SUPPLEMENTARY INFORMATION

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

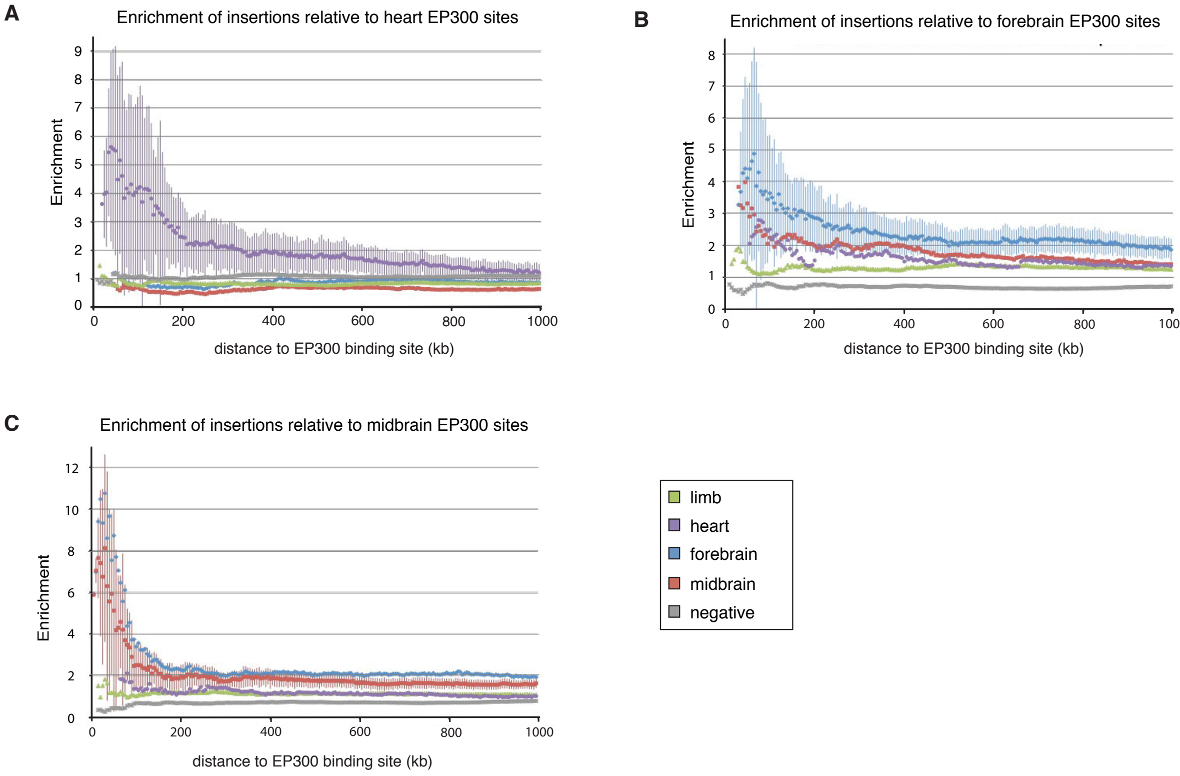
[#] Present address : New England Biolabs - Ipswich – MA - United States

