Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders


*Centro Andaluz de Biología del Desarrollo, Consejo Superior de Investigaciones Científicas/Universidad Pablo de Olavide, Seville 41013, Spain; †Centro de Biología Molecular Severo Ochoa, Consejo Superior de Investigaciones Científicas/Universidad Autónoma de Madrid, Madrid 28049, Spain; ‡Centro de Investigación Biomédica en Red de Enfermedades Raras, 28049 Madrid, Spain; §Department of Human Genetics, University of Chicago, IL 60637; and ¶Division of Biology, California Institute of Technology, Pasadena, CA 91125

Edited by Denis Duboule, University of Geneva, Geneva, Switzerland, and approved May 7, 2015 (received for review March 18, 2015)

Increasing evidence in the last years indicates that the vast amount of regulatory information contained in mammalian genomes is organized in precise 3D chromatin structures. However, the impact of this spatial chromatin organization on gene expression and its degree of evolutionary conservation is still poorly understood. The Six homeobox genes are essential developmental regulators organized in gene clusters conserved during evolution. Here, we reveal that the Six clusters share a deeply evolutionarily conserved 3D chromatin organization that predates the Cambrian explosion. This chromatin architecture generates two largely independent regulatory landscapes (RLs) contained in two adjacent topological associating domains (TADs). By disrupting the conserved TAD border in one of the zebrafish Six clusters, we demonstrate that this border is critical for preventing competition between promoters and enhancers located in separated RLs, thereby generating different expression patterns in genes located in close genomic proximity. Moreover, evolutionary comparison of Six-associated TAD borders reveals the presence of CCCTC-binding factor (CTCF) sites with diverging orientations in all studied deuterostomes. Genome-wide examination of mammalian HiC data reveals that this conserved CTCF configuration is a general signature of TAD borders, underscoring that common organizational principles underlie TAD compartmentalization in deuterostome evolution.

CTCF | TAD | Six cluster | regulatory landscapes | evolution

Noncoding DNA in animal genomes harbors regulatory information that controls the levels and the spatiotemporal activation of gene expression (1, 2). Information is precisely organized in the 3D chromatin favoring contacts between regulatory elements and target genes (3). Subdivision into topologically associating domains (TADs), megabase-scale domains within which sequences preferentially contact one another, represents a hierarchical level in chromatin 3D organization (4–9). The limited number of studies conducted to date show that a large fraction of TADs are invariant among cell types and largely conserved in mammals (4–6, 10), but the extent of their evolutionary conservation and functional organizer properties are largely unexplored.

Results and Discussion

We analyzed the 3D regulatory architecture of Six homeobox gene clusters, applying a bottom-up approach, from genomic organization to gene regulation and developmental patterning. Six homeobox genes, essential for the development of many embryonic structures, comprise three subfamilies—Six1/2, Six3/6, and Six4/5—tandemly arrayed in tight genomic clusters strongly conserved in several animal phyla (11). As a consequence of ancestral whole genome duplications, most vertebrate genomes contain two paralogous copies of the Six cluster: one containing Six2 and Six3 and the other Six1, Six4 and Six6. Teleost genomes contain four clusters (12) given the extra genome duplication at the base of their lineage. Thus, Six genes’ genomic organization remained strongly conserved, likely reflecting the presence of strong cis-regulatory constraints (11). Using high-resolution circular chromosome conformation capture (4C-seq) on zebrafish embryos, we identified genomic contacts of different genes located along the chromatin region encompassing the six3a-six2a cluster (Fig. 1A). The six3a-six2a cluster is split into two separate compartments defined by the genomic contacts of the two genes (Fig. 1 and Fig. S1A). The intergenic border constitutes a sharp border in which genomic contacts shift from one regulatory landscape to the other, as revealed by the distribution of the differential 3D interactions for the two promoters, obtained as the difference between the two Six 4C-seq signals (Fig. 1A). The precise location where this shift occurs corresponds to the genomic position with a maximal difference of accumulated contact reads for each gene along the entire locus (Fig. 1A, asterisk). Taking this position as a reference point (Fig. 1B, asterisk and dashed line), the preborder gene desert 5’ to six3a accounts for more than 80% of its 3D interactions; similarly six2a contacts are preferentially located at the opposite.

