# Functional and topological characteristics of mammalian regulatory domains

Orsolya Symmons, Veli Vural Uslu, Taro Tsujimura, et al.

*Genome Res.* published online January 7, 2014
Access the most recent version at doi:10.1101/gr.163519.113

# Functional and topological characteristics of mammalian regulatory domains

Orsolya Symmons [1], Veli Vural Uslu [1], Taro Tsujimura [1], Sandra Ruf [1], Sonya Nassari [1], Wibke Schwarzer [1], Laurence Ettwiller [2,#] and François Spitz [1,*]

[1] Developmental Biology Unit – European Molecular Biology Laboratory - Meyerhofstrasse 1 - 69117 Heidelberg – Germany

[2] Centre for Organismal Studies – University of Heidelberg – Germany

[#] Present address : New England Biolabs - Ipswich – MA - United States

[*] Corresponding author

François Spitz

email: spitz@embl.de

tel: +49 6221 387 8103

fax: +49 6221 387 8166

**Running title**: Characteristics of mammalian regulatory domains

**Keywords**: gene regulation, genomic organisation, long-range enhancers, boundary elements.

ABSTRACT

Long-range regulatory interactions have an important role in shaping gene expression programs. However, the genomic features that organize these activities are still poorly characterized. We conducted a large operational analysis to chart the distribution of gene regulatory activities along the mouse genome, using hundreds of insertions of a regulatory sensor. We found that enhancers distribute their activities along broad regions and not in a gene-centric manner, defining large regulatory domains. Remarkably, these domains correlate strongly with the recently described TADs, which partition the genome into distinct self-interacting blocks. Different features, including specific repeats and CTCF-binding sites, correlate with the transition zones separating regulatory domains, and may help to further organize promiscuously distributed regulatory influences within large domains. These findings support a model of genomic organisation where TADs confine regulatory activities to specific but large regulatory domains, contributing to the establishment of specific gene expression profiles.

## INTRODUCTION

Specificity of gene expression is key to tissue function and identity, and is in great part determined at the transcriptional level. Promoters, located proximally to transcriptional start sites, play an essential role in initiating gene expression, but their activity greatly depends on the action of more distal regulatory elements. Among these, enhancers remain the best characterized (Bulger and Groudine 2011; Ong and Corces 2011). They often show a specific chromatin signature, which has helped the annotation of enhancers active in human or mouse cell lines and tissues (Rada-Iglesias et al. 2011; Heintzman et al. 2009; Creyghton et al. 2010; The ENCODE Project Consortium 2012; Shen et al. 2012). Remarkably, a substantial fraction of these putative enhancers was found hundreds of kilobases from the nearest gene. Yet, how their activity is allocated to their target gene(s) across such large distances and sometimes several genes, remains unclear.

The evidence to date indicates that enhancer-promoter communication can be influenced by specific regulatory elements, which may contribute to enhancer-promoter interactions positively (promoter architecture (Ohtsuki et al. 1998), tethering elements (Calhoun and Levine 2003)) or negatively (eg. insulators/enhancer blockers (reviewed in (Gaszner and Felsenfeld 2006)). Also, the regulatory interactions between promoters and associated remote enhancers are usually associated with physical proximity (reviewed recently in (Bickmore and van Steensel 2013; Bulger and Groudine 2011). Several protein complexes, including CTCF, cohesin, and Mediator, have been proposed to have a role in organising these regulatory contacts (Kagey et al. 2010; Wendt et al. 2008; Parelho et al. 2008; Hadjur et al. 2009). Furthermore, recent studies suggest that the genome is organized into relatively cell-type invariant topological domains (TADs) characterized by preferential self-contacts

(Nora et al. 2012; Dixon et al. 2012; Sexton et al. 2012; Hou et al. 2012). Genes located in the same TAD show greater expression correlation than genes located in distinct ones (Nora et al. 2012), and TAD borders are enriched for factors implicated as insulator elements (Dixon et al. 2012; Hou et al. 2012). These observations suggested that TADs may form a backbone for tissue-specific regulatory interactions (Bickmore and van Steensel 2013; Gibcus and Dekker 2013; Nora et al. 2013).

However, despite a growing body of knowledge on individual regulatory elements and the three-dimensional organisation of the genome, the "rules of engagement" (Splinter and de Laat 2011) that determine the activity of remote *cis*-acting elements on the surrounding genes remain elusive, in part because a direct assessment of how such activities are distributed in the genomic environment is still lacking. Scrutinizing endogenous gene activity offers only a partial view of this distribution, since expression of endogenous genes also depends on their different promoters, and distinct post-transcriptional regulation. Furthermore, it does not provide information about the gene deserts that constitute about 25% of mammalian genomes and where many functionally important regulatory elements lie (Ovcharenko et al. 2004; Nobrega et al. 2003).

We have recently developed an efficient *in vivo* transposition system that allows the rapid production of mice with a single copy insertion of a regulatory sensor (Ruf et al. 2011). This regulatory sensor consists of a *lacZ* reporter gene driven by a weak promoter, and has minimal effect - if any - on the expression of the surrounding genes. The sensor is carried in a *Sleeping Beauty* transposon that integrates almost randomly in the genome (Horie et al. 2003; Liu et al. 2005). Importantly, the sensor is "naïve": it has not been subjected to the evolutionary selection that has shaped endogenous genes to favour or avoid regulatory influences. Therefore, a comparison of the

sensor's expression pattern at different insertions provides a simple and direct readout for the regulatory input acting on those positions.

By analysing the activity of the regulatory sensor in mouse embryos at several hundred insertion sites, we identified multiple factors that impinge on the action of *cis*-regulatory elements. We confirm circumstantial evidence that the influence of enhancers extends over several hundred kilobases, irrespective of the position of the normal target genes. By comparing adjacent insertions, we identified large regulatory domains as well as transition zones. We find that regulatory domains are included within TADs, and accordingly depleted in features associated with insulating activities, whereas TADs boundaries correspond to regulatory transitions. Thus, the influence of enhancers appears limited to the topological domain they are part of, providing direct support to the notion that TADs correspond to a functional subdivision of the genome into regions where otherwise promiscuous regulatory influences are confined.

## RESULTS

Our initial analysis of about 160 insertions of a regulatory sensor had suggested that tissue-specific regulatory activities are pervasively present throughout the genome (Ruf et al. 2011). From this, we inferred that with an increased number and density of insertions, one could map - in the natural genomic context – the intervals that were responsive to enhancers, and thus investigate the regulatory architecture of the mouse genome in greater detail (Figure 1A).

For this purpose, we generated and mapped more than thousand new integration sites using extensive transposition from single-copy starting points. We characterised a substantial subset of these (747 insertions) for *lacZ* expression (Supplementary Table 1, online TRACER database (Chen et al. 2013)). These insertions are distributed throughout the genome, with a subset clustered in domains of up to one megabase, owing to the local hopping of the *Sleeping Beauty* transposon (Horie et al. 2003). Overall, most insertions are localized far from endogenous transcriptional start sites (TSS) (Figure 1B), with two thirds more than 50 kilobases from the nearest TSS. Consistent with our first analysis (Ruf et al. 2011), approximately 55% of all insertions showed expression, with the vast majority expressed in a tissue-restricted manner (Symmons and Spitz 2013). Insertions far from TSS are more likely to show expression than the ones next to TSS (Figure 1B), probably due to local competition with endogenous gene promoters (Ruf et al. 2011). Importantly, independent insertions show extremely diverse expression patterns (Figure 1C): only nearby insertions or insertions around paralogous genes yielded expression patterns that were not clearly distinguishable.

**Binding of the enhancer-associated protein EP300 is predictive of tissue-specific expression over large distances.**

We first sought to obtain deeper insight into the range of action of enhancers. For this, we investigated whether insertions were more likely to show expression in a given tissue if located in the proximity of an active enhancer, and - if so - how far we could detect such enrichment. We compiled a list of 670 non-redundant viewpoints, annotated with a simple controlled vocabulary (Chen et al. 2013) (Supplementary Table 2). Since binding of the co-activator protein EP300 has been reported to demarcate active enhancers in a given tissue (Visel et al. 2009), we used EP300 binding sites from E11.5 mouse embryonic forebrain, midbrain, limb and heart (Visel et al. 2009; Blow et al. 2010). We identified pairs of sensor insertion/EP300 binding sites, and classified them as concordant, when the sensor was expressed in the tissue where the EP300 binding site was detected. By calculating the enrichment of such concordant pairs compared to a set of random insertion/EP300 site pairs, we revealed significant enrichment for concordant pairs up to a distance of 200 kilobases (Figure 2A and Supplementary Figure 1). When EP300 binding sites were paired with silent insertions or insertions expressed in different tissues, enrichment was weaker or not found. This analysis implied that enhancers - represented here by EP300-bound regions – exert their effects across large distances. A similar typical range has been observed when comparing the distance between enhancer regions and the TSS of endogenous genes (Chepelev et al. 2012; Blow et al. 2010; Li et al. 2012).

**Enhancer activity is not gene-centric and broadly distributed**

To refine this analysis, we made use of enhancers for which transgenic activity had been previously documented in E11.5 embryos, mostly from the Vista Enhancer

database (Visel et al. 2007). We carefully compared the spatial activity of these enhancers to that of the regulatory sensor when inserted in their vicinity (Figure 2B-E). Given the highly specific patterns of the different insertions and enhancers, this offered much greater confidence in a direct relationship between matching patterns. We defined pairs as "concordant" when the domain of enhancer activity was included in the one shown by the regulatory sensor; as "divergent" or "inactive", when the regulatory sensor was expressed elsewhere or not expressed (Supplementary Table 3). As shown in Figure 2B, all insertions positioned in close proximity (<10kb) of an enhancer showed expression patterns highly concordant with its autonomous activity. This ratio decreased with increasing distance, but even at 100-200kb, a third of insertions-enhancer pairs were concordant, compared to less than 7% for random associations of enhancers and insertions (Figure 2B, Fisher's Exact Probability Test, $p<3.305e-6$). Expression of our regulatory sensor was not dependent on its orientation (Supplementary Figure 2).

We then examined if enhancers distributed their activities preferentially towards their endogenous target gene. For this purpose, we considered a subset of enhancers for which we could confidently assign a – putative - target gene, using *in situ* gene expression data from the literature. We categorised these 107 enhancer-gene-insertion triplets (corresponding to 33 genomic loci) based on the relative order of their components, and determined whether the inserted sensor showed complete, partial or no overlapping expression with the enhancer/gene pair (Supplementary Table 4, Figure 3). Insertions located between an enhancer and its target gene showed overlap with the reported enhancer activity more frequently than insertions located either opposite to the gene or beyond the transcriptional start site (Fisher's exact test, $p=0.001457$ and $p=0.04238$ respectively). However, about half of the insertions

located opposite the target gene or beyond its promoter showed an expression pattern corresponding to the associated enhancer (Figure 3), indicating that enhancer activity is not exclusive to the promoter of their target gene. Furthermore, given the relatively small sample size, it is possible that the weak directionality of enhancer activity that we observed is due to the different distances associated with the different categories.