Supplementary Figures 1-8

Supplementary Table 6

Supplementary Note - Methods

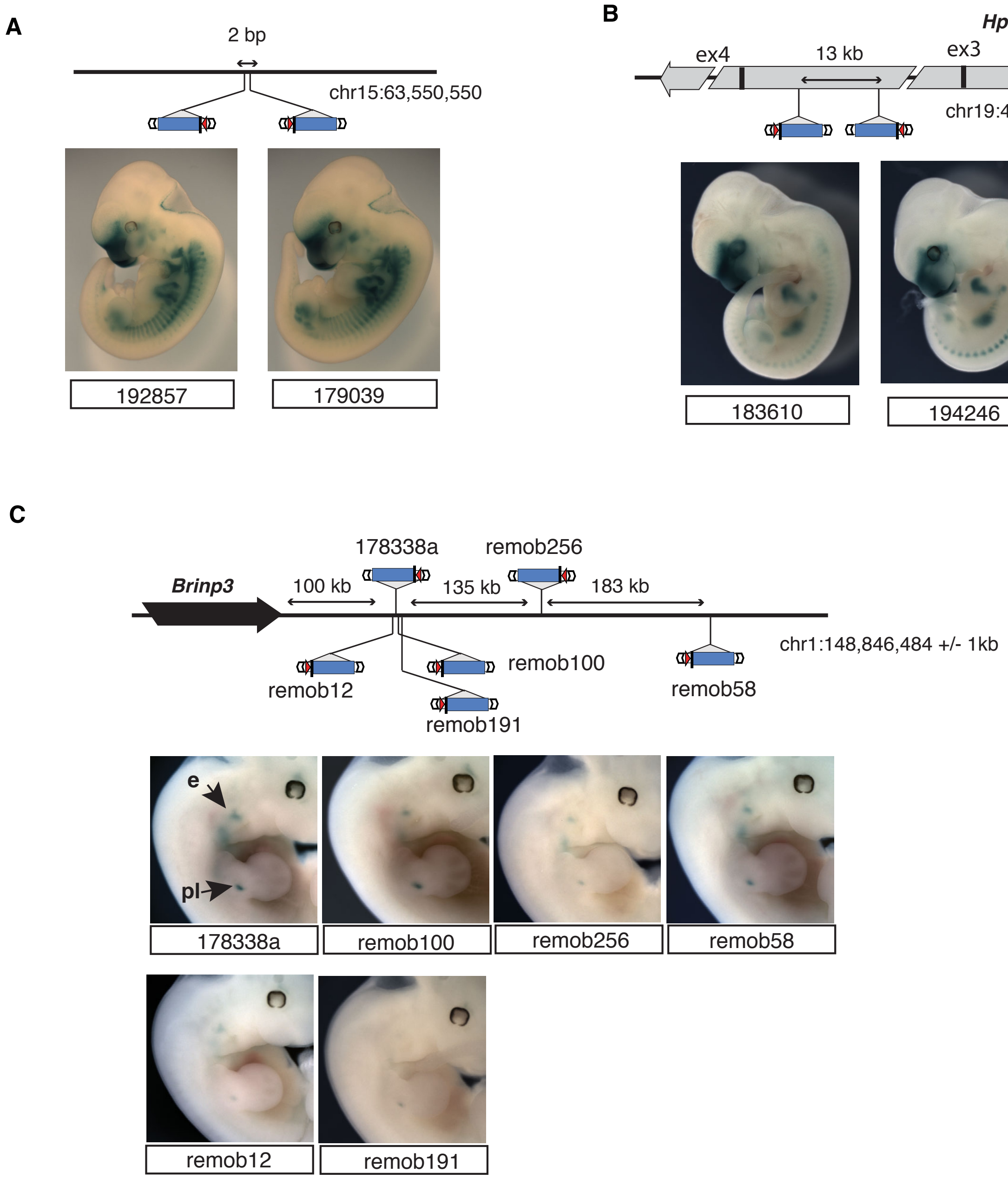
Supplementary Figure 1



Enrichment of insertions in the proximity of EP300 sites bound in the same tissue.

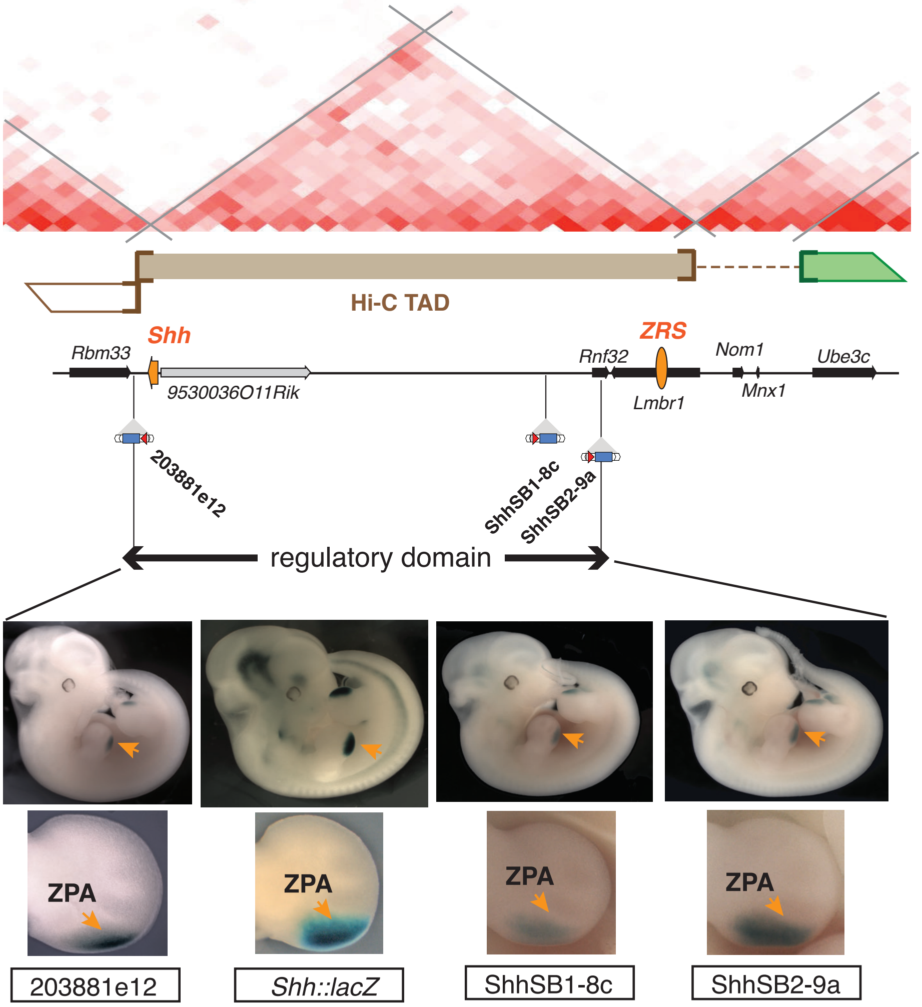
Enrichment of insertions with tissue-specific LacZ activity (compared to random insertions), at increasing distance (x-axis) from EP300 sites detected in heart (**A**), forebrain (**B**) and midbrain (**C**). Error bars represent one standard deviation from the mean. Colours indicate the tissue in which insertions were expressed (limb: green; heart: purple; forebrain: blue; midbrain: red; no LacZ activity: grey). EP300 data is taken from (Blow et al. 2010) and (Visel et al. 2009). The overlap and proximity between EP300 sites (Visel et al. 2009) may account for the enrichment of forebrain activity around midbrain EP300 sites, and vice versa.