Significance

Mammalian chromatin is compartmentalized in topologically associating domains (TADs), genomic regions within which sequences preferentially contact each other. This organization has been proposed to be essential to organize the regulatory information contained in mammalian genomes. We show that Six homeobox genes, essential developmental regulators organized in gene clusters across different animal phyla, share a deeply conserved chromatin organization formed by two abutting TADs that predates the Cambrian explosion. This organization is required to generate separate regulatory landscapes for neighboring genes within the cluster, resulting in very different gene expression patterns. Finally, we show that this extremely conserved 3D architecture is associated with a characteristic arrangement of CCCTC-binding factor (CTCF) binding sites in diverging orientations, revealing a genome-wide conserved signature for TAD borders.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequence reported in this paper has been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE66900).

1To whom correspondence should be addressed. Email: jlgomska@upo.es.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1505463112/-/DCSupplemental.
genes define one chromatin 3D territory opposite to six2a (red), and the six3a B

six2a six3a/six2a gene is expressed in apical neurogenic territories (red) and

six2a + six3a gene expression using Six4/5 June 16, 2015 7543

+ six2a and (dark blue), and Enhancer III (Enh III, light blue) viewpoints, respectively. The second track shows the interaction map remains constant throughout development, with some cases, corresponding clusters (Fig. 2, asterisks and dashed lines). In all cases, six genes’ contacts preferentially take place within the regulatory domains generated by the border region. Supporting the presence of these two regulatory landscapes in each cluster, the expression patterns of six3a, six3b, six6a, and six6b are dramatically different from those of six1/2 and six4, which instead largely share expression domains.

to determine whether a similar two-chromatin 3D territories organization is present in other vertebrates, we performed 4C-seq from the mouse Six gene promoters (Fig. 3 A and B), which again revealed a clear partition of both clusters, with developmentally invariant 4C-seq contacts (Fig. S4): the regulatory landscapes of Six3 and Six6 genes define one chromatin 3D territory opposite to those of Six1/2 and Six4. Similarly, the bipartite topology of the clusters correlates with the divergent expression patterns of the genes located at each side of the border regions. Furthermore, HiC data from mouse cells shows that the two different regulatory 3D domains defined by 4C-seq precisely match two separated TADs (Fig. 3 and Fig. S4), with the border region identified by 4C-seq located at the border of these TADs.

These results indicate that vertebrate Six clusters share an evolutionary conserved 3D chromatin architecture, possibly preceding whole genome duplications present in the vertebrate ancestor cluster. This possibility was verified by the analysis of the genome from gastrulating sea urchin embryos, a distant invertebrate deuterostome with an intact ancestral Six-tandem configuration, in which the Six3/6 gene is expressed in apical neurogenic territories with a pattern largely different from that of the Six1/2, as in vertebrates and other bilaterians (18–20). Consistently, the sea urchin Six cluster is divided into two regulatory landscapes: one defined by the Six3/6 promoter contacts and the other by Six1/2 and Six4/5 promoter contacts (Fig. 3C). Thus, a deeply conserved 3D chromatin architecture of the Six cluster composed of two different TADs has been preserved from deuterostome ancestors.

As these results suggest that strong long-lasting regulatory pressures constrained the evolution of this genomic organization, we attempted to determine the functional contribution of this evolutionary conserved TAD border to Six gene expression using BAC recombinering and zebrasfish transgenesis. We selected a 194-kb six3a/six2a-containing BAC spanning from +47 kb upstream of six2a to +120 kb upstream of six3a, including the intergenic region harboring the TAD border. In this BAC, we replaced the six2a and the six3a coding regions by GFP and mCherry, respectively. Transient transgenic zebrasfish embryos generated with this BAC showed strong mCherry but weak GFP expression in the brain (Fig. 3D, Left). These results are consistent with this BAC containing mainly six3a-associated regulatory regions. We then generated a deleted version of the BAC by removing an 18-kb fragment spanning the intergenic region containing the TAD border. This region is devoid of strong H3K27ac peaks, suggesting that no enhancers are present within the deletion. This deletion dramatically changed the reporters’ expression, and six2a-GFP was now strongly activated in the six3a

