A gene desert required for regulatory control of pleiotropic *Shox2* expression and embryonic survival

Samuel Abassah-Oppong[1,11,12], Brandon J. Mannion[2,9,11], Virginie Tissières[3], Eddie Rodríguez-Carballo[4,13], Anja Ljubojevic[1], Fabrice Darbellay[2], Tabitha A. Festa[1], Carly S. Sullivan[1], Guy Kelman[2], Riana D. Hunter[2], Catherine S. Novak[2], Ingrid Plajzer-Frick[2], Stella Tran[2], Jennifer A. Akiyama[2], Iros Barozzi[2,5], Guillaume Andrey[6], Javier Lopez-Rios[3], Diane E. Dickel[2], Axel Visel[2,7,8], Len A. Pennacchio[2,7,9], John Cobb[1,*] and Marco Osterwalder[2,10,*]

[1] Department of Biological Sciences, University of Calgary, 2500 University Drive N.W., Calgary, Alberta, T2N 1N4, Canada.

[2] Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

[3] Centro Andaluz de Biología del Desarrollo (CABD), CSIC-Universidad Pablo de Olavide-Junta de Andalucía, 41013 Seville, Spain.

[4] Department of Genetics and Evolution, University of Geneva, 1211 Geneva, Switzerland.

[5] Department of Surgery and Cancer, Imperial College London, London, UK.

[6] Department of Genetic Medicine and Development and iGE3, Faculty of Medicine, University of Geneva, 1211 Geneva, Switzerland.

[7] US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

[8] School of Natural Sciences, University of California, Merced, Merced, CA 95343, USA.

[9] Comparative Biochemistry Program, University of California, Berkeley, CA 94720, USA.

[10] Department for BioMedical Research (DBMR), University of Bern, 3008 Bern, Switzerland.

[11] These authors contributed equally.

[12] Present address: Department of Biochemistry, University of Cape Coast, Cape Coast, Ghana.

[13] Present address: Department of Molecular Biology, University of Geneva, 1211 Geneva, Switzerland.

[*] Correspondence: jacobb@ucalgary.ca (J.C.), marco.osterwalder@dbmr.unibe.ch (M.O.)

Keywords: *Shox2*, gene desert, enhancers, gene regulation, chromatin architecture, limb development, cardiac development, pacemaker, transcription factors.

ABSTRACT

The Shox2 homeodomain transcriptional regulator is known for its critical functions during mouse embryogenesis, enabling accurate development of limbs, craniofacial structures, neural populations and the cardiac conduction system. At the genomic level, the *Shox2* gene is flanked by an extensive gene desert, a continuous non-coding genomic region spanning over 500 kilobases that contains a multitude of evolutionarily conserved elements with predicted *cis*-regulatory activities. However, the transcriptional enhancer potential of the vast majority of these elements in combination with the biological necessity of the gene desert have not yet been explored. Using transgenic reporter assays in mouse embryos to validate an extensive set of stringent epigenomic enhancer predictions, we identify several novel gene desert enhancers with distinct tissue-specific activities in *Shox2* expressing tissues. 4C-seq chromatin conformation capture further uncovers a repertoire of gene desert enhancers with overlapping activities in the proximal limb, in a compartment essential for *Shox2*-mediated stylopod formation. Leveraging CRISPR/Cas9 to delete the gene desert region contained in the *Shox2* topologically associated domain (TAD), we demonstrate that this complex *cis*-regulatory platform is essential for embryonic survival and required for control of region-specific *Shox2* expression in multiple developing tissues. While transcription of *Shox2* in the embryonic limb is only moderately affected by gene desert loss, *Shox2* expression in craniofacial and cardiac domains is nearly abolished. In particular, *Shox2* transcripts in the sinus venosus (SV) encompassing the sinoatrial node (SAN) were depleted in embryos lacking the gene desert, likely accounting for the embryonic lethality due to *Shox2*-dependency of the SAN pacemaker. Finally, we discover a 1.5kb SV enhancer within the deleted gene desert region, which may act as a genomic module controlling the development of the cardiac conduction system. In summary, our results identify a gene desert indispensable for pleiotropic patterning and highlight the importance of these extensive regulatory landscapes for embryonic development and viability.

INTRODUCTION

The function of gene deserts has posed a considerable puzzle since these large noncoding regions were first shown to be a prominent feature of the human genome almost 20 years ago[1]. As further vertebrate genomes were sequenced, orthologous gene deserts that shared synteny were found[2]. Originally defined as gene-free chromosomal regions larger than 500 kilobases (kb), gene deserts frequently contain many interspersed, highly conserved sequences that function as transcriptional enhancers[3,4]. Not surprisingly, these extensive *cis*-regulatory landscapes are found enriched near genes with important developmental functions, such as transcription factors (TFs), suggesting a critical role for gene deserts in regulation of key developmental genes[2,3]. The first megabase-scale deletions of gene deserts surprisingly had no obvious effect on mouse development and only mildly affected the expression of nearby genes, suggesting that these chromosomal regions may be dispensable[5]. When chromosome-conformation-capture techniques were developed, it became possible to accurately predict the range and identity of specific *cis*-regulatory interactions within a given locus. For example, Montavon et al. applied these emerging technologies and genomic deletions to show that an 830kb gene desert containing a "regulatory archipelago" of limb enhancers was required for expression of *HoxD* genes in distal limbs[6]. Such an arrangement of dispersed enhancers within an extensive gene desert, or sometimes within gene-rich regions, has now emerged as a paradigm for understanding the control of tissue-specific transcription during development[7-9]. Thereby, the identification of topologically associating domains (TADs) as a unit of chromosomal organization has refined our understanding of how dispersed enhancers are integrated into a gene's regulatory architecture[10,11]. Since enhancer-promoter interactions are generally confined within a given TAD, deletions or inversions involving TAD boundaries can lead to a gain or loss of gene expression as regulatory interactions are redistributed within the reconfigured TADs[12,13]. Therefore, elucidating the regulatory activities in the vast non-coding segments of TADs can have profound implications for our understanding of the basis of human disease. To date, comprehensive studies of gene regulatory regions in mice involving chromosome conformation capture, transgenic reporter

assays and genomic deletions have been conducted on a restricted number of loci including *Shh*, *Pitx1*, *Epha4*/*Pax3*/*Ihh*, and the *HoxD* genes, and most commonly focusing on the developing limb[11,14-16].

In the current study, we focused on the mouse short stature homeobox 2 gene (*Shox2*) as an ideal model to study the *cis*-regulatory complement underlying pleiotropic gene expression and driving the development of multiple embryonic tissues. *Shox2* function is essential for the development of several discrete structures, including the proximal limb (the humerus and femur), craniofacial compartments, the facial motor nucleus of the hindbrain, and a subset of neurons of the dorsal root ganglia[17-21]. Most importantly, *Shox2* is required for cardiac pacemaker differentiation in the sinoatrial node (SAN) and therefore its inactivation leads to embryonic lethality due to bradycardia starting around embryonic day 11.5 (E11.5)[22,23]. We previously showed that the regulation of *Shox2* in limbs is controlled by multiple *cis*-regulatory modules and even the combined deletion of two *Shox2* proximal limb enhancers had relatively small effects on *Shox2* expression and limb morphology[24,25]. Here, we performed a more stringent test of the resilience of *Shox2* expression in multiple tissues by deleting the gene desert adjacent to *Shox2*, which encodes a plethora of genomic elements with developmental enhancer signatures. First, using a combination of epigenomic analysis, chromatin conformation capture and transgenic reporter assays, we identify numerous gene desert enhancers with distinct subregional activities in limbs, craniofacial compartments and neural cell populations, directly correlated with dynamic *Shox2* expression in mouse embryos. Our deletion analysis then uncovers a critical role of the gene desert in controlling *Shox2* expression not only in the proximal limb mesenchyme and craniofacial compartments, but also in the SAN-containing cardiac sinus venosus (SV). Finally, using open chromatin profiling from embryonic hearts we discover a SV enhancer likely involved in the essential *Shox2*-controlled regulation of the cardiac pacemaker system. Taken together, our results emphasize fundamental roles of a large *cis*-regulatory gene desert in transcriptional control of a key developmental gene.

RESULTS

The mouse *Shox2* transcription factor is located on chromosome 3 in a TAD spanning 1 megabase (Mb) of genomic sequence that contains the major fraction of a 675 kilobase (kb) gene desert[26] (**Fig. 1A**). This *Shox2*-TAD harbors one additional protein-coding gene (*Rsrc1*), while three other genes (*Mlf1*, *Veph1, Ptx3)* are found in neighboring chromatin domains (**Fig. 1A, S1A**). These *Shox2*-adjacent genes have not been involved in developmental patterning and show either near-ubiquitous (*Mlf1*, *Rsrc1*) or differential (*Veph*1, *Ptx3*) tissue-specific expression profiles (**Fig. S1A**). Transcription of *Shox2* is highly regulated around mid-gestation with prevalent expression domains in the developing limbs, craniofacial structures, the heart, neuronal populations of the mid- and hindbrain, and emerging facial nerves (**Fig. 1B, S1A**). This temporally dynamic and pleiotropic character suggests considerable complexity in the genomic regulatory landscape controlling *Shox2* activities. However, only a limited number of *Shox2*-associated transcriptional enhancers, with activities restricted to brain and limb sub-regions have been identified to date (Vista Enhancer Browser)[24,25,27].

To characterize the *cis*-regulatory complexity encoded in the extended *Shox2* TAD and specifically in the aforementioned gene desert, we established a map of stringent enhancer predictions using a combination of chromatin state profiles (ChromHMM) and H3K27ac ChIP-seq peak calls across sixty-six embryonic and perinatal tissue-stage combinations from ENCODE[28] (https://www.encode.project.org) (see **methods**). After excluding promoter regions, this analysis of the epigenome identified 30 genomic elements with robust enhancer signatures in at least one of the tissues and timepoints examined (**Figs. 1B, S1 and Table S1**). Remarkably, 16 of the 30 elements were located within the *Shox2* gene desert representing putative gene desert enhancers (GDEs). Indeed, the majority of GDEs showed dynamic spatiotemporal H3K27ac profiles including a combination of limb, craniofacial, cardiac or neuronal signatures (**Figs. 1B**). Collectively, these results suggest that the *Shox2* gene desert encodes a major fraction of the *cis*-regulatory modules controlling *Shox2* in a temporally and spatially-restricted manner in mouse embryos.