Supplementary Figure 2



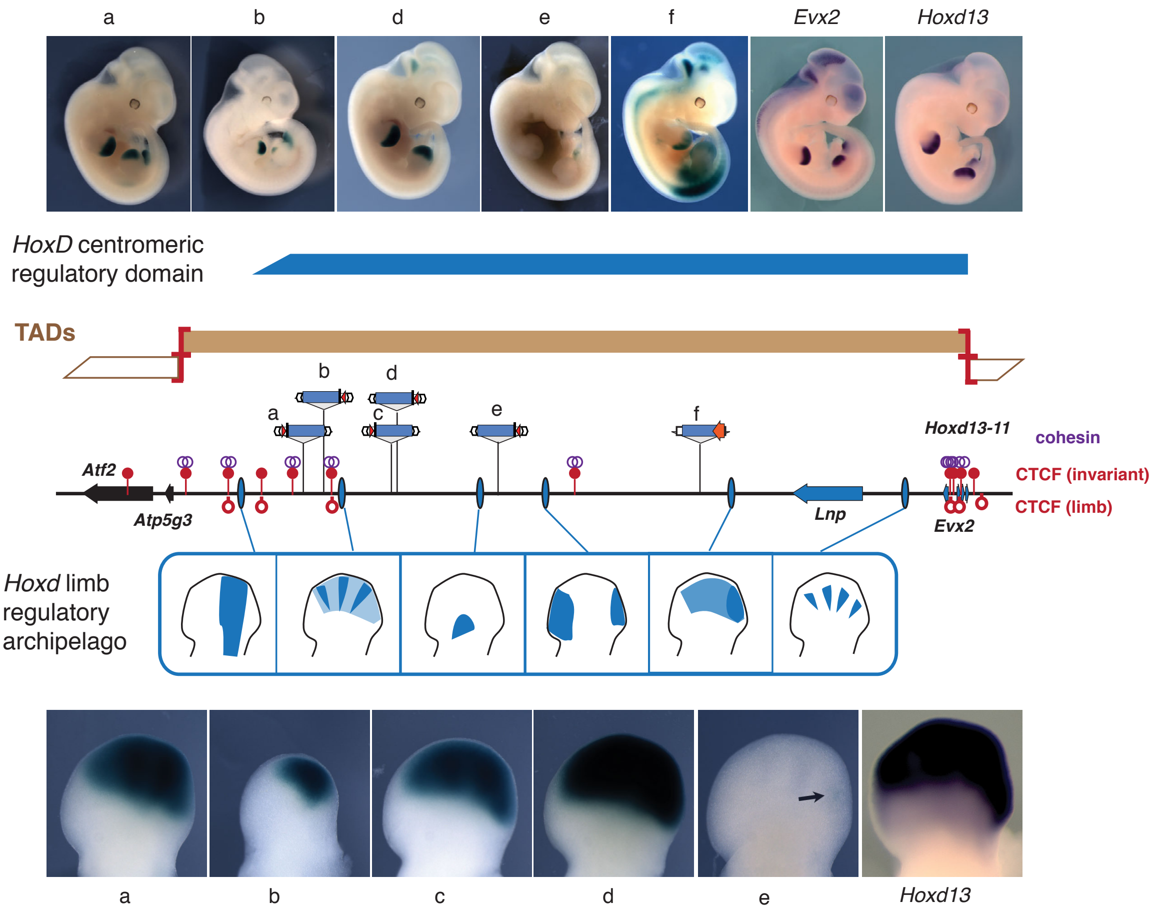
Orientation-independent activity of the regulatory sensor.
Precisely overlapping *lacZ* expression patterns obtained in E11.5 mouse embryos with adjacent insertions, indicating that the orientation of the insertion relative to the activating enhancer does not influence enhancer-sensor interactions. **(A)** Two insertions separated by 2 bp in a gene desert on mouse chromosome 15. 192857 was produced by remobilization from 179039, and inserted at the same position, into the duplicated TA scar. **(B)** Two independent insertions into an intron of *Hspe2*, on mouse chromosome 19. **(C)** Multiple insertions in a 300 kb region downstream of *Brinp3*, on mouse chromosome 1, inserted in different orientations and separated by distances from 6 bp to 183 kb. In each case, LacZ staining was detected in the developing ear (e) and proximal limb bud (pl). Further examples of orientation-independent activation are also shown in Supplementary Figures 4 and 8.

Supplementary Figure 3.



Long-range detection of a single limb enhancer at the *Shh* locus.
Expression of *Shh* (indicated with a *Shh::lacZ* allele, kindly provided by Andreas Kottmann (Gonzalez-Reyes et al. 2012)) in the zone of polarizing activity (ZPA, orange arrow) in the limb bud has been shown to arise from a single enhancer, the ZRS (orange oval), as indicated by loss-of-function experiments and enhancer-scanning of the entire gene desert (Jeong et al. 2006; Sagai et al. 2005). Insertions into this gene desert, as well as an insertion 30kb downstream of the gene recapitulate multiple aspects of the *Shh* expression pattern, including expression in the limb bud (orange arrow). This indicates that the ZRS – the only enhancer with limb activity in this region – can activate the regulatory sensor at long-distance. The large regulatory domain defined by these insertions largely overlaps with a topological domain defined by Hi-C, although the insertion downstream of *Shh* (203881e12) falls within the TAD boundary. A schematized version of the locus is shown, with genes indicated by arrows (*Shh* in orange, protein-coding genes in black, non-coding gene in grey). Hi-C data (Dixon et al. 2012) is represented by two-dimensional heat maps, and TADs are shown by white, brown and green bars, with the unstructured region indicated by a dashed line.