Overall, these multiple analyses show that enhancers act on their environment broadly and in a largely undiscriminating manner.

**Extended enhancer activity results in large regulatory domains**

Consistent with these findings, we noted that adjacent insertions frequently had similar expression, even when hundreds of kilobases apart. Large domains of co-regulation have been reported before, usually around developmental genes (Spitz et al. 2003; Zuniga et al. 2004; Kikuta et al. 2007), but with our extensive survey of the mouse genome, we could expand this list and refine the extent of such domains. Importantly, since the regulatory sensor is driven by the same promoter at each insertion site, changes in expression are not due to differently responsive promoters, but really outline the limits of enhancer action. We cannot formally exclude that co-expression of the sensor at different positions results from activation by different enhancers with overlapping activities, like those discovered around developmental genes (Uchikawa et al. 2003; Visel et al. 2013; Marinić et al. 2013; Carvajal et al. 2001; Hong et al. 2008). But individual enhancers usually have distinct spatial activities (Visel et al. 2007) and therefore, the spatially restricted expression patterns that we use to define co-expression provides confidence for genuine co-regulation. Supporting this further, insertions into loci where enhancers have been meticulously mapped allowed us to assess that they respond to the same enhancer(s) at multiple

distant positions (eg. *Shh*, Supplementary Figure 3 and *Hoxd*, Supplementary Figure 4).

We found 311 chromosomal intervals defined by the presence of 2 or more insertions within less than 2Mb. We assigned these intervals to different categories, depending on the expression patterns of the insertions (Figure 4A-C; Supplementary Figures 3-5, Supplementary Table 5). We defined *regulatory domains* (RDs; 46 regions; size from 3.8kb to 2.1Mb; median size: 359kb; total combined length: 22.3Mb) as the largest possible interval containing multiple insertions with shared expression. This is a conservative definition, and RDs likely extend beyond the insertions that define them. RDs are found on all chromosomes, except possibly Y, where we did not obtain insertions with expression. Within RDs, we occasionally saw quantitative variation or no detectable expression of the sensor at different positions. However, we never observed insertions with divergent expression between two insertions with the same activity.

Next, we defined *transition zones* (TZs) as regions separating insertions with distinct expression profiles (66 intervals, ranging from 14kb to 1.9Mb; median size: 734kb; total combined length: 53.8Mb). Other types of intervals (class A: two or more insertions without detectable expression; class B: one insertion with expression and one with no expression) were also annotated. B regions may be a type of transition, but we considered them separately, since several local features may lead to inactivity of the sensor (Ruf et al. 2011).

**Regulatory domains are included in topologically-associating domains**

Next, we looked if our operational subdivision of the genome matched any structural features of the genome. To this end, we compared RDs with the largely

cell-invariant self-associating "topologically-associating domains" (TADs) identified by Hi-C (Dixon et al. 2012). We found that the great majority of RDs (78%) were contained within one TAD (Figure 4D). The remaining ones extended into flanking unstructured regions (for which Hi-C data did not highlight specific compartmentalisation). In only one case, two adjacent insertions with partially overlapping – but not identical - pattern of expression were found in distinct adjacent TADs. Similarly, insertions mirroring the expression patterns of adjacent but remote genes were also found within the same TAD (Supplementary Figure 5C-E). In contrast, one third of TZs, including relatively short regions in gene-deserts, were located in different TADs (Figures 4C and Supplementary Figure 6). Since, on average, RDs were smaller than TZs, we performed several analyses to verify the significance of this different distribution relative to TADs. We randomly permutated RD, TZ, A and B regions, and found it was significant ($p<0.05$) that RDs almost never overlapped two or more TADs, unless we considered only extreme size ranges (below 400kb; above 1.5Mb). In particular, for intervals between 200kb and 1Mb, a range where the size distribution of the RD, TZ, A and B regions was not significantly different, the altered distribution of RDs and TZs relative to TADs was highly significant (Figure 4D). We also calculated the density of TAD ends for the real intervals and for 1000 random distributions of intervals of the same size in the genome, and found that RDs, but not the other categories, showed a significant depletion of TAD ends (Figure 4E).

**Distribution of regulatory activities and insulators.**

Next, we compared how insulators or elements with enhancer-blocking activities are distributed between the four types of operationally defined domains. We

considered CTCF sites, SINE B1 and SINE B2 elements, as well as the TSS of protein-coding and non-coding genes from the RefSeq collection (Supplementary Table 6). All four elements were clearly depleted from RDs (Figure 5A, Supplementary Figure 7), whereas type A and B control regions never showed depletion, and TZs showed no, or far less depletion. This depletion was largely maintained, even when we considered potentially confounding factors, such as the clustering of CTCF sites and TSS (Kim et al. 2007; Shen et al. 2012) or the overlap of CTCF sites and SINE B2 elements (Schmidt et al. 2012) (Supplementary Figure 7). The proportion of RDs containing at least one cell-invariant CTCF site was also lower than for TZ (39% compared to 70%), consistent with the proposed contribution of this protein in organising regulatory interactions (Phillips and Corces 2009). However, since TAD boundaries are enriched for the different elements associated with insulator activity (Dixon et al. 2012), we also restricted our analysis to intervals located within TADs. In this case, we still found lower than expected density of TSS and SINE B2. However, the depletion of CTCF binding sites and SINE B1 X35S in RDs was no longer statistically significant (Figure 5B, Supplementary Figure 7), suggesting that it correlated with the topological, rather than regulatory subdivision of the genome.

As CTCF (Phillips and Corces 2009), together with cohesin complexes (Hadjur et al. 2009; Kagey et al. 2010; Merkenschlager and Odom 2013), has been proposed to have a major role in organising regulatory interactions we examined their relationship to RDs in more detail, using available datasets (Supplementary Table 6). In agreement with previous analyses, we found a high degree of overlap between CTCF and cohesin-binding across mouse cell-types and tissues (Wendt et al. 2008; Parelho et al. 2008; Remeseiro et al. 2012). In several cases, cohesin and CTCF-

bound regions lie within regulatory TZs, often also corresponding to TAD boundaries, an observation consistent with their proposed role (Phillips-Cremins et al. 2013).

However, even if slightly less abundant in RDs (Figure 5C), many CTCF and cohesin binding sites are also interspersed within RDs, between co-expressed genes, insertions and their associated enhancers (Figure 6 and Supplementary Figures 4, 5 and 8). We noted that RDs can also be further subdivided in more “specialized” ones, each characterized by expression specificities additional to the ones defining the RD (Figure 5C, Supplementary Figure 5A,E). Some of these subdivisions, with the current resolution offered by the available insertions, might be outlined by the presence of CTCF/cohesin sites (Figure 5). However, by and large, it appeared difficult to match the mere distribution of the CTCF/cohesin sites with the distribution of regulatory activities, highlighting that, within TADs, the binding of cohesin/CTCF, defined by chromatin-immunoprecipitation, may not be a sufficient indicator of regulatory boundaries.

## DISCUSSION

In vertebrates, since many critical enhancers lie at considerable distances from the genes they influence, ensuring and controlling proper interactions between regulatory elements and promoters is essential (Splinter and de Laat 2011; Williamson et al. 2011; Bulger and Groudine 2011). Our mechanistic understanding of this process is derived from studies carried out on a limited number of model loci (Gaszner and Felsenfeld 2006; Montavon et al. 2011; Simonis et al. 2006; Jhunjhunwala et al. 2008; Marinić et al. 2013; Vernimmen et al. 2007; Tena et al. 2011; Amano et al. 2009). Here, we expanded these by probing the regulatory landscape of the entire mouse genome with a transposable, naïve *lacZ* sensor. This large-scale exploration of the genomic regulatory architecture identified - in an operational manner - the widespread presence of large regulatory domains (RDs), within which the sensor displayed highly similar expression patterns at multiple distant positions. These RDs largely overlapped with TADs, the sub-megabase-sized self-interacting intervals defined by chromosomal conformation capture analysis (Dixon et al. 2012; Nora et al. 2012). Insertions located in adjacent TADs almost systematically reported distinct regulatory activities, providing direct support for the suggested role of TADs as the basic building blocs of genomic regulatory architecture (Dixon et al. 2012; Nora et al. 2012; Hou et al. 2012; Gibcus and Dekker 2013). RDs are usually gene-poor and shared several other features with TADs, such as a depletion of various elements (SINE B2 repeats, constitutive CTCF-binding sites).

Our operational approach also shed some light on the organisation of these domains. Firstly, despite the overall inclusion of regulatory domains into TADs, the positions of the relative transitions were not always exactly superimposed, similarly to chromatin domains, which do not exactly correlate with TADs (Hou et al. 2012).

These differences may partially arise from the low resolution of Hi-C, which can locate topological transitions with only limited precision (20kb). Furthermore, topological transitions may not necessarily constitute absolute barriers, but act more like dampers (Andrey et al. 2013).

Within TADs, we found that enhancer activities are broadly distributed, and not targeted to specific regions (i.e. proximity of gene promoters or TAD borders). From their discovery, enhancers have been shown to act irrespectively of their orientation (Banerji et al. 1981). Genome-wide studies have further shown that enhancers can be found both 3' and 5' of their endogenous target genes (The ENCODE Project Consortium 2012; Shen et al. 2012; Li et al. 2012). Our data further stresses that orientation-independence is an intrinsic property of enhancers in their normal context: in addition to controlling of target genes, enhancers generally act pervasively throughout their regulatory domains.

This broad distribution of enhancer activities along large domains has several implications. It may account for the transcriptional "ripple" effect observed upon growth-factor stimulation (Ebisuya et al. 2008) and bystander gene activation (Spitz et al. 2003; Zuniga et al. 2004; Cajiao et al. 2004). Furthermore, widespread enhancer activities may have been evolutionarily advantageous for genes brought into new neighbourhood by chromosomal rearrangements (Cande et al. 2009; De et al. 2009) or retrotransposition (Vinckenbosch et al. 2006), or for emerging lncRNAs (Ponting et al. 2009; Kutter et al. 2012), facilitating the acquisition of new expression domains.

