the colocalization region is expanded to 50 kb. The G1 promoters are divided into two groups based on their positions relative to the closest CBSM. In detail, the group I promoters locate downstream of the CBSM, while group II locate upstream of the CBSM. The corresponding aggregated patterns show the enriched signals at different sides of the boundary between the SDA and the ISD (Fig4b and Fig4c). The pattern I (for group I G1 promoters) shows more signals on the left side of the boundary (within the SDA), whereas pattern II shows move signals on the right side of the boundary (within the ISD). In another word, we found the enriched signals is more related to the CBSM than the center G1 promoters. The neighboring CBSM may serve as the real insulator to stop the DNA extrusion, whereas the G1 promoters at the center do not. We finally reproduced the G1AP and G2AP by the corresponding DNA extrusion scenarios (Fig4d and FigS4b).

The above findings also indicate that DNA extrusion generally does not stop at the promoters. Therefore, most promoters do not contact distal enhancers directly.

3.5 The directional bias of G1AP and G2AP is determined by the imbalanced DNA extrusion activities

We noticed that in addition to the CBSM, G1 promoters also neighbor to the CBSP (Fig5a and FigS5b), suggesting that the CBSM alone can not explain the G1AP completely. In detail, about 60% of G1 promoters neighbors to CBSM and 33% neighbors to CBSP. The similar thing can be found for G2 promoters (FigS5a and FigS5c). The neighboring region is defined as the 50kb region flanking the promoters (Same for the below).

The CBSP and CBSM neighboring to G1 promoters could form the convergent pair or divergent pair. The distributions of these two configurations were shown in the joint probability heatmaps (Fig5b, See FigS5d for G2 promoters). For a dot with the coordination i (x-axis) and j (y-axis) in the heatmap, the color intensity indicates the probability to find the CBSP at position i and a CBSM at position j. The lower triangle of the heatmap (regions left to the green dash line in Fig5b) represents the probability for the convergent Ctcf pairs and the upper triangle represents the probability for the divergent Ctcf pairs. According to the distribution of high-intensity dots, we found that the neighboring CBSP and CBSM mainly form the divergent pairs and locate downstream of the G1 promoters. For G2 promoters, CBSP and CBSM also form the divergent configuration but locate upstream(FigS5d).

Therefore, the G1 and G2 promoters could be pulled upstream and downstream by the DNA extrusion activities associated with the divergent CTCF pairs. Why both promoters show obvious directional bias? As discussed above, these two DNA extrusion activities may have different strengths which could be checked by comparing the signals in SDA and SDB. As the regions neighboring to the G1 promoter generally contain multiple CBSP and CBSM(FigS5b and FigS5c), we chose the targets by two methods(FigS5e). First, we chose a pair of divergent CBS whose linear distance is the largest. Second, we chose all the CBSP or CBSM within the regions (300kb) flanking the center. Of note, the upstream DNA extrusion activities are represented by the signals within the SDA in the aggregated pattern centered on the CBSP(Green box in the right panel in Fig5c). On the other hand, the downstream DNA extrusion activities are represented by the signals within the SDB in the aggregated pattern centered on the CBSM (Green box in the left panel in Fig5c). We consistently observed that the upstream DNA activities are more active than the downstream ones for G1 promoters (method1: Fig5c, method2: FigS5f). For G2 promoters, the downstream DNA extrusion activities are more active than the upstream ones.(Fig5d).

Altogether, the directional bias of G1 and G2 promoters is the result of the imbalanced DNA extrusion activities associated with the local divergent CBS. The exact mechanism underlying the imbalanced DNA extrusion activities is not clear. But it suggested that the DNA extrusion initiation is under tight control.

3.6 DNA extrusion does not stop at the enhancers

A recent study proposed that PEIs are formed in a scenario in which the DNA extrusion stops or pauses at the enhancers or promoters, which will form the loops anchoring on promoters and enhancers[23]. However, our analysis indicated that one anchor of the DNA extrusion stops at the neighboring CBS but not G1 or G2 promoters. But where the other anchor stops is not revealed. Does it stop at the enhancers?

We try to answer this question by analyzing the aggregated maps centered on the enhancers. In detail, we collected the E1 enhancers which interact with the G1 promoters at a high intensity ($OE > 1.5$, see above for computing DBI) and the E2 enhancers which interact with G2 promoters at a high intensity. The aggregated patterns centered on E1 and E2 are named E1AP (Fig6a) and E2AP(Fig6b) respectively. E1AP and E2AP show the following features. First, the signals are enriched within the ISD. Second, the signals within ISD are not symmetrically distributed relative to the middle line. These two features are not observed for the other aggregated patterns discussed in this study.

What DNA extrusion scenario will generate a pattern in which the signals are en-

riched within the ISD? Based on the defined DNA extrusion behavior, the corresponding scenarios should have the following features. First, the center is the DSP (Fig6c). Second, a moving anchor cannot go back. Unfortunately, the genomic regions serving as the DSP have not been identified, which means the above scenario can not be validated by the aggregated map directly. Interestingly, we noticed that the aggregated HiC map centered on the middle regions between a pair of CTCF convergent binding sites shows the expected pattern(Fig6d), suggesting that the middle regions could be the potential DSPs.

On the other hand, the asymmetrical distribution of signals within E1AP and E2AP can be explained by assuming that the two anchors move at different velocities. In terms of the probability distribution, the different velocities of anchor moving can be simulated in the scenario in which the insulators have different densities at two sides of the center DSP. According to that, We reproduced the E1AP in the scenario in which there are more CBSP locating at upstream than DBSM locating at downstream(Fig6e). For E2AP, the corresponding scenario should be reversed, that is, more CBSM downstream than CBSP upstream(Fig6f). Consistently, we found that the E1 enhancers have more upstream CBSP(Fig6g), whereas E2 enhancers have more downstream CBSM (Fig6h).

The above analysis indicates that DNA extrusion does not stop or pause at the enhancers. The observed PEIs may reflect a scenario in which the convergent CTCF loops bring promoters and enhancers in proximity. If this is true, we should find enhancers tend to neighbor to one anchor while promoters neighbor to the other anchor. For G1 promoters, we identified the closest CBSM and the closest upstream CBSP and examine how enhancers are distributed within this convergent CTCF domain (FigS6a). For three defined regions (50kb) that flank the upstream CBSP (Region 1), the middle position (Region 2), and the downstream CBSM, we found enhancers are more enriched within region 1 and 3 but relatively depleted within region 2(FigS6b). A similar finding is observed for G2 promoters(FigS6c). Therefore, the distribution of enhancers supports the hypothesis that PEIs rely on convergent CTCF loops.

4 Discussion

In this study, we introduced a new method that can reconstruct the DNA extrusion scenarios from the aggregated pattern. We proposed that the HiC signals directly relevant to DNA extrusion are very limited. However, the aggregation will enrich the DNA extrusion signals while average those generated by local specific DNA interactions. Our modeling has successfully reproduced several aggregated patterns suggesting that the underlying assumptions are reasonable.

Currently, two views are explaining how PEIs are formed by DNA extrusion. First, the DNA extrusion will pause or stop at the promoter and enhancers, which forms the loops anchoring on the promoter and enhancer directly[23]. Second, DNA extrusion will form the convergent CTCF loops which bring the promoters and enhancers in proximity. Furthermore, the communication between promoters and enhancers relies on the diffusion of transcriptional factors[24]. We demonstrated that the DNA extrusion stopping at promoters or enhancers should not be a frequent event. In contrast, the promoters and relevant enhancers should be enclosed in a convergent CTCF loop by DNA extrusion, which is also manifested as promoters searching enhancers in a specific direction. Altogether, our finding is consistent with the second view.

Our study highlighted that the DSP plays an important role within a specific DNA

extrusion scenario. The mechanism which regulates the DSP at both sides of the G1 and G2 promoters could explain why the imbalanced DNA extrusion activities exist. According to the previous studies, DSP could be mediated by the loading factor NIPBL-MAU2, the local transcription, and the releasing factor WAPL[8][9]. DSP may also contain the consensus sequences which will be the important targets for future genome editing studies. It will be interesting to know whether regulating the distribution of DSPs could affect PEIs, DNA recombination, and other genome folding events.