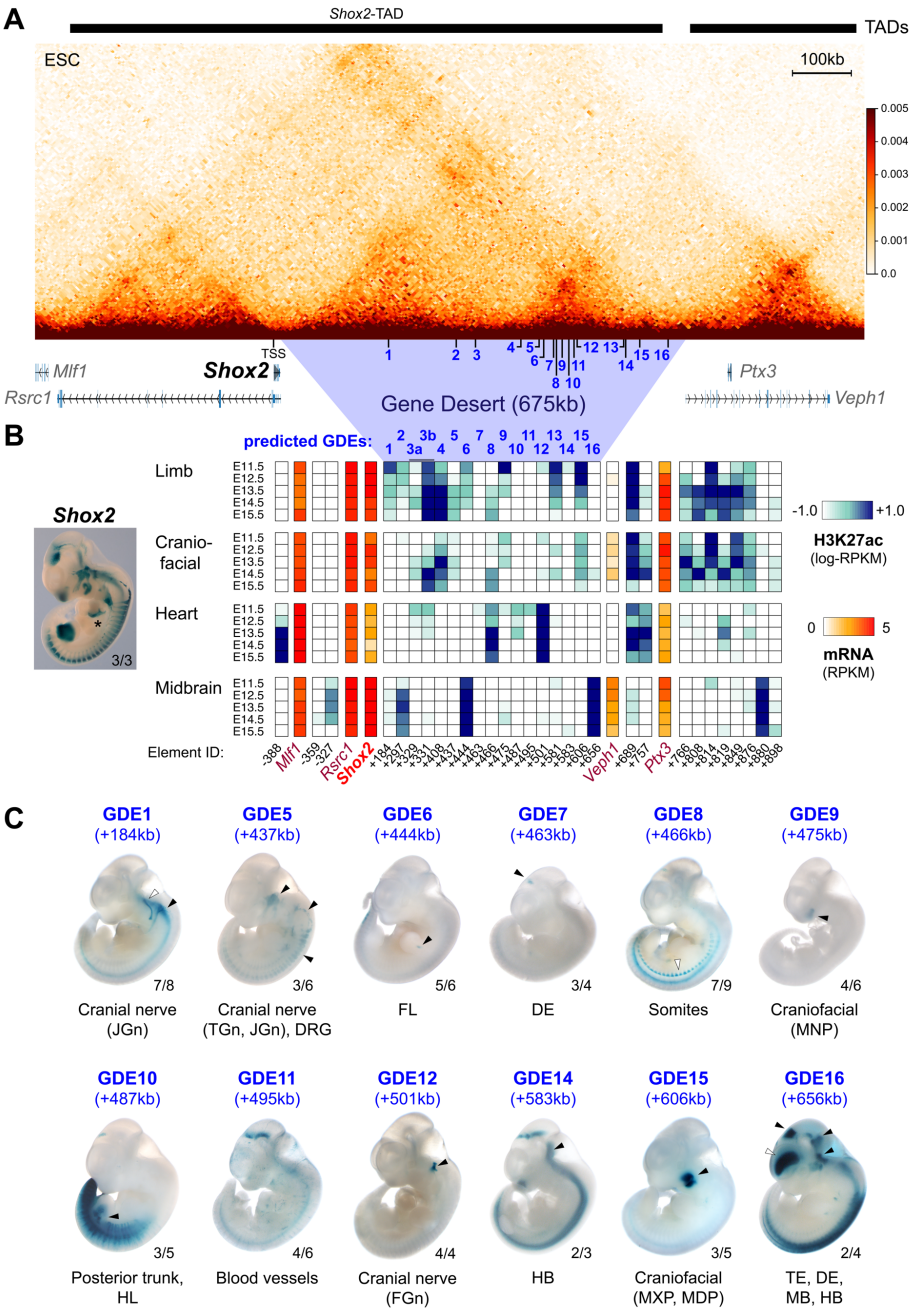

Figure 1. *Cis*-regulatory potential of the *Shox2*-adjacent gene desert in regulation of pleiotropic *Shox2* expression during embryogenesis. (**A**) Reprocessed high-resolution interaction heatmap (see **methods**) from Hi-C data of mouse embryonic stem cells (ESC)[26] including the *Shox2* TAD (chr3:66337001-67337000) and flanking genes. The gene desert (blue shade) and predicted gene desert enhancer elements (GDEs) are indicated. TSS: *Shox2* transcriptional start site. (**B**) Left: *LacZ*-stained embryo heterozygous for the *Shox2-LacZ* reporter knock-in allele[20]. The forelimb was removed for visibility of the heart (*). Right: Heatmap illustrating ChromHMM-filtered and H3K27ac-predicted enhancer regions (Element IDs) and their temporal activities in tissues with critical *Shox2* functions (see **Table S1 and methods**). The full set of tissues is shown in **Fig. S1**. Blue shades represent H3K27ac enrichment and red shading illustrates mRNA expression profiles of protein-coding genes present in the region. E, embryonic day. (**C**) Identification of embryonic enhancer activities in 12/16 GDEs at E11.5 using *in vivo* transgenic *LacZ* reporter assays. Arrowheads: Reproducible enhancer activities with (black) or without (white) overlap to *Shox2* expression domains. Numbers on the bottom right of each embryo represent the reproducibility of *LacZ* patterns (reproducible tissue-specific staining vs. number of embryos with any *LacZ* staining). JGn, TGn, FGn: jugular, trigeminal and facial ganglion, respectively. DRG, dorsal root ganglia. FL, Forelimb. HL, Hindlimb. MNP, medial nasal process. MXP-MDP, maxillary-mandibular region. TE, Telencephalon. DE, Diencephalon. MB, Midbrain. HB, Hindbrain. The genomic distance from the *Shox2* TSS (+, downstream; -, upstream) is indicated for all element IDs.

While H3K27 acetylation represents the primary epigenomic mark used to predict active transcriptional enhancers genome-wide[28,29], these predictions are not always congruent with cell-type or tissue-specific activities *in vivo*[30,31]. Therefore, to determine the relevant developmental enhancer activities of predicted GDE elements we conducted *LacZ* transgenic reporter assays in mouse embryos at embryonic day 11.5 (E11.5) (**Fig. 1C and Table S2**), a stage characterized by wide-spread and functionally relevant *Shox2* expression[25]. This analysis led to the identification of a battery of novel *in vivo* enhancers with distinct tissue-restricted activities, many closely overlapping subregional *Shox2* expression domains in craniofacial compartments, cranial nerve or brain regions (**Fig. 1B, C**). However, while multiple GDEs showed elevated H3K27ac signatures in developing limbs, transgenic screen identified only one element (GDE6) able to drive reporter activity in forelimbs (**Fig. 1C**). Notably, two GDEs (GDE9 and GDE15) displayed elevated H32K27ac in both limb and craniofacial tissues, but drove *LacZ* reporter expression exclusively in *Shox2*-overlapping craniofacial domains in the medial nasal (MNP) and maxillary-mandibular (MXP, MDP) processes, respectively (**Fig. 1C**). Our analyses also revealed multiple enhancers (GDE1, 5 and 12) with activities in cranial nerve tissue, including the trigeminal (TGn), facial (FGn) and jugular (JGn) ganglia, as well as the dorsal root ganglia (DRG) (**Fig. 1C**). *Shox2* is expressed in all these neural crest-derived tissues, but a functional requirement has only been observed for the development of the FGn and the mechanosensory neurons of the DRG[20,21]. Interestingly, while no H3K27ac profiles for cranial nerve populations were available from ENCODE, both GDE5 and GDE12 elements showed elevated H3K27ac in craniofacial tissue at E11.5, potentially mirroring the common neural-crest origin of cranial nerve and a subset of craniofacial cell populations[32]. At mid-gestation, *Shox2* is also expressed in the diencephalon (DE), midbrain (MB) and hindbrain (HB), and is specifically required for cerebellar development[33]. In accordance, our gene desert enhancer screen also revealed a set of novel brain enhancers (GDE7, 14 and 16) overlapping *Shox2* domains in the DE, MB or HB (**Fig. 1C**). In contrast, despite the presence of strong cardiac enhancer signatures in a subset of the tested gene desert elements, none of these predicted *cis*-regulatory modules drove reproducible reporter

expression in the heart at E11.5 (**Fig. 1B, C**). Taken together, our results uncover the potential of the *Shox2* gene desert to regulate a significant portion of the pleiotropic *Shox2* expression pattern and emphasize the importance of validating tissue-specific epigenomic predictions *in vivo* using transgenic reporter assays.

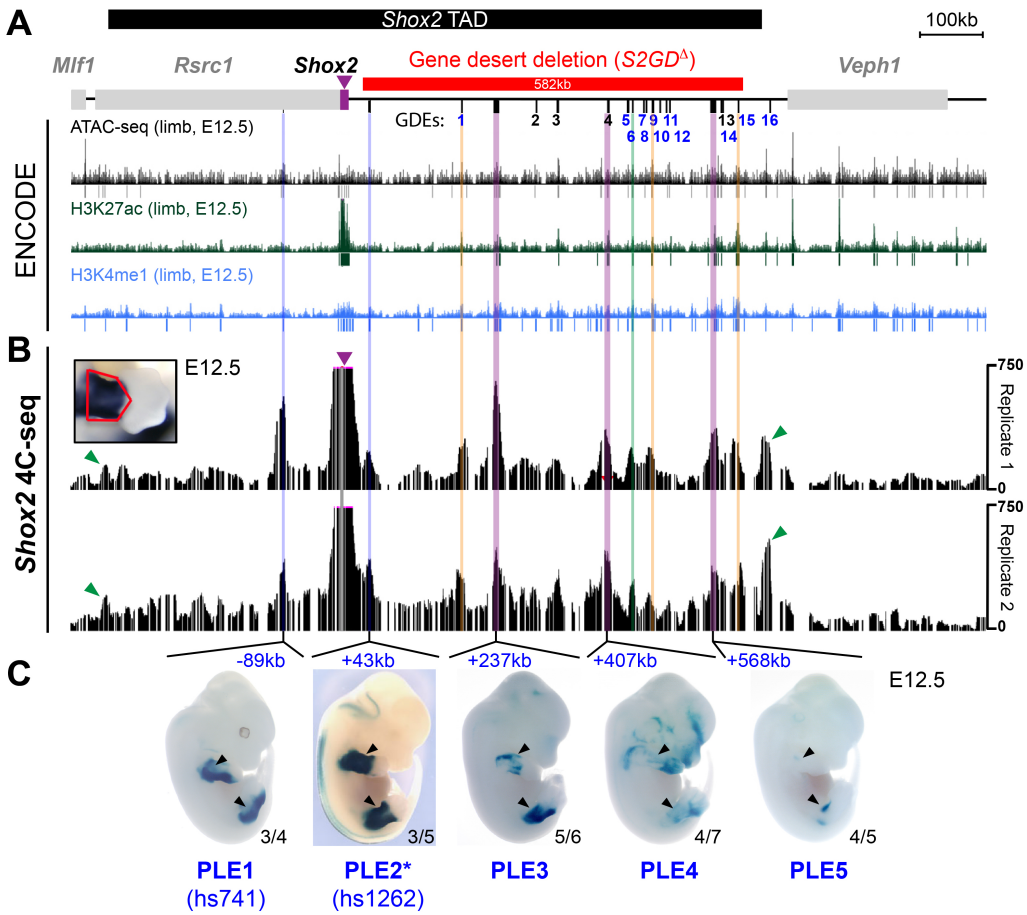

Figure 2. Chromatin confirmation capture identifies the gene desert as a hub for *Shox2*-interacting limb enhancers. (**A**) Map of the 1.4 Mb extended *Shox2*-TAD (see **Fig. 1A**) and the locations of the predicted gene desert enhancers (GDEs) 1-16, in blue those with confirmed *in vivo* enhancer activities at E11.5 (**Fig. 1B**, **1C and Table S1**). Mm10 UCSC browser tracks from ENCODE[28] indicate additional putative limb enhancer elements at E12.5 (bars below each track indicate peak calls). The extension of the CRISPR/Cas9-introduced gene desert deletion (*S2*GD$^{\Delta}$) is represented by the red bar. (**B**) 4C-seq interaction profiles from two biologically independent proximal limb samples at E12.5 are shown. The 4C-seq viewpoint is located within exon 1/intron 1 of the mouse *Shox2* gene (purple arrowhead). The inset displays the *Shox2* expression domain at E12.5 (*in situ* hybridization) and the region dissected for 4C-seq (red outline). Green arrowheads indicate CTCF-interacting regions localized at the boundaries of the *Shox2*-TAD (**Fig. S2A, B**). (**C**) *In vivo* transgenic *LacZ* reporter validation of predicted proximal limb enhancers (PLEs) based on enrichment in 4C-seq replicates. PLE1 is the mouse ortholog of the human hs741 enhancer sequence and PLE2 corresponds to the LHB-A/hs1262 enhancer[24,25](*). PLE3, 4 and 5 represent novel limb enhancers identified from 4C-seq profiles (purple lines). The GDE6 limb enhancer identified in **Fig. 1C** also shows 4C-seq enrichment (green line). The remaining regions enriched in 4C-seq replicates (orange lines) represent GDE elements 1, 9 and 15 with non-limb activities (see **Fig. 1C**). The embryos shown are representatives from stable transgenic *LacZ* reporter lines (see **Fig. S3**). The reproducibility is indicated by the number of original transgenic mouse lines (per construct injected) with similar *LacZ* staining in the limb vs. the number of mouse lines harboring insertions of the transgene (per construct injected).

