Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes

Yinxiu Zhan, Luca Mariani, Iros Barozzi, et al.

Genome Res. published online January 5, 2017
Access the most recent version at doi:10.1101/gr.212803.116

Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes

Yinxiu Zhan[1,2], Luca Mariani[3§], Iros Barozzi[4], Edda G. Schulz[3#], Nils Bluthgen[5,6], Michael Stadler[1,7], Guido Tiana[8], Luca Giorgetti[1]*

[1] Friedrich Miescher Institute for Biomedical Research, Basel, CH-4053, Switzerland

[2] University of Basel, CH-4003 Basel, Switzerland.

[3] Institut Curie, PSL Research University, CNRS UMR3215, INSERM U934, 26 Rue d'Ulm, 75248 Paris Cedex 05, France

[4] Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[5] Institute of Pathology, Charité - Universiätsmedizin Berlin, 10117 Berlin, Germany

[6] Interdisciplinary Research Institute for the Life Sciences, Humboldt University, 10115 Berlin, Germany

[7] Swiss Institute of Bioinformatics, CH-4058 Basel, Switzerland.

[8] Department of Physics and Center for Complexity and Biosystems, University of Milano and Istituto Nazionale di Fisica Nucleare, 20133, Milano, Italy

§ current address: Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

current address: Max-Planck-Institute for Molecular Genetics, Otto-Warburg-Laboratorium, Max Planck Research Group "Regulatory Networks in Stem Cells"

* correspondence: luca.giorgetti@fmi.ch

Running title: TADs are a functionally privileged folding scale

Keywords: Chromosome conformation, topologically associating domains, transcriptional regulation

Abstract

Understanding how regulatory sequences interact in the context of chromosomal architecture is a central challenge in biology. Chromosome conformation capture revealed that mammalian chromosomes possess a rich hierarchy of structural layers, from multi-megabase compartments to sub-megabase topologically associating domains (TADs) and sub-TAD contact domains. TADs appear to act as regulatory microenvironments by constraining and segregating regulatory interactions across discrete chromosomal regions. However, it is unclear whether other (or all) folding layers share similar properties, or rather TADs constitute a privileged folding scale with maximal impact on the organization of regulatory interactions. Here we present a novel algorithm named CaTCH that identifies hierarchical trees of chromosomal domains in Hi-C maps, stratified through their reciprocal physical insulation, which is a single and biologically relevant parameter. By applying CaTCH to published Hi-C datasets, we show that previously reported folding layers appear at different insulation levels. We demonstrate that although no structurally privileged folding level exists, TADs emerge as a functionally privileged scale defined by maximal boundary enrichment in CTCF and maximal cell-type conservation. By measuring transcriptional output in embryonic stem cells and neural precursor cells, we show that the likelihood that genes in a domain are co-regulated during differentiation is also maximized at the scale of TADs. Finally, we observe that regulatory sequences occur at genomic locations corresponding to optimized mutual interactions at the same scale. Our analysis suggests that the architectural functionality of TADs arises from the interplay between their ability to partition interactions and the specific genomic position of regulatory sequences.

Introduction

Characterizing the three-dimensional organization of chromosomes in mammalian cells is a central challenge, especially in the light of determining how regulatory sequences such as enhancers and promoters interact and ensure precise control of gene expression during development. Methods based on chromosome conformation capture (3C) and notably 4C, 5C and Hi-C, which measure physical interaction frequencies of genomic loci in the three-dimensional nuclear space, have revealed that mammalian chromosomes possess a rich hierarchy of structural layers (Gibcus and Dekker 2013). Each chromosome is partitioned in multi-megabase 'A' and 'B' compartments, reflecting the associations of alternating large regions of active and inactive chromatin (Lieberman-Aiden et al. 2009). Compartments are further subdivided into topologically associating domains (TADs), contiguous sub-megabase genomic regions within which the chromatin fiber preferentially associates (Dixon et al. 2012; Nora et al. 2012), which are further partitioned into smaller sub-structures and 'contact domains' (Berlivet et al. 2013; Phillips-Cremins et al. 2013; Rao et al. 2014). Finally, as a further level of complexity, TADs also interact with each other into "meta-TAD" trees that extend up to several Mb (Fraser et al. 2015). Given the cell population-averaged nature of 3C-based experiments, the observed nested hierarchies of interaction domains may arise as statistical patterns resulting from an average over millions of alternative conformations of the chromatin fiber (Fudenberg and Mirny 2012; Giorgetti et al. 2014; Junier et al. 2015).

Although more than one mechanism might give rise to TADs and sub-TAD structures, CTCF (CCCTC-binding factor) and the cohesin complex appear to be largely responsible for the establishment and maintenance of TADs and sub-TAD structures and boundaries. Indeed, CTCF and cohesin are enriched at TAD boundaries (Van Bortle et al. 2014; Dixon et al. 2012), but they also bind pervasively within TADs and are involved in the formation of sub-TAD structure (Rao et al. 2014; Sanborn et al. 2015; de Wit et al. 2015) although the molecular mechanisms that lead to structure formation are unclear (Merkenschlager and Nora 2016). In addition, open chromatin and active transcription positively correlate with the presence of

TADs and sub-TAD structure (Hou et al. 2012; Phillips-Cremins et al. 2013; Ulianov et al. 2015) and active histone modifications are enriched at TAD boundaries (Dixon et al. 2012), suggesting that interactions between active regulatory sequences may contribute to establish chromosomal architecture. However, transcription does not seem to be strictly needed for maintaining TAD boundaries (Nora et al. 2012).

Irrespective of the mechanisms underlying their formation, genetic evidence suggests that TADs contribute to establish correct interactions patterns between enhancers and promoters (Lupiáñez et al. 2015; Symmons et al. 2014; Franke et al. 2016). Consistent with this, transcriptional co-regulation of neighboring genes is favored within TADs during differentiation (Nora et al. 2012) and upon transcriptional responses to external stimuli (Dily et al. 2014). TADs are thought to act on the one hand by increasing the chances that regulatory elements meet each other in the three-dimensional space within a single domain; and on the other hand, by segregating physical interactions across boundaries, thus decreasing the probability that deleterious interactions occur. Hence, the degree at which each TAD is insulated with respect to its neighbors may be an important parameter in the establishment of the correct regulatory connections. It is however unclear whether the functional attributes that have been observed at the level of TADs (namely the ability to constrain enhancer-promoter interactions and promote transcriptional co-regulation) are specific to the folding layer of TADs themselves; and if so, why those properties emerge at this particular folding scale.

A comprehensive analysis that considers all previously identified topological levels simultaneously, and compares them to one another in terms of their functional and physical properties, is currently lacking. A small number of algorithms that identify hierarchies of topological domains are available (Filippova et al. 2014; Shavit et al. 2016; Weinreb and Raphael 2015). However, none of them provides a quantitative description of how the various layers of domains differ from one another. In addition, these algorithms define hierarchies of interaction domains depending on one or more parameters that do not have a clear biological or structural interpretation. To overcome these limitations, we developed a novel algorithm

called CaTCH (Caller of Topological Chromosomal Hierarchies) that identifies nested topologies of structural domains in Hi-C datasets based on a single parameter, the reciprocal physical insulation between domains, which is a simple and biologically relevant measure. Here we describe the CaTCH algorithm, and report the results of comparing the structural and functional properties of domains across the folding hierarchy of the mouse genome.

Results

CaTCH: An algorithm to detect and stratify nested hierarchies of topological domains

In order to comprehensively describe the multi-scale organization of chromosomal folding hierarchies, we developed an algorithm that segments Hi-C interaction maps into multiple alternative sets of domains, and stratifies them according to a single parameter. We adopted a thermodynamic interpretation of Hi-C datasets (Fudenberg and Mirny 2012) in which the Hi-C signal between a pair of loci is proportional to the *probability* of detecting them in proximity across the cell population. For any pair of adjacent chromosomal domains A and B, we then defined their reciprocal insulation (RI) as

$$RI(A,B)=[\ P_{in}(A)+P_{in}(B)-P_{out}(A,B)\ ]/\ [P_{in}(A)+P_{in}(B)]*100 \quad (1)$$

where P_{in} and P_{out} are the average Hi-C counts within a domain and across two adjacent domains respectively (**Figure 1a**; see Methods section). Small (large) values of *RI* thus correspond to domains that are poorly (strongly) insulated from their first neighbors. For example, 70% reciprocal insulation means that the average Hi-C counts across the boundaries of two adjacent domains are 70% smaller than the average counts within the two domains.