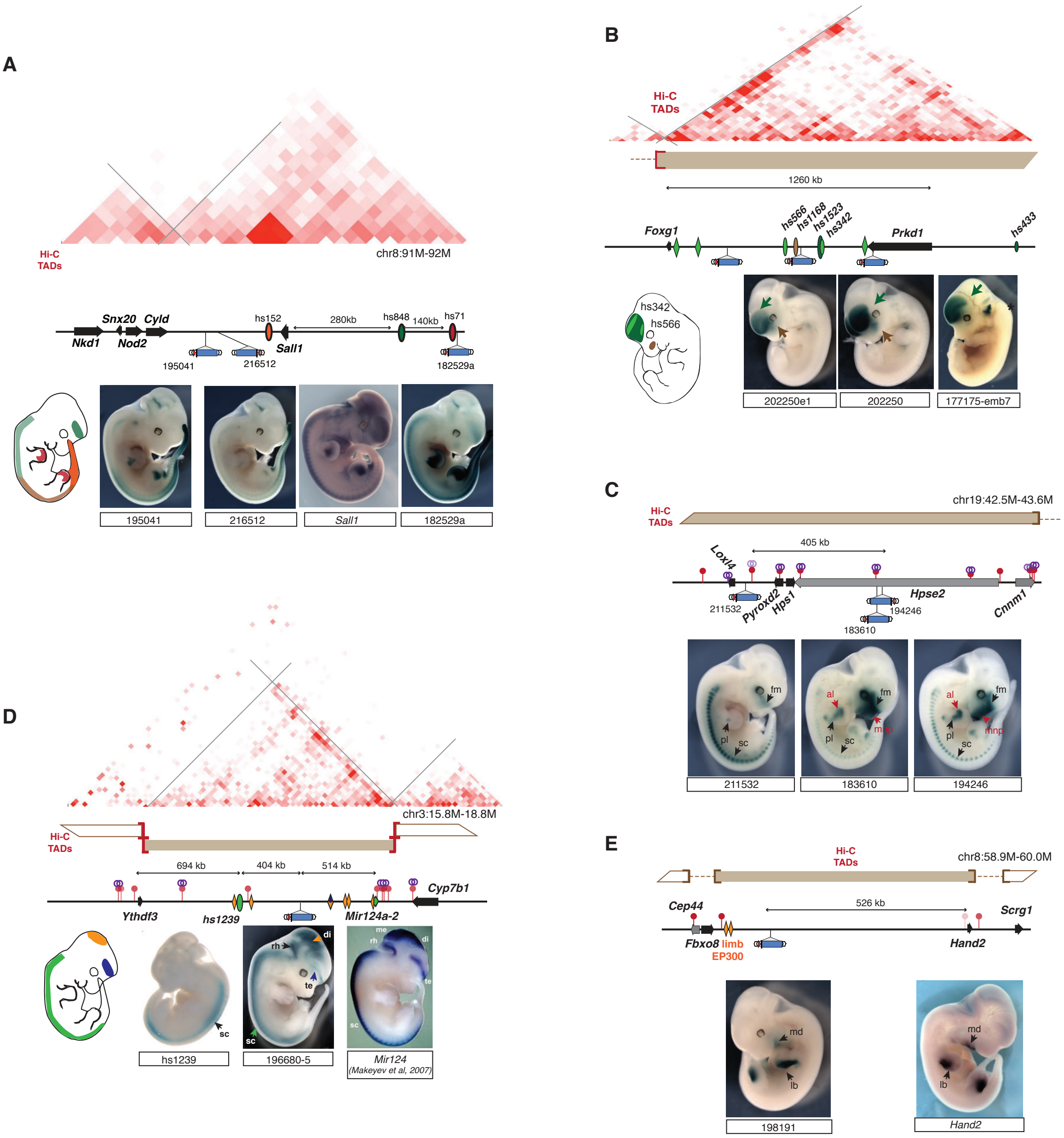
Supplementary Figure 4.



The regulatory sensor integrates the input from several non-redundant enhancers at multiple positions of the *Hoxd* regulatory domain.

Multiple enhancers with diverse spatially restricted activities (blue ovals, patterns of activity represented on a limb outline) have been identified in a large gene-poor region centromeric to the *Hoxd* cluster (Spitz et al. 2003; Gonzalez et al. 2007; Montavon et al. 2011), forming a *regulatory archipelago*. Multiple insertions have been obtain within this interval (a-e: SB sensor, this work and (Ruf et al. 2011); f: *Hoxd9*lac transgene, rel5 (Montavon et al. 2011)). Expression patterns in E11.5 (E11 for b) embryos are shown, including details of the forelimb (anterior: left; posterior: right). At all insertions (except e, which showed only extremely weak expression in the posterior limb, arrow), the reporter gene showed the typical expression pattern resulting from the integrated ouput of the archipelago, and not only the specific activity of the nearby enhancer element. CTCF (red lollipops = cell-invariant sites detected by ChIP, red lollipops with white center = ChIP-chip data from embryonic limbs (Soshnikova et al. 2010)) and cohesin-complex (purple rings) (including embryonic limb data from (DeMare et al. 2013)) are interspersed within the relatively homogeneous regulatory domain. The regulatory domain (extended here to include the endogenous genes) and the corresponding TAD are shown as colored bars.