Shox2 exerts a crucial role during limb development in controlling the formation of the humerus and femur via direct chondrogenic and osteogenic patterning mechanisms[17,34-36]. Although multiple elements with limb enhancer potential were identified in the gene desert (**Fig. 1B**), our transgenic screen of GDE elements only uncovered a single enhancer with forelimb activity at E11.5 (GDE6). Another, previously characterized limb enhancer (LHB-A/hs1262)[24,25] located 43kb downstream of the *Shox2* transcriptional start site (TSS) was not selected by our epigenomic profiling analysis, as a result of an earlier activation pattern and differential temporal enhancer signatures[28] (**Fig. 2A, S2A, B**). Therefore, to better define the ensemble of limb enhancers interacting with *Shox2* and relevant for limb chondrogenesis and/or osteogenesis, we performed circular chromosome conformation capture (4C-seq) from proximal limbs at E12.5 (**Fig. 2B, S2C**). We conducted two independent 4C-seq experiments using a viewpoint directly adjacent to the *Shox2* promoter (**Fig. 2A, B and Table S3**). The two replicates displayed reproducible interaction profiles revealing discrete regions with high interaction frequencies with the *Shox2* promoter (**Fig. 2B)**. Notably, the vast majority of these regions was located within the gene desert and also marked by open chromatin, H3K4me1 and/or H3K27 acetylation (**Fig. 2A**), indicative of *cis*-regulatory modules[6,28]. In accordance, five of these preferentially interacting regions mapped to GDE elements, including limb (GDE 6) and craniofacial enhancers (GDE 9, 15) (**Fig. 1C, S1, 2A, 2B**). In addition, our 4C-seq results confirmed interactions between *Shox2* and the previously identified proximal limb enhancers (PLEs) m741/hs741 (termed here PLE1) and LHB-A/hs1262 (termed PLE2) located upstream (-89kb) and downstream (+43kb) of the *Shox2* TSS, respectively (**Fig. 2B, C and Table S4**)[24,25,36]. Finally, our 4C-seq analysis identified three *Shox2*-contacting gene desert modules (+237kb, +407kb and +568kb) with limb enhancer signatures (**Fig. 2A, B**). And indeed, subsequent transgenic analysis in mouse embryos at E12.5 revealed that each of these elements (termed PLE3, 4 and 5) on its own was able to drive transgenic reporter expression in the proximal limb (**Fig. 2B, C and Table S4**). While both, PLE3 and PLE4 displayed activities co-localizing with skeletal progenitors from E11.5 to E13.5, PLE5 activity was restricted to the proximal-anterior limb mesenchyme and apparent at

later stages (E12.5 and E13.5), predominantly in the hindlimb (**Fig. S3**). In a last step, using 4C-seq we assessed the 3D interaction profiles of selected individual enhancers (PLE2 and PLE4) (**Fig. S2C**). These experiments corroborated the specific interactions observed between both enhancers and the *Shox2* promoter (**Fig. S2C**). Interestingly, while PLE2 shows no interaction with other enhancers, PLE4 is establishing contacts with two other proximal limb enhancers (PLE1 and PLE3) (**Fig. S2C**). This finding suggests that several 3D conformations co-exist in the limb at the *Shox2* locus, each one involving a different enhancer subset contacting the *Shox2* promoter. In summary, our results unveil a proximal limb enhancer (PLE) repertoire encoded in the *Shox2* gene desert and suggest a significant role of the gene desert in controlling limb-specific *Shox2* expression.

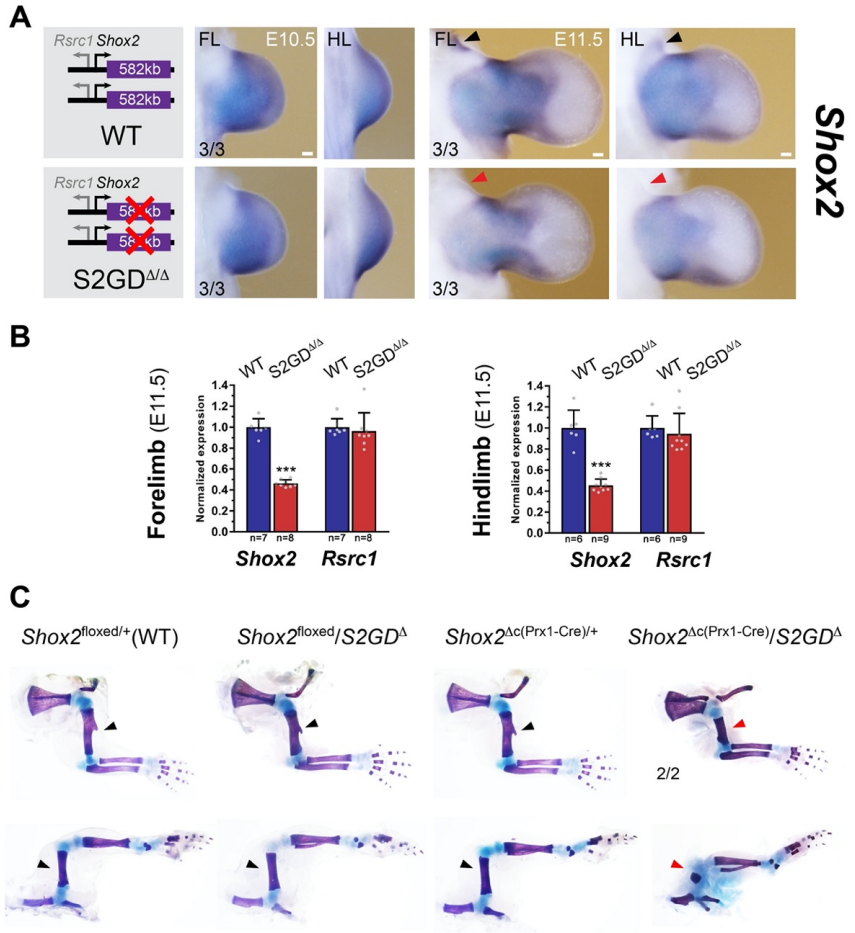

Figure 3. The gene desert controls quantitative *Shox2* expression in limbs as part of a resilient regulatory architecture. (**A**) ISH revealing spatial *Shox2* expression in fore- and hindlimb buds of embryos homozygous for the gene desert deletion (*S2*GD$^{\Delta/\Delta}$) at E10.5 and E11.5. Red arrowhead indicates loss of a proximal-anterior *Shox2* expression domain in absence of the 582kb gene desert. Scale bar, 100um. (**B**) qPCR gene expression profiling shows reduction of *Shox2* expression in fore- and hindlimbs of embryos lacking the gene desert at E11.5. Expression of *Rsrc1* remains unchanged. Bar graphs indicate mean and standard deviation (error bars). Dots represent individual data points. ***, $P < 0.001$ (two-tailed, unpaired *t*-test). (**C**) The gene desert is required for proximal limb development in a "sensitized" genetic background with conditionally reduced limb-specific *Shox2* gene dosage (due to the *Prx1*-Cre transgene). Skeletal

preparations of limbs from control (*Shox2*$^{floxed/+}$, *Shox2*^{floxed}/S2GD$^{\Delta}$, *Shox2*$^{\Delta c(Prx1\text{-}Cre)/+}$) and sensitized gene desert knockout (*Shox2*$^{\Delta c(Prx1\text{-}Cre)}$/S2GD$^{\Delta}$) newborn mice are shown. Red arrowheads point to severely reduced stylopod elements in fore- and hindlimbs of sensitized gene desert knockout mice, as opposed to normal stylopod morphology in control mice (black arrowheads). Chondrogenic skeletal elements are stained blue, ossified structures red. For each genotype, the number of independent biological replicates with similar results is indicated.

Next, to determine the functional necessity of this limb enhancer repertoire and the regulatory relevance of the *Shox2* gene desert as a whole, we used CRISPR/Cas9 in mouse zygotes to delete the gene desert region (582kb) located within the *Shox2*-TAD and encompassing PLE2-5 as well as GDE1-15 elements (**Figs. 2A, S4A and Tables S5, S6**). Heterozygous F1 mice with clean deletion breakpoints (*S2*GD$^{\Delta/+}$) (**Fig. S4A**) were born at expected Mendelian ratios and showed no impaired viability and fertility. However, following intercross of F1 heterozygotes, no mice homozygous for the gene desert deletion were born, and *S2*GD$^{\Delta/\Delta}$ embryos displayed lethality between E11.5 and E13.5 (**Fig. S4C**), reminiscent of the lethality observed in *Shox2*-deficient embryos due to cardiac pacemaker defects[22,23]. Assessment of *Shox2* expression in fore- and hindlimbs of *S2*GD$^{\Delta/\Delta}$ embryos at mid-gestation revealed surprising resilience of the spatial *Shox2* transcript domain (**Fig. 3A**), despite the loss of multiple PLEs (**Fig. 2C, S3**). Instead, *Shox2* transcript levels in the limb were reduced by approximately half in absence of the gene desert, indicating significant quantitative contributions of the PLE elements (**Fig. 3B and Table S7**). To circumvent embryonic lethality and to study the cumulative phenotypic requirement of the gene desert enhancers for limb skeletal morphology, we used a *Prx1*-Cre conditional approach[37] allowing allelic reduction of *Shox2* specifically in the limb (**Fig. 3C**). Remarkably, loss of the gene desert in a sensitized genetic background (defined by reduced *Shox2* gene dosage due to *Prx1*-Cre-mediated *Shox2* inactivation on one allele) revealed severe shortening of the stylopod in both limb types, most pronounced in the hindlimb (**Fig. 3C**). Together, these results indicate that telomeric (upstream) limb enhancers (including hs741) act largely autonomously in controlling spatial *Shox2* expression, while the centromeric (downstream) gene desert limb enhancers have a role in conferring transcriptional and phenotypic robustness in a predominantly quantitative manner.

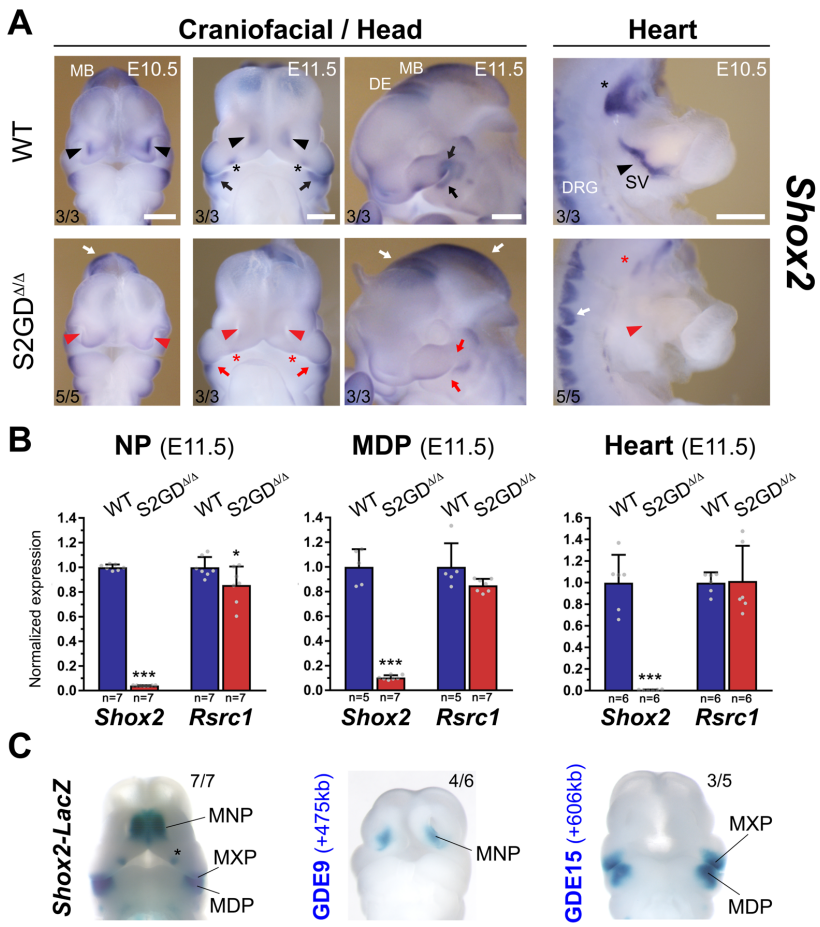

Figure 4. The gene desert controls pleiotropic *Shox2* expression with a predominant impact in craniofacial and cardiac domains. (**A**) RNA *in situ* hybridization (ISH) revealing severely reduced *Shox2* transcripts in craniofacial and cardiac expression domains in embryos homozygous for the gene desert deletion (*S2*GD$^{\Delta/\Delta}$). Left: downregulated *Shox2* expression in the medial nasal process (MNP, arrowheads), anterior portion of the palatal shelves (asterisk) and proximal maxillary (MXP) and mandibular (MDP) processes (arrows) in *S2*GD$^{\Delta/\Delta}$ embryos. Right: absence of *Shox2* expression in the cardiac sinus venosus (SV, arrowhead) and the nodose ganglion of the vagus nerve (asterisk) in *S2*GD$^{\Delta/\Delta}$ embryos at E10.5. White arrows mark tissues in which *Shox2* expression is retained. Scale bar, 500um. (**B**) Quantitative real-time PCR (qPCR) showing severe downregulation of *Shox2* transcripts in craniofacial and cardiac tissues at E11.5. Expression of *Rsrc1* remains unchanged in cardiac and mandibular tissues and is only minimally altered in the nasal process (NP). Bar graphs show mean and standard deviation (error bars). Dots indicate individual data points. ***, $P < 0.001$; *, $P < 0.05$ (two-tailed, unpaired *t*-test). (**C**) Ventral view of GDE9 and GDE15 craniofacial *LacZ* reporter activities (identified in **Fig. 1C**) accurately overlapping *Shox2* expression in the MNP and MXP/MDP. Asterisk marks anterior palatal shelf for which no enhancer could be identified. DE, Diencephalon. MB, Midbrain. DRG, Dorsal root ganglia.