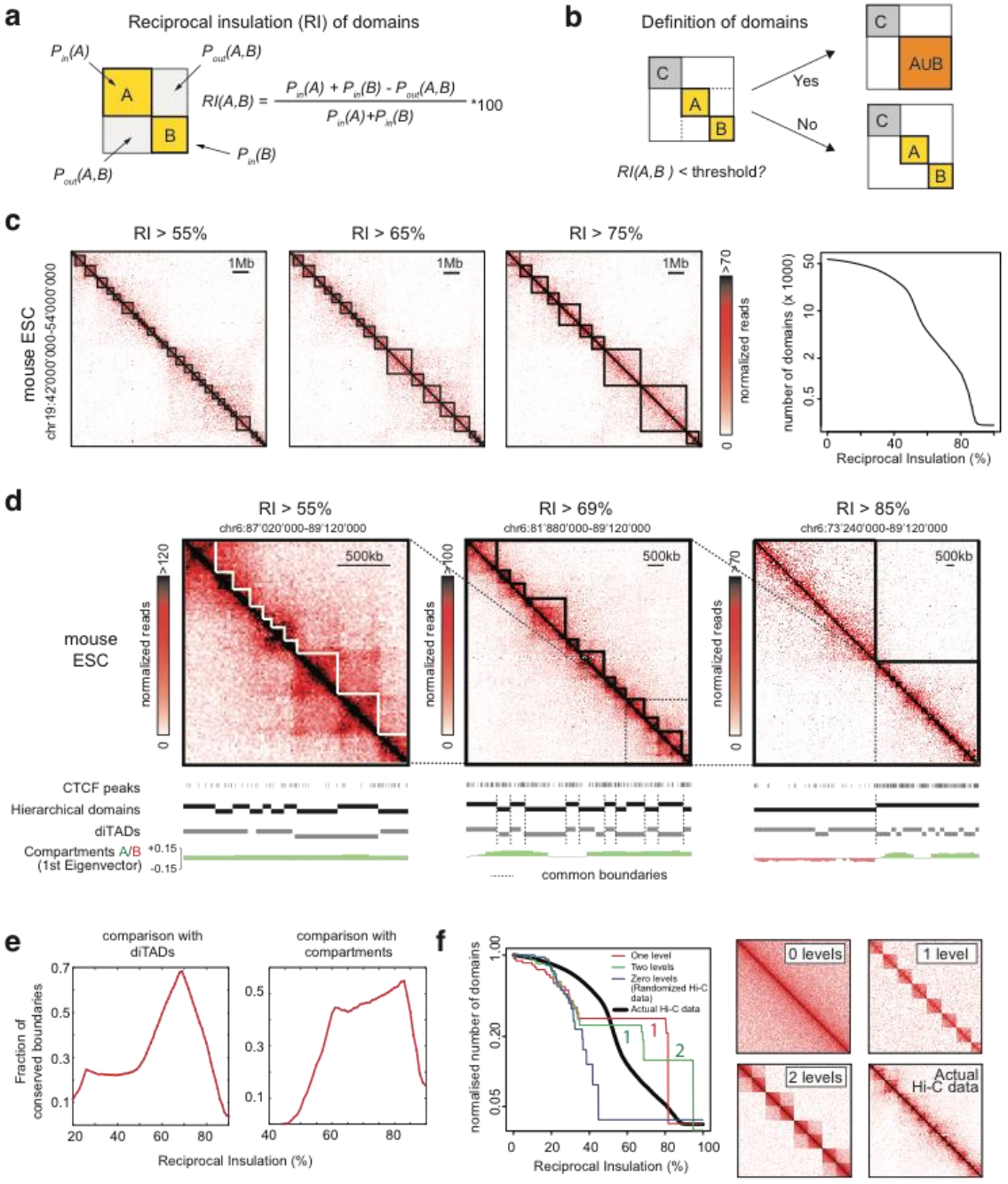

Figure 1. Schematic description of reciprocal insulation and the domain-calling algorithm

a. Schematic representation of reciprocal insulation (RI) between two fictitious domains A and B in Hi-C data.

b. The CaTCH algorithm merges two adjacent domains if their reciprocal insulation is smaller than a given threshold.

c. Left: Examples of sets of domains defined in mouse ESCs Hi-C data (20 kb binning) imposing different threshold on RI. Right: Number of domains detected in ESC as a function of RI.

d. Sub-TAD contact domains (left), directionality-index based TADs (middle) and A/B compartments (right) are identified at different RI values.

e. Fraction of boundaries of diTAD (left) and compartments (right) overlapping with boundaries of domains identified by CaTCH as a function of RI.
f. Left: Number of domains detected by CaTCH as a function of RI in the real genome (black line), or in computationally generated contact maps with zero (blue), one (red) or two preferential folding levels (green). The corresponding heat maps are shown in the right panels. Numbers of domains were normalised to the initial step (0% insulation) to allow comparison.

Given a certain degree of reciprocal insulation, the algorithm merges all consecutive domains whose reciprocal insulation is lower than the chosen threshold (**Figure 1b**; see Methods section), similarly to what commonly performed by agglomerative hierarchical clustering (Hastie et al. 2009). Thus, for any reciprocal insulation threshold, detected domains are *at least* insulated by the threshold value. By smoothly increasing the threshold on the insulation, the algorithm detects a set of domains that are increasingly more insulated, larger and containing previous domain layers. This results in a nested hierarchy of differentially insulated domains (**Figure 1c**). We dubbed this algorithm CaTCH, for Caller of Topological Chromosomal Hierarchies.

A key property of CaTCH is that it does not rely on the tuning of any free parameter to identify one particular folding scale. The only parameter in the algorithm is the reciprocal insulation threshold itself, which is systematically varied to define and stratify the entire hierarchy of domains, rather than tuned to identify a single domain set. Moreover, unlike parameters in existing approaches to identify multi-scale domain structures in Hi-C datasets (Filippova et al. 2014; Weinreb and Raphael 2015a; Shavit et al. 2016), the reciprocal insulation is a biologically relevant measure estimating how efficiently a domain is physically insulated from its immediate neighbors. CaTCH is provided as an R package at https://github.com/zhanyinx/CaTCH_R (source code can be found in Supplemental Methods).

Sub-TAD contact domains, TADs and compartments emerge at different levels in the folding hierarchy

We first applied CaTCH to published Hi-C datasets from female mouse ESCs (Giorgetti et al. 2016) binned at 20-kb resolution. As expected, when increasing the reciprocal insulation parameter, the algorithm detected increasingly larger and fewer topological domains (**Figure 1c**), with 5% changes in reciprocal insulation translating into ~30% changes in the number and size of domains (**Supplemental Figure S1a**). We found a similar trend when analyzing other cell types, notably neural precursor stem cells (NPCs) derived from the same ESCs line (Giorgetti et al. 2016), and the mouse B-cell lymphoma CH12 cell line (Rao et al. 2014) (**Supplemental Figure S1b**). In ESC, below 40% reciprocal insulation domains are too small (<100 kb on average) to be characterized with data at 20-kb resolution. At higher insulation values however we detected domains with a size (180 kb on average) in the range of sub-TAD structures and 'contact domains' identified in previous studies (Berlivet et al. 2013; Phillips-Cremins et al. 2013; Rao et al. 2014) (**Figure 1d** left and **Supplemental Figure S1c**). More than 60% of domain boundaries identified at 55% reciprocal insulation contain at least a CTCF peak identified in a published ChIP-seq dataset (Cheng et al. 2014), consistent with the notion that sub-TAD structures are highly correlated with CTCF binding (Phillips-Cremins et al. 2013). In addition, although the resolution of the Hi-C dataset is not high enough to distinguish most of the CTCF-associated 'loop' signals as in ref. (Rao et al. 2014), we noticed that ~45% of domains at this scale have at least one CTCF peak at both boundaries (**Supplemental Figure S1d**). Of the CTCF-delimited domains however, only 35% had convergent CTCF sites (compared to 98% of 'loop domains', defined as contact domains with strong interaction between boundaries in Rao *et al.* (Rao et al. 2014)). This is largely due to the fact that the domains identified in the latter study are a subset of domains detected by CaTCH at 55% reciprocal insulation (see below); however a direct comparison between the two domain sets is not possible, due to the lack of ESCs Hi-C datasets in the study by Rao and colleagues (Rao et al. 2014).