Constraining enhancer activity through the formation of distinct topological compartments may contribute to gene regulation in two important ways. It may help specify gene-enhancer interactions, and restrict ectopic "enhancer adoption" to accidental disruption of existing topologies, for example through chromosomal

rearrangements (Niedermaier et al. 2005; Spitz et al. 2003; Kantaputra et al. 2010; Gostissa et al. 2009; Kokubu et al. 2003; Marinić et al. 2013). Alternatively, it may help integrate the activity of multiple regulatory elements spread along large intervals (Sanyal et al. 2012; Shen et al. 2012; Li et al. 2012; Marinić et al. 2013; Montavon et al. 2011; Delpretti et al. 2013; Carvajal et al. 2001; Uchikawa et al. 2003; Visel et al. 2013) into the coherent regulatory units that have been described as regulatory archipelagos, holo-enhancers or chromatin-hubs (Montavon et al. 2011; Palstra et al. 2003; Marinić et al. 2013). Our observation that in intact endogenous loci, randomly inserted sensors report the complete integrated output provided to normal target genes (and do not decompose it into the individual activities of the closest enhancers) (e.g. *Hoxd*, Supplementary Figure 4; *Foxg1*, Figure 4B; *Twist1* (Birnbaum et al. 2012; Ruf et al. 2011)), further supports the notion that regulatory control is exerted by coordinated, and not individual action, of enhancers. It indicates that this integration is orchestrated at the level of the regulatory domain, not at endogenous gene promoters (Marinić et al. 2013).

Along a given regulatory domain, the expression level detected by the sensor can vary quantitatively and reveal subdomains, including few positions apparently refractory to activation. We speculate that these “cold-spots” within otherwise permissive domains may shield genes from the influence of surrounding enhancers (Marinić et al. 2013), adding to other mechanisms of specificity such as core promoter sequences (Ohtsuki et al. 1998). This fine-scale organisation of activities may be due to the positions of the corresponding enhancers, to local epigenetic modifications (Arnold et al. 2013; Akhtar et al. 2013), or to different tissue-specific three-dimensional organisation within domains. Indeed, restructuring of physical interactions within TADs has been shown during lineage-commitment (Phillips-

Cremins et al. 2013) and in different cell-types (Dixon et al. 2012). We suggest that such changes in local interactions may correlate with the different sub-regulatory domains that we observed within TADs. It will be an exciting avenue to explore what proteins (Aragon et al. 2013; Merkenschlager and Odom 2013; Kagey et al. 2010; Remeseiro et al. 2012; Hadjur et al. 2009) and nuclear substructures (Gibcus and Dekker 2013; Bickmore and van Steensel 2013; Meuleman et al. 2013) may be involved. Even though we found frequent co-occurrence of CTCF bound-regions with transitions between regulatory domains, our current analyses also highlight the difficulty of inferring the structure of these domains from the mere presence of CTCF/cohesin binding sites detected by chromatin-immunoprecipitation. This reinforces the growing perception that CTCF may have a versatile role (Sanyal et al. 2012; Phillips-Cremins et al. 2013; Handoko et al. 2011), maybe different within topological domains and at their borders, and that analysing these activities in the context of a 3D genome will be essential (Handoko et al. 2011; DeMare et al. 2013). Our data also emphasizes the need of direct functional approaches, as the one developed here, to map regulatory interactions and to compare them to physical conformations or chromatin maps. The confrontation of these approaches will be essential to understand the mechanistic basis and principles underlying the organisation of the genome. Ultimately, such an understanding will be crucial to predict whether and how variants (such as structural variants) affect topological and regulatory organisation, and consequently influence gene expression and phenotype (Weischenfeldt et al. 2013).

**METHODS.**

Mouse lines and embryos with SBlac insertions.

Insertions sites of the transposon were generated, mapped, and *lacZ* expression analysis was performed as described previously (Ruf et al. 2011). Information on insertions is in Supplementary Table 1 and details are on the TRACER website (tracedatabase.embl.de) (Chen et al. 2013). Whole-mount *in situ* hybridisations were carried out as described in Ruf et al. 2011. Templates for mRNA antisense probe synthesis were described in Spitz et al. 2003 (*Hoxd13*), Ruf et al. 2011 (*Sall1*), or produced by PCR on mouse cDNA using the following primers (Foxg1_2F: GCCAAGCTGGCCTTTAAGC; Foxg1_2R: ATTCTCCCACATTGCACCTC; Nr2f2_3Sp6: CATTTAGGTGACACTATAGCCACATGGGCTACATCAGAC; Nr2f2_5: GGGCGGAGGAACCTGAGCTACAC; Hand2_F: AGGACTCAGAGCATCAACAGC; Hand2-R: AGCGGATGCTCAAAGGTG).

Mouse experiments were conducted in accordance with the principles and guidelines in place at European Molecular Biology Laboratory, as defined and overseen by its Institutional Animal Care and Use Committee, in accordance with the European Convention 18/3/1986 and Directives 86/609/EEC and 2010/63/EU.

Genomic datasets and resources used for the study.

Genomic datasets were obtained from public sources and are summarised in Supplementary Table 6.

SB-EP300 comparison

We considered SB insertions with expression in heart, limb, midbrain or forebrain, where EP300 binding data has been generated (Blow et al. 2010; Visel et

al. 2009), and merged insertions with the same expression pattern <5kb apart into one (Supplementary Table 2). In a given tissue, we calculated the enrichment of insertions with expression in that tissue, compared to the same number of randomly selected insertions with no expression in that tissue (200 randomisations). Enrichment was calculated for increasing distances (with steps of 5000 bp from 0 to 1 Mb, provided more than 5 insertions were found within). To test if this enrichment was specific to the relevant tissue, we repeated the analysis using insertions expressed in other tissues or not expressed. For this, we removed EP300 sites if they were within 10kb of a EP300 site in the first tissue considered, to reduce the possible confounding signal arising from clusters of different tissue-specific EP300 sites (Visel et al. 2009).

Comparison of insertions, autonomous enhancer activity and endogenous genes

We complemented the set of enhancers active at E11.5 from the VISTA enhancer browser (Visel et al. 2007) with additional enhancers from the literature. We visually inspected expression overlap between enhancer-insertion pairs (where insertions were within 200kb of these enhancers; Supplementary Table 3), and annotated pairs as concordant, discordant or inactive. As random control datasets (Supplementary Table 3) we first performed random permutation of the positions of the insertions and enhancers, and extracted new adjacent pairs. Secondly, we randomly drew insertion – enhancer pairs and compared their expression patterns.

For the Enhancer-Regulatory Sensor-Gene triplets, we compared the expression of characterized enhancers to reported *in situ* hybridization gene expression of either the first two flanking genes or genes less than 2Mb away using gene expression resources (Supplementary Table 6). We considered sensor insertions lying within the same genomic interval, and annotated their expression as concordant, partially

overlapping or divergent (no or different expression) with the enhancer-gene pair. These enhancer-gene-insertion triplets were then categorized depending on the position of the insertion relative to the enhancer and gene TSS. For neighbouring insertions that showed the same annotation/relative position for a given enhancer, we kept only one triplet. Informative triplets are listed in Supplementary Table 4.

Defining regions with characteristic expression patterns.

We functionally defined different classes of genomic intervals, by comparing the expression of adjacent insertions <2Mb apart. Intervals were both insertions showed no expression were annotated as class A. Class B was formed by pairs where one insertion was expressed whereas the other was not. Intervals defined by two insertions with discordant (non-overlapping) expression patterns formed transition zones. Intervals were insertions showed overlapping *lacZ* expression were grouped as regulatory domains (RDs). To define a RD, the observed spatial pattern had to be identical, not only correspond to the same broad anatomical domain: patterns as in Figure 1B are considered divergent. In cases, when multiple insertions within a 2Mb window were concordantly expressed, we considered the most centromeric and telomeric insertions as the boundaries of the RD, but only if no insertion in between were discordantly expressed. Thus, we merged adjacent RDs with same activity together, sometimes including small B regions. Similarly, we determined the minimal extent of transition zones as the smallest region flanked by divergently expressed insertions (*i.e.* where none of the expression patterns overlapped), even if they were disrupted by insertions with no expression. We did not further group insertion pairs, where one or both insertions are not expressed, except when three or more

consecutive insertions showed no expression, and one pair was <10kb apart. The extended annotated regions are listed in Supplementary Table 5.

Comparison to topological domains and other genomic landmarks.

We compared the occurrence of certain features in our functionally defined regions to a random model. To estimate the significance (p-value) of the frequency of certain features (*eg.* topological boundaries, CTCF sites, SINE B2 elements) for each category (*ie*. all A, B, RD and TZ domains, or those found intra-TAD only) we applied bootstrap following recommendations from (Phipson and Smyth 2010). For details see Supplementary Note 1. For TAD boundaries we also performed random permutation, as described in the text and Supplementary Note 1. For the analysis we used the datasets described in Supplementary Table 6.

## DATA ACCESS

All insertions generated for this study are documented in the TRACER database (Chen et al. 2013) (http://tracerdatabase.embl.de; or http://www.ebi.ac.uk/panda-srv/tracer/index.php).

## ACKNOWLEDGEMENTS.

We thank Angel-Carlos Roman for SINEB1-X35S coordinates. We thank the members of the EMBL Laboratory Animal Resources Facility for animal welfare and husbandry, especially Silke Feller and Michaela Wesch. For computational analysis we received invaluable support from Charles Girardot and the EMBL Genome Biology Computational Support Group. We also thank members of the Spitz lab and

colleagues from EMBL for helpful discussion and sharing reagents, and Laura Panavaite for providing the in situ data for *Hand2*. O.S. and V.V.U were supported by PhD fellowships from the Louis-Jeantet Foundation and Jeff Schell Darwin Trust, respectively. T.T. was supported successively by postdoctoral fellowships awarded by the Uehara Memorial Foundation and Japan Society for the Promotion of Science. This work was supported by the European Molecular Biology Laboratory, the European Commission-FP7 (grant Health 223210/CISSTEM) and the Human Frontier Science Program (grant RGY0081/2008-C) (to F.S.).

## DISCLOSURE DECLARATION

The authors declare no conflict of interest.

## AUTHORS' CONTRIBUTIONS.

F.S. and O.S. designed the experiments. O.S., V.V.U., T.T., S.R., S.N. and W.S. performed the experiments. O.S., L.E. and F.S. analysed the data. O.S. and F.S. wrote the manuscript with input from the other authors.

## REFERENCES

Akhtar W, de Jong J, Pindyurin AV, Pagie L, Meuleman W, de Ridder J, Berns A, Wessels LFA, van Lohuizen M, van Steensel B. 2013. Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* **154**: 914–927.

Amano T, Sagai T, Tanabe H, Mizushina Y, Nakazawa H, Shiroishi T. 2009. Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. *Dev Cell* **16**: 47–57.