A recent study demonstrated that loops colliding at the CBS could form enhancer hubs that bring promoters and multiple enhancers together[25]. However, our analysis indicated that this scenario should not happen frequently as it will generate two strips in CPAP and CMAP. For the same reason, we proposed that the tandem CTCF plus or minus binding sites should be associated with a dominant DNA extrusion event but not with multiple ones. This is consistent with the DNA extrusion scenario reported for the protocadherin alpha gene. In this gene locus, only one promoter can interact with the distal enhancer region suggesting that a dominant DNA extrusion event is responsible for the PEI and no hub structure is formed[26]. This scenario may be valid for other genomic loci containing the tandem CBS, such as the IGH region in which the V(D)J recombination is implemented[27][28].

Finally, it is important to note that our conclusion is based on analyzing the aggregated patterns which reflect the average behaviors of many DNA extrusion events. In addition to the generalized model, we expect to see more details about the specific DNA extrusion scenario which underlies the local PEIs or other important processes.

5 Methods

5.1 HiC data processing and constructing the aggregated map

HiC data were obtained from GEO as shown in table I. The sequence data were aligned to hg19 or mm10 and the low mapping quality reads were removed. The data were further filtered for the duplicats, the same fragment reads, and the danglings. The raw matrices were represented at the 10kb resolution and normalized by the method described in[29]. The Observed/Expected matrices were obtained using the methods described in [1]. The aggregated maps were generated on the O/E matrices using the method described in [5].

5.2 Modeling DNA extrusion by markov chain

DNA extrusion is simulated by a markov chain model. The transition matrix represents the probability distribution for the DNA extrusion moving from one status (x_i, y_i) to another status (x_{i+1}, y_{i+1}), which is sampled from a uniform distribution if the moving is not relevant to the insulators or boundaries. We defined two type of insulators corresponding to the CBSM and CBSP, which will block the anchors from one sepcific direction respectively. The equilibrium distribution of the markov chain model is finally converted to the simulated HiC matrix. Refer to the supplementary materials, section 1 for more details.

5.3 Definition of PEIs

PEIs are the HiC interactions between a promoter region and an enhancer region and the O/E strength (Observed/Expected) is above 1.5. A promoter region is a 3kb genomic region center on an annotated promoter. An enhancer region is characterized by the histone marker H3K27Ac.

6 Tables

Table 1: Datasets

HiC	
GM12878 HiC	GSE63525
Mouse NPC	GSE96107
Mouse Neuron	GSE96107
Chip_seq	
GM12878 CTCF	GSM935611
GM12878 H3K27Ac	GSM733771
RNA_seq	
GM12878 RNAseq	GSM5330715

References

[1] 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., Sandstrom,R., Bernstein,B., Bender,M.A., Groudine,M., Gnirke,A., Stamatoyannopoulos,J., Mirny,L.A., Lander,E.S., and Dekker,J. (Oct, 2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science,* **326**(5950), 289–293.

[2] Dekker,J. and Mirny,L. (Mar, 2016) The 3D Genome as Moderator of Chromosomal Communication. *Cell,* **164**(6), 1110–1121.

[3] Lupiáñez,D.G., Kraft,K., Heinrich,V., Krawitz,P., Brancati,F., Klopocki,E., Horn,D., Kayserili,H., Opitz,J.M., Laxova,R., Santos-Simarro,F., Gilbert-Dussardier,B., Wittler,L., Borschiwer,M., Haas,S.A., Osterwalder,M., Franke,M., Timmermann,B., Hecht,J., Spielmann,M., Visel,A., and Mundlos,S. (May, 2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell,* **161**(5), 1012–1025.

[4] Luppino,J.M., Park,D.S., Nguyen,S.C., Lan,Y., Xu,Z., Yunker,R., and Joyce,E.F. (08, 2020) Cohesin promotes stochastic domain intermingling to ensure proper regulation of boundary-proximal genes. *Nat Genet,* **52**(8), 840–848.

[5] 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., and Aiden,E.L. (Dec, 2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell,* **159**(7), 1665–1680.

[6] Rowley,M.J. and Corces,V.G. (12, 2018) Organizational principles of 3D genome architecture. *Nat Rev Genet,* **19**(12), 789–800.

[7] Davidson,I.F. and Peters,J.M. (Mar, 2021) Genome folding through loop extrusion by SMC complexes. *Nat Rev Mol Cell Biol,*.

[8] Busslinger,G.A., Stocsits,R.R., van der Lelij,P., Axelsson,E., Tedeschi,A., Galjart,N., and Peters,J.M. (04, 2017) Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. *Nature,* **544**(7651), 503–507.

[9] Haarhuis,J.H.I., van der Weide,R.H., Blomen,V.A., Yanez-Cuna,J.O., Amendola,M., van Ruiten,M.S., Krijger,P.H.L., Teunissen,H., Medema,R.H., van Steensel,B., Brummelkamp,T.R., de Wit,E., and Rowland,B.D. (05, 2017) The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell,* **169**(4), 693–707.

[10] Ganji,M., Shaltiel,I.A., Bisht,S., Kim,E., Kalichava,A., Haering,C.H., and Dekker,C. (04, 2018) Real-time imaging of DNA loop extrusion by condensin. *Science,* **360**(6384), 102–105.

[11] Banigan,E.J., van den Berg,A.A., Brandão,H.B., Marko,J.F., and Mirny,L.A. (04, 2020) Chromosome organization by one-sided and two-sided loop extrusion. *Elife,* **9**.

[12] Kim,E., Kerssemakers,J., Shaltiel,I.A., Haering,C.H., and Dekker,C. (03, 2020) DNA-loop extruding condensin complexes can traverse one another. *Nature,* **579**(7799), 438–442.

[13] Fudenberg,G., Abdennur,N., Imakaev,M., Goloborodko,A., and Mirny,L.A. (2017) Emerging Evidence of Chromosome Folding by Loop Extrusion. *Cold Spring Harb Symp Quant Biol,* **82**, 45–55.

[14] Zhang,X., Branciamore,S., Gogoshin,G., Rodin,A.S., and Riggs,A.D. (11, 2017) Analysis of high-resolution 3D intrachromosomal interactions aided by Bayesian network modeling. *Proc. Natl. Acad. Sci. U.S.A.,* **114**(48), E10359–E10368.

[15] Bonev,B., Mendelson Cohen,N., Szabo,Q., Fritsch,L., Papadopoulos,G.L., Lubling,Y., Xu,X., Lv,X., Hugnot,J.P., Tanay,A., and Cavalli,G. (Oct, 2017) Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell,* **171**(3), 557–572.

[16] Rao,S.S.P., Huang,S.C., Glenn St Hilaire,B., Engreitz,J.M., Perez,E.M., Kieffer-Kwon,K.R., Sanborn,A.L., Johnstone,S.E., Bascom,G.D., Bochkov,I.D., Huang,X., Shamim,M.S., Shin,J., Turner,D., Ye,Z., Omer,A.D., Robinson,J.T., Schlick,T., Bernstein,B.E., Casellas,R., Lander,E.S., and Aiden,E.L. (Oct, 2017) Cohesin Loss Eliminates All Loop Domains. *Cell,* **171**(2), 305–320.

[17] Hsieh,T.S., Cattoglio,C., Slobodyanyuk,E., Hansen,A.S., Rando,O.J., Tjian,R., and Darzacq,X. (05, 2020) Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. *Mol Cell,* **78**(3), 539–553.

[18] Vian,L., P?kowska,A., Rao,S.S.P., Kieffer-Kwon,K.R., Jung,S., Baranello,L., Huang,S.C., El Khattabi,L., Dose,M., Pruett,N., Sanborn,A.L., Canela,A., Maman,Y., Oksanen,A., Resch,W., Li,X., Lee,B., Kovalchuk,A.L., Tang,Z., Nelson,S., Di Pierro,M., Cheng,R.R., Machol,I., St Hilaire,B.G., Durand,N.C., Shamim,M.S., Stamenova,E.K., Onuchic,J.N., Ruan,Y., Nussenzweig,A., Levens,D., Aiden,E.L.,

and Casellas,R. (05, 2018) The Energetics and Physiological Impact of Cohesin Extrusion. *Cell,* **173**(5), 1165–1178.

[19] Fudenberg,G., Abdennur,N., Imakaev,M., Goloborodko,A., and Mirny,L.A. (2017) Emerging Evidence of Chromosome Folding by Loop Extrusion. *Cold Spring Harb. Symp. Quant. Biol.,* **82**, 45–55.