Supplementary Figure 5

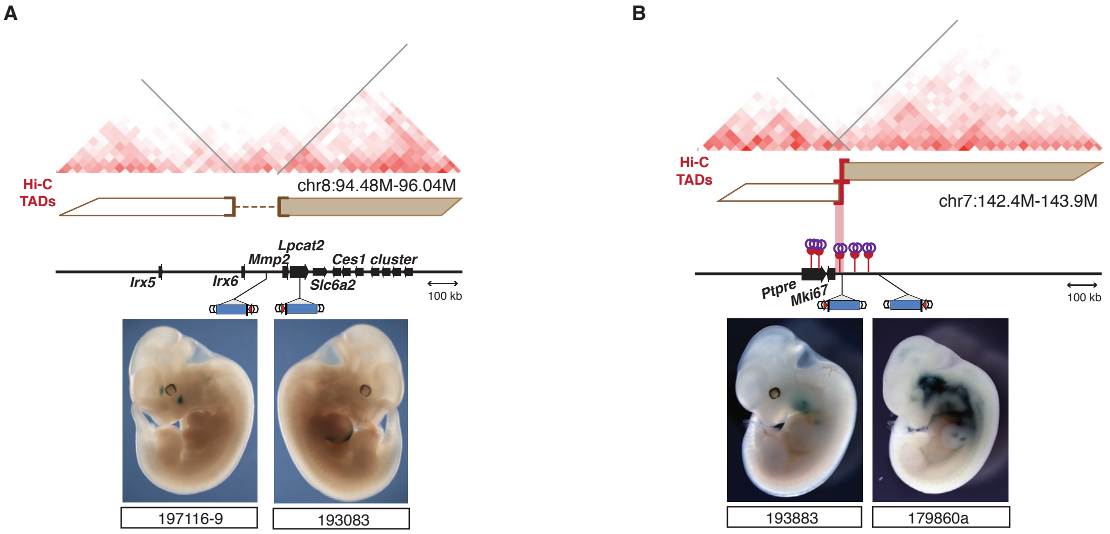


Overlap between the functional and topological subdivision of the genome

Extended regions of co-expression, outlined by insertions of the regulatory sensor (defining a regulatory domain, **A-C**) or by an endogenous gene and flanking insertions (**D-E**) are contained within a single topological domain. The different loci (coordinates given are based on NCBIM37/mm9 mouse genome assembly) are represented schematically, with the extent of the RD and distances between genes (black boxes) and insertions indicated. The constitutive topological domains (Hi-C TADs as defined by (Dixon et al. 2012) (data shown from mouse ES cells) are represented as two-dimensional heat maps and as brown and white coloured bars, with unstructured regions shown as dashed lines. CTCF and cohesin binding sites are represented by red lollipops and purple circles, respectively, with the intensity indicating degree of tissue-invariance. Enhancers (ovals) and EP300-bound regions (diamonds) are indicated. Outlines of embryos show corresponding colors in the regions where enhancer were found active (data from (Visel et al. 2007)), or from which the EP300-

ChIP data was generated (Visel et al. 2009; Blow et al. 2010). **(A).** Multiple insertions in the gene desert surrounding *Sall1* define a RD based on expression at the midbrain-hindbrain boundary, in the posterior limb bud and the tail. The RD is contained within a single TAD. **(B).** Insertions in the large gene-desert flanking *Foxg1* (as shown in Figure 4B-adapted from Chen et al. 2013) define a large RD that is included in a TAD containing several forebrain EP300 sites and confirmed enhancers (Visel et al. 2013), the activities of which correspond to the ones detected by the sensor (dark green=active broadly in the forebrain; light green=restricted activity to sub-regions of the forebrain; brown=cranio-facial) **(C).** Expression in the facial mesenchyme (fm), proximal limb bud (pl) and sclerotome of three insertions on chromosome 19 outlines a RD, which is included in a single TAD. Within the RD, the two insertions in an intron of *Hpse2* (183610 and 194247) show additional expression in the anterior limb bud (al) and the medial nasal process (mnp), indicating the possible presence of substructures with the RD. **(D).** The miRNA gene *Mir124-2* is located at the boundary of a gene desert. The entire gene desert is part of a single TAD, and a distant insertion into this gene desert captured the expression domain of the miRNA gene (*in situ* panel reprinted from Molecular Cell, Makeyev *et al.* 2007, with permission from Elsevier) in the spinal cord (sc), rhombencephalon (rc), diencephalon (dl) and telencephalon (te). One putative enhancer with expression in the spinal cord, also located in the gene desert, has been documented in the VISTA enhancer browser. **(E).** The 198191 insertion half a megabase away from the *Hand2* gene shares expression with the gene in the posterior limb bud (lb) and the mandibulary process (md), indicating a possible co-regulation. The interval is contained within a single TAD.

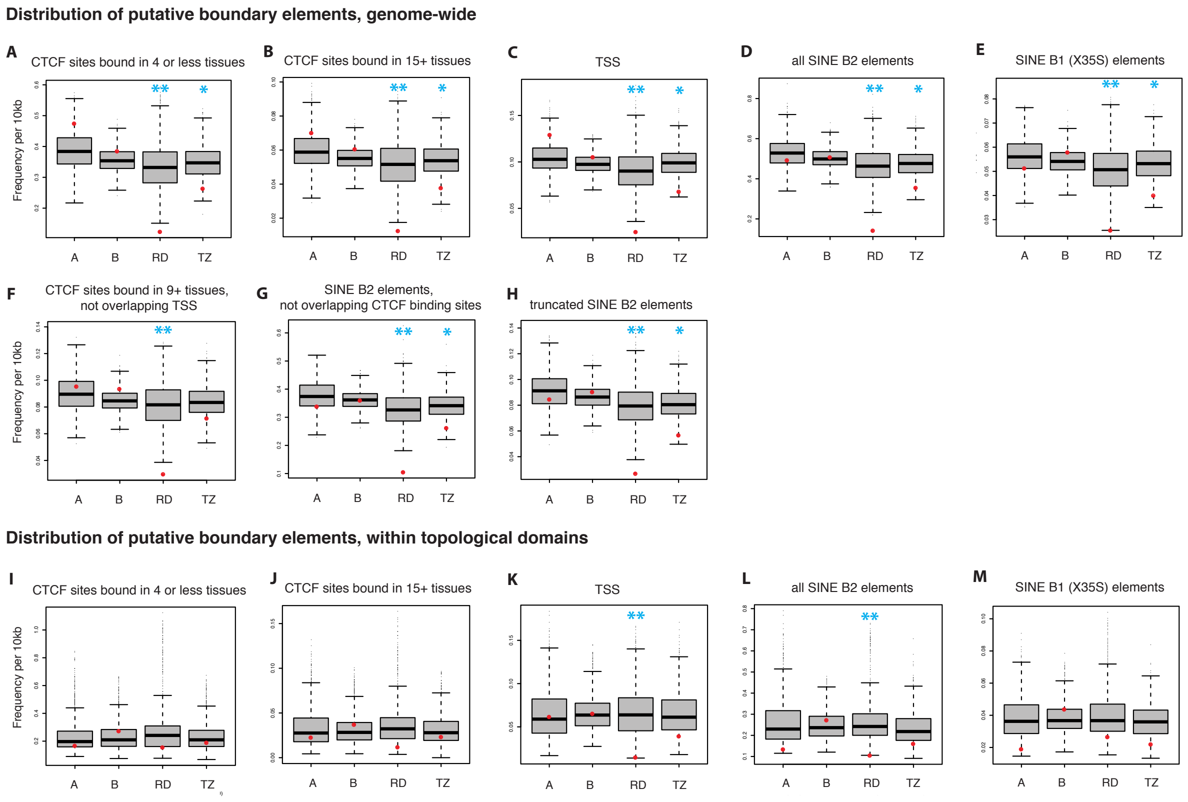
Supplementary Figure 6.