Andrey G, Montavon T, Mascrez B, Gonzalez F, Noordermeer D, Leleu M, Trono D, Spitz F, Duboule D. 2013. A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science* **340**: 1234167.

Aragon L, Martinez-Perez E, Merkenschlager M. 2013. Condensin, cohesin and the control of chromatin states. *Curr Opin Genet Dev* **23**: 204–211.

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**: 1074–1077.

Banerji J, Rusconi S, Schaffner W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**: 299–308.

Bickmore WA, van Steensel B. 2013. Genome architecture: domain organization of interphase chromosomes. *Cell* **152**: 1270–1284.

Birnbaum RY, Clowney EJ, Agamy O, Kim MJ, Zhao J, Yamanaka T, Pappalardo Z, Clarke SL, Wenger AM, Nguyen L, et al. 2012. Coding exons function as tissue-specific enhancers of nearby genes. *Genome Res* **22**: 1059–1068.

Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2010. ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* **42**: 806–810.

Bulger M, Groudine M. 2011. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**: 327–339.

Cajiao I, Zhang A, Yoo EJ, Cooke NE, Liebhaber SA. 2004. Bystander gene activation by a locus control region. *EMBO J* **23**: 3854–3863.

Calhoun VC, Levine M. 2003. Long-range enhancer-promoter interactions in the Scr-Antp interval of the Drosophila Antennapedia complex. *Proc Natl Acad Sci USA* **100**: 9878–9883.

Cande JD, Chopra VS, Levine M. 2009. Evolving enhancer-promoter interactions within the tinman complex of the flour beetle, Tribolium castaneum. *Development* **136**: 3153–3160.

Carvajal JJ, Cox D, Summerbell D, Rigby PW. 2001. A BAC transgenic analysis of the Mrf4/Myf5 locus reveals interdigitated elements that control activation and

maintenance of gene expression during muscle development. *Development* **128**: 1857–1868.

Chen C-K, Symmons O, Uslu VV, Tsujimura T, Ruf S, Smedley D, Spitz F. 2013. TRACER: a resource to study the regulatory architecture of the mouse genome. *BMC Genomics* **14**: 215.

Chepelev I, Wei G, Wangsa D, Tang Q, Zhao K. 2012. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res* **22**: 490–503.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* **107**: 21931–21936.

De S, Teichmann SA, Babu MM. 2009. The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. *Genome Res* **19**: 785–794.

Delpretti S, Montavon T, Leleu M, Joye E, Tzika A, Milinkovitch M, Duboule D. 2013. Multiple Enhancers Regulate Hoxd Genes and the Hotdog LncRNA during Cecum Budding. *Cell Reports* **5**: 137–150.

DeMare LE, Leng J, Cotney J, Reilly SK, Yin J, Sarro R, Noonan JP. 2013. The genomic landscape of cohesin-associated chromatin interactions. *Genome Res* **23**: 1224–1234.

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. http://www.nature.com/nature/journal/vaop/ncurrent/full/nature11082.html.

Ebisuya M, Yamamoto T, Nakajima M, Nishida E. 2008. Ripples from neighbouring transcription. *Nat Cell Biol* **10**: 1106–1113.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Gaszner M, Felsenfeld G. 2006. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* **7**: 703–713.

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

Gostissa M, Yan CT, Bianco JM, Cogné M, Pinaud E, Alt FW. 2009. Long-range oncogenic activation of Igh-c-myc translocations by the Igh 3' regulatory region. *Nature* **462**: 803–807.

Gray PA, Fu H, Luo P, Zhao Q, Yu J, Ferrari A, Tenzen T, Yuk D-I, Tsung EF, Cai Z, et al. 2004. Mouse brain organization revealed through direct genome-scale TF expression analysis. *Science* **306**: 2255–2257.

Hadjur S, Williams LM, Ryan NK, Cobb BS, Sexton T, Fraser P, Fisher AG,

Merkenschlager M. 2009. Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature* **460**: 410–413.

Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CWH, Ye C, Ping JLH, Mulawadi F, et al. 2011. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* **43**: 630–638.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**: 108–112.

Hong J-W, Hendrix DA, Levine MS. 2008. Shadow enhancers as a source of evolutionary novelty. *Science* **321**: 1314.

Horie K, Yusa K, Yae K, Odajima J, Fischer SEJ, Keng VW, Hayakawa T, Mizuno S, Kondoh G, Ijiri T, et al. 2003. Characterization of Sleeping Beauty transposition and its application to genetic screening in mice. *Mol Cell Biol* **23**: 9189–9207.

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.

Jhunjhunwala S, van Zelm MC, Peak MM, Cutchin S, Riblet R, van Dongen JJM, Grosveld FG, Knoch TA, Murre C. 2008. The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. *Cell* **133**: 265–279.

Jukkola T, Lahti L, Naserke T, Wurst W, Partanen J. 2006. FGF regulated gene-expression and neuronal differentiation in the developing midbrain–hindbrain region. *Developmental Biology* **297**: 141–157.

Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, et al. 2010. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**: 430–435.

Kantaputra PN, Klopocki E, Hennig BP, Praphanphoj V, Le Caignec C, Isidor B, Kwee ML, Shears DJ, Mundlos S. 2010. Mesomelic dysplasia Kantaputra type is associated with duplications of the HOXD locus on chromosome 2q. *Eur J Hum Genet* **18**: 1310–1314.

Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engström PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K, et al. 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res* **17**: 545–555.

Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B. 2007. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**: 1231–1245.

Kokubu C, Wilm B, Kokubu T, Wahl M, Rodrigo I, Sakai N, Santagati F, Hayashizaki Y, Suzuki M, Yamamura K-I, et al. 2003. Undulated short-tail deletion mutation in the mouse ablates Pax1 and leads to ectopic activation of

neighboring Nkx2-2 in domains that normally express Pax1. *Genetics* **165**: 299–307.

Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC. 2012. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet* **8**: e1002841.

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, et al. 2012. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**: 84–98.

Liu G, Geurts AM, Yae K, Srinivasan AR, Fahrenkrug SC, Largaespada DA, Takeda J, Horie K, Olson WK, Hackett PB. 2005. Target-site preferences of Sleeping Beauty transposons. *J Mol Biol* **346**: 161–173.

Marinić M, Aktas T, Ruf S, Spitz F. 2013. An Integrated Holo-Enhancer Unit Defines Tissue and Gene Specificity of the Fgf8 Regulatory Landscape. *Dev Cell* **24**: 530–542.

McGlinn E, Richman JM, Metzis V, Town L, Butterfield NC, Wainwright BJ, Wicking C. 2008. Expression of the NET family member Zfp503 is regulated by hedgehog and BMP signaling in the limb. *Dev Dyn* **237**: 1172–1182.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**: 495–501.

Merkenschlager M, Odom DT. 2013. CTCF and cohesin: linking gene regulatory elements with their targets. *Cell* **152**: 1285–1297.

Meuleman W, Peric-Hupkes D, Kind J, Beaudry JB, Pagie L, Kellis M, Reinders M, Wessels L, van Steensel B. 2013. Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Res* **23**: 270–280.

Montavon T, Soshnikova N, Mascrez B, Joye E, Thevenet L, Splinter E, de Laat W, Spitz F, Duboule D. 2011. A regulatory archipelago controls hox genes transcription in digits. *Cell* **147**: 1132–1145.

Niedermaier M, Schwabe GC, Fees S, Helmrich A, Brieske N, Seemann P, Hecht J, Seitz V, Stricker S, Leschik G, et al. 2005. An inversion involving the mouse Shh locus results in brachydactyly through dysregulation of Shh expression. *J Clin Invest* **115**: 900–909.

Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302**: 413.

Nora EP, Dekker J, Heard E. 2013. Segmental folding of chromosomes: a basis for structural and regulatory chromosomal neighborhoods? *Bioessays* **35**: 818–828.

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.

Ohtsuki S, Levine M, Cai HN. 1998. Different core promoters possess distinct regulatory activities in the Drosophila embryo. *Genes Dev* **12**: 547–556.

Ong C-T, Corces VG. 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* **12**: 283–293.

Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L. 2004. Evolution and functional classification of vertebrate gene deserts. *Genome Res* **15**: 137–145.

Palstra R-J, Tolhuis B, Splinter E, Nijmeijer R, Grosveld F, de Laat W. 2003. The beta-globin nuclear compartment in development and erythroid differentiation. *Nat Genet* **35**: 190–194.

Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, Jarmuz A, Canzonetta C, Webster Z, Nesterova T, et al. 2008. Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**: 422–433.

Phillips JE, Corces VG. 2009. CTCF: master weaver of the genome. *Cell* **137**: 1194–1211.

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.

Phipson B, Smyth GK. 2010. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol* **9**: Article39.

Ponting CP, Oliver PL, Reik W. 2009. Evolution and functions of long noncoding RNAs. *Cell* **136**: 629–641.

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**: 279–283.

Remeseiro S, Cuadrado A, Gómez-López G, Pisano DG, Losada A. 2012. A unique role of cohesin-SA1 in gene regulation and development. *EMBO J* **31**: 2090–2102.

Ruf S, Symmons O, Uslu VV, Dolle D, Hot C, Ettwiller L, Spitz F. 2011. Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nat Genet* **43**: 379–386.

Saba R, Johnson JE, Saito T. 2005. Commissural neuron identity is specified by a homeodomain protein, Mbh1, that is directly downstream of Math1. *Development* **132**: 2147–2155.

Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of

gene promoters. *Nature* **489**: 109–113.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**: 335–348.

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.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, et al. 2012. A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**: 116–120.

Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W. 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* **38**: 1348–1354.

Spitz F, Gonzalez F, Duboule D. 2003. A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* **113**: 405–417.

Splinter E, de Laat W. 2011. The complex transcription regulatory landscape of our genome: control in three dimensions. *EMBO J* **30**: 4345–4355.

Symmons O, Spitz F. 2013. From remote enhancers to gene regulation: charting the genome's regulatory landscapes. *Philos Trans R Soc Lond, B, Biol Sci* **368**: 20120358.

Tena JJ, Alonso ME, la Calle-Mustienes de E, Splinter E, de Laat W, Manzanares M, Gómez-Skarmeta JL. 2011. An evolutionarily conserved three-dimensional structure in the vertebrate Irx clusters facilitates enhancer sharing and coregulation. *Nat Commun* **2**: 310.

Uchikawa M, Ishida Y, Takemoto T, Kamachi Y, Kondoh H. 2003. Functional analysis of chicken Sox2 enhancers highlights an array of diverse regulatory elements that are conserved in mammals. *Dev Cell* **4**: 509–519.