To determine the actual overlap between domains identified by CaTCH and contact domains described in ref. (Rao et al. 2014), we analyzed the 10kb-resolution Hi-C data that were

obtained in CH12 cells in the same study. Maximal overlap between the two domain sets occurred at 62% reciprocal insulation in CH12 (**Supplemental Figure S1e**), where 78% of boundaries of previously identified contact domains are also detected by CaTCH. However, CaTCH detects more domains than those identified in ref. (Rao et al. 2014) (**Supplemental Figure S1f**), which explains the lower proportion of domains delimited by convergent CTCF sites in our dataset. Thus, sub-TAD contact domains are detected by CaTCH as relatively lowly insulated regions.

We next sought to identify the scale in the folding hierarchy where domains detected by CaTCH most closely resemble TADs. Since directionality index analysis (Dixon et al. 2012) has been used to define TAD boundaries in a number of previous studies, here we adopted this benchmark definition of TADs and refer to these domains as 'diTADs' (directionality index TADs). It is important to point out that the set of diTADs identified in a Hi-C experiment depends on the value of two tunable parameters, one setting a limit to the maximal genomic distance over which Hi-C interactions are evaluated (**Supplemental Figure S1g**), and the other defining the minimum acceptable size of domains. We set these parameter to 2 Mb and 80 kb, respectively, as used previously (Dixon et al. 2012) to build a reference set of diTADs. This resulted in the identification of 2220 diTADs with a median size of 840kb, compatible with earlier analyses in mouse ES cells (Dixon et al. 2012). The best overlap between hierarchical domains detected by CaTCH and diTADs occurred at around 69% reciprocal insulation (**Figure 1d**, center), where ~70% of diTAD boundaries coincide with hierarchical domain boundaries (**Figure 1e,** left) and their size distributions are very similar (**Supplemental Figure S1h**). Domains detected by our algorithm at this scale are slightly (although not significantly) smaller than diTADs (median size 760 kb vs. 840 kb, **Supplemental Figure S1h-i**). Most (74%) CaTCH boundaries not corresponding to TADs indeed divide diTADs in smaller domains (**Supplemental Figure S1j**). Thus, diTADs are detected by CaTCH as domains that are more robustly insulated than contact domains.

At even higher reciprocal insulation, hierarchical domains detected by CaTCH correspond to regions of increasingly longer-range associations between TADs, in the range of meta-TADs described in Ref. (Fraser et al. 2015), themselves contained into even larger domains occurring at very high insulation (around 85%, **Figure 1d** right). These domains largely overlap with consecutive stretches of genomic sequence belonging to either the 'A' or 'B' compartments (Lieberman-Aiden et al. 2009), as detected by eigenvector analysis (Imakaev et al. 2012) on the same ESCs Hi-C dataset (**Figure 1d** right, **Figure 1e** right). Consistent with the notion that A/B compartments represent predominantly active/inactive chromatin, using publicly available ChIP-seq datasets in ESCs (**Supplemental_Table_S1)** we found that the difference in histone modification patterns within vs. across domain boundaries is maximized at this scale (**Supplemental Figure S1k**).

Thus, CaTCH identifies a continuous spectrum of nested self-interacting chromosome domains, stratified as contiguous genomic regions with differential reciprocal insulation levels. Previously described chromosomal structures such as sub-TAD contact domains, TADs, and groups of TADs emerge at different scales within the nested folding hierarchy, and are characterized by increasing reciprocal insulation levels.

A continuous nested hierarchy of topological associating structures

We then sought to determine whether one or more privileged reciprocal insulation levels exist among the folding hierarchy, and correspond to any of the previously reported folding layers. If such level(s) existed, some simple fundamental quantities, such as the number or size of domains detected by the CaTCH algorithm, would have a discontinuous behavior as a function of the reciprocal insulation parameter. To exemplify this concept, we computationally generated simplified control contact maps by artificially imposing the presence of zero, one or two scales of domains, separated by sharp transitions in contact probabilities between consecutive layers (see Methods section, **Figure 1f**). For these controls, CaTCH detected a

number of plateaus in the size (or number) of domains equal to the number of distinct hierarchical levels (**Figure 1f**, left), irrespectively of the genomic size of the domains (**Supplemental Figure S1l**). Compared to theses controls, the ESC genome does not exhibit any structurally privileged scale, at least for domains defined using reciprocal insulation as a measure (black line in **Figure 1f**), irrespective of whether the entire genome is considered or selectively regions that belong to either the A (active) or B (inactive) compartment (**Supplemental Figure S1b)**. A similar trend can be observed in NPCs and CH12 cells (**Supplemental Figure S1b**), suggesting that no obvious privileged structural scale exists in ESC and differentiated cell types. As a notable consequence, TADs do not appear as a natural intrinsic structural scale in the nested hierarchy of domains. This prompted us to investigate whether functional properties that have been previously attributed to TADs specifically emerge at the TAD scale, or are rather widespread among the folding hierarchy.

Enrichment in active histone marks is maximized at the scale of TADs

TAD boundaries have been shown to be enriched in histone modifications associated with active transcription (Dixon et al. 2012). We therefore analyzed publicly available ChIP-seq datasets in ESCs (**Supplemental_Table_S1)** and computed the enrichment for distinct histone marks at the boundaries of the domains across all the scales in the folding hierarchy. Marks associated with active transcription showed a steady increase in enrichment as a function of reciprocal insulation, and reached a plateau at the level of diTADs (~69%, **Supplemental Figure S2a**). Thus, although active histone marks show widespread enrichment across the folding hierarchy, they are maximally enriched at the scale of TADs and TAD aggregates (meta-TADs and compartments). Consistent with previous results (Dixon et al. 2012) the H3K9me3 repressive mark was found depleted at many levels in the folding hierarchy, and notably at the level of diTADs (**Supplemental Figure S2a**).

CTCF clustering at boundaries is maximized at the scale of TADs

Consistent with its putative role in establishing and/or maintaining chromosomal structure, CTCF is enriched at boundaries of contact domains (Berlivet et al. 2013; Phillips-Cremins et al. 2013; Rao et al. 2014), TADs (Dixon et al. 2012) and meta-TAD trees (Fraser et al. 2015). We therefore computed the enrichment in the number of CTCF ChIP-seq peaks (Cheng et al. 2014) at domain boundaries at all folding scales. As expected, CTCF binding is enriched at boundaries of every level across the folding hierarchy; however, CTCF enrichment it is maximized at the scale of TADs, and in particular at ~65% reciprocal insulation (**Figure 2a**) corresponding to domains that are slightly less insulated than diTADs detected using standard directionality index parameters. Identical results were found by using the input-normalized CTCF ChIP-seq signal per boundary, rather than the number of ChIP-seq peaks (**Supplemental Figure S2b**). We noticed that maximal CTCF enrichment is due to both a maximal number of boundaries containing at least one CTCF peak, and a maximal average number (~1.9) of CTCF peaks per boundary (**Supplemental Figure S2c**), which are mostly found within the 40kb upstream or downstream from the boundary (**Supplemental Figure S2d)**.

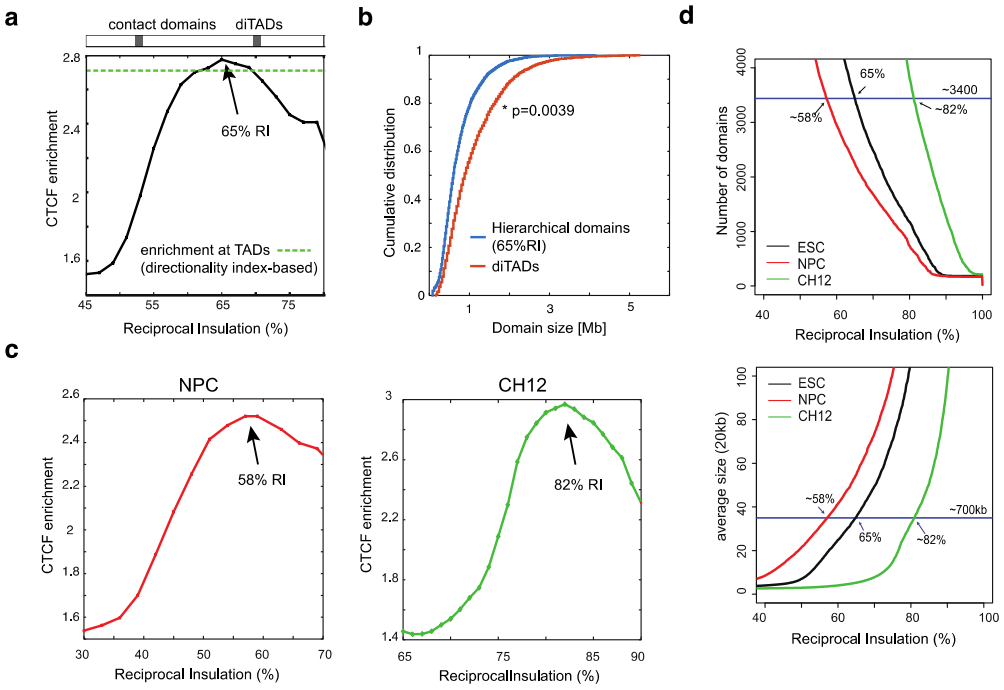

Figure 2. CTCF clustering at domains boundaries is maximal at the scale of TADs
a. CTCF enrichment at domains boundaries is widespread among the folding hierarchy in mouse ES cells. However, maximal enrichment occurs at 65% RI, where it is slightly higher compared to TADs identified by directionality index analysis (diTADs).
b. Domains at 65% are slightly smaller than diTADs identified on the same dataset.
c. CTCF enrichment at domains boundaries in NPCs and CH12 cells shows a similar trend as in ESCs, with maxima located at 58% and 82% RI in NPCs and Ch12, respectively.
d. The number and size of domains defined by maximal CTCF enrichment at boundaries are similar in ESCs, NPCs and CH12 cells.