Overlap between functional transition zones and topological boundaries

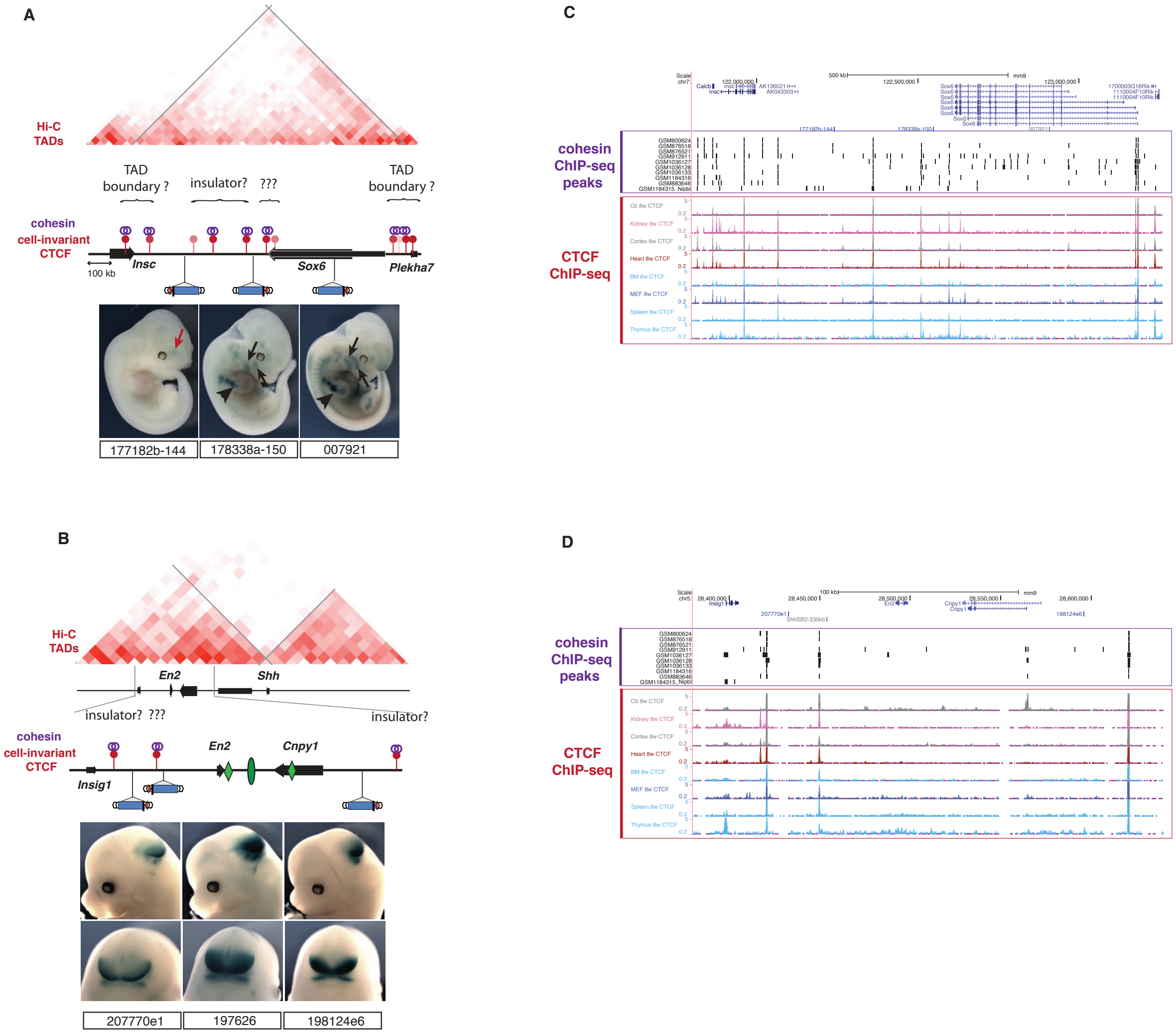
Insertions bounding a transition zone are separated by a topological boundary. The different loci (coordinates given are based on NCBIM37/mm9 mouse genome assembly) are represented schematically. The constitutive topological domains (Hi-C TADs as defined by (Dixon et al. 2012), data shown from mouse ES cells) are represented as two-dimensional heat maps and as brown and white coloured bars, with unstructured regions shown as dashed lines. CTCF and cohesin binding sites are represented by red lollipops and purple circles, respectively **(A).** Two insertions downstream of *Irx6* show strikingly different expression, outlining a TZ. The TZ overlaps a TAD boundary. **(B)** A TZ defined by two insertions in a gene desert upstream of *Mki67* overlaps the boundary between two adjacent TADs, as well as the presence of multiple tissue-invariant CTCF and cohesin binding sites.