Fig. 1. The six3a/six2a cluster in divided in two developmentally stable 3D compartments. (A) The first, third, and fourth tracks show 4C-seq data on whole zebrafish embryos at 24 hpf from six2a (red), six2a (dark blue), and Enhancer III (Enh III, light blue) viewpoints, respectively. The second track shows the difference between the number of reads in the 4C-seqs from the six3a and six2a viewpoints. Black and gray signals indicate positive or negative differences, respectively. The fifth track shows accumulative reads along the region for six3a (red) and six2a (blue). Differences of these accumulative reads are shown in the track below. The asterisk marks the border region. (B) 4C-seq on whole zebrafish embryos at dome, 80% epiboly, 24 and 48 hpf from the six3a and six2a viewpoints (black triangles). Contact percentages for each gene on the two 3D compartments are indicated.

The 3D chromatin structures of the remaining zebrafish six clusters (Fig. 2 and Figs. S2 and S3) were always developmentally conserved with two largely different regulatory landscapes for each cluster: one detected from the six3/six promoters’ side (six3a, six3b, six6a, and six6b) and the other from the promoters at the opposite side of the cluster (six1/2 and six4) (Fig. 2). Indeed, a border isolated six3a, six3b, six6a, and six6b from the other genes of the corresponding clusters (Fig. 2, asterisks and dashed lines). In all cases, six genes’ contacts preferentially take place within the regulatory domains generated by the border region (Fig. 2). Supporting the presence of these two regulatory landscapes in each cluster, the expression patterns of six3a, six3b, six6a, and six6b are dramatically different from those of six1/2 and six4, which instead largely share expression domains.

Gómez-Marín et al.
brain-associated domain. We conclude that the deeply evolutionary conserved TAD border is critical in preventing competition between promoters for enhancers located in opposite regulatory landscapes. This border, in turn, helps generating largely different expression patterns of genes located in the same genomic cluster.

To explore the possible determinants of the highly conserved 3D architecture of the Six clusters, we searched for conserved signatures associated with the TAD borders that bisect them in all deuterostomes examined. The DNA binding protein CCCTC-binding factor (CTCF) is instrumental for the 3D organization of the chromatin. CTCF, which is highly enriched at TAD borders (21), facilitates the formation of DNA loops that form between CTCF binding sites because of their attracting and interacting converging (head-to-head) orientations (10). Consistently, CTCF ChIP-seq data from human and mouse cells revealed several mammalian conserved CTCF sites at the deeply conserved Six TAD borders (Fig. 4A and Fig. S5A–C), which notably are disposed in a diverging (tail-to-tail) orientation. We checked whether this diverging orientation was also present in zebrafish and scanned Six clusters to detect CTCF sites. As CTCF ChIP-seq data are not available in this species, we used our recently generated ATAC-seq data from zebrafish embryos at 24 hpf (22), because this technique allows the identification of open chromatin regions at high resolution, including CTCF sites (23). This technique revealed several ATAC-seq footprints at Six border regions containing high score CTCF motifs that, as in the case of mammals, were also disposed in the same tail-to-tail diverging orientation and located in equivalent conserved syntenic positions (Fig. 4B, Figs. S5 D–F and S6, and Table S1). A divergent CTCF arrangement was found also in the intergenic region in which the Six border of the sea urchin cluster is located (Fig. 4C).

These results suggest that this configuration of CTCF sites could be a conserved and genome-wide used signature of TAD borders. We thus scanned available CTCF ChIP-seq data for the orientation of all CTCFs binding sites within 50 kb of mouse and human TAD borders (5). We found a very pronounced shift in the orientation of the CTCF sites from to the other side of mammalian borders, with a clear diverging configuration (Fig. 4D and E). Moreover, 72% and 48% of the mouse and human borders, respectively, contained at least one pair of CTCF sites with this tail-to-tail configuration. This shift in the orientation of the CTCF sites was not observed in regions around promoters, which are also enriched in CTCF sites (Fig. S5 G and H).