Domains at 65% relative insulation are 20% smaller compared to 'standard' diTADs (600 kb vs. 840 kb median size, **Figure 2b**) and frequently originate from the splitting of one diTAD in two or more smaller domains (**Supplemental Figure S2e**). The majority (~70%) of these 'new' boundaries have at least one occupied CTCF site, which explains the slightly higher enrichment in CTCF compared to standard diTADs. However, by systematically varying the values of parameters in the directionality index algorithm, we identified alternative sets of diTADs where CTCF enrichment is higher than standard diTADs and comparable to (although slightly lower than) 65% RI domains (**Supplemental Figure S2f**). Importantly, these alternative directionality index domains correspond to domains detected by CaTCH in the 65%-70% relative insulation range (**Supplemental Figure S2f**, arrows). This confirms that the

TAD scale is characterized by maximal CTCF enrichment at boundaries compared to other folding levels. We will hereafter refer to domains identified by CaTCH at 65% minimal reciprocal insulation simply as TADs, since they constitute the set of domains with maximal CTCF enrichment.

Reciprocal orientation of CTCF binding sites has been shown to be highly predictive of strong long-range 'looping' interactions (Rao et al. 2014; Vietri Rudan et al. 2015; de Wit et al. 2015; Guo et al. 2015). We therefore assessed the orientation of the two most internal CTCF motifs on either side of each domain and found that at the scale of TADs the fraction of domains where CTCF motifs were convergent was maximal (**Supplemental Figure S2g**, left), with ~22% of domains possess convergent binding sites. Thus, both CTCF clustering and head-to-head orientation of the most internal CTCF motifs are maximized at the scale of TADs. Using available CTCF ChIP-Seq datasets (Phillips-Cremins et al. 2013; Cheng et al. 2014), we found that in both NPCs and CH12 cells, CTCF enrichment at boundaries showed a similar trend as in ESCs, with a peak around 58% and 82% reciprocal insulation in NPCs and CH12, respectively (**Figure 2c**). The fraction of domains with convergent CTCF motifs peaked at the same RI values (**Supplemental Figure S2g**). Importantly, despite the difference in absolute reciprocal insulation values, the number and size of domains at maximal CTCF enrichment were extremely similar across the three cell types (**Figure 2d**). In addition, conservation of boundaries across the three cell types was also found to be maximal at the same scale, with ~70% of boundaries conserved between any two cell types (**Supplemental Figure S2h**). Thus, the scale of TADs appears in the entire folding hierarchy not only as the domain scale that maximizes CTCF enrichment at boundaries, but also as the scale where domains are most conserved across cell types.

We next sought to determine any confounding effect on the determination of the optimal RI value due to experimental factors, such as different sequencing depth of Hi-C libraries or different versions of the Hi-C protocol. To study the effect of sequencing coverage, we performed CaTCH and CTCF enrichment analysis on a down-sampled ESC Hi-C dataset

obtained by reducing by half the total number of sequenced reads. CTCF enrichment at domain boundaries was maximized at very similar reciprocal insulation value than in the full dataset (67% vs. 65%, **Supplemental Figure S2i**), largely corresponding to the same set of domains (**Supplemental Figure S2k**). Thus, sequencing depth is not likely to have a strong impact on the reciprocal insulation values where TADs appear. Next, to understand the impact of using Hi-C datasets obtained using different experimental protocols, we performed a comparative analysis of two datasets obtained in mouse fetal liver cells (Nagano et al. 2015) using either the 'dilution' (Lieberman-Aiden et al. 2009; Belton et al. 2012) or the 'in situ' ligation protocols (Nagano et al. 2013; Rao et al. 2014). Using a published CTCF dataset (Cheng et al. 2014) we found that maximal CTCF enrichment occurred at different reciprocal insulation values (**Supplemental Figure S2j-k**), with the dilution protocol leading to smaller values compared to the in situ protocol (70% vs. 77%). This is consistent with the lower insulation values where TADs appear in NPC and ESC (where Hi-C was performed with the dilution protocol (Giorgetti et al. 2016)) compared to CH12 cells (in situ protocol), and is compatible with the previous observation that the in situ protocol leads to sharper TAD boundaries (Nagano et al. 2015). These results point at Hi-C protocol variants as a main determinant of reciprocal insulation, and suggest that the scale of TADs occurs in the 58-70% range (64% ± 6%) in dilution Hi-C datasets, and in the 77-82% range (80% ± 3%) in the in situ experiments that were analyzed.

Transcriptional co-regulation during differentiation is maximal at the scale of TADs

Motivated by the finding that CTCF and active histone marks enrichment at boundaries is maximal at the scale of TADs, we set out to determine whether domains at this scale encompass maximally co-regulated genes, which is a further important functional attribute proposed for TADs (Dily et al. 2014; Nora et al. 2012). For this, we performed strand-specific RNA-seq on total RNA from the ESC and NPC lines in which the Hi-C had been performed

(Giorgetti et al. 2016). Strand specificity allowed us to unambiguously assign up- or downregulated transcripts in the case of two overlapping transcriptional units. For all levels in the folding hierarchy, we then set out to determine how many domains are transcriptionally co-regulated during the differentiation from ESCs to NPCs.

We defined a domain to be down- (up-) co-regulated at the empirical p<=0.05 level if the number of down- (up-) co-regulated genes in the domain is larger than in 95% of cyclically permutated genomes (see Methods section). For each insulation level and the corresponding domain set, we then calculated a Z-score as the difference between the number of co-regulated domains that were observed in the real genome and the mean number of co-regulated domains detected in 2,000 randomizations of the genome (**Figure 3a**; see Supplemntal Methods), weighed by its standard deviation. Interestingly, at all insulation levels, the subset of domains that we detected to be up- or downregulated at the level of p<=0.05 level show maximal transcriptional changes during development (**Figure 3b** and **Supplemental Figure S3a**). Thus, domains with high level of transcriptional co-regulation largely overlap with those where the most dramatic changes in gene expression occur during differentiation.

At the level of TADs in ESC (65% insulation), we detected 114 co-regulated domains, accounting for ~4% of the total number of TADs and ~10% of those exhibiting expression changes during differentiation (>=2 up- or downregulated genes). This represents a >2.5-fold enrichment relative to the values expected in randomized genomes. Moreover, the number of co-regulated TADs (65% reciprocal insulation) is very similar to that observed at the level of TADs in the context of the acute transcriptional response to progesterone in a human breast cancer cell line (Dily et al. 2014).