Supplementary Figure 7



Distribution of boundary elements relative to operationally defined domains. Tissue-specific (bound in 4 or less tissues; (**A**)) and highly tissue-invariant (bound in 15 or more tissues; **(B)**) CTCF sites**,** as well as transcriptional start sites (TSS; (**C**)), SINE B2 **(D)** and a subclass of SINEB1 elements (X35S; **(E)**) are depleted within regulatory domains ("RD") when compared to a randomized dataset. The same features are less or not depleted in transition zones ("TZ") and control regions (class "A" and "B").. This specific depletion was also observed when we considered only CTCF sites not overlapping TSS **(F)**, or SINE B2 elements not overlapping CTCF binding sites **(G)** or truncated SINE B2 elements, which contain neither CTCF, nor PolII-PolIII binding sites, which have been suggested to contribute to SINE B2 boundary activity (Lunyak et al. 2007) **(H)**. When considering only intra-TAD regions TSS **(K)** and SINE B2 elements **(L)** maintained depletion. Tissue-specific CTCF sites **(I)**, and X35S elements were also no longer depleted when comparing only intraTAD regions, whereas highly tissue-invariant CTCF sites show lower density, albeit with barely statistical significance (p=0.086), in RDs **(J)**, **(M)**. For all panels the frequency distribution of the randomized dataset is shown by a grey box-plot, while the true observed measurement is represented by a red dot. Cases where the difference between the real observation and the randomization is statistically significant (P<0.05) are indicated by a single blue asterisk, and cases where it is highly statistically significant (P<0.01) is indicated by two asterisks.

Supplementary Figure 8



Constitutive CTCF sites within extended regulatory domains.

(A). Schematic representation of a large TAD extending from *Insc* to *Plekha7* and encompassing the *Sox6* gene. Insertions showed both shared expression in the developing limb and facial chondrocytes (black arrowheads and arrows, respectively), and specific ones (e.g. forebrain, red arrow). Constitutive CTCF sites (dark red lollipops) or more cell-type variable ones (light red lollipops) are dispersed throughout the region. Constitutive CTCF sites are largely co-bound by cohesin (purple circles). Whereas some of these CTCF-cohesin sites may correlate with topological or regulatory boundaries, the role and function of others is more elusive. **(B).** An extended regulatory domain encompasses the *En2-Cnpy1* genes, with multiple insertions showing LacZ staining at the midbrain-hindbrain boundary in E13 embryos (expression is also seen in E11.5 embryos – see Figure 3). This expression matches the activity of a documented enhancer (green oval), and several additional EP300 binding sites in the midbrain have been identified in this region (green diamonds). Three constitutive CTCF sites (red lollipops) that overlap tissue-invariant cohesin binding sites are found interspersed in this region. **(C-D)** Screenshots of the *Insc-Plekha7* and *En2-Cnpy1* regions in the UCSC browser, showing extensive overlaps between cell-invariant cohesin and CTCF peaks. CTCF and cohesin data assembled from published resources (Supplementary Table 6). Heat maps for Hi-C data taken from (Dixon et al. 2012).

Supplementary Tables

Table S1. List of all transposon insertions used for large-scale comparative analysis. Positions are based on the NCBI37/mm9 mouse genome assembly. Details available on the TRACER database (Chen et al. 2013).

Table S2. List of insertions (after clustering of neighboring ones) for expression analysis, annotated for activity in the limb, midbrain, forebrain and heart (Y=expression detected) in E11-12 embryos.

Table S3. Neighbouring enhancer-transposon insertions pairs (within 200kb), with comparison of their activity/expression patterns. Enhancer names and references (PMID) are given. Two additional tables list the results of the randomisations performed on the data with the corresponding annotations of the pairs

Table S4. List of the enhancer-gene-transposon insertions triplets. Relative positions were calculated by taking the enhancer as a reference (position 1 = middle of the enhancer fragment), and adapting orientations to have the target gene TSS located at a "plus" position. Loci where the enhancers (can) regulate multiple target genes are indicated in comments.

Table S5. List of regulatory domains and transitions zones, and other intervals with regulatory information. The IDs and position of the insertions determining the ends of the intervals are given. Their type is annotated as described in the Experimental Methods. Their position with respect to overlapping topological structures was defined using data from (Dixon et al. 2012): "intra-TADs" labels interval that are contained within a single TAD; "interrupted" is for intervals which span two or more adjacent TADs; intervals that are extend from a TAD into an adjacent unstructured domain are labelled as "unstructured".

Table S6. – Source of the datasets used for comparative analysis.

Type of Data	Reference	Web / details
SB sensor expression data in mouse embryos	(Chen et al. 2013)	http://tracerdatabase.embl.de or http://www.ebi.ac.uk/panda-srv/tracer/index.php
Hi-C, mouse (ES, cortex)	(Dixon et al. 2012)	http://chromosome.sdsc.edu/mouse/hi-c/download.html
p300 ChIP peaks [1]	(Blow et al. 2010)	
CTCF ChIP peaks [2]	(Shen et al. 2012)	http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeLicrTfbs/
Cohesin [3] ChIP peaks	(Remeseiro et al. 2012) (Seitan et al. 2013) (DeMare et al. 2013) (Phillips-Cremins et al. 2013) (The ENCODE Project Consortium et al. 2012)	GSM800624: SA1; MEFs GSM876521: SMC3; MEFs GSM876518: SMC1; MEFs GSM1184316: Rad21; thymocytes GSM1184315: Nipbl; thymocytes GSM1036133: SMC1_cortex_e14.5 GSM1036128: SMC1_limb_e11.5 GSM1036127: SMC1_limb_e11.5 GSM883646: SMC1; NPC GSM912935: Rad21; MEL GSM912911: Rad21; CH12
TSS	UCSC genome browser	http://genome.ucsc.edu/index.html Mm9; from RefSeq Track
SINEB2 [4]	UCSC genome browser	http://genome.ucsc.edu/index.html Mm9; from Repeat Masker Track (repClass=SINE; repFamily=B2)
SINEB1 X35S	(Roman et al. 2008)	provided by Angel Carlos Roman.
Enhancers	(Visel et al. 2007)	http://enhancer.lbl.gov/
In situ gene expression data	Emage (Richardson et al. 2010) EMBRYS (Yokoyama et al. 2009) GXD (Finger et al. 2011) MAMEP (Geffers et al. 2012)	http://www.emouseatlas.org/emage/ http://embrys.jp/embrys/html/MainMenu.html http://www.informatics.jax.org/expression.shtml http://mamep.molgen.mpg.de/

Notes

[1] For EP300 sites the position considered was the centre of the peak.