In summary, Six genes have maintained their general 3D chromatin organization throughout deuterostome evolution to generate different regulatory landscapes in the same gene cluster. This chromatin configuration may be even more ancient than that reported here, dating back to the ancestor of all bilaterians or even eumetazoans, because the Six cluster is present in some lophotrochozoans and cnidarian genomes. Our results demonstrate that specific 3D chromatin architectures can be evolutionarily preserved across different animal phyla and divergent body plans. Our data also show the power of evolutionary comparison to identify sequence signatures associated with deeply conserved borders. This strategy reveals that CTCF binding sites positioned in diverging orientation, likely by promoting loops toward opposite
directions, may be essential for TAD border formation/maintenance along deuterostome evolution.

Materials and Methods

4C-seq. Mouse and zebrafish experiments were approved by the Ethic Committee of the University Pablo de Olavide and the Consejo Superior de Investigaciones Científicas in accordance with national and European regulations.

4C-seq assays were performed as previously reported (24–27). Whole zebrafish (at 48 hpf, 24 hpf, 80% epiboly, and dome stages) and mouse (at 9.5 and 14.5 d of development) embryos were processed as previously described (28) to get ∼10 million isolated cells. For sea urchin samples, 10 million 48-hpf embryos were dissociated into ice-cold dissociation buffer (1 M glycine, pH 8, 2 mM EDTA, and protease inhibitors). After several washes in Ca++ free artificial sea water and one in PBS, cells were cross-linked for 10 min in 2% (wt/vol) paraformaldehyde in PBS at room temperature. The reaction was stopped by adding glycine to a final concentration of 125 mM for 5 min at room temperature, and the cells were then collected and treated equally as mouse and zebrafish cross-linked cells. Briefly, cells were treated with lysis buffer [10 mM Tris·HCl, pH 8, 10 mM NaCl, 0.3% IGEPAL (octylphenoxypolyethoxyethanol) CA-630 (Sigma-Aldrich; I8896), and 1× protease inhibitor mixture (Complete; Roche; 11697498001)]. Nuclei were digested with DpnII endonuclease (New England Biolabs; R0543M) and ligated with T4 DNA ligase (Promega; M1804). Subsequently, Csp6I endonuclease (Fermentas, Thermo Scientific; FD0214) was used in a second round of digestion, and the DNA was ligated again. Specific primers (Table S2) were designed close to gene promoters or regulatory elements with primer3 (bioinfo.ut.ee/primer3-0.4.0/primer3). Illumina adaptors were included in the primer sequences (29). Eight 25-μL PCRs were performed with

Fig. 3. Conserved subdivision of Six clusters in two 3D compartments along deuterostome evolution. (A–C) 4C-seq in whole mouse (A and B) or sea urchin (C) embryos from the different Six genes as viewpoints. Border regions are indicated by an asterisk, and the two 3D compartments are shaded in red and blue. Contact percentages for each gene on the two 3D compartments is shown. In murine clusters, HiC data from mouse ES cells is shown below. HiC data show that clusters are divided into the two TADs also defined by 4C-seq. Border regions identified by 4C-seq coincide in both clusters with TADs borders. (Right) Expression patterns of the different genes. (D) Transient transgenic fish embryos injected with the full six2a-GFP/six3a-mCherry BAC (Left) or a TAD-deleted version (Δ18kb) of the same BAC (Right) at 72 hpf. Arrowheads point to the brain expression domain characteristic of the six3a gene.