Shox2 also displays important tasks in assuring normal craniofacial development, involving a requirement of *Shox2* for palatogenesis as well as formation of the temporomandibular joint (TMJ) required for jaw functionality in mammals[18,19]. These tasks are dependent on embryonic *Shox2* expression in distinct craniofacial domains, such as the anterior part of the palatal shelves and the maxillary-mandibular junction, respectively[18,19]. Notably, at E11.5, *S2*GD$^{\Delta/\Delta}$ embryos revealed *Shox2* downregulation in precisely the anterior portion of the palatal shelves as well as the proximal maxillary (MXP) and mandibular (MDP) processes (**Figs. 4A, 4B**). Furthermore, *Shox2* expression

in the medial nasal process (MNP) was severely downregulated at E10.5 and E11.5 (**Figs. 4A, 4B**). Hereby, the reduction of *Shox2* expression in the MXP-MDP domain and MNP of *S2*GD$^{\Delta/\Delta}$ embryos suggests an essential functional contribution of the two craniofacial enhancers (GDE9 and GDE15) identified in our transgenic screen based on epigenomic predictions and located in the deleted gene desert region (**Figs. 1B, 2A**). Importantly, GDE9 and GDE15 show activity patterns that closely overlap *Shox2* in the maxillary-mandibular (MXP-MDP) and MNP compartments, respectively (**Fig. 4C**). In addition, transgenic validation of other predicted GDEs identified multiple brain and cranial nerve activities (**Figs. 1B, C**), but with the exception of the nodose ganglion no obvious alterations in spatial *Shox2* expression in these tissues were observed in *S2*GD$^{\Delta/\Delta}$ embryos (**Fig. 4A**). Hereby, the presence of multiple brain enhancers with overlapping activities in the diencephalon, midbrain and hindbrain, both inside and outside the gene desert (**Fig. S5**, Vista Enhancer Browser), suggests that removal of brain-specific enhancers might be buffered by redundant enhancer interactions[38]. Strikingly, *in situ* hybridization (ISH) analysis in *S2*GD$^{\Delta/\Delta}$ embryos at E10.5 revealed absence of *Shox2* transcripts in the sinus venosus (SV) myocardium comprising the SAN region (**Fig. 4A**). Quantitative expression profiling in *S2*GD$^{\Delta/\Delta}$ embryonic hearts at E11.5 furthermore revealed severe downregulation of cardiac *Shox2* transcripts (**Fig. 4B**). Together, these findings indicate that the embryonic lethality observed in *S2*GD$^{\Delta/\Delta}$ embryos is a result of depleted *Shox2* in the SV myocardium encompassing SAN pacemaker cells[22,39], potentially due to the deletion of a cardiac SV enhancer located in the gene desert. However, rather surprisingly, our previous transgenic validation of epigenomic predictions did not reveal regulatory modules driving reproducible reporter activity in cardiac tissues (**Fig. 1B, C**).

At E11.5, Shox2 protein is specifically localized in the sinus venosus (SV) myocardium which includes the venous valves and the SAN pacemaker cell population[40] (**Fig 5A, B**). In accordance with the absence of *Shox2* transcripts in the SV at E10.5 (**Fig. 4A**), we found that in E11.5 *S2*GD$^{\Delta/\Delta}$ embryos Shox2 is largely depleted in cells of the SV comprising the SAN pacemaker myocardium marked by Hcn4[40], while it is retained to some degree in the mandible (**Fig. 5A, B**). As

Shox2 gene inactivation leads to embryonic lethality due to a SAN pacemaker defect[22,23], our results suggest that in *S2*GD$^{\Delta/\Delta}$ embryos SAN-specific loss of *Shox2* is responsible for the observed embryonic lethality phenotype (**Fig. S4C**), indicating the presence of one (or multiple) critical SV enhancers in the deleted gene desert region. In search of a cardiac enhancer located in the gene desert we then conducted ATAC-seq[41] from embryonic hearts at E11.5 to define genome-wide open chromatin signatures including potential *cis*-regulatory modules with cardiac and consequently SV enhancer activity at E11.5 (**Fig. 5C, D**). ATAC-seq peak calling analysis uncovered 10 elements within the deleted gene desert region which were significantly enriched for open chromatin (**Fig. 5C and Table S8**). Four of these elements co-localized with regions enriched for H3K27ac in the heart at E11.5 and were identified as part of our initial epigenomic analysis (GDE7, GDE10, GDE11 and GDE12) (**Fig. 1B**). As none of these elements drove reproducible *LacZ* reporter activity in cardiac regions at E11.5 (**Fig. 1C**), we also validated the remaining six gene desert elements with significant open chromatin signatures (+224kb, +283kb, +326kb, +389kb, +405kb, +520kb) using transgenic reporter assays at E11.5 (**Fig. 5C**). Strikingly, the element located 326kb downstream of the *Shox2* TSS was the only one to drive reproducible *LacZ* reporter expression in the heart and indeed its activity co-localized with *Shox2* in the SV myocardium (**Fig. 5B, C, E, F**). To refine the genomic sequence driving SV enhancer activity we then also validated a second element (termed +325kb) partially overlapping the +326kb enhancer in a block of conserved sequence marked by low ATAC-seq signal (**Fig. 5E**). Remarkably, the +325kb region showed identical reporter activity overlapping *Shox2* expression in the SV at E11.5, indicating that SV enhancer activity is restricted to the 1.5kb region of overlap (**Fig. 5E**). Interestingly also, the conserved sequence in the region of overlap harbors a binding motif of the Tbx5 transcription factor ($p<0.001$, JASPAR CORE vertebrates collection, based on PWMScan[43]) (**Fig. S6**), a presumptive upstream regulator of *Shox2* in SAN pacemaker cells[39]. Together, these results identify a gene desert enhancer with specific activity in the

SV, whose absence in *S2*GD$^{\Delta/\Delta}$ embryos potentially accounts for the embryonic lethal loss of *Shox2* expression in cardiac SAN pacemaker cells.