[20] de Wit,E., Vos,E.S., Holwerda,S.J., Valdes-Quezada,C., Verstegen,M.J., Teunissen,H., Splinter,E., Wijchers,P.J., Krijger,P.H., and de Laat,W. (Nov, 2015) CTCF Binding Polarity Determines Chromatin Looping. *Mol Cell,* **60**(4), 676–684.

[21] Riggs,A.D. (Jan, 1990) DNA methylation and late replication probably aid cell memory, and type I DNA reeling could aid chromosome folding and enhancer function. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.,* **326**(1235), 285–297.

[22] 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., Geeting,K.P., Gnirke,A., Melnikov,A., McKenna,D., Stamenova,E.K., Lander,E.S., and Aiden,E.L. (Nov, 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**(47), E6456–6465.

[23] Krietenstein,N., Abraham,S., Venev,S.V., Abdennur,N., Gibcus,J., Hsieh,T.S., Parsi,K.M., Yang,L., Maehr,R., Mirny,L.A., Dekker,J., and Rando,O.J. (05, 2020) Ultrastructural Details of Mammalian Chromosome Architecture. *Mol Cell,* **78**(3), 554–565.

[24] El Khattabi,L., Zhao,H., Kalchschmidt,J., Young,N., Jung,S., Van Blerkom,P., Kieffer-Kwon,P., Kieffer-Kwon,K.R., Park,S., Wang,X., Krebs,J., Tripathi,S., Sakabe,N., Sobreira,D.R., Huang,S.C., Rao,S.S.P., Pruett,N., Chauss,D., Sadler,E., Lopez,A., N?brega,M.A., Aiden,E.L., Asturias,F.J., and Casellas,R. (08, 2019) A Pliable Mediator Acts as a Functional Rather Than an Architectural Bridge between Promoters and Enhancers. *Cell,* **178**(5), 1145–1158.

[25] Allahyar,A., Vermeulen,C., Bouwman,B.A.M., Krijger,P.H.L., Verstegen,M.J.A.M., Geeven,G., van Kranenburg,M., Pieterse,M., Straver,R., Haarhuis,J.H.I., Jalink,K., Teunissen,H., Renkens,I.J., Kloosterman,W.P., Rowland,B.D., de Wit,E., de Ridder,J., and de Laat,W. (08, 2018) Enhancer hubs and loop collisions identified from single-allele topologies. *Nat Genet,* **50**(8), 1151–1160.

[26] Guo,Y., Monahan,K., Wu,H., Gertz,J., Varley,K.E., Li,W., Myers,R.M., Maniatis,T., and Wu,Q. (Dec, 2012) CTCF/cohesin-mediated DNA looping is required for protocadherin alpha promoter choice. *Proc Natl Acad Sci U S A,* **109**(51), 21081–21086.

[27] Zhang,Y., Zhang,X., Ba,Z., Liang,Z., Dring,E.W., Hu,H., Lou,J., Kyritsis,N., Zurita,J., Shamim,M.S., Presser Aiden,A., Lieberman Aiden,E., and Alt,F.W. (09, 2019) The fundamental role of chromatin loop extrusion in physiological V(D)J recombination. *Nature,* **573**(7775), 600–604.

[28] Lin,S.G., Ba,Z., Alt,F.W., and Zhang,Y. (2018) RAG Chromatin Scanning During V(D)J Recombination and Chromatin Loop Extrusion are Related Processes. *Adv Immunol,* **139**, 93–135.

[29] Knight,P. and Ruiz,D. (2012) A Fast Algorithm For Matrix Balancing.. *MA Journal of Numerical Analysis,*.

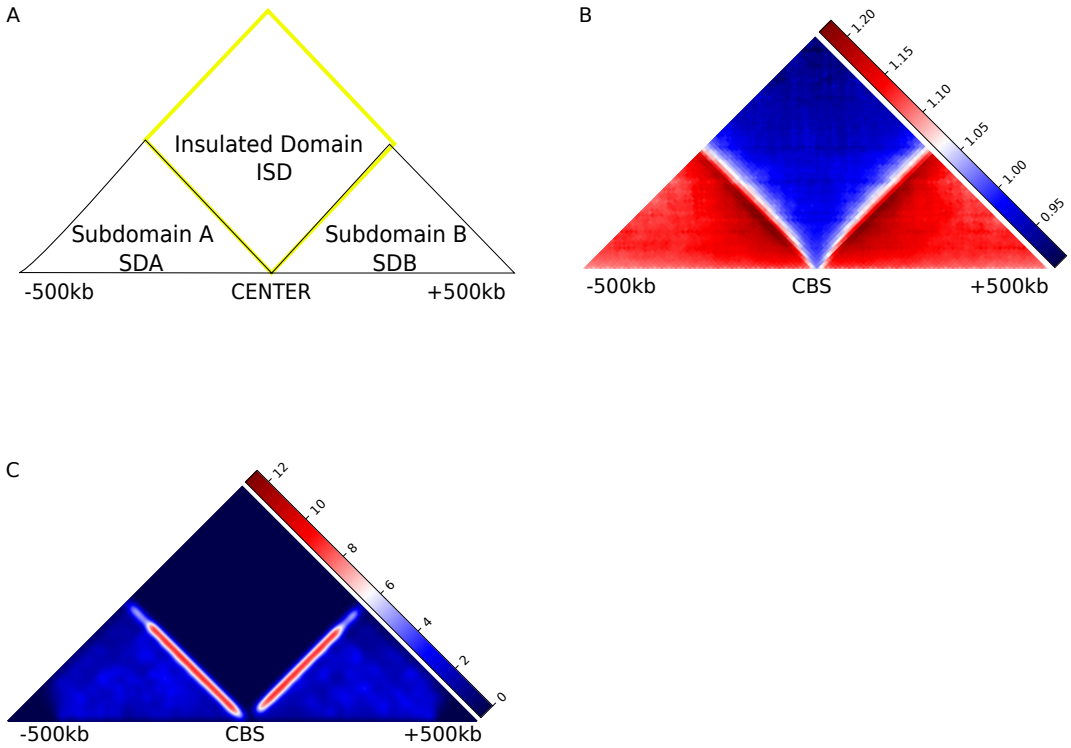


Figure 1: Modeling the CAP. (A)The whole aggregated map can be divided into three areas, SDA, ISD, and SDB. (B) The aggregated HiC map centered on the genome-wide CBS (Total 25849) in GM12878 cells. The unit is 10kb which is the same for the rest aggregated images in this study. (C) The simulated aggregation map corresponding to CAP, which is obtained by aggregating 2000 simulated images. Each image is generated by a DNA extrusion scenario with a CBS at the center.