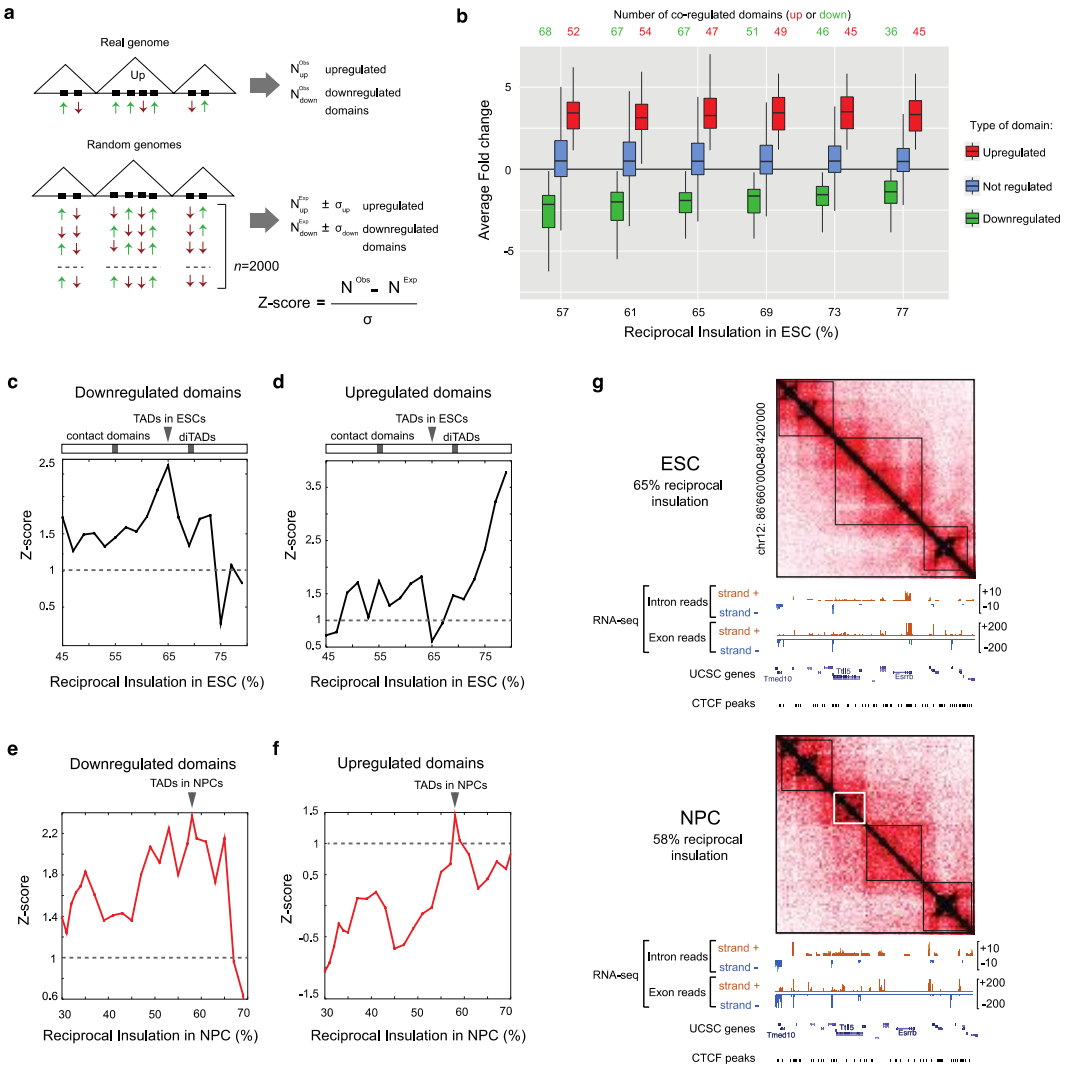

Figure 3. Transcriptional co-regulation defines a functional privileged scale

a. Schematic representation of the definition of statistical enrichment in the number of co-regulated domains. A domain is down- (up-) co-regulated if its number of down- (up-) co-regulated genes is larger than in 95% of cyclic permutated genomes (empirical p<=0.05). A Z-score is calculated as the difference between the number of co-regulated domains detected in the real genome (N^{obs}) and the mean number of co-regulated domains detected in 2000 randomized genomes (N^{exp}), weighed by its standard deviation σ^{exp}.

b. Distribution of average fold changes in expression level for domains at different RI values. For each RI value, the number of domains that are either up- or downregulated during differentiation (at the p<=0.05 level) is also shown in the upper part of the graph. Box: 25%-75% range (black line: median).

c. The statistical enrichment in the number of down-regulated domains is plotted as a function of the RI threshold. Transcriptional co-regulation is significant at any level below ~70% RI, but maximal at 65%.

d. Same as panel b for up-regulated domains.

e. Same analysis as in panel b, when using domains based on Hi-C data in NPCs.

f. Same as panel d for up-regulated domains.

g. Example of domains that were created *de novo* during differentiation and detected only in the set of NPC TADs (58% RI).

For genes that are downregulated during differentiation, the Z-score is maximum at the scale of TADs (**Figure 3c**). To check the robustness of the analysis against stochastic fluctuation of expression changes, we studied the behavior of z-scores upon randomly reshuffling (n=1000) 10% of genes. For 66% of these partially reshuffled genomes, the maximum Z-score was found to be located within a 4% interval around 65% reciprocal insulation (63%-66%) (**Supplemental Figure S3b**), supporting the robustness of the result. This analysis suggests that TADs in ESCs constitute a functionally privileged scale, maximizing the co-regulation of genes that are downregulated during the differentiation into neural precursor stem cells.

The behavior of upregulated genes was remarkably different, with low (if any) enrichment in transcriptional co-regulation within domains below 75% reciprocal insulation (**Figure 3d**) and maximal enrichment at the scale of A/B compartments (>80%). We reasoned that this could be due to the fact that not all TADs identified in ESCs are predictive for transcriptional co-regulation of genes that become activated during differentiation. We thus performed the same analysis on domains identified in NPCs, and found that co-regulation of both down- and upregulated genes is maximized within domains defined in NPCs around 58% reciprocal insulation (**Figure 3e-f**). This is the set of TADs defined in NPCs by maximal CTCF clustering at their boundaries (see **Figure 2c**). We verified that these results are not affected by the presence of an inactive X chromosome in NPCs, as maximal co-regulation was observed at the scale of TADs even when expression changes of X-linked genes (excluding genes that escape X inactivation in this clone (Giorgetti et al. 2016)) were corrected to account for their mono-allelic expression in NPCs (**Supplemental Figure S3c** and Supplemental Methods).

Thus, TADs defined in the initial developmental stage (ESCs) are the scale where transcriptional co-regulation of downregulated genes is maximal, whereas the set of domains that better favors the co-regulation of upregulated genes corresponds to TADs defined in the

final state (NPCs). This can be largely explained by the fact that although most TAD boundaries (~70%) are conserved between ESCs and NPCs, a significant fraction (~30%) is not. In particular, although most upregulated TADs are conserved and detected in both ESCs and NPCs, we detected 20% more upregulated TADs in NPCs than in ESCs, corresponding to domains that were defined *de novo* during differentiation in parallel with a significant increase in their genomic activity (**Figure 3g** and **Supplemental Figure S3d**). This might suggest that TADs in NPCs are more predictive of transcriptional co-regulation of up-regulated genes because domains that are transcriptionally active in the final state of differentiation can only be precisely detected when they are active (i.e. in the final state), but do not appear as defined in the set of TADs in the initial state. These domains represent extreme cases that illustrate that increased genomic activity is associated with increased structural complexity, and in this case with *de novo* formation of local structures. This is reminiscent of what observed on the inactive X chromosome (Giorgetti et al. 2016), where the presence of TAD-like structures is only observed in the context of gene activation.

Enhancer-promoter communication is optimized at the scale of TADs

The finding that TADs emerge as the folding scale that maximizes transcriptional co-regulation, but is not an intrinsically defined structural level (cf. Figure 1) prompted us to ask whether TADs specifically favors enhancer-promoter communication in ESCs. We analyzed available ChIP-seq datasets (**Supplemental_Table_S1**) to identify enhancers based on H3K27ac, H3K4me1 and H3K4me3 patterns (see Supplemental Methods); active promoters were identified using the strand-specific total RNA-seq datasets generated in this study.

To check whether the presence of domains at each level in the hierarchy corresponds to gain (or loss) in interactions, we considered pairs of Hi-C bins containing enhancers and promoters. We then calculated the ratio between their Hi-C counts vs. the genomic average for loci separated by the same genomic distance (**Figure 4a**). We observed substantial enrichment

in interactions between enhancers and promoters within the same (active) domain up to ~65% reciprocal insulation (**Figure 4b**, red curve). Thus, TADs appear to be within the uppermost scales in the folding hierarchy where enhancer-promoter contacts are maximally enriched within domains. On the other hand, enhancer-promoter interactions are also enriched *across* boundaries with the two neighboring domains, until slightly below the scale of TADs (**Figure 4c**, red curve). This reflects the fact that domains up to TADs are detected as increasingly bigger units, which are defined by the union of smaller sub-domains found at lower insulation values where enhancer-promoter interactions are strongly enriched (cf. **Figure 4b**). At higher reciprocal insulation, interactions across boundaries are depleted. CTCF-bond loci showed a similar pattern with even higher levels of enrichment within domains, and lower enrichment across domain boundaries (**Figure 4b-c**, black curves). This result is obtained irrespectively of the reciprocal orientation of pair of CTCF motifs, although enrichments are globally higher for convergent CTCF sites (**Supplemental Figure S4a**). Importantly, when we considered all pairs of loci within the same active domains where enhancers and promoters were identified, or random interactions drawn from the same distribution of distances as enhancer-promoter pairs, we observed much lower increase in interactions inside domains (**Figure 4b**, green and blue curves). Moreover, interactions across domains were also depleted at low insulation levels (**Figure 4c**). We obtained very similar results in NPCs and the CH12 cell line (**Supplemental Figure S4b-c**).