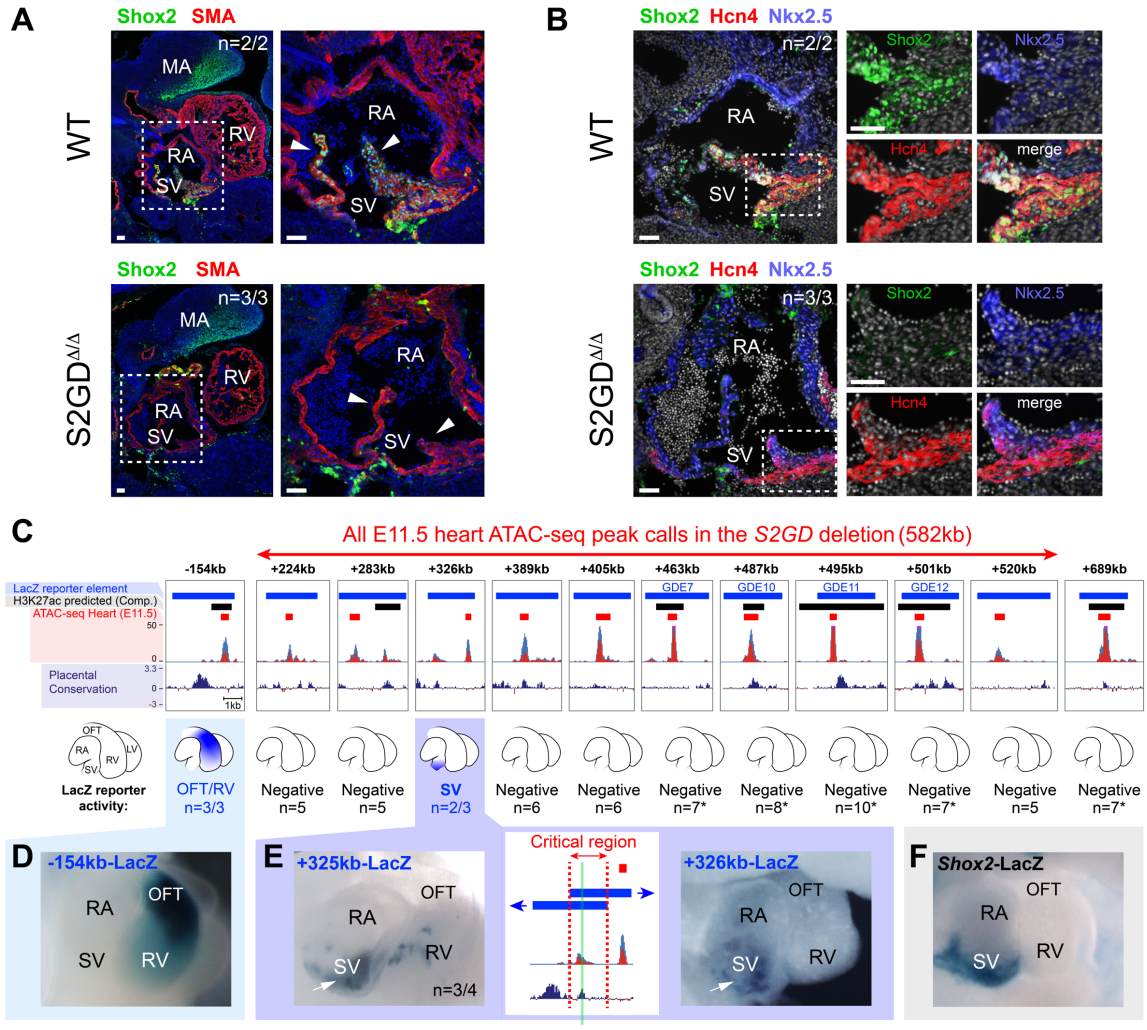

Figure 5. Identification of a sinus venosus (SV) gene desert enhancer implicated in critical regulation of *Shox2* in sinoatrial cardiac pacemaker cells. (**A**) Compared to abundant Shox2 protein (green) in the SV of wildtype (WT) hearts, Shox2 is depleted in myocardial cells of the SV including the venous valves (white arrowheads) in *S2*GD$^{\Delta/\Delta}$ embryos at E11.5. In contrast, Shox2 remains present, although reduced, in the mandibular arch (MA). Smooth muscle actin marks the myocardium (red). White arrowheads point to the venous valves. Nuclei are stained blue. (**B**) Co-localization of Shox2 (green), Hcn4 (red) and Nkx2-5 (blue, marker of myocardial progenitors) in hearts of WT and S2GD$^{\Delta/\Delta}$ embryos at E11.5. Shox2 is lost from the Hcn4-marked SAN pacemaker myocardium in S2GD$^{\Delta/\Delta}$ embryos (dashed outline shown at higher magnification). Nuclei are shown gray. "n" indicates number of embryos per genotype analyzed, with similar results. Scale bars, 50µm. (**C**) Top: UCSC browser schemes of all gene desert elements containing cardiac ATAC-seq peaks (red bars) at E11.5. Read enrichment of replicate samples is shown in a stacked configuration (blue: replicate 1; red: replicate 2). Black bars indicate putative cardiac enhancer elements enriched for H3K27ac in hearts at E11.5[31,42]. Blue bars represent elements used for *LacZ* reporter transgenesis including at least one flanking region of conserved genomic sequence (as indicated by the Placental Mammal base-wise conservation track by PhyloP). Distance of each genomic region from the *Shox2* TSS is indicated (-, upstream; +, downstream). Bottom: Schematics of the mouse embryonic heart at E11.5 (side view) illustrating reproducible LacZ reporter activities (blue). "n" indicates the fraction of transgenic embryos with reproducible staining in the heart over the total number of transgenic embryos analyzed. Single numbers represent transgenic embryos without (reproducible) staining in the heart. Asterisk indicates enhancers with reproducible activities in non-heart tissues (**Fig. 1C, S5**). Identified cardiac *LacZ* enhancer activities (**D**, **E**) are compared to cardiac *Shox2* expression (*Shox2-LacZ*) localized in the SV (white arrows) (**F**). The region shared between the +325kb and +326kb SV enhancer elements is delineated by red lines, shows reduced conservation and ATAC-seq signal, and harbors a significant Tbx5 motif (green line) (**Fig. S6**). RA, right atrium, RV, right ventricle, LV, left ventricle, OFT, outflow tract.

DISCUSSION

The majority of gene deserts located in the vicinity of developmental regulators are considered evolutionarily ancient and stable, and typically harbor a large number of conserved elements with predicted *cis*-regulatory signatures[2]. Assessment of extensive *cis*-regulatory regions flanking a number of developmental genes, such as the cluster of *HoxD* genes or *Sox9*, has demonstrated the biological relevance of gene deserts and non-coding chromatin domains in regulation of developmental gene expression[6,13,44]. Nevertheless, the precise functional contributions of gene deserts near a majority of critical developmental regulators remains unexplored. Here, we characterize the *cis*-regulatory output and functions of a gene desert downstream of the *Shox2* transcriptional regulator. Our results reveal the *cis*-regulatory complexity underlying transcriptional orchestration of a key developmental gene with important implications for functional interpretation of enhancer-gene interactions and of the evolution of gene deserts into pleiotropic expression control units.

A reservoir of transcriptional enhancers essential for pleiotropic *Shox2* expression.

Enhancers with tissue- and stage-specific biological functions typically exhibit restricted temporal activity windows[31]. To pinpoint the robust *cis*-regulatory activities embedded in the gene desert and involved in the regulation of *Shox2*, we chose an unbiased approach based on the presence of the active enhancer mark H3K27ac across a range of embryonic stages[28]. While it remains challenging to predict precise temporal and spatial enhancer activities from bulk tissues *in vivo*, the stringent and unbiased nature of our analysis identified 12 novel gene desert enhancers (from 16 predictions) with specific subregional activities in *Shox2*-expressing tissues, such as limb, craniofacial compartments, cranial nerve and brain cell populations. In addition, our 4C-seq chromatin conformation capture from limb in combination with subsequent transgenic analysis starts to delineate the likely critical cluster of limb enhancers orchestrating *Shox2*-mediated stylopod formation. This cluster is reminiscent of a multipartite enhancer ensemble, such as the one regulating the *Indian Hedgehog* (*Ihh*) gene, or the *HoxD* cluster genes, in multiple tissues and due to additively acting enhancers with

partially overlapping activities[45,46]. While many developmental enhancers with overlapping counterparts are known to exert specific tasks, they also exhibit partially redundant functions serving as a regulatory buffer to ensure phenotypic robustness[24,47,48]. We observe similar transcriptional resilience of spatial *Shox2* expression following CRISPR-mediated removal of the gene desert, in particular in the limb and brain. The functional significance of the gene desert for limb development is corroborated by quantitative reduction of *Shox2* in absence of this regulatory landscape, leading to severely affected stylopod development in a genetically sensitized background. The cumulative removal of enhancers via deletion of the gene desert further allowed functional assessment of fundamental *cis*-regulatory activities in other tissues. Most notably, in absence of the gene desert, we observed a depletion of *Shox2* transcripts in the sinus venosus (or inflow tract), comprising the SAN pacemaker population and most likely cause of the observed embryonic lethality phenotype[22,49]. Furthermore, our results demonstrate that craniofacial *Shox2* expression and in particular *Shox2* transcripts in the mandibular and nasal processes critically depend on the presence of the gene desert. A recent study uncovered that human (and mouse) extreme long-range enhancers located in a large gene desert upstream of *Sox9* are acting across nearly 1.5 Mb to regulate Sox9 expression in craniofacial regions, such as the nasal, maxillary and mandibular processes[50]. Similarly, our study identifies gene desert enhancers with activities in nasal and maxillary-mandibular regions, the latter likely critical for the formation of the temporomandibular joint[18].

Cis-regulatory control of cardiac _Shox2_ essential for embryonic viability.

Alongside other TFs, such as *Isl1* or *Tbx3*, *Shox2* in mice is a key regulator of cardiac pacemaker cells of the sinoatrial node (SAN), the primary pacemaker of the heart[39]. While the genetic hierarchies and transcriptional cell states orchestrating cardiac pacemaker development have been characterized, the genomic *cis*-regulatory modules underlying this process have remained largely unexplored. Here we demonstrate an essential regulatory requirement of the gene desert for embryonic viability at mid-gestation by maintaining *Shox2* transcription in the cardiac sinus venosus (SV) encompassing the SAN pacemaker myocardium. In a very recent, independently published study, van Eif et al. report

complementary observations at the same locus[49]. In this study, they performed ATAC-seq on SAN-like pacemaker cells differentiated from human pluripotent embryonic stem cells (hESC) and Hcn4^{+} SAN cells of newborn mice to delineate the *cis*-regulatory modules controlling the expression of TFs promoting cardiac pacemaker cell fate, such as *TBX3*, *ISL1* and *SHOX2*. While the authors initially focused on human *cis*-regulatory landscapes near these genes, they used CRISPR/Cas9 deletions to investigate the function of homologous SAN-specific accessible chromatin regions in the *Shox2* and *Tbx3* loci in mouse embryos[49]. In particular, within the 582kb gene desert domain deleted here, their study narrows the critical space down to a ~250kb region. Consistent with our observations in embryonic hearts of *S2*GD$^{\Delta/\Delta}$ embryos (**Fig. 4A, 5A, B**), Van Eif et al. confirm the embryonic lethality phenotype in their embryos lacking the 250kb region and show that the lethality is likely a result of a hypoplastic SAN (and venous valves) due to loss of Shox2 protein in the SV[49]. In addition, through our targeted exploration we now define a 1.5kb element located within this 250kb window and driving transcriptional activity specifically in the *Shox2* domain of the SV (**Fig. 6C, E**), potentially acting as critical enhancer controlling *Shox2* in SAN pacemaker cells. Further enhancer deletion analyses will uncover whether *Shox2* transcription in the SAN is controlled by a single *cis*-regulatory unit or is shielded by multiple enhancers as it could be the case in human embryos[49].

A blueprint for disease-relevant enhancer repertoires controlling human *SHOX*.

Together, our findings significantly expand on former analyses that identified a panel of mouse (and human) *Shox2* enhancers with activities mostly restricted to limb and hindbrain (Vista Enhancer Browser)[25,27]. Interestingly, such tissue-specific activities were also found to be conserved in distinct elements of the similar-sized gene desert flanking the human *SHOX* gene[25,51]. Disruption of enhancers within the gene desert downstream of *SHOX* represents the likely mechanistic cause of Léri-Weill dyschondrosteosis (LWD) and idiopathic short stature (ISS) syndromes in a significant fraction of cases[52] and *SHOX* haploinsufficiency is directly associated with the skeletal abnormalities observed in Turner syndrome and LWD[53,54]. One study has also found a link between neurodevelopmental disorders and microduplications at the *SHOX* locus, suggesting that such

perturbations may alter neural development or function[55]. In humans, *SHOX2* represents the closely related paralog of *SHOX* and is encoded in all vertebrate genomes. However, while many functional aspects of human *SHOX2* remain unknown, a link between heterozygous *SHOX2* mutations and SAN dysfunction as well as familial/early onset atrial fibrillation has recently been demonstrated[56,57]. Rodents have lost their *SHOX* gene in the course of evolution and therefore entirely rely on the function of *Shox2*, which features an identical DNA-interacting homeodomain and is replaceable by human *SHOX* in a mouse knock-in line[58]. Thus, in light of the overlapping expression patterns and critical functions of mouse *Shox2* and human *SHOX,* as well as the presence of a gene desert downstream of both genes, our results provide a blueprint for the investigation of the regulatory control of pleiotropic *SHOX* expression, especially in those tissues where both genes are expressed during development: the hindbrain, thalamus, pharyngeal arches and limbs[59,60]. It will be particularly interesting to determine whether "orthologous" cardiac, craniofacial, neural and/or limb enhancers exist, and whether human *SHOX* enhancers share motif content or other enhancer grammar characteristics[61] with mouse *Shox2* enhancers. Indeed, human and mouse orthologs of a highly conserved enhancer located 160kb/47kb downstream of human *SHOX* and mouse *Shox2*, respectively, were found to drive overlapping activities in the hindbrain[25]. Such enhancers presumably originate from a single ancestral *SHOX* locus, preceding the duplication of *SHOX* and *SHOX2* paralogs and are therefore considered evolutionary ancient. Within this context, future comparative studies should search for deeply conserved orthologs of *SHOX* and *SHOX2* enhancers in basal chordates such as amphioxus, which express their single *Shox* gene in the developing hindbrain[62]. The recent identification of orthologous *Islet* gene enhancers in sponges and vertebrates[63] demonstrate the promise of such an approach.

MATERIALS AND METHODS

Experimental Design

All animal work at Lawrence Berkeley National Laboratory (LBNL) was reviewed and approved by the LBNL Animal Welfare Committee. Knockout and transgenic mice were housed at the Animal Care Facility (the ACF) at LBNL. Mice were monitored daily for food and water intake, and animals were inspected weekly by the Chair of the Animal Welfare and Research Committee and the head of the animal facility in consultation with the veterinary staff. The LBNL ACF is accredited by the American Association for the Accreditation of Laboratory Animal Care International (AAALAC). Transgenic mouse assays and enhancer knock-outs at LBNL were performed in *Mus musculus* FVB strain mice. Animal work at the University of Calgary involving the production, housing and analysis of transgenic mouse lines shown in **Figs. 2 and S3**, as well as breeding and skeletal analysis of S2GD mice, was approved by the Life and Environmental Sciences Animal Care Committee (LESACC). All experiments with mice were performed in accordance with Canadian Council on Animal Care guidelines as approved by the University of Calgary LESACC, Protocol # AC13-0053. The following developmental stages were used in this study: embryonic day E10.5, E11.5, E12.5, E13.5 and newborn mice (the latter only for skeletal preparations). Animals of both sexes were used in these analyses. Sample size selection and randomization strategies were conducted as follows:

Transgenic mouse assays. Sample sizes were selected empirically based on our previous experience of performing transgenic mouse assays for >3,000 total putative enhancers (VISTA Enhancer Browser: https://enhancer.lbl.gov/). Mouse embryos were excluded from further analysis if they did not encode the reporter transgene or if the developmental stage was not correct. All transgenic mice were treated with identical experimental conditions. Randomization and experimenter blinding were unnecessary and not performed.