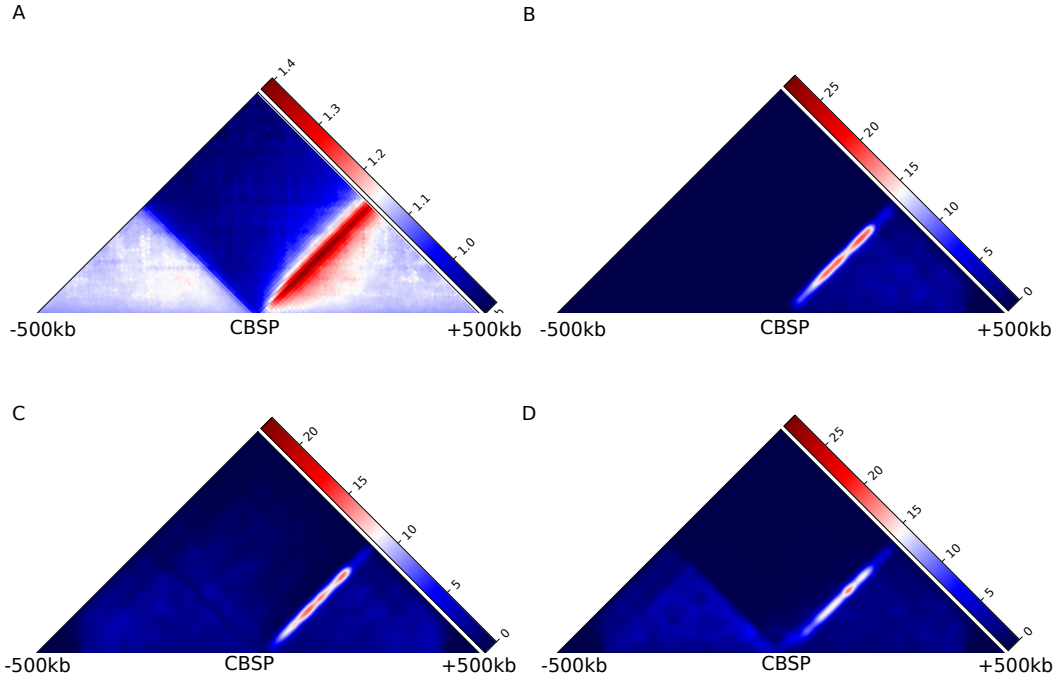


Figure 2: Modeling the CPAP. (A)The aggregated HiC map is centered on the genome-wide CBSP. The pattern includes two boundaries and one strip between ISD and SDB. (B) The simulated pattern is generated by the DNA extrusion scenario in which no DNA extrusion events are initiated upstream of the center. Note that this pattern does not have signals in SDA which is different from the pattern in (A). (C) The simulated pattern is generated by assuming that DNA extrusion initiated upstream can traverse the CBSP at the center. Note that it loses the boundary between SDA and ISD which is different from (A). (D) This pattern is generated by assuming that the upstream DNA extrusion is blocked by a CBSM locating upstream the center CBSP. It reproduced the major features in (A), including the two boundaries, the strip between ISD and SDB, and the overall structure.

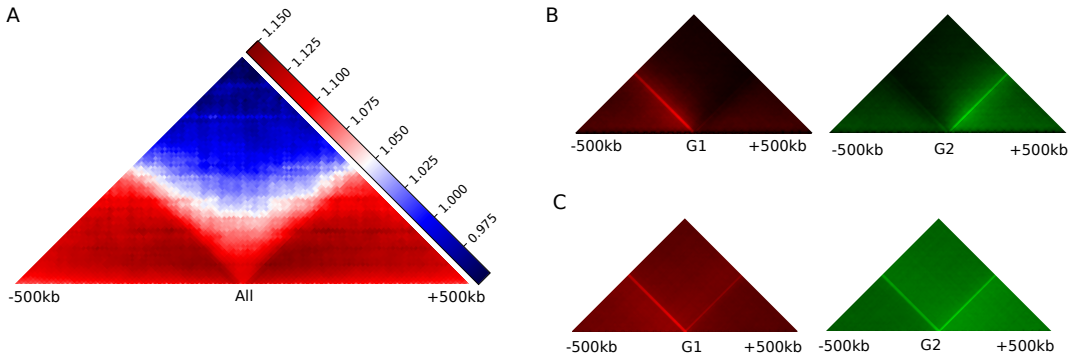


Figure 3: Promoters (G1 and G2) interact with enhancers in a specific direction. (A) The aggregated HiC map centered on all promoters in GM12878. (B) The directional bias of PEIs for G1 promoters(left, red color) and G2 promoters (right, green color). Note that only the PEI signals are aggregated. For G1 promoters, the left signals are 5.4 fold more than the right signals. (C) The distribution of the enhancer density for G1 promoters (left, red signal) and G2 promoters (right, green signal). For G1 promoters, the left enhancers are 1.4 more than the right enhancers.

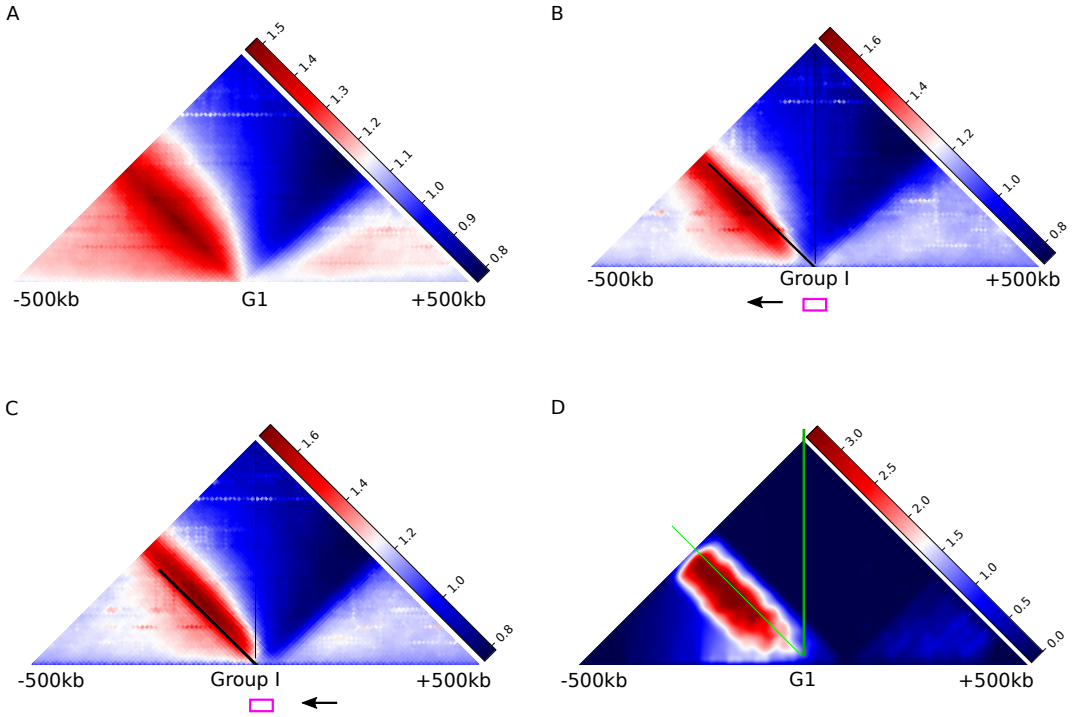


Figure 4: G1AP is related to the neighboring CBSM. (A) The aggregated HiC pattern centered on the G1 promoters (G1AP). Note that G1AP has a strip more diffusive than that in CMAP (FigS2a). (B) Aggregated HiC map centered on the Group I G1 promoters (Pattern I). The Group I G1 promoters (represented by the magnet box) locate downstream of the CBSM (represented by the black arrow). Note that the signals are enriched within the SDA, which is left to the boundaries (Black line)(C) Aggregated HiC map centered on the Group II G1 promoters (Pattern II). The Group II G1 promoters (represented by the magnet box) locate upstream of the CBSM (represented by the black arrow). Note that the signals are enriched within the ISD, which is right to the boundaries (Black line)(C)(D) The simulated reuslt for G1AP. Note that the diffusive strip is formed by the neighboring CBSM, not by the center G1 promoters