Thus, TADs occur in the folding range where enhancer-promoter communication might be ‘optimal’, i.e. enhancer-promoter contacts are maximally enriched within domains, but begin to be depleted across domain boundaries.

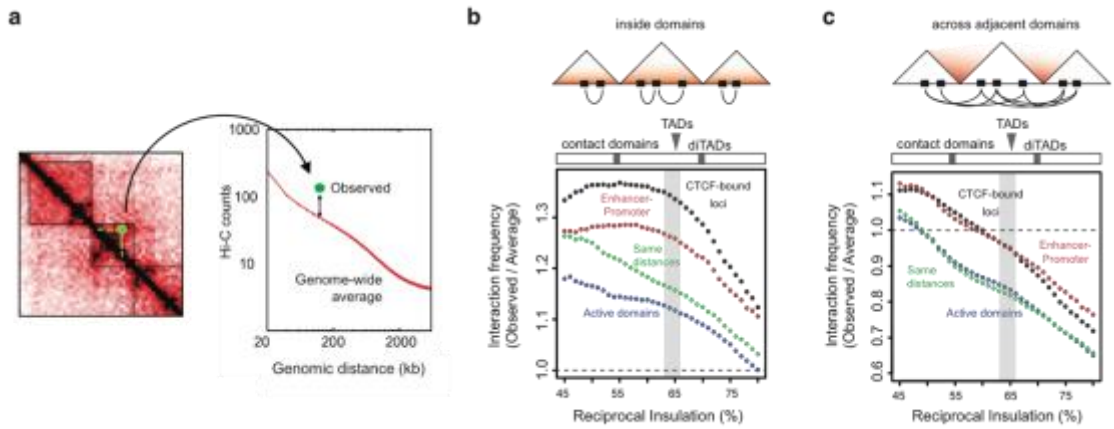

Figure 4. TADs define a scale where promoter-enhancer communication is optimal in ESCs

a. Schematics of contact enrichment analysis. For each pair of loci, we calculated the ratio between observed Hi-C counts and the genome-wide average counts for loci located at the same genomic distance.

b. Enrichment in interactions between pairs of loci belonging to the same domain, as a function of reciprocal insulation. Colors refer to random loci within active TADs (blue), enhancer-promoter pairs (red), random loci with the same distance distribution as enhancer-promoter pairs (green) and CTCF-containing loci (black). Median enrichment over all pairs of considered loci are plotted. Grey shaded area indicates the 63%-66% confidence interval where maximal co-regulation of genes occurs in partially reshuffled genomes (cf. Supplemental Figure S3b).

c. Enrichment (or depletion) in interactions between pair of loci defined as in panel a, but located across consecutive domains. Grey shaded area as in panel b.

The local complexity in chromosomal folding correlates with transcriptional activity and CTCF binding

We next used the CaTCH algorithm quantify local chromosome folding complexity within each TAD, and correlate it to the level of local transcriptional activity. To this aim, we first computed the number of hierarchical sub-lebvels that can be identified within a domain (see Methods) as a measure for local folding complexity. We then used the RNA-seq profiles to assign transcripts to domains based on the genomic position of their promoters. We did not limit our analysis to the exonic signal (corresponding to mature mRNA), but we also considered the intronic reads, the latter being a more reliable measure of transcriptional activity (see Methods section). We found that at the level of single TADs, a quantitative correlation exists between the number of sub-levels, and both total (exonic) and unspliced mRNA reads per domain

(**Figure 5a-b** and **Supplemental Figure S5a**). The number of sub-levels also correlates with the mean transcriptional level per gene (**Supplemental Figure S5b**), and with the number of transcribed promoters in the domain (**Supplemental Figure S5c**). We also observed that smaller TADs tend to be denser in actively transcribed genes (**Figure 5c**), and are globally more active than larger domains (**Supplemental Figure S5d**). In addition, the number of sub-levels correlates with the density of CTCF ChIP-seq peaks *within* the domain (**Figure 5d**).

These observations would predict that during differentiation from ESCs to NPCs, local changes in transcriptional activities should correspond at least in part to changes in local folding complexity. To verify this hypothesis, we considered the set of TADs defined in ESCs and studied the changes in the number of sub-levels in the same regions in NPCs. We found indeed that domains where transcriptional activity increases during differentiation increase their internal structural complexity and vice versa, as exemplified by the average change in the number of sublevels in the most dynamic TADs (**Figure 5e** and **Supplemental Figure S5e**).

Finally, given that the local transcriptional activity and CTCF occupancy modulate folding complexity within single domains, we reasoned that sharp transitions in these quantities across domain boundaries could also contribute to domain segregation. By definition, each domain level in the folding hierarchy (including TADs) is defined by the *minimal* reciprocal insulation of its constituent domains. Thereby, each TAD in ESC is insulated from its neighbors by at least 65%. Interestingly, we found that at the level of single TADs, reciprocal insulation correlates with the difference in transcriptional activity and CTCF occupancy *within* vs. *across* its borders (**Supplemental Figure S5f-g**). Similar results were found when considering all other levels in the hierarchy, either at lower or higher levels of insulation compared to TADs (**Supplemental Figure S5f**). Thus, sharper transitions in the genomic density of CTCF binding sites and transcribed genes correspond to stronger boundaries between adjacent domains.

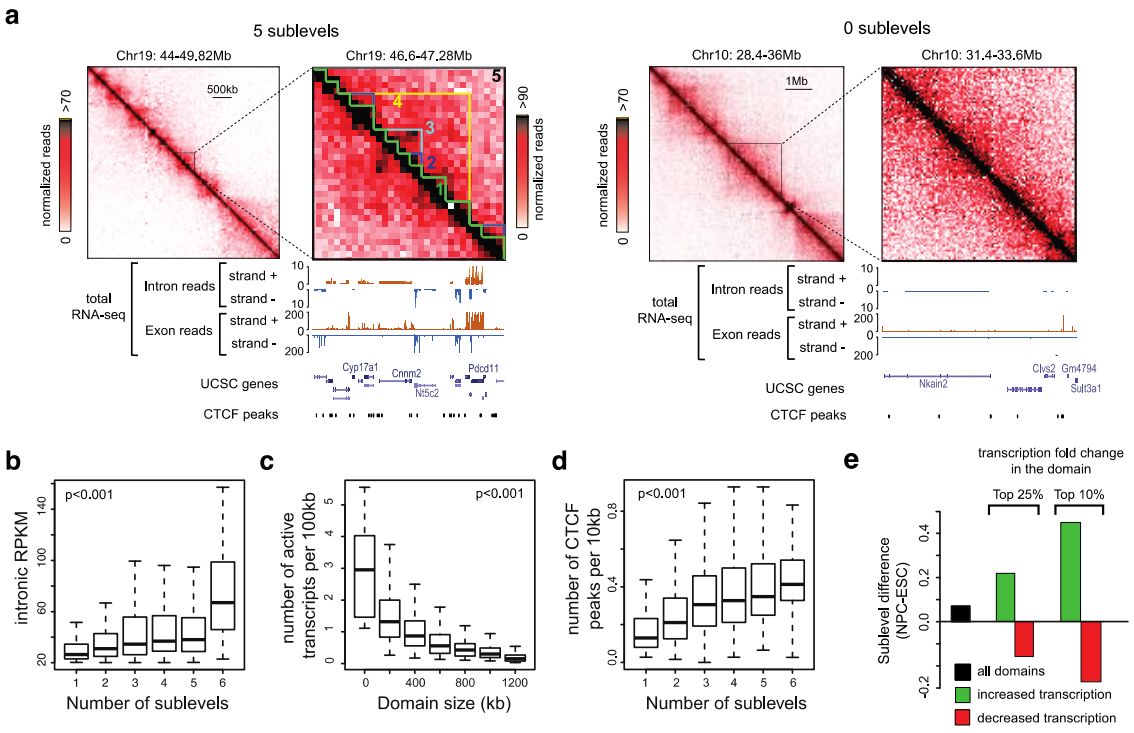

Figure 5. Local (changes in) folding complexity correlate with transcriptional activity in ESCs and during differentiation

a. Examples of regions with different level of local folding complexity and correlated transcriptional activities.

b. The number of sub-levels in a domain correlates with the transcriptional activity within the domain (shown for domains at 65% RI in ESC). p-value: t-test associated to Spearman correlation.

c. Smaller domains tend to be denser in actively transcribed genes and therefore globally more active than larger domains (shown for domains at 65% RI).

d. The number of sub-levels in a domain correlates with the density of CTCF bound sites within the domain (shown for domains at 65% RI).

e. Local changes in transcriptional activities during differentiation from ESCs to NPCs correspond to changes in local hierarchical complexity (see Methods). Differences in the number of hierarchical sublevels are shown for the 25% and 10% most up- or down-regulated domains identified at 65% RI in ESC.