Gómez-Marín et al. PNAS | June 16, 2015 | vol. 112 | no. 24 | 7545
the Expand Long Template PCR System (Roche; 11759060001) for each viewpoint, pooled together, and purified using the High Pure PCR Product Purification Kit (Roche; 11732668001). The Quanti-iT PicoGreen dsDNA Assay Kit (Invitrogen; P11496) was used to measure sample concentration and then deep sequenced. 4C-seq data were analyzed, with some changes, as described in ref. 26. In brief, raw sequencing data were de-multiplexed and aligned using the Mouse July 2007 (NCBI37/mm9), Zebrafish December 2009 (Zv9/danRer7), or Sea Urchin June 2011 (BCM-HGSC Spur_v3.1/strPur4) assemblies as the reference genomes. Reads located in fragments flanked by two restriction sites of the same enzyme or in fragments smaller than 40 bp were filtered out. Mapped reads were then converted to reads-per-first-enzyme-fragment-end units and smoothed using a 30 DpnII fragments mean running window algorithm. For comparative analysis, we took the genomic region containing the major part of the regulatory landscape of every gene present in a particular cluster and normalized all datasets by total weight. To predict the position of putative topological borders, the accumulated contacts of each viewpoint along the studied region were calculated, and then the position of maximum difference between accumulative contacts of different viewpoints located in the same cluster was determined. Furthermore, by calculating signal lineal ratio between genes supposed to be allocated in different TADs, we estimate whether this predicted region corresponds to the transition point in which contact from one gene start to decline considerably at the same time that contacts from the other begin to drastically increase. Once border position was calculated, we determined the percentage of reads aligned in both sides of the putative border for each gene of the cluster.

**Zebrafish Whole-Mount In Situ Hybridization.** Antisense RNA probes labeled with digoxigenin were prepared from cDNAs, and specimens were fixed, hybridized, and stained as previously described (30, 31). In situ probes were developed with nitro blue tetrazolium (NBT)/5-bromo-4-chloro-3-indolyl-phosphate (BCIP).

**Mouse Whole-Mount In Situ Hybridization.** Six2 and Six3 probes were kindly provided by Pascal Maire (Institute of Biomedical Research, Paris) and Tristán Rodríguez (Faculty of Medicine, Imperial College London, London), respectively. The Six6 probe has been described previously (32). Six1 and Six4 probes were generated by reverse amplification of whole mouse embryo RNA (stage 9.5 dpc) using the following primer pairs: Six1-F: CTCGGTCCTTCTGCTCCA, Six1-R: TCGCTGCTACCCTAACCG, Six4-F: CCTTTATGCTGTTAGTCCAGGG, Six4-R: CGGCCTTAGAACGTTTGTGAC, as described by the Allen Brain Atlas (www.brain-map.org). Amplified fragments were isolated and cloned into the pSpark vector.
vector (Canvax Biotech). Riboprobes were prepared using T7 RNA polymerase following manufacturer’s instructions (Roche Diagnostics) on linearized cDNA clones. Whole-mount in situ hybridization was carried out on Cba X C57BB mouse embryos using digoxigenin-labeled riboprobes as described previously (33) using an InstuPro robot (Intavis).

Sea Urchin Whole-Mount in Situ Hybridization. Whole-mount in situ hybridization in Strongylocentrotus purpuratus has been described previously (34). Briefly, sea urchin embryos were fixed in glutaraldehyde solution. The fixed embryos were incubated in the hybridization buffer (50% (vol/vol) formamide, 5× SSC, 1× Denhardt’s, 1 mg/ml yeast tRNA, 50 ng/ml heparin, and 0.1% Tween-20) with 0.5 ng/μl digoxigenin- and fluorescein-labeled RNA probe(s) at 60 °C for 18 h. Posthybridization washes were performed in hybridization buffer, 2× SSCT (2× SSC and 0.1% Tween-20), 0.2× SSCT, and 0.1× SSC, each for 20 min at 60 °C. Subsequently, the antibody incubations were performed at room temperature with 1:1,000 diluted anti-DIG Fab (Roche). The embryos were extensively washed before staining reaction, including six times with MABT buffer (0.1 M maleic acid, 0.15 M NaCl, and 0.1% Tween-20) and twice with AP buffer [100 mM Tris Cl (pH 9.5), 100 mM NaCl, 50 mM MgCl₂, and 1 mM levamisole]. BCIP and NBT were used for staining. Probes for six1/2 and six3/6 genes were obtained previously in refs. 19 and 35.

Generation of Recombinant BACs. BAC clone (number 7482) from the DaniaKey zebrafish BAC library containing the six2a and six3a genes was purchased from Source BioScience. Recombinant BACs were generated as previously described (36, 37). For further details, see SI Materials and Methods.