Vernimmen D, De Gobbi M, Sloane-Stanley JA, Wood WG, Higgs DR. 2007. Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *EMBO J* **26**: 2041–2051.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci USA* **103**: 3220–3225.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**: 854–858.

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser-

-a database of tissue-specific human enhancers. *Nucleic Acids Res* **35**: D88–92.

Visel A, Taher L, Girgis H, May D, Golonzhka O, Hoch RV, McKinsey GL, Pattabiraman K, Silberberg SN, Blow MJ, et al. 2013. A high-resolution enhancer atlas of the developing telencephalon. *Cell* **152**: 895–908.

Wang D, Chang PS, Wang Z, Sutherland L, Richardson JA, Small E, Krieg PA, Olson EN. 2001. Activation of cardiac gene expression by myocardin, a transcriptional cofactor for serum response factor. *Cell* **105**: 851–862.

Weischenfeldt J, Symmons O, Spitz F, Korbel JO. 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* **14**: 125–138.

Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T, et al. 2008. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**: 796–801.

Williamson I, Hill RE, Bickmore WA. 2011. Enhancers: from developmental genetics to the genetics of common human disease. *Dev Cell* **21**: 17–19.

Zuniga A, Michos O, Spitz F, Haramis A-PG, Panman L, Galli A, Vintersten K, Klasen C, Mansfield W, Kuc S, et al. 2004. Mouse limb deformity mutations disrupt a global control region within the large regulatory landscape required for Gremlin expression. *Genes Dev* **18**: 1553–1564.

## FIGURE LEGENDS

**Figure 1 – Mapping the distribution of regulatory activities along the genome.**

**(A).** Insertion of a regulatory sensor (drawing of sensor) at different distances from an enhancer (blue oval) and different positions relative to its target gene (blue arrow) can report on the domains of action of the enhancer. **(B)**. 734 insertions were characterized for expression in mid-gestation mouse embryos. About 55% of these reported regulatory activities. Insertions with (blue) or without (white) expression were broadly distributed around endogenous gene transcriptional start sites (distribution made with GREAT (McLean et al. 2010)). **(C).** Examples of the diverse expression patterns obtained in E11.5 embryonic forelimb (left) or forebrain (right). Numbers refer to insertion identifiers used in TRACER database (Chen et al. 2013).

**Figure 2 – Expression of the regulator sensor is correlated with surrounding enhancers, up to large distances.**

**(A).** Enrichment of insertions showing LacZ activity in a given tissue relative to limb EP300 binding sites. Enrichment for insertions with expression in the limb (green) compared to random insertions is calculated at increasing distance from the nearest EP300 site (x-axis). Error bars represent one standard deviation from the mean. Enrichments of insertions with activity in other tissues but not in the limb (heart: purple; forebrain: blue; midbrain: red) or with no LacZ activity (grey) are also displayed. Results for EP300 sites from other tissues are shown in Supplementary Figure 1. **(B).** Comparison of enhancer and sensor activity. Different groups were considered, according to the relative distance between the insertions and the enhancers (number of insertion-enhancer pairs indicated above each bar). The two

random datasets are described in the Material and Methods section. The proportions of concordant enhancer-insertions pairs in different groups were compared using Fisher's exact test. **(C-E)**. Examples of concordant enhancer-insertion pairs. The different loci are schematized (enhancer: blue oval with VISTA reference; sensor: drawing of transposon; endogenous genes: arrows or grey bars with black exons), with putative target genes of enhancers indicated by labelling the gene the same colour as the enhancer. Photos of representative embryos of the *in vivo* enhancer assays are from the Vista Enhancer Browser (Visel et al. 2007) **(C)**. The sensor reported the activity of an intronic diencephalon/midbrain enhancer, which likely contributes to the regulation of the distant *Lhx2* gene (Gray et al. 2004). **(D)**. Heart-specific expression of the sensor when inserted adjacent to a heart-specific enhancer, possibly regulating the adjacent *Myocd* gene (Wang et al. 2001). The eye expression shown on the representative transgenic embryo (star) is ectopic. **(E).** The sensor, inserted next to *Znf503* (McGlinn et al. 2008), showed expression in the posterior forelimb, which overlapped with the activity of a distant enhancer (fl, blue/white arrow). The enhancer was also active in the neural tube, but the sensor was not expressed in that region.

**Figure 3 – Non-gene-centric enhancer activities are detected across large distances**

**(A).** The number of insertions that correctly (blue) or partially (light blue) reported the activity of a neighbouring tissue-specific enhancer, or showed a different activity (orange). Insertions were grouped depending on their position relative to the enhancer/target gene, as schematized below the chart. **(B-D).** Examples of expression detected with the regulatory sensor (photos) in non-gene-centred situations. Gene

(arrows) and enhancer (ovals) activities are colour-coded and shown on the embryo outline. **(B).** An insertion between the *En2* and *Cnpy1* genes matches their expression at the mid/hindbrain junction (Jukkola et al. 2006), as well as the activity of an enhancer on the far side of *En2* (see also Supplementary Figure 8) **(C).** *Barhl2* expression in the midbrain and diencephalon requires remote enhancers (Saba et al. 2005), and a diencephalon enhancer (hs612 (Visel et al. 2007)) is present upstream of this gene. Enhancer activity extends to a downstream insertion. **(D).** *Sall1* gene expression is controlled by multiple enhancers spread in the two surrounding gene deserts, and insertions flanking the gene display overlapping expression patterns.

**Figure 4 – Extended domains of co-regulation correlate with the subdivision of the genome into topological domains.**

**(A-C).** Outlines of loci, with genes displayed as arrows and insertions as drawing of transposon. Regulatory domains and transition zones are labelled, TADs (identified by Hi-C in mouse ES cells) are indicated by green and brown bars and unstructured regions by dashed lines**.** Hi-C interaction frequencies are represented as a two-dimensional heat map (from (Dixon et al. 2012)). **(A).** Multiple insertions in the chr3:7.3-8.3M interval outlined an extended regulatory domain characterized by shared expression in the facial and trunk mesenchyme, and in neural crest derivatives. This domain extends into the adjacent unstructured region (insertion 201179e9), but two telomeric insertions, located in a different TAD, showed different patterns (the proximal limb expression of 181912bc-133 is anterior, whereas insertions in the flanking RD have a more medial expression), defining two transition zones. **(B).** Multiple insertions in the vicinity of the *Foxg1* gene display the typical forebrain (fb) expression of the gene (adapted from Chen et al. 2013). Expression in the ear (*) is

due to another insertion also present in 177175-emb7. The regulatory domain defined by the gene and the insertions is contained within a single TAD. A more detailed version of this panel is shown in Supplementary Figure 5E. **(C).** Two insertions in a gene desert between *Kcnt2* and *Cdc73* are divergently expressed in the limb bud (lb) and forebrain (fb), delineating a transition zone. This coincides with the respective insertions being located in different TADs. **(D)** Size distribution (y axis) and relationship with TADs (colour-code) of functionally defined intervals. Only a single regulatory domain (RD) overlaps a TAD boundary. Random permutation of regions 200kb-1Mb in length (boxed area), where the size distribution of the different functional categories is not statistically different (Kolmogorov-Smirnov test, p=0.8560 for RD vs TZ+A+B; p=0.9240 for RD vs TZ), showed that the RDs are significantly under-represented in the "separated TADs" category. **(E).** Unlike control regions (classes A, B and transition zones (TZ)), RDs show depletion in topological boundaries compared to equally sized, randomly distributed fragments. Grey box-plots represent the results of randomization; red dots the position of the real data. The depletion is statistically significant (p=0.009), as indicated by the blue star.

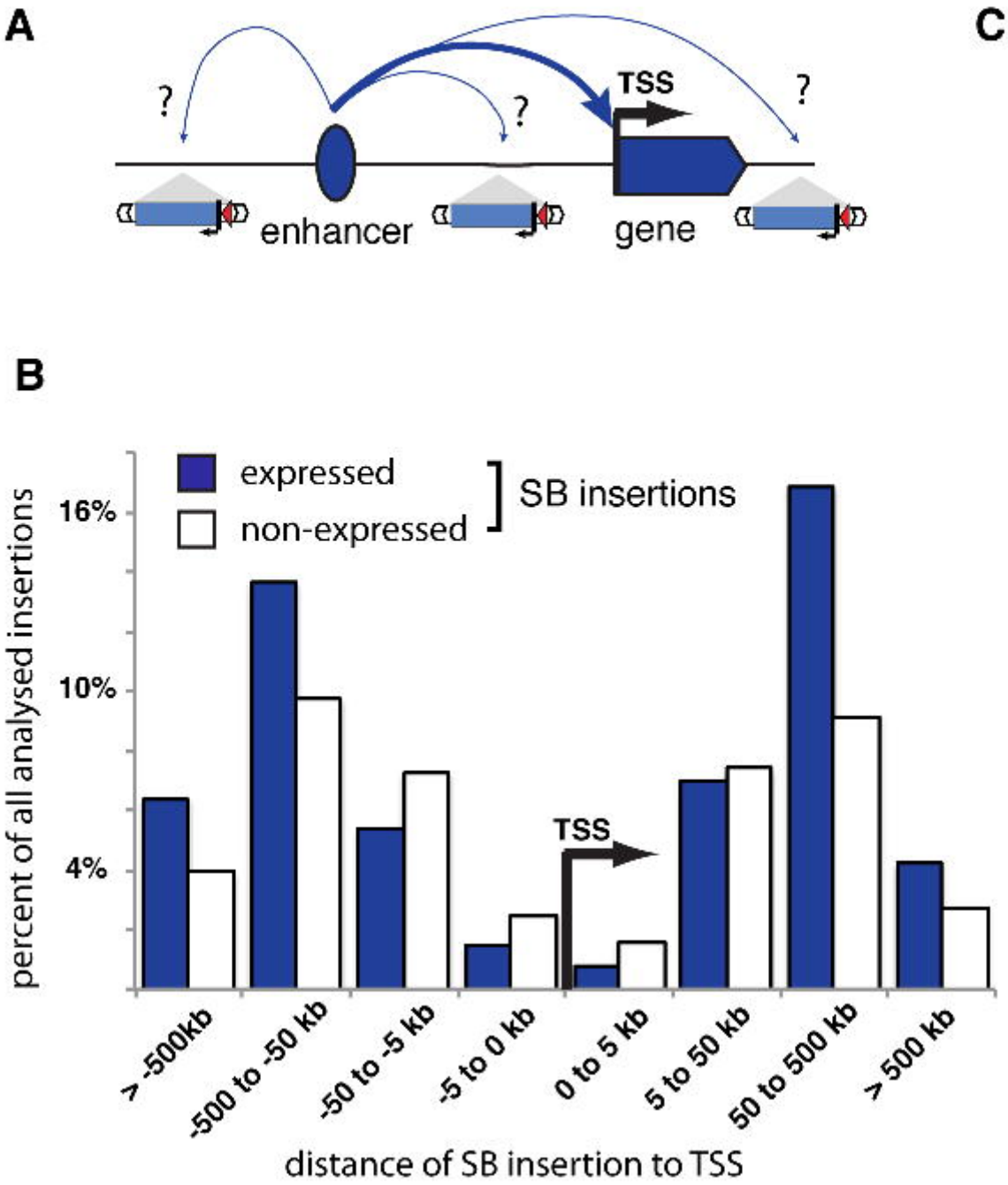
**Figure 5 – Depletion of CTCF and cohesin in regulatory domains reflects the topological segmentation of the genome.**