Knockout mice. Sample sizes were selected empirically based on our previous studies[24,38]. All phenotypic characterization of knockout mice employed a matched littermate selection strategy. Analyzed *S2GD* knockout embryos and mice described in this paper resulted from crossing

heterozygous gene desert deletion (*S2GD*$^{\Delta/+}$) mice together to allow for the comparison of matched littermates of different genotypes. Embryonic samples used for *in situ* hybridizations and quantitative gene expression profiling were dissected and processed blind to genotype.

Hi-C data re-analysis

Raw reads from Hi-C on mouse embryonic stem cells (mESCs) from Bonev *et al*., 2017, available on GEO (GSE96107), were reprocessed using HiCUP v.0.6.1. Valid pairs used to generate the Hi-C map in **Fig. 1A** are available on GEO (GSE161259) and the code used to generate the representation of the extended *Shox2* TAD is available on https://github.com/lldelisle/Hi-C_reanalysis_Bonev_2017. The matrix heatmaps were plotted using pygenometracks[64].

In vivo transgenic *LacZ* reporter analysis

For all elements tested, except PLEs, transgenic mouse *LacZ* reporter assays were conducted as previously described[31,65] and the related primer sequences and genomic coordinates are listed in **Tables S2 and S8**. Predicted enhancer elements were PCR-amplified from mouse genomic DNA (Clontech) and cloned into an Hsp68-*LacZ* expression vector[31]. PLE elements were amplified via PCR from bacterial artificial chromosomes containing the appropriate mouse genomic DNA (**Table S4)** then cloned into the *βlacz* plasmid, which contains a minimal human *β*-globin promoter-*LacZ* cassette, as described[25]. Due to their large size, PLE3 (10,351 bp) and PLE5 (9,473 bp) were amplified with the proofreading polymerase in the SequalPrep™ Long PCR Kit (Invitrogen). Permanent transgenic lines (**Fig. S3**) were produced at the University of Calgary Centre for Mouse Genomics by pronuclear injection of DNA constructs into CD-1 single-cell stage embryos as described[66]. Male founder animals (or male F1 progeny produced from transgenic females) were crossed to CD-1 females to produce transgenic embryos which were stained with X-gal by standard techniques[65].

4C-seq

For each of two biological replicates, proximal forelimbs were dissected in PBS from 10-12 E12.5 CD-1 embryos using the cutting pattern shown in the inset of **Fig. 2B**. Tissue was prepared for 4C-

seq as described[67]. Cells were dissociated by incubating the pooled tissue in 250µl PBS supplemented with 10% fetal fetal calf serum (FCS) and 1 mg/ml collagenase (Sigma) for 45 minutes at 37° C with shaking at 750 rpm. The solution was passed through a cell strainer (Falcon) to obtain single cells which were fixed in 9.8 ml of 2% formaldehyde in PBS/10% FCS for 10 minutes at room temperature, and lysed and 4C-seq performed[68]. Libraries were prepared by overnight digestion with NlaIII (New England Biolabs (NEB)) and ligation for 4.5 hours with 100 units T4 DNA ligase (Promega, #M1794) under diluted conditions (7 ml), followed by de-crosslinking overnight at 65°C after addition of 15ul of 20mg/ml proteinase K. After phenol/chloroform extraction and ethanol precipitation the samples were digested overnight with the secondary enzyme DpnII (NEB) followed again by phenol/chloroform extraction and ethanol precipitation purification, and ligated for 4.5 hours in a 14 ml volume. The final ligation products were extracted and precipitated as above followed by purification using Qiagen nucleotide removal columns. For each viewpoint, libraries were prepared with 100 ng of template in each of 16 separate PCR reactions using the Roche, Expand Long Template kit with primers incorporating Illumina adapters. Viewpoint and primer details are presented in **Table S3**. PCR reactions for each viewpoint were pooled and purified with the Qiagen PCR purification kit and sequenced with the Illumina HiSeq to generate single 100bp reads. Demultiplexed reads were mapped and analyzed with the 4C-seq module of the HTSstation pipeline as described[69]. Results are shown in UCSC browser format as normalized reads per fragment after smoothing with an 11-fragment window and mapped to mm10 (**Figs. 2B, S2C**). Raw and processed (bedgraph) sequence files are available under GEO accession number GSE161194.

Generation of gene desert knock-out mice using CRISPR/Cas9

Mouse strains encoding the 582kb gene desert deletion centromeric to the *Shox2* gene body were engineered using *in vivo* CRISPR/Cas9 editing, as previously described with minor modifications[24]. Pairs of single guide RNAs (sgRNAs) targeting genomic sequence 5' and 3' of the gene desert were designed using CHOPCHOP[70] (see **Table S5** for sgRNA sequences and coordinates). To generate the deletion a mix containing Cas9 mRNA (final concentration of 100 ng/ul) and two sgRNAs (25

ng/ul each) in injection buffer (10 mM Tris, pH 7.5; 0.1 mM EDTA) was injected into the cytoplasm of single-cell FVB strain mouse embryos. Founder (F0) mice were genotyped via PCR utilizing High Fidelity Platinum Taq Polymerase (Thermo Fisher) to identify the desired deletion breakpoints generated via NHEJ (see **Fig. S4A** and **Table S6** for genotyping strategy, primer sequences and PCR amplicons). Sanger sequencing was used to identify and confirm deletion breakpoints in F0 and F1 mice (**Fig. S4A**).

In situ hybridization

For assessment of spatial gene expression changes in mouse embryos, whole mount *in situ* hybridization using digoxigenin-labeled antisense riboprobes was performed as previously described[71]. At least three independent embryos were analyzed for each genotype. Embryonic tissues were imaged using a Leica MZ16 microscope coupled to a Leica DFC420 digital camera.

Quantitative real-time PCR (qPCR)

Isolation of RNA from microdissected embryonic tissues at E11.5 was performed using the Ambion RNAqueous Total RNA Isolation Kit (Life Technologies) according to the manufacturer's protocol. RNA was then subjected to RNase-free DNase (Promega) treatment and reverse transcribed using SuperScript III (Life Technologies) with poly-dT priming according to manufacturer instructions. qPCR was conducted on a LightCycler 480 (Roche) using KAPA SYBR FAST qPCR Master Mix (Kapa Biosystems) according to manufacturer instructions. qPCR primers (*Shox2*, *Rsrc1, Actb*) were described previously[24]. Relative gene expression levels were calculated via the $2^{-\Delta\Delta C_T}$ method, normalized to the *Actb* housekeeping gene, and the mean of wild-type control samples was set to 1.

Skeletal preparations

Euthanized newborn mice were eviscerated, skinned and fixed in 1 % acetic acid in EtOH for 24 hours. Cartilage was stained overnight with 1 mg/mL Alcian blue 8GX (Sigma) in 20% acetic acid in EtOH. After washing in EtOH for 12 hours and treatment with 1.5 % KOH for three hours, bones were stained in 0.15 mg/mL Alizarin Red S (Sigma) in 0.5 % KOH for four hours, followed by de-staining in 20 % glycerol, 0.5 % KOH.

ENCODE H3K27ac ChIP-seq and mRNA-seq analysis

To establish a heatmap revealing putative enhancers and their temporal activities within the *Shox2* TAD interval, a previously generated catalog of strong enhancers identified using ChromHMM[72] across mouse development was used[28]. Briefly, calls across 66 different tissue-stage combinations were merged and H3K27ac signals quantified as log2-transformed RPKM. Estimates of statistical significance for these signals were associated to each region for each tissue-stage combination using the corresponding H3K27ac ChIP-seq peak calls. These were downloaded from the ENCODE Data Coordination Center (DCC) (http://www.encodeproject.org/, see **Table S1**, *sheet 3* for the complete list of sample identifiers). To this purpose, short reads were aligned to the mm10 assembly of the mouse genome using bowtie (ref), with the following parameters: *-a -m 1 -n 2 -l 32 -e 3001*. Peak calling was performed using MACS v1.4, with the following arguments: *--gsize=mm --bw=300 --nomodel --shiftsize=100*[73]. Experiment-matched input DNA was used as control. Evidence from two biological replicates was combined using IDR (https://www.encodeproject.org/data-standards/terms/). The *q*-value provided in the replicated peak calls was used to annotate each putative enhancer region defined above. In case of regions overlapping more than one peak, the lowest *q*-value was used. RNA-seq raw data was downloaded from the ENCODE DCC (http://www.encodeproject.org/, see **Table S1**, *sheet 3* for the complete list of sample identifiers).

Immunofluorescence (IF)

IF was performed as previously described[24]. Briefly, mouse embryos at E11.5 were isolated in cold PBS and fixed in 4% PFA for 2–3h. After incubation in a sucrose gradient and embedding in a 1:1 mixture of 30% sucrose and OCT compound, sagittal 10µm frozen tissue sections were obtained using a cryostat. Selected cryo-sections were then incubated overnight with the following primary antibodies: anti-Shox2 (1:300, Santa Cruz JK-6E, sc-81955), anti-SMA-Cy3 (1:250, Sigma, C6198), anti-Hcn4 (1:500, Thermo Fisher, MA3-903) and anti-Nkx2.5 (1:500, Thermo Fisher, PA5-81452). Goat-anti mouse, goat anti-rabbit and donkey anti-rat secondary antibodies conjugated to Alexa Fluor 488, 568, or 647 (1:1,000, Thermo Fisher Scientific) were used for detection. Hoechst 33258 (Sigma-

Aldrich) was utilized to counterstain nuclei. A Zeiss AxioImager fluorescence microscope in combination with a Hamamatsu Orca-03 camera was used to acquire fluorescent images.

ATAC-seq and data processing

ATAC-seq was performed as described[74] with minor modifications. Per replicate, pairs of wildtype mouse embryonic hearts at E11.5 were micro-dissected in cold PBS and cell nuclei were dissociated in Lysis buffer using a douncer. Approx. 50'000 nuclei were then pelleted at 500 RCF for 10 min at 4°C and resuspended in 50 µL Transposition reaction mix containing 25 µL Nextera 2x TD buffer and 2.5 µL TDE1 (Nextera Tn5 Transposase; Illumina) (cat. no. FC-121-1030) followed by incubation for 30 minutes at 37°C with shaking. The reaction was purified using the Qiagen MinElute PCR purification kit and amplified using defined PCR primers[41]. ATAC-seq libraries were purified using the Qiagen MinElute PCR purification kit (ID: 28004), quantified by the Qubit Fluorometer with the dsDNA HS Assay Kit (Life Technologies) and quality assessed using the Agilent Bioanalyzer high sensitivity DNA analysis assay. Libraries were pooled and sequenced using single end 50 bp reads on a HiSeq 4000 (Illumina).

ATAC-seq data analysis from wild-type heart replicate samples at E11.5 followed ENCODE2 specifications (May 2019, https://www.encodeproject.org/atac-seq): CASAVA v1.8.0 (Illumina) was utilized to demultiplex data, and reads with CASAVA 'Y' flag (purity filtering) were discarded. Adaptor trimming (cutadapt_v1.1) (https://cutadapt.readthedocs.io/) was used with parameter '-e 0.1 -m 5'. For read mapping and peak calling, bowtie2 was used[75] (version 2.2.6) with parameters '-X2000 --mm –local'. bowtie2 aligned 66% of the reads uniquely, and 35% to more than one location. Reads were aligned to both GRCm38/mm10 and NCBI37/mm9 reference genomes with GENCODE annotations, allowing for multi-mapped reads. Unmapped failed reads, duplicates, and low-quality reads (MAPQ = 255) were removed using SAMtools[76] (v1.7) and Picard (https://broadinstitute.github.io/picard) (v1.126). For each sample, 20-25 million reads were retrieved after all quality checks. Peak calling was then performed using MACSv2[73,77] (v2.1.0) with p-value<0.01, and a smoothing window of 150bp. Finally, peaks were filtered in two steps and

resulted in 100-200k peaks per sample: (a) excluding the 164 blacklisted coordinates from ENCODE[78] mm10 (ENCFF547MET), and (b) overlap across replicates and pseudo replicates. To visualize signal obtained for each of the replicates a UCSC track hub was generated for the mm9 and mm10 genomes in the Genome Browser (GSE160127).