CTCF Motif Distribution Along TAD Borders. For calculating the average position of the CTCF motifs in each strand within the TAD borders, we searched the motifs located in CTCF Chip-seq peaks (38) common in at least five different tissues in mice or in constitutive CTCF peaks (39) in humans, which overlap with windows of 100 kb around each border. We used Clover software (40) with a score threshold of 10 for this search. We then plotted the motif distribution for each strand within these 100-kb windows around TAD borders using horizontal boxplots. For the direct comparison of the motif density at each position in these windows, we divided them in 20 bins of 5 kb each and plotted the distribution of motifs in both strands. As a random control for this analysis, we used 1,000 genomic windows of 100 kb centered in promoters that are known to be enriched in CTCF binding sites but are not involved in chromatin loop formation.

ACKNOWLEDGMENTS. We thank Dr. Eric Davidson for support with the sea urchin BAC clone (number 74B2) from the Nature 295(5558):1306 also supported J.C. and M.S. The mouse embryos used were purchased from Source BioScience (38x499) was purchased from Source BioScience DanioKey zebrafish BAC library containing the 19 and 35.

The fixed embryos were incubated in the hybridization buffer [50% (vol/vol) formamide, 5× SSC, 1× Denhardt’s, 1 mg/ml yeast tRNA, 50 ng/ml heparin, and 0.1% Tween-20] with 0.5 ng/μl digoxigenin- and fluorescein-labeled RNA probe(s) at 60 °C for 18 h. Posthybridization washes were performed in hybridization buffer, 2× SSCT (2× SSC and 0.1% Tween-20), 0.2× SSCT, and 0.1× SSC, each for 20 min at 60 °C. Subsequently, the antibody incubations were performed at room temperature with 1:1,000 diluted anti-DIG Fab (Roche). The embryos were extensively washed before staining reaction, including six times with MABT buffer (0.1 M maleic acid, 0.15 M NaCl, and 0.1% Tween-20) and twice with AP buffer [100 mM Tris Cl (pH 9.5), 100 mM NaCl, 50 mM MgCl₂, and 1 mM levamisole]. BCIP and NBT were used for staining. Probes for six1/2 and six3/6 genes were obtained previously in refs. 19 and 35.

Generation of Recombinant BACs. BAC clone (number 7482) from the DaniaKey zebrafish BAC library containing the six2a and six3a genes was purchased from Source BioScience. Recombinant BACs were generated as previously described (36, 37). For further details, see SI Materials and Methods.

CTCF Motif Distribution Along TAD Borders. For calculating the average position of the CTCF motifs in each strand within the TAD borders, we searched the motifs located in CTCF Chip-seq peaks (38) common in at least five different tissues in mice or in constitutive CTCF peaks (39) in humans, which overlap with windows of 100 kb around each border. We used Clover software (40) with a score threshold of 10 for this search. We then plotted the motif distribution for each strand within these 100-kb windows around TAD borders using horizontal boxplots. For the direct comparison of the motif density at each position in these windows, we divided them in 20 bins of 5 kb each and plotted the distribution of motifs in both strands. As a random control for this analysis, we used 1,000 genomic windows of 100 kb centered in promoters that are known to be enriched in CTCF binding sites but are not involved in chromatin loop formation.

ACKNOWLEDGMENTS. We thank Dr. Eric Davidson for support with the sea urchin experiments; Dr. W. De Laat for providing the 4C-seq method before publication; Fernando Casares for fruitful discussions; and Drs. Pascal Maire (Instituto de Biomédical Research) and Tristán Rodríguez (Imperial College London) for providing reagents. This study was supported by the Spanish and Andalusian Governments (BFU2013-41322-P and Proyecto de Excelencia BIO-396 to J.L.-S., BFU2013-45213-P to P.B., BFU2014-55738-REDT to J.L.-S. and P.B., and BFU2011-22928 to J.C. and M.L.-M.). EU-FP7-PEOPLE-2011-CIG Grant 303904 “EPAXIALMYF5KO” also supported J.C.