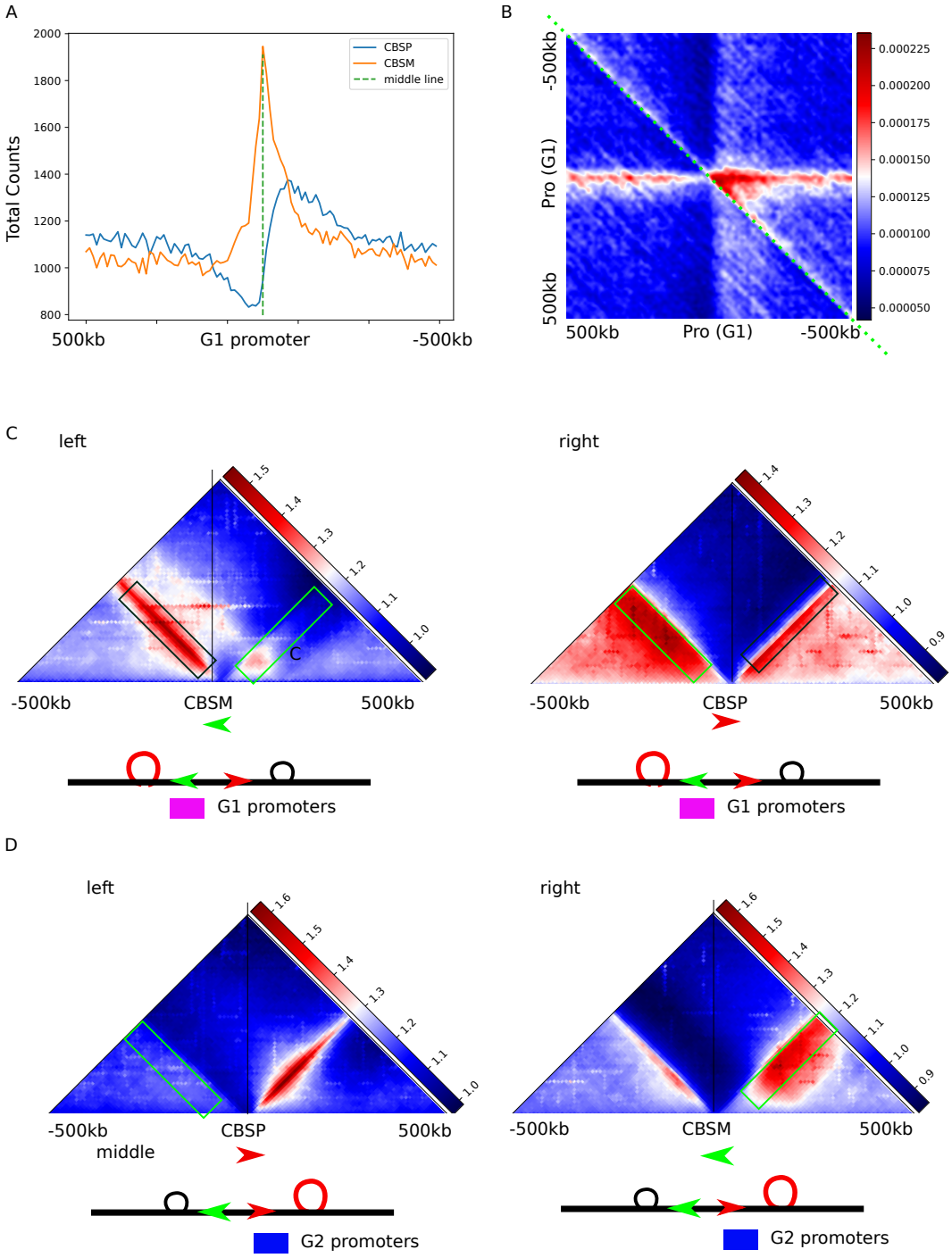
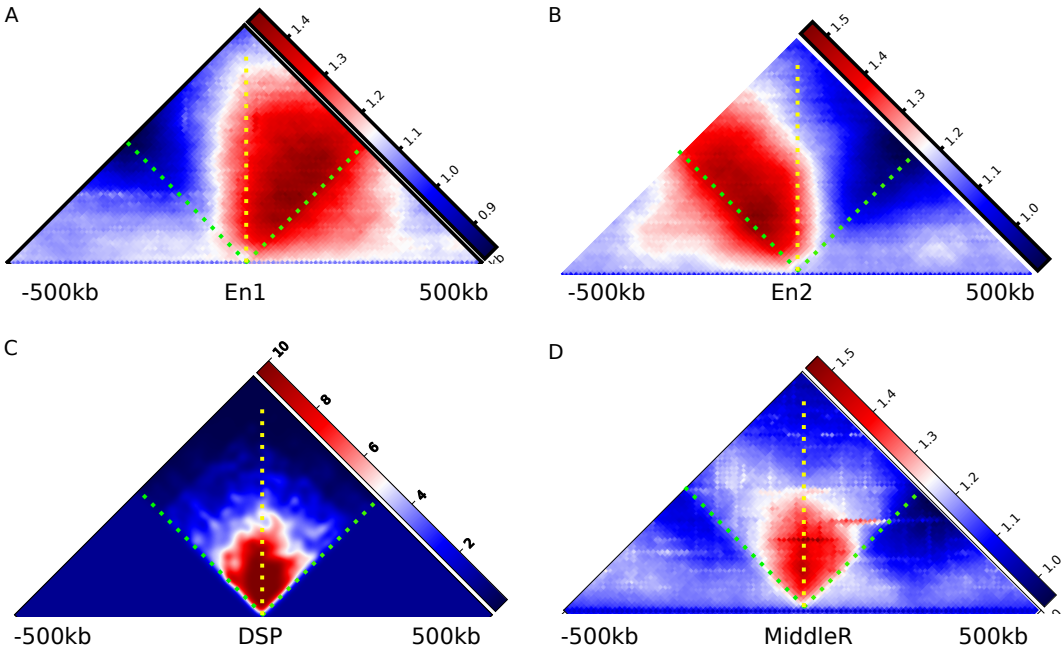


Figure 5: G1 promoters are pulled toward upstream by the stronger upstream DNA extrusion activities. (A) The distribution of CBSM and CBSP relative to the G1 promoters. Note that the number of CBSM (yellow line) is about twofold more than that of the CBSP (blue line). (B) The joint probability heatmap shows the distribution of the divergent or convergent CBS. The enriched high-intensity dots (red signal) locate downstream of the G1 promoters (Center) and represent the divergent configuration (Locate at the upper triangle). (C) The upstream DNA extrusion is more active than the downstream one. The upstream DNA extrusion activities are represented by the signals in the green box in the right panel. The downstream DNA extrusion activities are represented by the signals in the green box in the left panel. (D) Same as in (C) but for G2 promoters. The downstream DNA extrusion activities are stronger than the upstream ones.


A
1.4
1.3
1.2
1.1
1.0
0.9
-500kb
En1
500kb
B
1.5
1.4
1.3
1.2
1.1
1.0
-500kb
En2
500kb
C
10
8
6
4
2
-500kb
DSP
500kb
D
1.5
1.4
1.3
1.2
1.1
1.0
-500kb
MiddleR
500kb

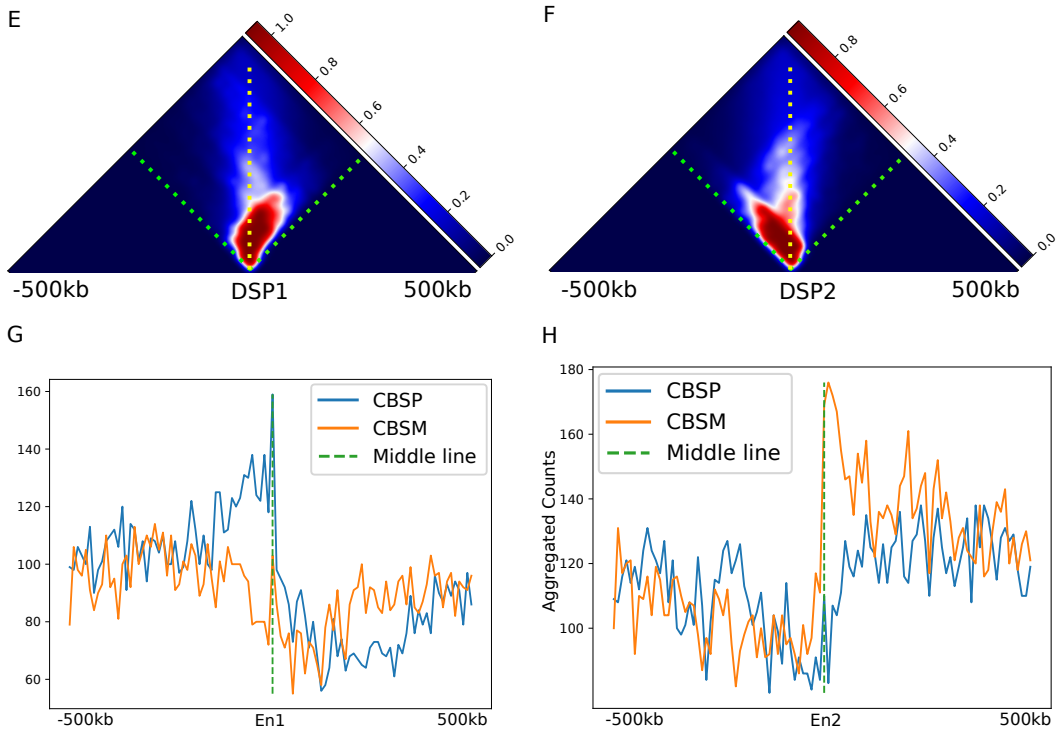


Figure 6: DNA extrusion does not stop at enhancers. (A) The aggregated HiC map centered on the E1 enhancers. (B) The aggregated HiC map centered on the E2 enhancers. Note that for both (A) and (B) the signals are enriched within the ISD and asymmetrical to the middle line (yellow dash line). (C) The simulated aggregated pattern centered on the DNA extrusion starting points (DSP). (D) The aggregated pattern centered on the middle region (middleR) of the convergent CBS associated with the G1 promoters. Refer to FigS6a. Note that the similarity between (C) and (D) suggests that some middle regions could be the DSPs. (E) The simulated result for the DSPs which have more upstream CBSP than downstream CBSM. (F)The simulated result for the DSPs which have more downstream CBSM than upstream CBSP. Note that the signals in (E) are more at the right side of the middle line (yellow dash line), whereas the signals in (F) are more at the left side. Regarding to the asymmetrical signal distribution, (E) is like (A), and (F) is like (B). (G). The distribution of CBSP and CBSM relative to the E1 enhancers. (H). Same as (G) but for En2 enhancers.