Tissue-invariant CTCF sites (bound in >9 tissues) are significantly depleted (but not absent) in regulatory domains when considering all functionally defined regions **(A)**. However, the statistical significance of this depletion becomes marginal if the comparison is limited to intervals included in TADs, and not compared against the overall genome **(B)**. Constitutive cohesin complex-binding sites (not shown), or

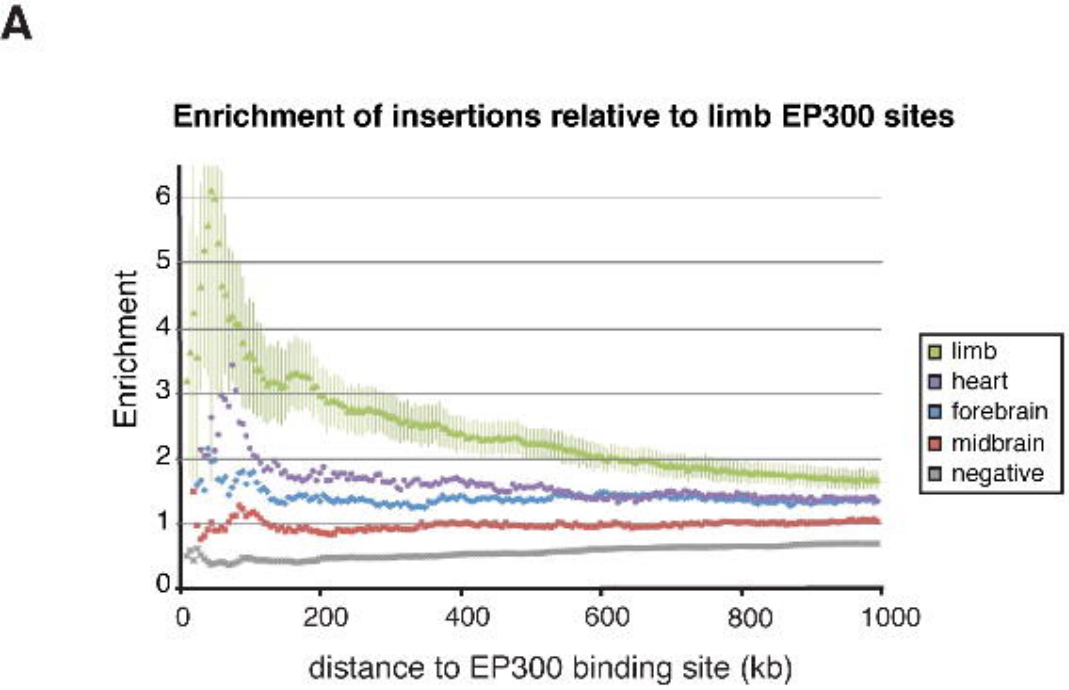
CTCF/cohesin co-occupied regions **(C)** showed also only a slightly reduced density in RDs, when compared to the part of the genome that is included in TADs.

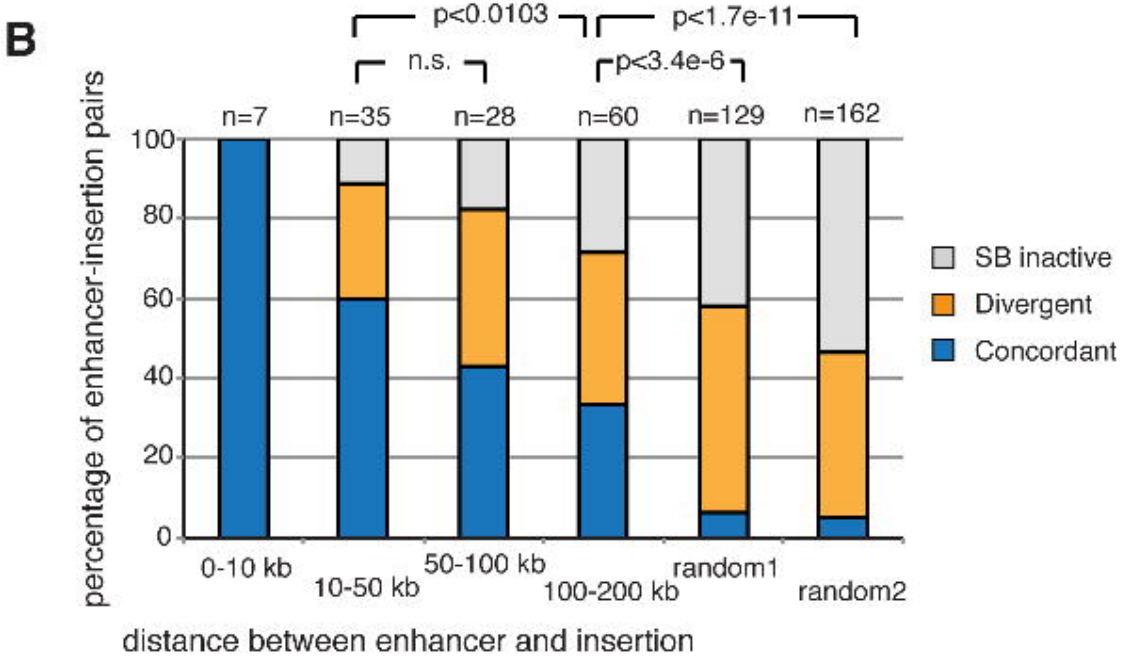
**Figure 6 – CTCF and cohesin sites are interspersed in regulatory domains.** Schematic representation of the large topological/regulatory domain on chr7:75.5M-77.8M. The two genes (*Arrdc4* and *Nr2f2*) are represented as arrows. The corresponding TAD is represented by a two-dimensional heat map (Dixon et al. 2012). Several constitutive CTCF sites (red lollipop, colour intensity proportional to cell-invariance), largely co-bound by cohesin (purple rings), are interspersed in this interval. Insertions spread across almost 2 Mb showed highly overlapping patterns in the proximal limb (blue arrow, upper panel), face (blue arrow, middle panel) and at the midbrain/diencephalon boundary (blue arrow, lower panel), forming a large regulatory domain. This large domain can be subdivided into smaller tissue-specific landscapes (green, purple and brown) based on expression patterns displayed by only a subset of the insertions and quantitative differences in LacZ staining intensity. These different regulatory influences overlap with *Nr2f2* expression, detected by whole-mount *in situ* hybridisation. *In situ* hybridisation with *Arrdc4* probes did not reveal specific expression in E11.5 embryos. Embryos 183036-emb4 and 176069-emb50 were described previously (Ruf et al. 2011).

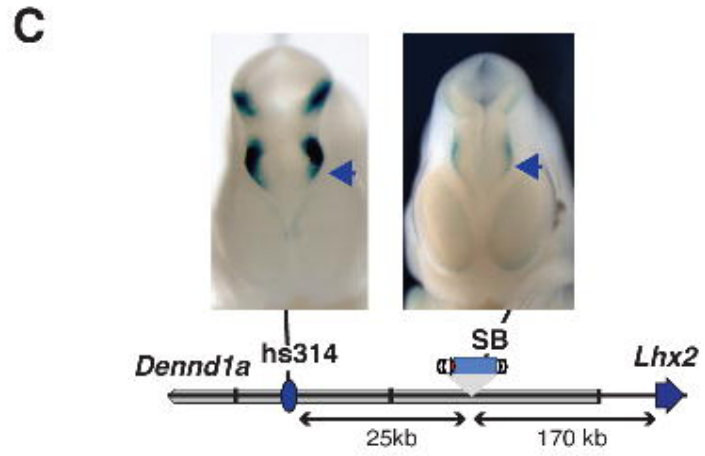
# Figure 1

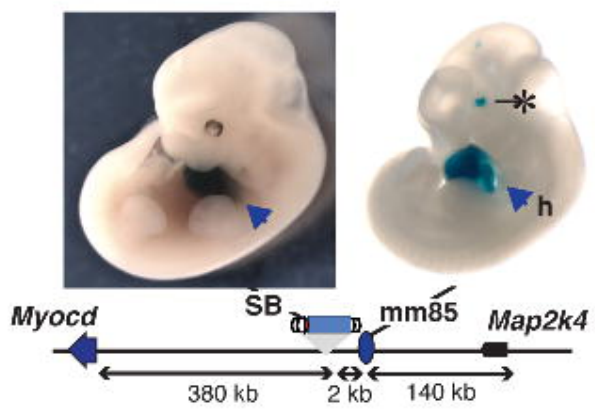

A
?
?
?
TSS
enhancer
gene
B
expressed
non-expressed
SB insertions
16%
10%
4%
percent of all analysed insertions
TSS
> -500kb
-500 to -50 kb
-50 to -5 kb
-5 to 0 kb
0 to 5 kb
5 to 50 kb
50 to 500 kb
> 500 kb
distance of SB insertion to TSS
C
183036-14
179860b
194640-7
202157
201763
204198
196247
179868d

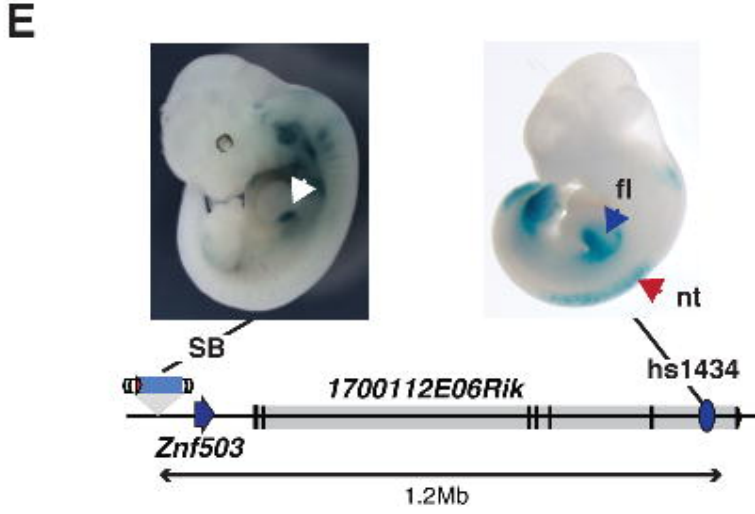
# Figure 2


A
Enrichment of insertions relative to limb EP300 sites
Enrichment
0
1
2
3
4
5
6
0
200
400
600
800
1000
distance to EP300 binding site (kb)
limb
heart
forebrain
midbrain
negative