[2] For sites bound by CTCF in different tissues we merged sites (taking the position of their median as new position) when they overlapped by at least 1 bp.

[3] We define as a cohesin binding sites any region identified as a peak in ChIP-seq experiments targeting any protein of the cohesin complex. For the analyses carried out, we restricted to “invariant” regions, occupied by the different proteins in the different tissues and cell-types assessed. Importantly, we found a very large overlap between the different datasets.

[4] From the SINE B2 list, we further distinguished the elements that were truncated to only the last 90 bases (or less) of the consensus sequence, and therefore did not contain the PolII, PolIII and CTCF binding site.

Supplementary Note 1.

Enrichment analysis for EP300-SB comparison

Enrichment was calculated as the mean of

$$N_{\text{specific},x}/N_{\text{random},x},$$

with $N_{\text{specific},x}$ being the number of true insertions and $N_{\text{random},x}$ being the number of random insertions located within x bp of the nearest EP300 site. $N_{\text{random},x}$ was set to 1 if no random insertions were found within x. Only x with $N_{\text{specific},x}$ of 5 or more were considered. We tested enrichment at a distance from 0 to 1 Mb with window sizes increasing by 500bp increments.

Randomisation was repeated 200 times for each enrichment analysis.

Since EP300 sites from different tissues often cluster together, particularly for midbrain and forebrain (Visel et al. 2009) ,we removed EP300 sites, which were less than 10kb away from a EP300 site in the second tissue considered, but the effect is still appreciable for the brain tissues.

Calculating the (p-value) of the frequency of certain features (*eg.* topological boundaries, CTCF sites, SINE B2 elements) for a given sequence set RSA (eg "regulatory domains", class C). 1000 sequence sets were randomly generated, each of which contained non-overlapping sequences mimicking sequences of RS_A: for each original region of RS_A, a region of the same size was randomly sampled from the chromosome of the original RS_A region restricted to the allowed genome space. The allowed genome space considered for sampling random regions was matched to RS_A properties and was either the whole genome or restricted to the TAD domains. In the later case, only the subset of the region set that was intra-TAD was used as RS_A. For each random sequence set, we determined the feature frequency in random regions. The observed feature frequency in RS_A (F_{obs}) is ranked within the 1000 random feature frequencies and the p-value is estimated given the formula: $rank_{obs}/(N_{rdm}+1)$, where $rank_{obs}$ is the rank of F_{obs} and N_{rdm} is the number of random dataset used in the simulation (always 1000 in this study). Note that with 1000 random sets, 0.000999001 is the best p-value one can obtain.

References for Supplementary Information

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.

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.

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.

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

Finger JH, Smith CM, Hayamizu TF, McCright IJ, Eppig JT, Kadin JA, Richardson JE, Ringwald M. 2011. The mouse Gene Expression Database (GXD): 2011 update. *Nucleic Acids Res* **39**: D835–41.

Geffers L, Herrmann B, Eichele G. 2012. Web-based digital gene expression atlases for the mouse. *Mamm Genome* **23**: 525–538.

Gonzalez F, Duboule D, Spitz F. 2007. Transgenic analysis of Hoxd gene regulation during digit development. *Dev Biol* **306**: 847–859.

Gonzalez-Reyes LE, Verbitsky M, Blesa J, Jackson-Lewis V, Paredes D, Tillack K, Phani S, Kramer ER, Przedborski S, Kottmann AH. 2012. Sonic hedgehog maintains cellular and neurochemical homeostasis in the adult nigrostriatal circuit. *Neuron* **75**: 306–319.

Jeong Y, El-Jaick K, Roessler E, Muenke M, Epstein DJ. 2006. A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. *Development* **133**: 761–772.

Lunyak VV, Prefontaine GG, Núñez E, Cramer T, Ju B-G, Ohgi KA, Hutt K, Roy R, García-Díaz A, Zhu X, et al. 2007. Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* **317**: 248–251.

Makeyev EV, Zhang J, Carrasco MA, Maniatis T. 2007. The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Mol Cell* **27**: 435–448.

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.

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.

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.

Richardson L, Venkataraman S, Stevenson P, Yang Y, Burton N, Rao J, Fisher M, Baldock RA, Davidson DR, Christiansen JH. 2010. EMAGE mouse embryo spatial gene expression database: 2010 update. *Nucleic Acids Res* **38**: D703–9.

Roman AC, Benitez DA, Carvajal-Gonzalez JM, Fernandez-Salguero PM. 2008. Genome-wide B1 retrotransposon binds the transcription factors dioxin receptor and Slug and regulates gene expression in vivo. *Proc Natl Acad Sci U S A* **105**: 1632–1637.

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.

Sagai T, Hosoya M, Mizushina Y, Tamura M, Shiroishi T. 2005. Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* **132**: 797–803.

Seitan V, Faure A, Zhan Y, McCord R, Lajoie B, Ing-Simmons E, Lenhard B, Giorgetti L, Heard E, Fisher A, et al. 2013. Cohesin-based chromatin interactions enable regulated gene expression within pre-existing architectural compartments. *Genome Res*.

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.

Soshnikova N, Montavon T, Leleu M, Galjart N, Duboule D. 2010. Functional analysis of CTCF during mammalian limb development. *Dev Cell* **19**: 819–830.

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

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.

Yokoyama S, Ito Y, Ueno-Kudoh H, Shimizu H, Uchibe K, Albini S, Mitsuoka K, Miyaki S, Kiso M, Nagai A, et al. 2009. A systems approach reveals that the myogenesis genome network is regulated by the transcriptional repressor RP58. *Dev Cell* **17**: 836–848.