7 Appendix

7.1 Abbreviation

CBS: CTCF binding sites

CBSM: CTCF binding sites with minus motif

CBSP: CTCF binding sites with plus motif

CPAP: Ctcf plus aggregated pattern

CMAP: Ctcf minus aggregated pattern

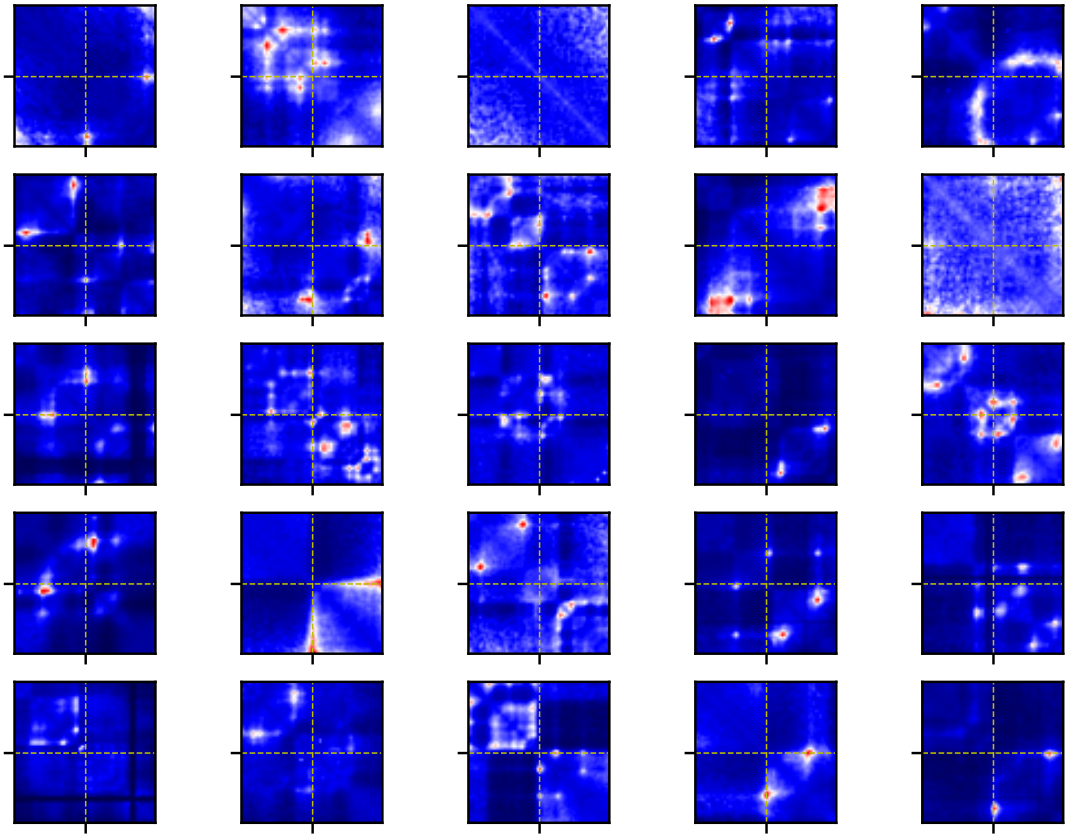
G1AP: G1 promoter aggregated pattern

G2AP: G2 promoter aggregated pattern

PEI: promoter enhancer interaction

7.2 Supplementary Figures

A



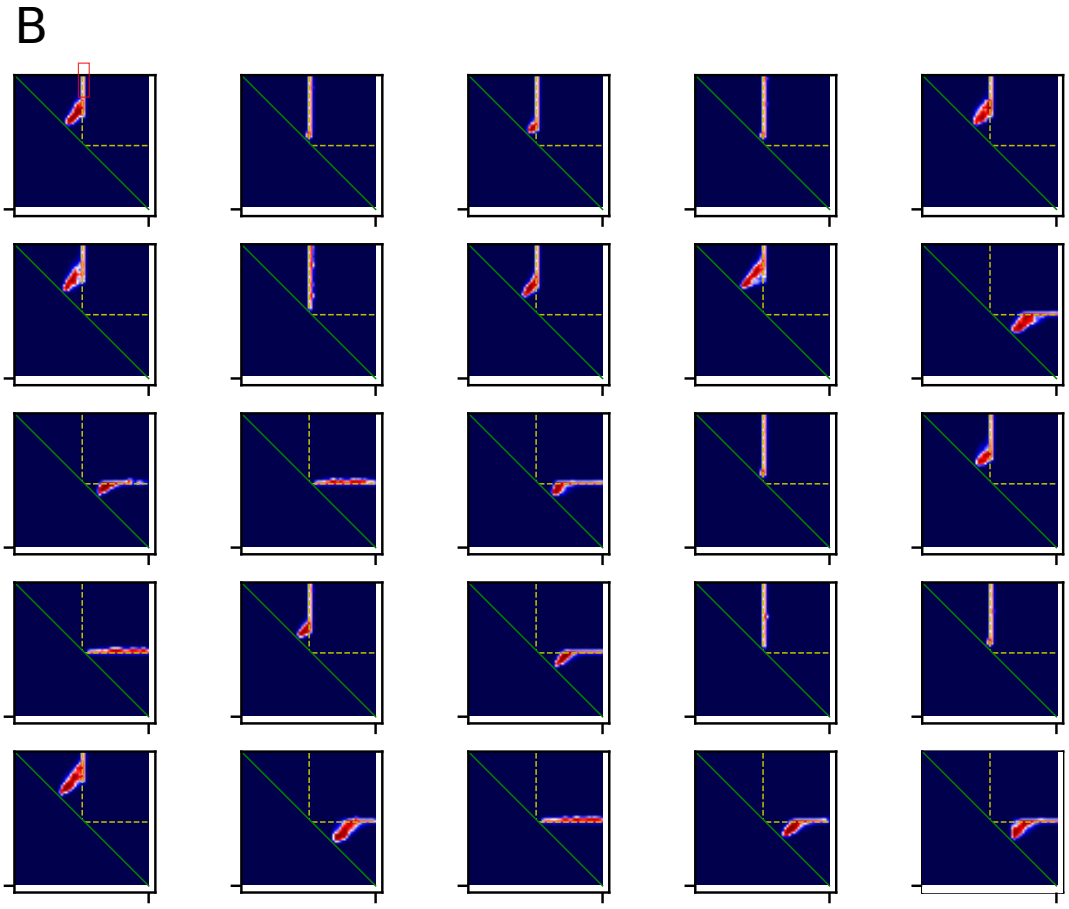


Figure S1: (A) The randomly selected component images which are aggregated to generate CAP(Fig1b). Note that the CAP can not be found in any of these images. (B) The randomly selected images generated by the compatible DNA extrusion scenarios. Note that most of the images have signals enriched to form part of the strip. See the signals within the red box in the top-left figure, which is an example about forming part of the strip (The vertical yellow dash line). The center CBS prevent the DNA extrusion from moving into the ISD (Between two yellow dash line) which will generate the different signal level between SDA/SDB and ISD, which is manifested as the boundaries.