Discussion

Determining how enhancers exert their regulatory functions on distal promoters critically depends upon our level of understanding of the three-dimensional organization of chromatin. Several studies have provided evidence on the fundamental role of compartmentalization into TADs to instruct enhancer-promoter communication (Franke et al. 2016; Lupiáñez et al. 2015;

Symmons et al. 2014; Nora et al. 2012), but they remain elusive on what makes TADs 'special' as compared to other chromosomal folding layers, such as sub-TADs and notably contact domains or meta-TADs. In this study we present a new domain-calling algorithm that is able to segment Hi-C interaction maps into nested sets of topologically associating domains, based on their reciprocal physical insulation. Our approach to partition the genome into nested sets of domains has two main advantages over existing hierarchical TAD callers (Filippova et al. 2014; Shavit et al. 2016; Weinreb and Raphael 2015): 1) The CaTCH algorithm does not rely on any free parameters, except reciprocal insulation itself that is used to stratify the domains; 2) Different from other methods that identify hierarchies of domains, where parameters have an unclear structural or biological interpretation, reciprocal insulation estimates how well a domain is segregated from its neighbors. CaTCH is fast and requires small computing power: identifying a whole hierarchy of domains on a single 100-Mb chromosome takes less than 4 minutes on a single CPU, starting from mouse Hi-C data at 20-kb resolution. We note that reciprocal insulation is conceptually similar to the 'local contrast' measure introduced in Ref. (Van Bortle et al. 2014); here, however, we used the parameter to define a full hierarchical tree of domains, rather than employing it to characterize the strength of boundaries of a given set of domains.

By applying CaTCH to published Hi-C datasets, we were able to show that previously reported topological structures are detected by the algorithm as differentially insulated levels within a continuous hierarchy of nested folding layers (**Figure 1**). This gave us the possibility to compare all levels simultaneously in terms of their structural and functional properties. Based on purely structural characteristics of the domains detected over the entire mouse genome, we found that none of these sets constitutes an intrinsically privileged scale. However, we observed that the scale of TADs emerges as a privileged functional one, where fundamental properties previously associated to TADs and notably related to their role in long-range transcriptional regulation are maximized.

CTCF clustering at domain boundaries has been repeatedly reported as one of the hallmarks of topological domains across species (Van Bortle et al. 2014; Dixon et al. 2012; Sexton et al. 2012; Vietri Rudan et al. 2015). In agreement with that, we show that maximal CTCF clustering at boundaries is highly predictive of the set of domains with the most conserved boundaries across cell types (**Figure 2**). In fact, finding hierarchical levels with approximately 3,400 domains seems to provide a sufficient operational criterion to identify the TAD scale when using CaTCH (**Figure 2d**) even in the absence of matched CTCF ChIP-seq datasets.

The resolution of our dataset (20 kb) does not enable the detecting of looping interactions between single CTCF sites that can be found in very high-resolution Hi-C (Rao et al. 2014) or ChIA-PET experiments (Tang et al. 2015), and it is therefore not possible to assess the precise reciprocal orientation of CTCF site clusters that occur within domain boundaries. However, between 15% and 22% of the most internal CTCF site pairs at the boundaries of TADs are convergent, which represents a maximum across the entire folding hierarchy (**Supplemental Figure S2f**).

Although boundary-associated CTCF might play an important role in defining domains and in particular TADs, CTCF also pervasively binds within domains. Within a given hierarchical level and TADs in particular, domains that are more reciprocally insulated tend to have a higher imbalance in the number of CTCF-bound sites within vs. across their boundaries. Notably, regions that are highly bound by CTCF and are flanked by low-occupancy domains are highly insulated from the flanking regions (see for example **Supplemental Figure S5g**, right). In addition, the density of CTCF-bound sites within a domain correlates with the hierarchical complexity of topological domains at all scales, including TADs (**Figure 5**). Together with the fact that the hierarchical complexity also correlates with the overall transcriptional activity of a domain, this is in line with earlier findings that sub-TAD structures are strongly associated with CTCF bound sites and active regulatory sequences (Phillips-Cremins et al. 2013). However, our results also suggest that interactions mediated by CTCF (and possibly additional factors associated with active regulatory sequences) *within* transcriptionally active domains play an

important role in modulating the strength of boundaries between adjacent domains. Strong asymmetry in CTCF occupancy and transcriptional activity across boundaries can arise as a consequence of marked transitions in gene density and/or number of regulatory sequences. In addition, asymmetry can occur corresponding to cell-type specific transitions in genomic activity between adjacent TADs (cf. **Supplemental Figure S5g**, right panel). This in turn might be driven by transitions in the enrichment for cell-type specific regulatory sequences (such as binding sites for lineage-determining transcription factors) across the boundary between the two TADs.

TADs appear in the uppermost layers in the folding hierarchy where interactions *within* active domains specifically, and between enhancers and promoters in particular, are strongly enriched compared to the genome-wide average interactions (**Figure 4**). On the other hand, interactions *across* the boundaries of active TADs start to be depleted as compared to genome-wide averages. TADs thus appear to belong to the domain scale where a trade-off is established between maximizing interactions within the interior of domains, and not enriching interactions across domain boundaries. In this light, it is remarkable that TADs emerges as the set of domains maximizing the co-regulated during differentiation is maximal (**Figure 4**). Although the precise mechanisms that govern enhancer action on promoters is still unknown, it is tempting to speculate that rather than absolute interaction frequency, the balance between interactions within and across domains determines the genomic range of action of enhancers, and this could contribute at least in part to establishing higher transcriptional co-regulation at the level of TADs.

Methods

Hi-C datasets

ESCs and NPCs Hi-C datasets were obtained in Ref. (Giorgetti et al. 2016). Reads from 129Sv and Cast/EiJ alleles were combined to increase read depth, and data were binned at 20 kb resolution. CH12 data are from Rao *et al.* (Rao et al. 2014), binned at 10 kb. Mouse fetal liver Hi-C data are from Nagano *et al.* (Nagano et al. 2015), binned at 25 kb. ESC, NPC and liver Hi-C were normalized with iterative correction (Imakaev et al. 2012). CH12 were normalized with VC-SQRT method (Rao et al. 2014).

The CaTCH algorithm

The algorithm takes a normalized Hi-C matrix as an input, binned at an arbitrary resolution r. The genome is first partitioned into domain seeds of size 2*r, which are progressively merged into larger domains. Reciprocal insulation (RI) is defined as in Eq. (1) in the main text. Given a threshold on RI, two consecutive domains are merged into one if their RI is smaller than the threshold. Increasing the RI threshold from 0% to 100% in steps of 0.1% results in increasingly larger domains. To lose memory of the initial partitioning of the genome into domain seeds, small shifts (2 genomic bins) in domain boundaries are allowed at each step. Finally, to avoid that the discrete increase in RI threshold (0.1% steps) results in a final domain tree that depends on the order of mergings and is therefore not unique, we impose a rule on merging order: a domain can be merged with either the one that precedes or the one that follows it along the genome, the pair with lowest RI is merged first (see Supplemental Methods).

Computationally generated contact maps with preferential folding levels

Control contact maps with one or two folding levels were created by generating a power law decreasing contact map for each level, to which Gaussian random noise is added (see Supplemental Methods). The contact map with zero folding layers was generated by replacing the actual Hi-C counts in the contact map for chr19 in ESCs with the average genome-wide counts for loci with the same genomic distance, and adding Gaussian noise.

Cell culture

Culture of the female mouse ES cell line F121.6 (129Sv-Cast/EiJ) and NPC clone analyzed in Ref. (Giorgetti et al. 2016) was performed as previously described (Gendrel et al. 2014; Giorgetti et al. 2016). All cells used in this study were characterized for absence of mycoplasma contamination.