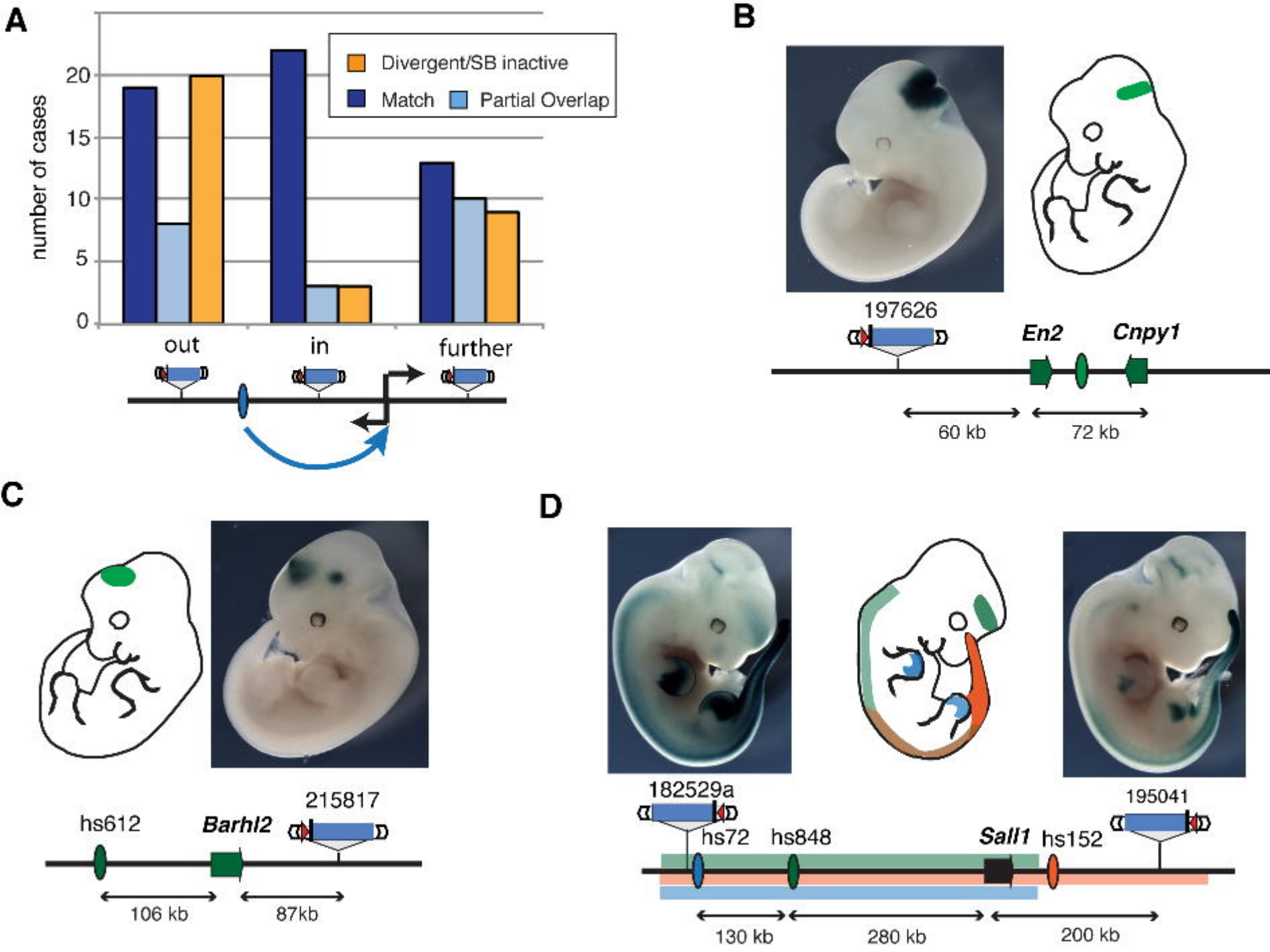

B
p<0.0103
p<1.7e-11
n.s.
p<3.4e-6
n=7
n=35
n=28
n=60
n=129
n=162
percentage of enhancer-insertion pairs
0
20
40
60
80
100
0-10 kb
10-50 kb
50-100 kb
100-200 kb
random1
random2
distance between enhancer and insertion
SB inactive
Divergent
Concordant


C
Dennd1a
hs314
SB
Lhx2
25kb
170 kb

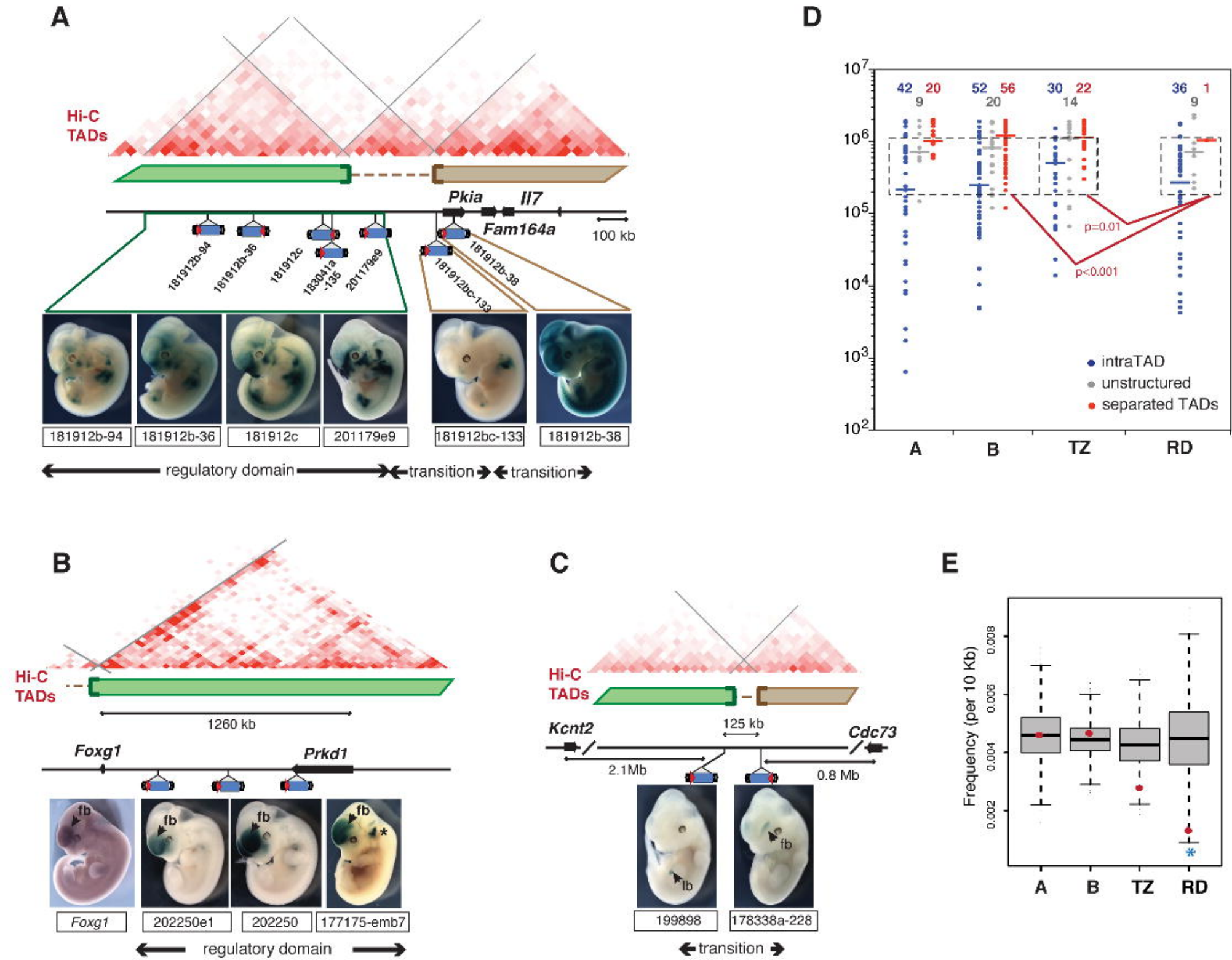

D
h
Myocd
SB
mm85
Map2k4
380 kb
2 kb
140 kb


E
fl
nt
SB
hs1434
1700112E06Rik
Znf503
1.2Mb

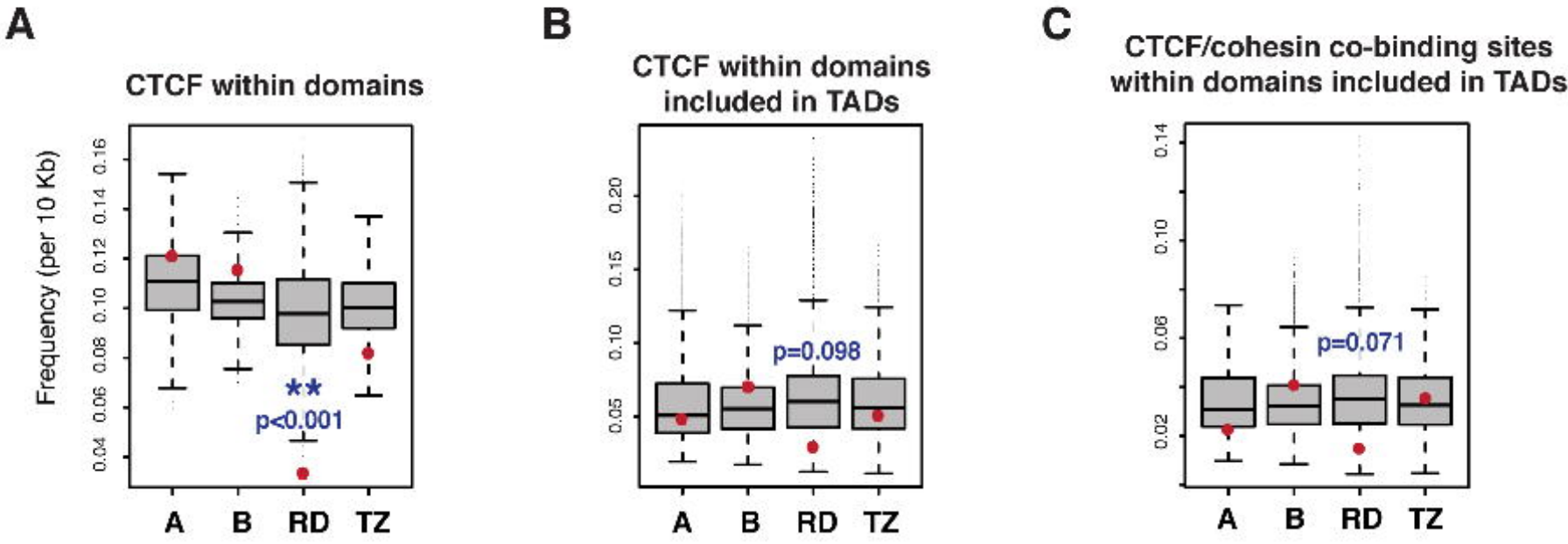
# Figure 3


A
number of cases
0
5
10
15
20
Divergent/SB inactive
Match
Partial Overlap
out
in
further
B
197626
En2
Cnpy1
60 kb
72 kb
C
hs612
Barhl2
215817
106 kb
87kb
D
182529a
hs72
hs848
Sall1
hs152
195041
130 kb
280 kb
200 kb