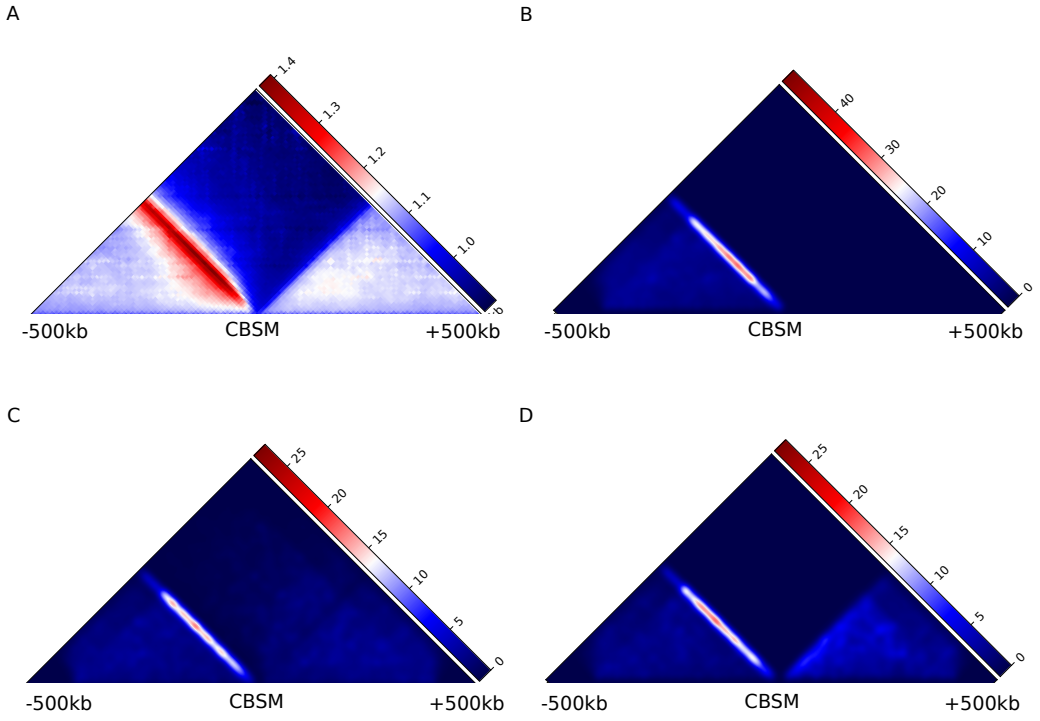


Figure S2: Modeling the CMAP. Refer to the main figures from Fig2a to Fig2d. (A) The aggregated map centered on the genome-wide CBSM. Note that this pattern has a strip between SDA and ISD, which is different from CPAP (Fig2a). (B) The simulated pattern is generated by the DNA extrusion scenario in which no DNA extrusion events are initiated downstream of the center. Note that this pattern does not have signals in SDB. (C) The simulated pattern is generated by assuming that DNA extrusion initiated downstream can traverse the CBSM at the center. Note that it loses the boundary between SDB and ISD. (D) This pattern shows the major features as in CMAP, including the two boundaries, the strip between ISD and SDA, and the overall structure.

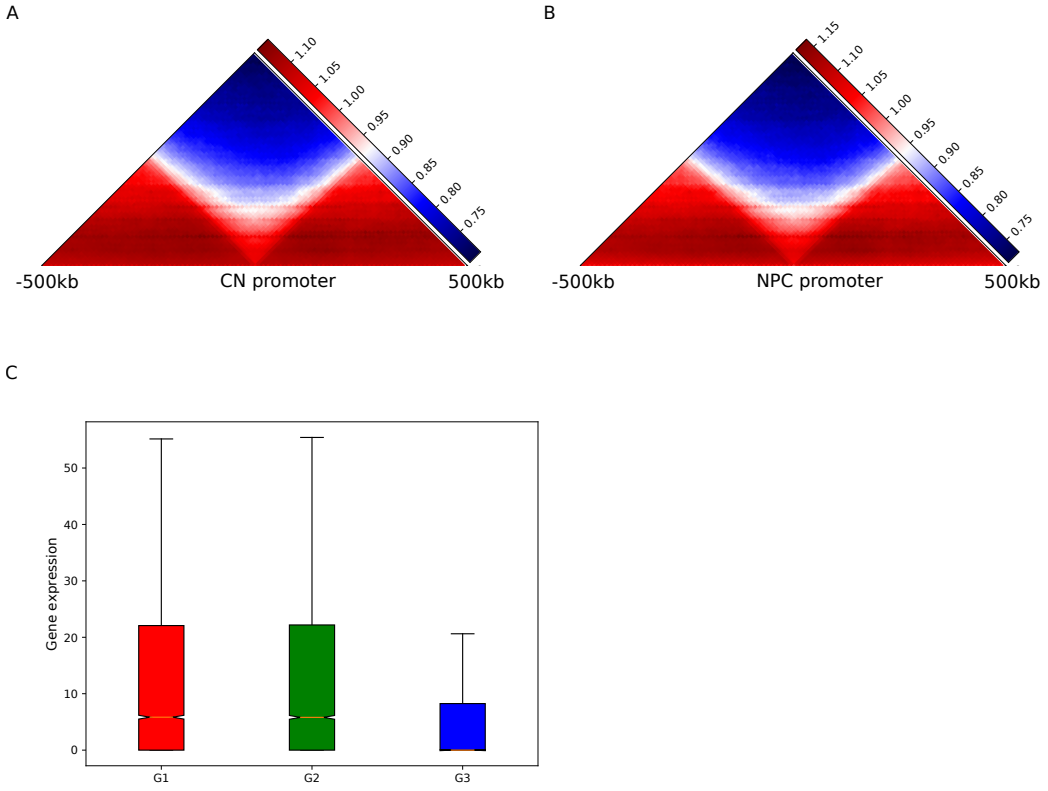


Figure S3: (A) Aggregated HiC map centered on promoters in mouse central neuron cells. (B) Aggregated HiC map centered on promoters in mouse neural progenitor cells. (C) The box plot figure shows the gene expression level associated with G1, G2, and G3 promoters in GM12878 cells.

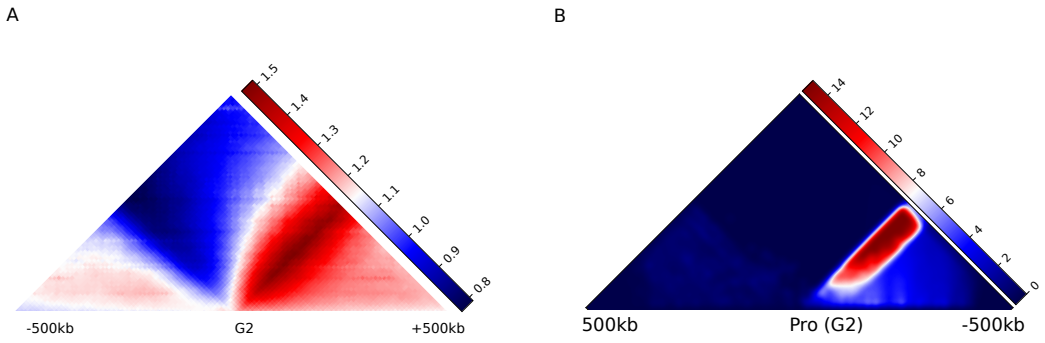
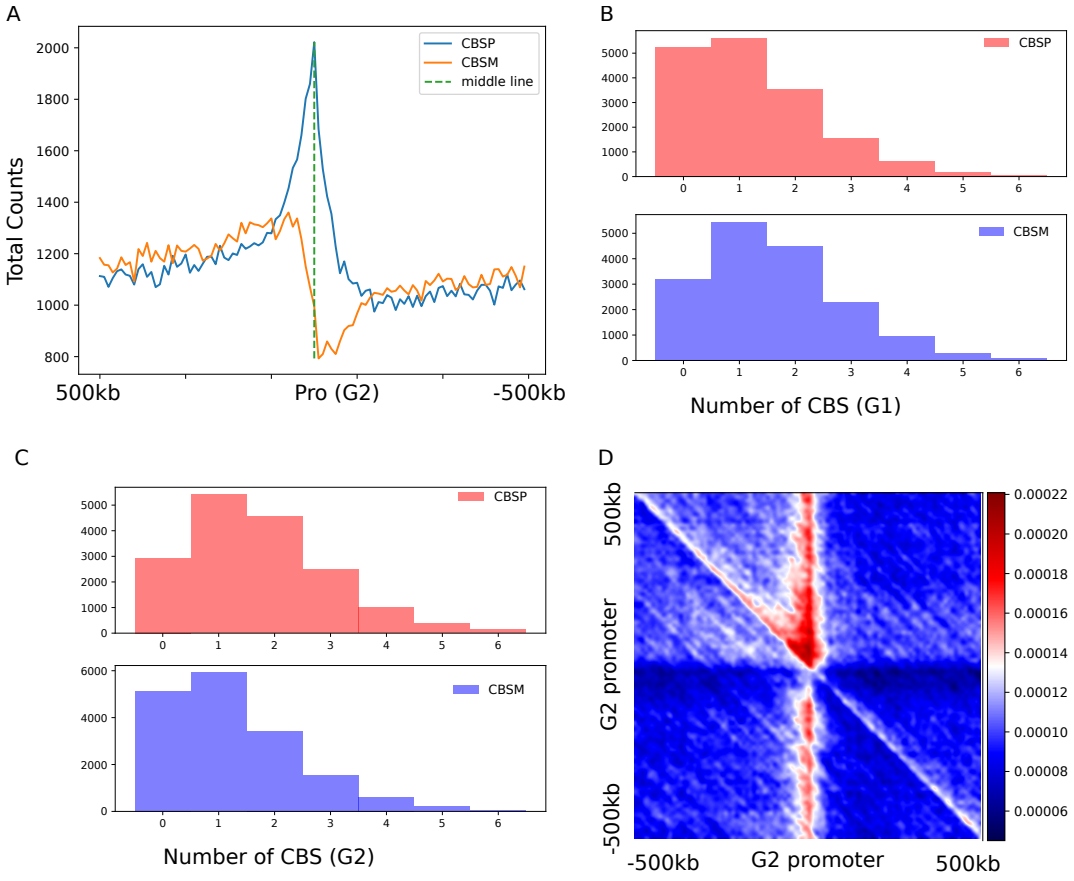