RNA-seq data and analysis, other analyses of genomic data

Strand-specific total RNA-seq libraries from two biological replicates of ESCs and NPCs were prepared with the ScriptSeq v2 kit (Illumina) and sequenced on an Illumina HiSeq 2000 for a total of ~30 million uniquely aligned reads per sample. Samples were aligned to mouse mm9 For details on the RNA-seq and ChIP-seq analysis, CTCF motif assignment and enhancer calling, please refer to Supplemental Methods.

Definition of hierarchical sub-levels

A sub-region within a domain at any scale in the folding hierarchy was defined as a sub-level, if it is detected as a domain over more than 5% of the preceding reciprocal insulation thresholds. P-values in transcription and CTCF content vs. number of sublevels (**Figure 5**) were obtained using the function cor.test in R (Spearman method) and represent the results of t-tests on the Spearman's correlation coefficient.

Analysis of structural re-organization during differentiation

We focused on TADs defined in ESCs and defined the number of sub-levels detected in NPCs in the corresponding regions, using NPC domains below 58% since those are the domains that best match domains at 65% in ESCs (see **Supplemental Figure S2g**). We estimated the local amount of structural re-organisation as the change in the number of sub-levels between ESCs and NPCs.

Analysis of enhancer-promoter interactions

Genomic 20-kb (ESCs and NPCs) and 10-kb (CH12) bins were assigned to 'enhancer', 'promoter' or 'CTCF' categories if they contain at least one of these elements (see

Supplemental Methods). If a bin shows multiple classifications, it was assigned to all categories. In the analysis for **Figure 4**, in order to avoid including under-sampled interactions due to limited Hi-C coverage at large genomic distances, we only considered pairs of loci separated by less than 2Mb in ESCs and NPCs, and 1 Mb in CH12 cells. Cutoffs were chosen to exclude genomic distances where average Hi-C counts are dominated by experimental noise (**Supplemental Figure S4d**).

Data access

The sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE84724.

CaTCH is provided as an R package at https://github.com/zhanyinx/CaTCH_R. Source code can be found in Supplemental Methods.

Acknowledgements

Research in the Giorgetti lab was supported by the Novartis Research Foundation. Initial analyses for this study were conceived and established in Edith Heard's laboratory, Institut Curie and PSL (Paris) where L.M.'s salary was paid for by an EMBO ASTF 563-2012 Fellowship to L.M. and SyBoSS 62012 grant to Edith Heard. We would like to thank Federico Comoglio for assistance on code development and for critically reading the manuscript, Stéphane Thiry and Tim Roloff for assistance on RNA-seq library preparation and sequencing, Edith Heard for critically reading the manuscript, NIBR computing resources and Stefan Grzybek for help with cluster and server supports and Mikael Attia for cell culture. We acknowledge The ENCODE Project Consortium and in particular the Ren and Hardison laboratories for ChIP-Seq datasets in ESC and CH12, and the Myers laboratory for ChIP-seq datasets in fetal liver cells.

Author contributions

Y.Z. wrote the code and performed all analyses with assistance from G.T, L.M., M.S.; L.M., E.S. and N.B. set up the preliminary analyses and discussed the results; I.B. performed enhancer calling and discussed the results; G.T. assisted with data analysis. L.G. prepared RNA samples. L.G. designed the study. L.G. and Y.Z. wrote the paper.

Disclosure declaration

All authors declare no conflict of interest.

References

Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. 2012. Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods* **58**: 268–276.

Berlivet S, Paquette D, Dumouchel A, Langlais D, Dostie J, Kmita M. 2013. Clustering of Tissue-Specific Sub-TADs Accompanies the Regulation of HoxA Genes in Developing Limbs. *PLOS Genet* **9**: e1004018.

Chen J, Hero AO, Rajapakse I. 2016. Spectral identification of topological domains. *Bioinformatics* **32**: 2151–2158.

Cheng Y, Ma Z, Kim B-H, Wu W, Cayting P, Boyle AP, Sundaram V, Xing X, Dogan N, Li J, et al. 2014. Principles of regulatory information conservation between mouse and human. *Nature* **515**: 371–375.

de Wit E, Vos ESM, Holwerda SJB, Valdes-Quezada C, Verstegen MJAM, Teunissen H, Splinter E, Wijchers PJ, Krijger PHL, de Laat W. 2015. CTCF Binding Polarity Determines Chromatin Looping. *Mol Cell* **60**: 676–684.

Dily FL, Baù D, Pohl A, Vicent GP, Serra F, Soronellas D, Castellano G, Wright RHG, Ballare C, Filion G, et al. 2014. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev* **28**: 2151–2162.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376–380.

Filippova D, Patro R, Duggal G, Kingsford C. 2014. Identification of alternative topological domains in chromatin. *Algorithms Mol Biol* **9**: 14.

Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, Kraft K, Kempfer R, Jerković I, Chan W-L, et al. 2016. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**: 265–269.

Fraser J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, Barbieri M, Moore BL, Kraemer DC, Aitken S, et al. 2015. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol Syst Biol* **11**: 852–852.

Fudenberg G, Mirny LA. 2012. Higher-order chromatin structure: bridging physics and biology. *Curr Opin Genet Dev* **22**: 115–124.

Gendrel A-V, Attia M, Chen C-J, Diabangouaya P, Servant N, Barillot E, Heard E. 2014. Developmental Dynamics and Disease Potential of Random Monoallelic Gene Expression. *Dev Cell* **28**: 366–380.

Gibcus JH, Dekker J. 2013. The Hierarchy of the 3D Genome. *Mol Cell* **49**: 773–782.

Giorgetti L, Galupa R, Nora EP, Piolot T, Lam F, Dekker J, Tiana G, Heard E. 2014. Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription. *Cell* **157**: 950–963.

Giorgetti L, Lajoie BR, Carter AC, Attia M, Zhan Y, Xu J, Chen CJ, Kaplan N, Chang HY, Heard E, et al. 2016. Structural organization of the inactive X chromosome in the mouse. *Nature* **535**: 575–579.

Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, 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.

Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning*. Springer New York, New York, NY http://link.springer.com/10.1007/978-0-387-84858-7 (Accessed July 11, 2016).

Hou C, Li L, Qin ZS, Corces VG. 2012. Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains. *Mol Cell* **48**: 471–484.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9**: 999–1003.

Junier I, Spill YG, Marti-Renom MA, Beato M, le Dily F. 2015. On the demultiplexing of chromosome capture conformation data. *FEBS Lett* **589**: 3005–3013.

Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S. 2014. Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics* **30**: i386–i392.

Lieberman-Aiden E, Berkum NL van, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**: 289–293.

Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, et al. 2015. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* **161**: 1012–1025.

Merkenschlager M, Nora EP. 2016. CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. *Annu Rev Genomics Hum Genet* **17**: 17–43.

Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P. 2013. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**: 59–64.

Nagano T, Várnai C, Schoenfelder S, Javierre B-M, Wingett SW, Fraser P. 2015. Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol* **16**: 175.

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, et al. 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**: 381–385.

Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, Ong C-T, Hookway TA, Guo C, Sun Y, et al. 2013. Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. *Cell* **153**: 1281–1295.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2015. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **162**: 687–688.

Sanborn AL, Rao SSP, Huang S-C, Durand NC, Huntley MH, Jewett AI, Bochkov ID, 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* **112**: E6456–E6465.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell* **148**: 458–472.

Shavit Y, Walker BJ, Lio' P. 2016. Hierarchical block matrices as efficient representations of chromosome topologies and their application for 3C data integration. *Bioinformatics* **32**: 1121–1129.

Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, Zhou XJ. 2015. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res* gkv1505.

Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, Ettwiller L, Spitz F. 2014. Functional and topological characteristics of mammalian regulatory domains. *Genome Res* **24**: 390–400.

Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B, et al. 2015. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163**: 1611–1627.

Ulianov SV, Khrameeva EE, Gavrilov AA, Flyamer IM, Kos P, Mikhaleva EA, Penin AA, Logacheva MD, Imakaev MV, Chertovich A, et al. 2015. Active chromatin and

transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res* gr.196006.115.

Van Bortle K, Nichols MH, Li L, Ong C-T, Takenaka N, Qin ZS, Corces VG. 2014. Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biol* **15**: R82.

Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S. 2015. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Rep* **10**: 1297–1309.

Weinreb C, Raphael BJ. 2015a. Identification of hierarchical chromatin domains. *Bioinformatics* btv485.