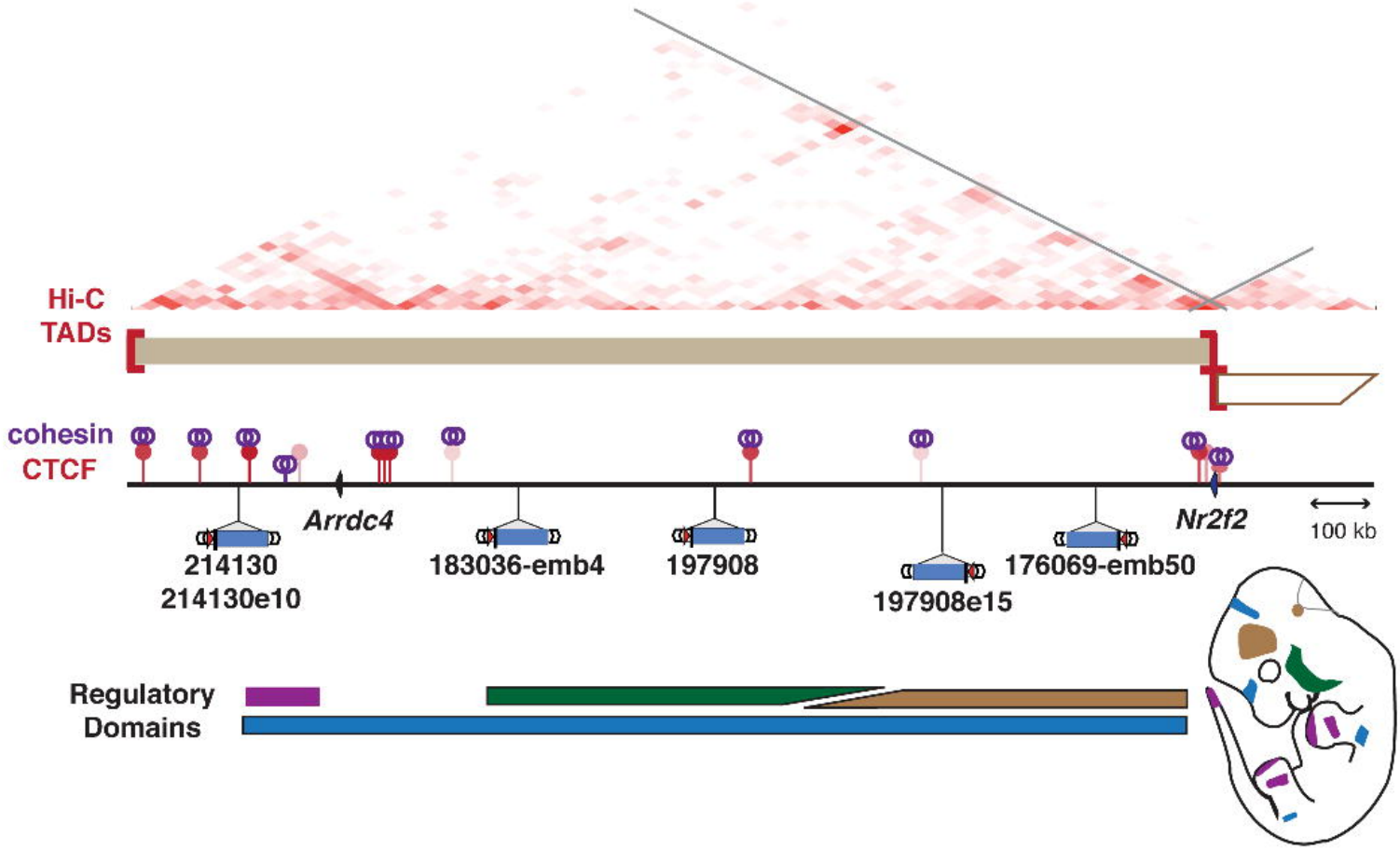
# Figure 4

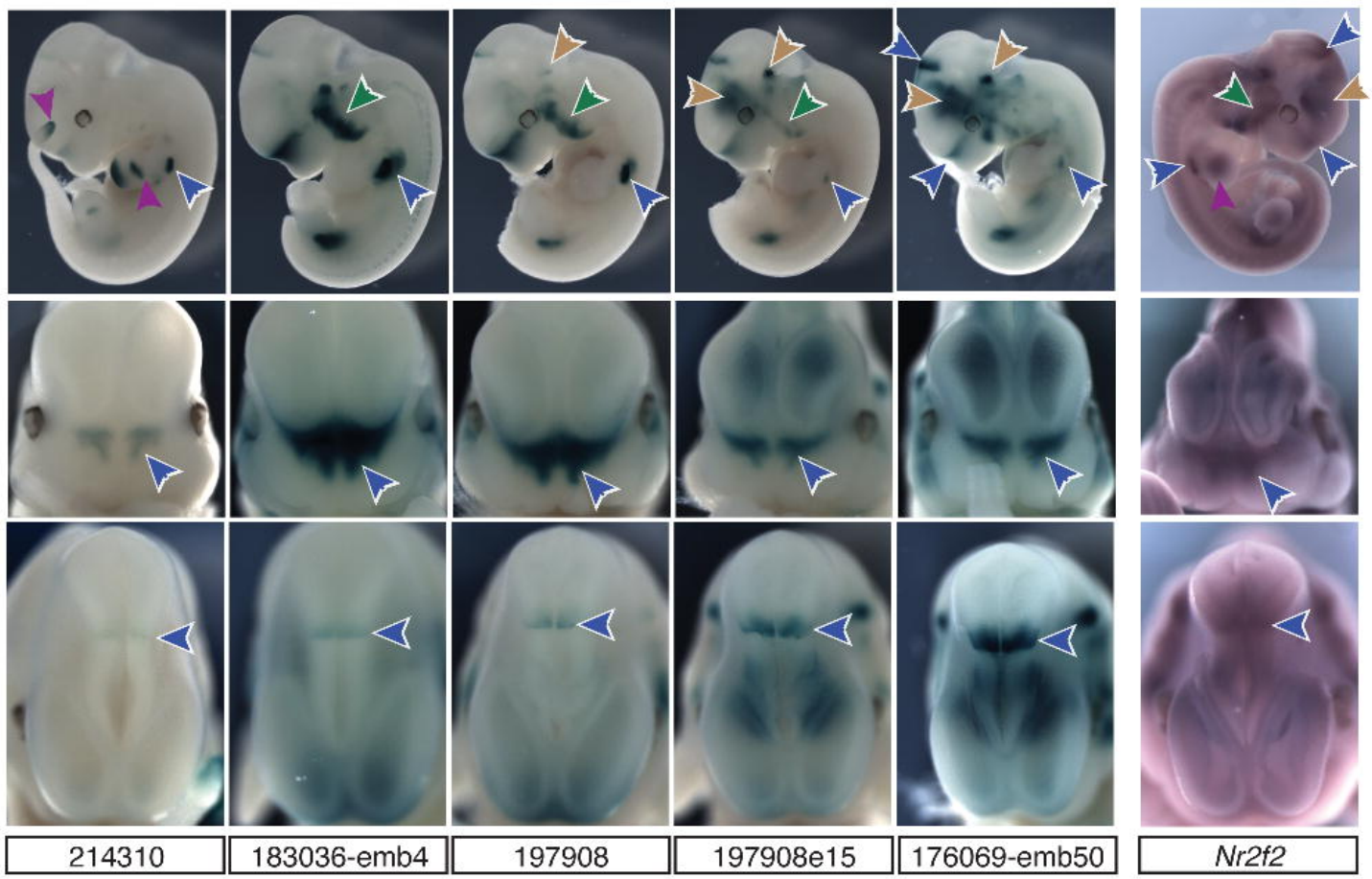

A
Hi-C
TADs
Pkia
Il7
Fam164a
100 kb
181912b-94
181912b-36
181912c
183041a
-135
201179e9
181912bc-133
181912b-38
181912b-94
181912b-36
181912c
201179e9
181912bc-133
181912b-38
regulatory domain
transition
transition
D
42 20
9
52 56
20
30 22
14
36 1
9
p=0.01
p<0.001
intraTAD
unstructured
separated TADs
A
B
TZ
RD
B
Hi-C
TADs
1260 kb
Foxg1
Prkd1
fb
Foxg1
202250e1
202250
177175-emb7
regulatory domain
C
Hi-C
TADs
Kcnt2
125 kb
Cdc73
2.1Mb
0.8 Mb
fb
199898
178338a-228
transition
E
Frequency (per 10 Kb)
0.002
0.004
0.006
0.008
A
B
TZ
RD

# Figure 5


A
CTCF within domains
Frequency (per 10 Kb)
0.04
0.06
0.08
0.10
0.12
0.14
0.16
**
p<0.001
A
B
RD
TZ
B
CTCF within domains
included in TADs
0.05
0.10
0.15
0.20
p=0.098
A
B
RD
TZ
C
CTCF/cohesin co-binding sites
within domains included in TADs
0.02
0.06
0.10
0.14
p=0.071
A
B
RD
TZ

# Figure 6


Hi-C
TADs
cohesin
CTCF
Arrdc4
Nr2f2
100 kb
214130
214130e10
183036-emb4
197908
197908e15
176069-emb50
Regulatory
Domains


214310
183036-emb4
197908
197908e15
176069-emb50
Nr2f2