Figure S4: (A) The aggregated HiC pattern centered on G2 promoters. (B) The simulated G2AP.



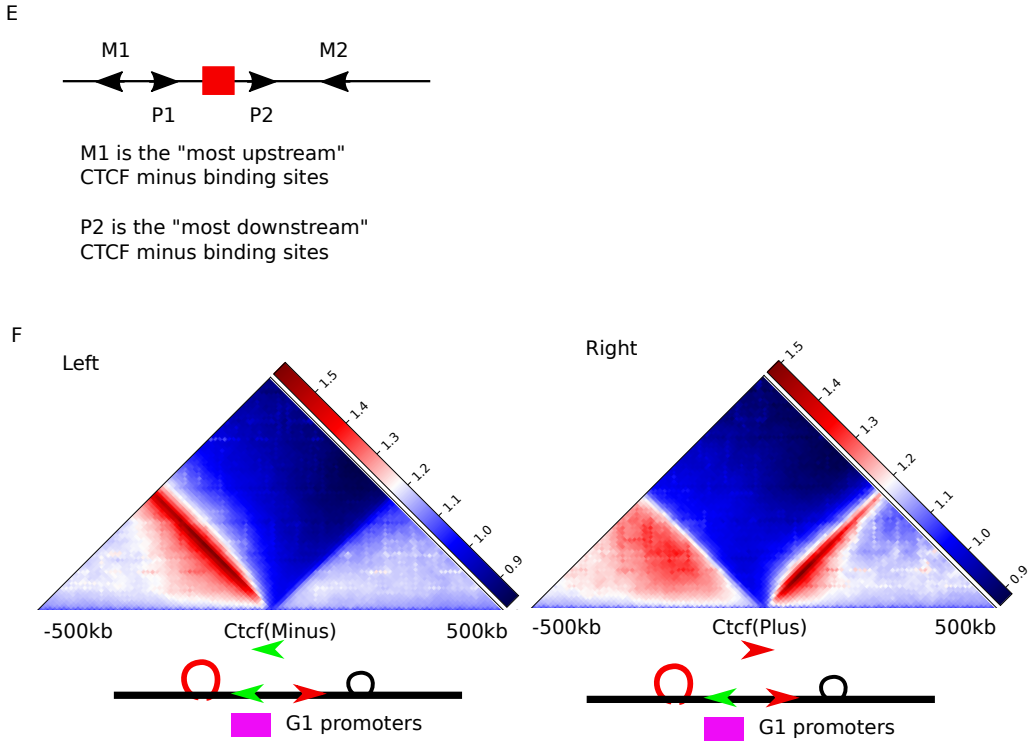


Figure S5: (A) The distribution of CBSP and CBSM relative to the G2 promoter. (B) The histogram shows the number of CBSM and CBSP in the 300kb regions flanking the G1 promoters. (C) Same as (B) but for the G2 promoters. (D) Same as the Fig5b but for G2 promtoers. Note that the high-intensity dots locate upstream of the G2 promoters. (E) The two methods of choosing CBSP and CBSM for aggregation. The first method will choose M1 for CBSM aggregation and P2 for CBSP aggregation as they form the largest divergent pair. (F) The aggregated results for the CBSM (left) and CBSP (right). The CBSM and CBSP are chosen according to the second methods (See Fig5c, which is based on the first method)

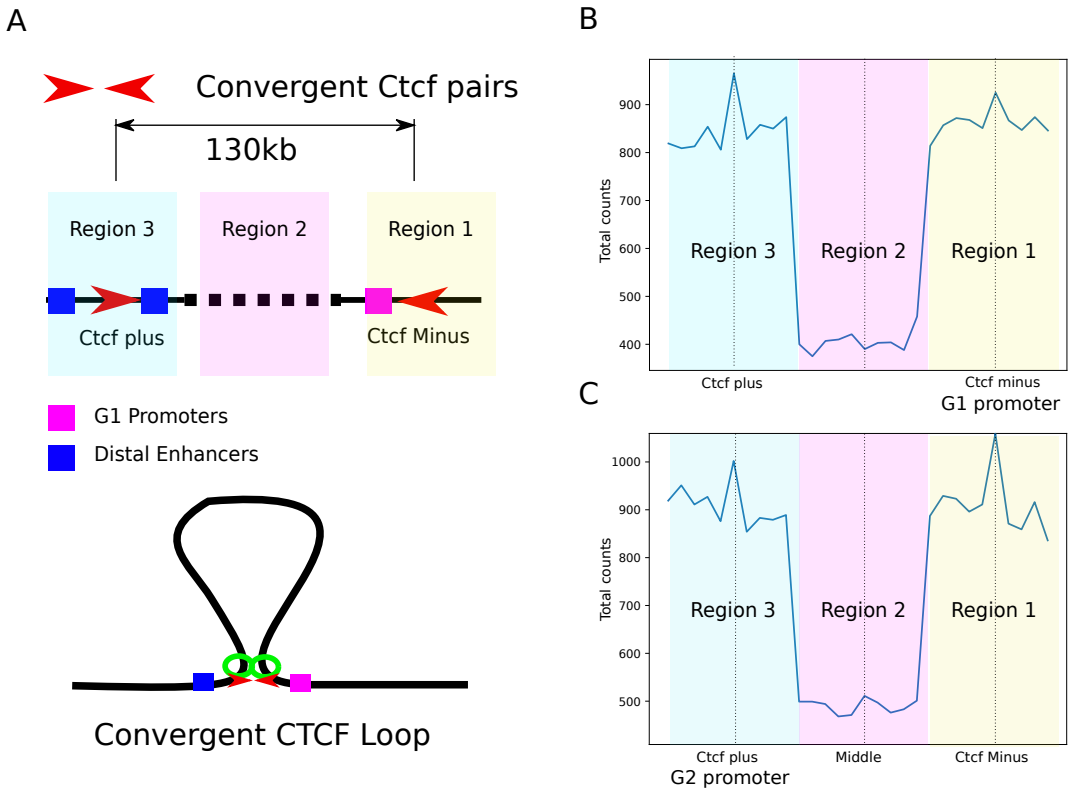


Figure S6: (A) The convergent CBS pairs associated with the G1 promoters (magnet box). The G1 promoters and the distal enhancers (blue box) are brought in proximity by DNA extrusion which forms the convergent CTCF loops. (B)The enhancers are more enriched within region 1 and 3 than that within region 2. Note that region 1 flanks the G1 promoter. (C) Same as (B) but for G2 promoters. Note that region 3 flanks the G2 promoter.