Data availability

Limb 4C-seq and heart ATAC-seq datasets are available in the NCBI GEO database with the accession codes GSE161194 and GSE160127, respectively. All relevant transgenic *in vivo* enhancer data is available at the Vista Enhancer Browser (https://enhancer.lbl.gov) (see **Table S8** for Vista Enhancer IDs). Correspondence and requests for materials should be addressed to J.C. (jacobb@ucalgary.ca) or M.O. (marco.osterwalder@dbmr.unibe.ch).

Competing interests

The authors declare no competing financial interests.

Acknowledgements

This work was supported by Swiss National Science Foundation (SNSF) grant PCEFP3_186993 (to M.O.), a Discovery Grant (RGPIN/355731-2013) from the Natural Sciences and Engineering Research Council of Canada (to J.C.) and National Institutes of Health grants R01HG003988, U54HG006997, R24HL123879 and UM1HL098166 (to A.V. and L.A.P.). J.L-R. is supported by the MICINN grants BFU2017-82974-P and MDM-2016-0687 (Unidad de Excelencia María de Maeztu institutional grant). GA is supported by Swiss National Science Foundation Grant PP00P3_176802. F.D. is supported by a SNSF postdoc.mobility fellowship (P400PB_194334). We thank L. Lopez-Delisle for sharing re-analyzed Hi-C data and D. Duboule for hosting and supporting 4C-Seq experiments as well as training in his laboratory. We are grateful to C. Fielding at the Clara Christie Centre for Mouse Genomics for pronuclear injections conducted at the University of Calgary. We thank the members of the L.A.P., A.V., and D.E.D. group for technical advice and useful comments on the manuscript. Research at the E.O. Lawrence Berkeley National Laboratory was performed under Department of Energy Contract DE-AC02-05CH11231, University of California.

Author contributions

S.A.-O., B.J.M, J.C. and M.O. conceived the study. S.A.-O., B.J.M, F.D., R.H., J.A.A., J.C. and M.O. performed experimental work including transgenic analyses and genome editing experiments. S.A.-

O., E.R-C., A.L., G.A. and J.C. conducted 4C-seq experiments and analysis. V.T. performed the *in situ* hybridization analysis under the supervision of J.L.-R. E.R.-C., G.K. and I.B. performed bioinformatic analyses. T.A.F and C.S.S. performed skeletal phenotyping. C.S.N, I.P.-F. and S.T. conducted pro-nuclear injections. D.E.D., A.V. and L.A.P. provided project funding and support. J.C. and M.O. provided project funding and wrote the manuscript with input from the remaining authors.

REFERENCES

1. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
2. Ovcharenko, I. *et al.* Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15,** 137–145 (2005).
3. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302,** 413–413 (2003).
4. Catarino, R. R. & Stark, A. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev.* **32,** 202–223 (2018).
5. Nobrega, M. A., Zhu, Y., Plajzer-Frick, I., Afzal, V. & Rubin, E. M. Megabase deletions of gene deserts result in viable mice. *Nature* **431,** 988–993 (2004).
6. Montavon, T. *et al.* A regulatory archipelago controls Hox genes transcription in digits. *Cell* **147,** 1132–1145 (2011).
7. Robson, M. I., Ringel, A. R. & Mundlos, S. Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3D. *Mol. Cell* **74,** 1110–1122 (2019).
8. Marinić, M., Aktas, T., Ruf, S. & Spitz, F. An integrated holo-enhancer unit defines tissue and gene specificity of the Fgf8 regulatory landscape. *Dev. Cell* **24,** 530–542 (2013).
9. Schoenfelder, S. & Fraser, P. Long-range enhancer-promoter contacts in gene expression control. *Nat. Rev. Genet.* **55,** 5 (2019).
10. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).
11. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161,** 1012–1025 (2015).
12. Lupiáñez, D. G., Spielmann, M. & Mundlos, S. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends in Genetics* **32,** 225–237 (2016).
13. Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538,** 265–269 (2016).
14. Symmons, O. *et al.* The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. *Dev. Cell* **39,** 529–543 (2016).
15. Kragesteen, B. K. *et al.* Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. *Nat. Genet.* **50,** 1463–1473 (2018).
16. Rodríguez-Carballo, E. *et al.* The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. *Genes Dev.* **31,** 2264–2281 (2017).
17. Cobb, J., Dierich, A., Huss-Garcia, Y. & Duboule, D. A mouse model for human short-stature syndromes identifies Shox2 as an upstream regulator of Runx2 during long-bone development. *Proc. Natl Acad. Sci. USA* **103,** 4511–4515 (2006).
18. Gu, S., Wei, N., Yu, L., Fei, J. & Chen, Y. Shox2-deficiency leads to dysplasia and ankylosis of the temporomandibular joint in mice. *Mech. Dev.* **125,** 729–742 (2008).
19. Yu, L. *et al.* Shox2-deficient mice exhibit a rare type of incomplete clefting of the secondary palate. *Development* **132,** 4397–4406 (2005).
20. Rosin, J. M., Kurrasch, D. M. & Cobb, J. Shox2 is required for the proper development of the facial motor nucleus and the establishment of the facial nerves. *BMC Neurosci.* **16,** 39 (2015).
21. Scott, A. *et al.* Transcription factor short stature homeobox 2 is required for proper development of tropomyosin-related kinase B-expressing mechanosensory neurons. *J Neurosci.* **31,** 6741–6749 (2011).
22. Blaschke, R. J. *et al.* Targeted mutation reveals essential functions of the homeodomain transcription factor Shox2 in sinoatrial and pacemaking development. *Circulation* **115,** 1830–1838 (2007).
23. Espinoza-Lewis, R. A. *et al.* Shox2 is essential for the differentiation of cardiac pacemaker cells by repressing Nkx2-5. *Dev. Biol.* **327,** 376–385 (2009).
24. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554,** 239–243 (2018).
25. Rosin, J. M., Abassah-Oppong, S. & Cobb, J. Comparative transgenic analysis of enhancers from the human SHOX and mouse Shox2 genomic regions. *Hum. Mol. Genet.* **22,** 3063–3076 (2013).
26. Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171,** 557–572.e24 (2017).
27. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35,** D88–92 (2007).
28. Gorkin, D. U. *et al.* An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583,** 744–751 (2020).
29. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–283 (2011).
30. Akerberg, B. N. *et al.* A reference map of murine cardiac transcription factor chromatin occupancy identifies dynamic and conserved enhancers. *Nat. Commun.* **10,** 4907–16 (2019).

31. Nord, A. S. *et al.* Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155,** 1521–1531 (2013).
32. Méndez-Maldonado, K., Vega-López, G. A., Aybar, M. J. & Velasco, I. Neurogenesis from Neural Crest Cells: Molecular Mechanisms in the Formation of Cranial Nerves and Ganglia. *Front. Cell Dev. Biol.* **8,** 635 (2020).
33. Rosin, J. M. *et al.* Mice lacking the transcription factor SHOX2 display impaired cerebellar development and deficits in motor coordination. *Dev. Biol.* **399,** 54–67 (2015).
34. Yu, L. *et al.* Shox2 is required for chondrocyte proliferation and maturation in proximal limb skeleton. *Dev. Biol.* **306,** 549–559 (2007).
35. Bobick, B. E. & Cobb, J. Shox2 regulates progression through chondrogenesis in the mouse proximal limb. *J. Cell. Sci.* **125,** 6071–6083 (2012).
36. Ye, W. *et al.* A unique stylopod patterning mechanism by Shox2-controlled osteogenesis. *Development* **143,** 2548–2560 (2016).
37. Logan, M. *et al.* Expression of Cre recombinase in the developing mouse limb bud driven by aPrxl enhancer. *Genesis* **33,** 77–80 (2002).
38. Dickel, D. E. *et al.* Ultraconserved Enhancers Are Required for Normal Development. *Cell* **172,** 491–499.e15 (2018).
39. van Eif, V. W. W., Devalla, H. D., Boink, G. J. J. & Christoffels, V. M. Transcriptional regulation of the cardiac conduction system. *Nat. Rev. Cardiol.* **15,** 617–630 (2018).
40. Christoffels, V. M., Smits, G. J., Kispert, A. & Moorman, A. F. M. Development of the Pacemaker Tissues of the Heart. *Circ. Res.* **106,** 240–254 (2010).
41. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10,** 1213–1218 (2013).
42. Dickel, D. E. *et al.* Genome-wide compendium and functional assessment of in vivo heart enhancers. *Nat. Commun.* **7,** 12923 (2016).
43. Ambrosini, G., Groux, R. & Bucher, P. PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics* **34,** 2483–2484 (2018).
44. Andrey, G. *et al.* A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science* **340,** 1234167 (2013).
45. Will, A. J. *et al.* Composition and dosage of a multipartite enhancer cluster control developmental expression of Ihh (Indian hedgehog). *Nat. Genet.* **49,** 1539–1545 (2017).
46. Rodríguez-Carballo, E. *et al*. Chromatin topology and the timing of enhancer function at the Hoxd locus. *biorxiv.org* doi:10.1101/2020.07.12.199109
47. Frankel, N. *et al.* Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466,** 490–493 (2010).
48. Perry, M. W., Boettiger, A. N., Bothma, J. P. & Levine, M. Shadow enhancers foster robustness of Drosophila gastrulation. *Curr. Biol.* **20,** 1562–1567 (2010).
49. van Eif, V. W. *et al.* Genome-Wide Analysis Identifies an Essential Human TBX3 Pacemaker Enhancer. *Circ. Res.* cell**115,** 1921 (2020).
50. Long, H. K. *et al.* Loss of Extreme Long-Range Enhancers in Human Neural Crest Drives a Craniofacial Disorder. *Cell Stem Cell* **27,** 765–783.e14 (2020).
51. Skuplik, I. *et al.* Identification of a limb enhancer that is removed by pathogenic deletions downstream of the SHOX gene. *Sci. Rep.* **8,** 14292 (2018).
52. Chen, J. *et al.* Enhancer deletions of the SHOX gene as a frequent cause of short stature: the essential role of a 250 kb downstream regulatory domain. *J. Med. Genet.* **46,** 834–839 (2009).
53. Rao, E. *et al.* Pseudoautosomal deletions encompassing a novel homeobox gene cause growth failure in idiopathic short stature and Turner syndrome. *Nat. Genet.* **16,** 54–63 (1997).
54. Shears, D. J. *et al.* Mutation and deletion of the pseudoautosomal gene SHOX cause Leri-Weill dyschondrosteosis. *Nat. Genet.* **19,** 70–73 (1998).
55. Tropeano, M. *et al.* Microduplications at the pseudoautosomal SHOX locus in autism spectrum disorders and related neurodevelopmental conditions. *J. Med. Genet.* **53,** 536–547 (2016).
56. Li, N. *et al.* A SHOX2 loss-of-function mutation underlying familial atrial fibrillation. *Int. J Med. Sci.* **15,** 1564–1572 (2018).
57. Hoffmann, S. *et al.* Functional Characterization of Rare Variants in the SHOX2 Gene Identified in Sinus Node Dysfunction and Atrial Fibrillation. *Front. Genet.* **10,** 648 (2019).
58. Liu, H. *et al.* Functional redundancy between human SHOX and mouse Shox2 genes in the regulation of sinoatrial node formation and pacemaking function. *J. Biol. Chem.* **286,** 17029–17038 (2011).
59. Clement-Jones, M. *et al.* The short stature homeobox gene SHOX is involved in skeletal abnormalities in Turner syndrome. *Hum. Mol. Genet.* **9,** 695–702 (2000).
60. Durand, C. *et al.* Alternative splicing and nonsense-mediated RNA decay contribute to the regulation of SHOX expression. *PLoS ONE* **6,** e18115 (2011).

61. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167,** 1170–1187 (2016).
62. Jackman, W. R. & Kimmel, C. B. Coincident iterated gene expression in the amphioxus neural tube. *Evol. Dev.* **4,** 366–374 (2002).
63. Wong, E. S. *et al.* Deep conservation of the enhancer regulatory code in animals. *Science* **370,** (2020).
64. Lopez-Delisle, L. *et al.* pyGenomeTracks: reproducible plots for multivariate genomic data sets. *Bioinformatics* (2020). doi:10.1093/bioinformatics/btaa692
65. Kothary, R. *et al.* Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice. *Development* **105,** 707–714 (1989).
66. Nagy, A., Gertsenstein, M., Behringer, R. R. & Vintersten, K. Manipulating the Mouse Embryo. (Cold Spring Harbor, N.Y. : *Cold Spring Harbor Laboratory Press*, 2003).
67. Noordermeer, D. *et al.* Temporal dynamics and developmental memory of 3D chromatin architecture at Hox gene loci. *eLife* **3,** e02557 (2014).
68. Noordermeer, D. *et al.* The dynamic architecture of Hox gene clusters. *Science* **334,** 222–225 (2011).
69. David, F. P. A. *et al.* HTSstation: a web application and open-access libraries for high-throughput sequencing data analysis. *PLoS ONE* **9,** e85879 (2014).
70. Labun, K. *et al.* CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res.* **8,** 302 (2019).
71. Tissières, V. *et al.* Gene Regulatory and Expression Differences between Mouse and Pig Limb Buds Provide Insights into the Evolutionary Emergence of Artiodactyl Traits. *Cell Rep.* **31,** 107490 (2020).
72. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9,** 215–216 (2012).
73. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
74. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109,** 21.29.1–21.29.9 (2015).
75. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
76. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
77. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protocols* **7,** 1728–1740 (2012).
78. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **9,** 9354–5 (2019).
79. Monti, R. *et al.* Limb-Enhancer Genie: An accessible resource of accurate enhancer predictions in the developing limb. *PLoS Comput. Biol.* **13,** e1005720 (2017).

Supplementary Figures

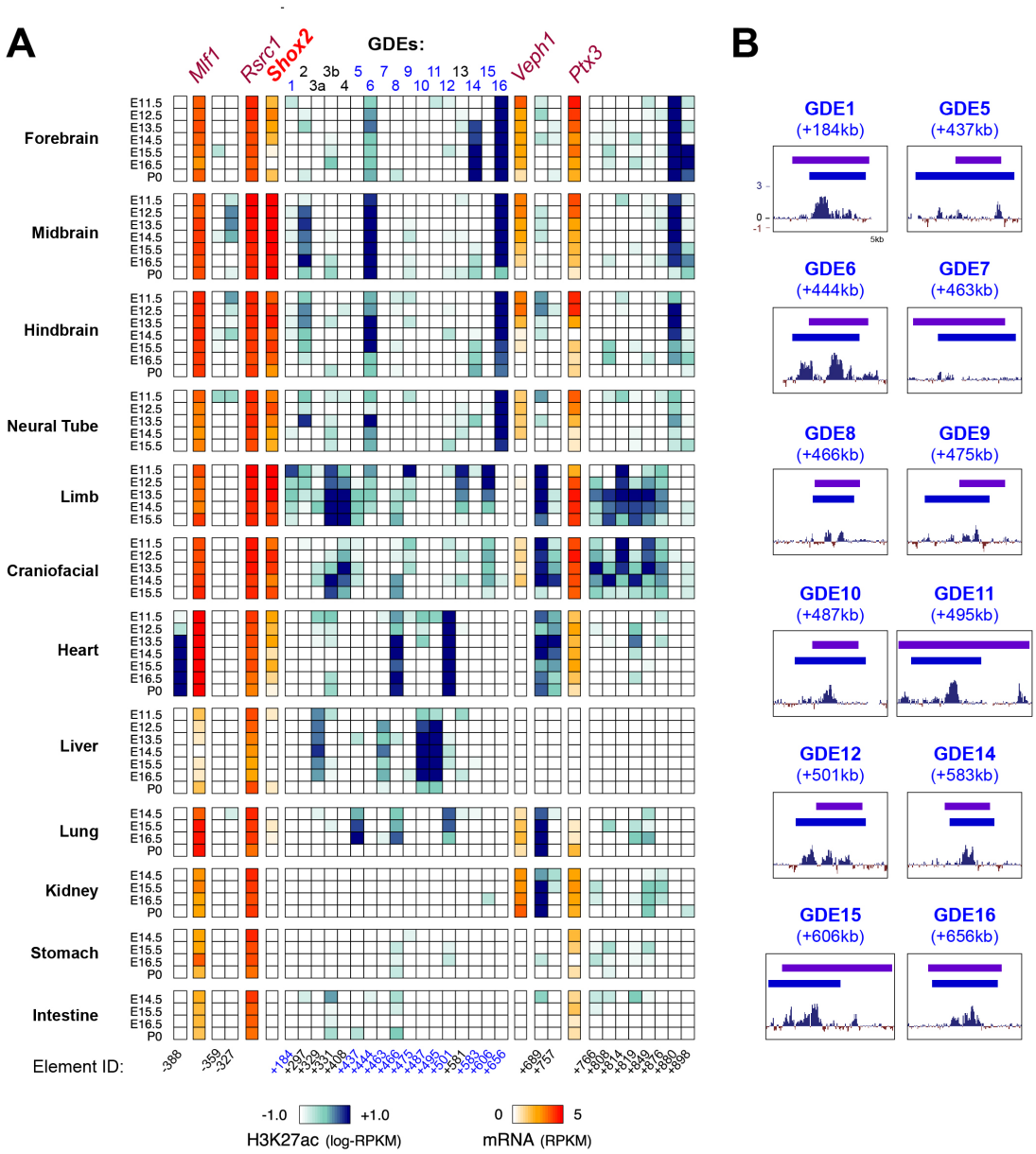

Figure S1. Prediction of spatio-temporal enhancer activities in the extended *Shox2* regulatory domain. (**A**) Complete heat map listing predicted enhancer elements (Element IDs) within the extended *Shox2* TAD region[10] across different time-points and tissues (see also **Fig. 1**), based on H3K27ac marks[28] and ChromHMM filtering (see **methods**). Blue shading indicates levels of H3K27ac ChIP-seq enrichment and red shades illustrate transcript levels (ENCODE RNA-seq datasets) of genes located in the region (**Table S1**). Predicted gene desert enhancer elements (GDEs) are indicated and those with confirmed enhancer activities at E11.5 (**Fig. 1C**) are marked blue. (**B**) Base-wise conservation track by PhyloP (Placental mammals) for each GDE with validated tissue-specific enhancer activities at E11.5 is shown (**Fig. 1C and Table S2**). Distance (in *kb*) from the *Shox2* TSS is indicated for each Element ID (-, upstream; +, downstream).

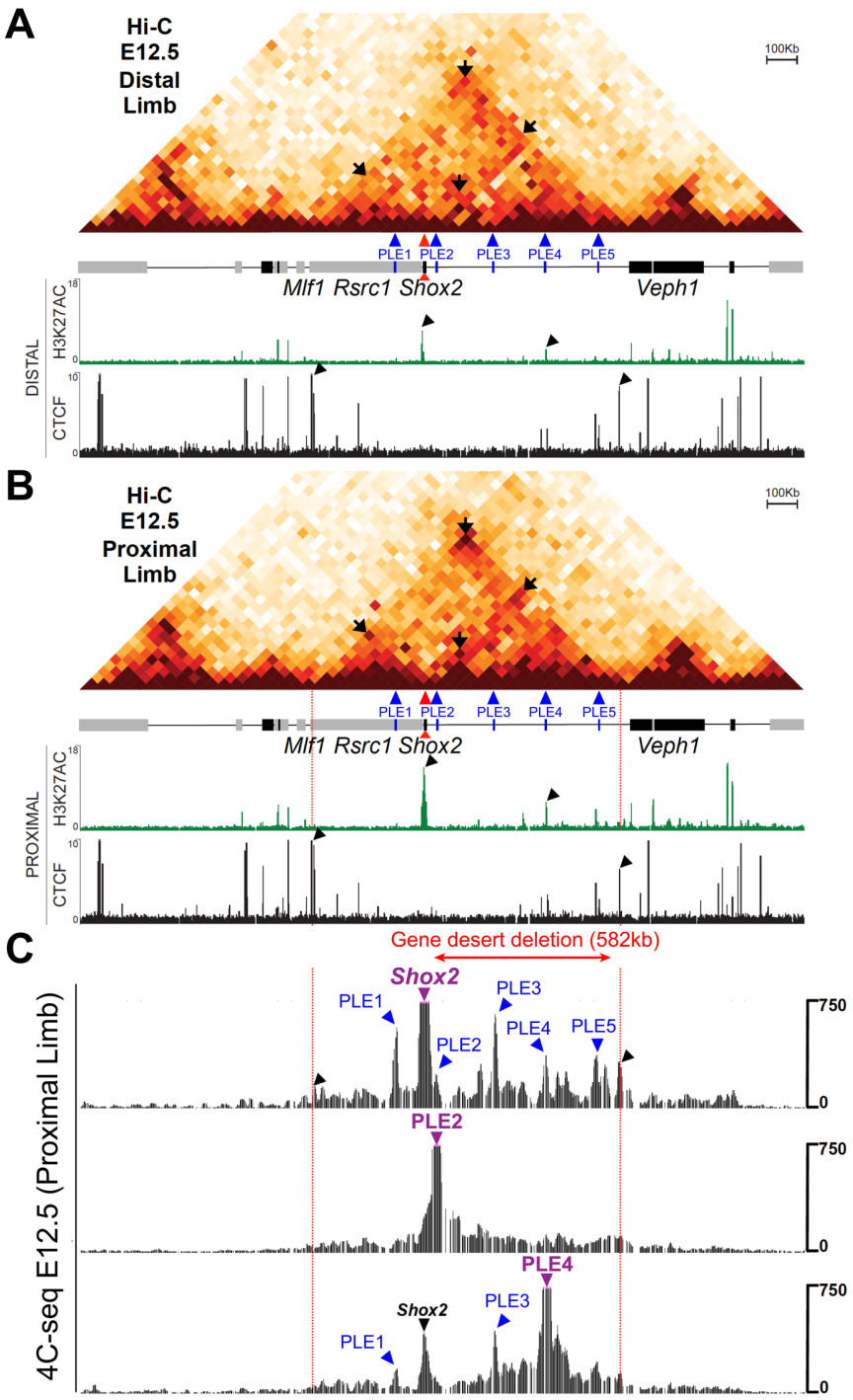

Figure S2. 3D-chromatin interactions within the *Shox2* regulatory domain in developing limbs. Comparison of chromatin interactions within the 2.4Mb region centered around *Shox2* (mm10 coordinates, chr3:65,720,000-68,120,000) in distal (**A**) and proximal (**B**) limbs at E12.5, including Hi-C, H3K27Ac and CTCF ChIP-seq tracks from Rodriguez-Carballo *et al.*, 2017. The Hi-C plot and CTCF profiles delineate the *Shox2*-TAD boundaries in limbs (red lines, see also **Fig. 2**) and indicate that several interactions are stronger in proximal limb cells as compared to those in distal limb progenitors (black arrows within Hi-C heatmap). Of the four highlighted interaction points, two indicate strong *Shox2* interactions with the contact domain boundaries (left and right arrows), while the upper arrow represents a strong interaction of the boundaries with themselves, representing a corner peak. The bottom arrow indicates a stronger interaction of *Shox2* with PLE3 in the proximal limb as compared to distal limbs. The CTCF profiles show strong peaks (black arrowheads) at the contact domain boundaries (red lines) in both proximal and distal E12.5 limbs. Arrowheads in the H3K27ac tracks indicate peaks at the location of *Shox2* and PLE4, which are stronger in proximal limbs. Genes are shown as rectangles in the maps, with black indicating genes transcribed from left to right (telomeric

to centromeric) and gray indicating genes transcribed in the opposite direction. (**C**) 4C-seq interaction profiles from *Shox2* (see also **Fig. 2B**), PLE2 (hs1262/LHB-A) and PLE4 viewpoints (purple arrowheads) (**Table S3**). One of two biologically independent replicates with similar results is shown. Most interactions obtained with each viewpoint are located within the 1Mb *Shox2*-TAD. Black arrowheads indicate interactions of TAD boundaries (red lines) with *Shox2*.

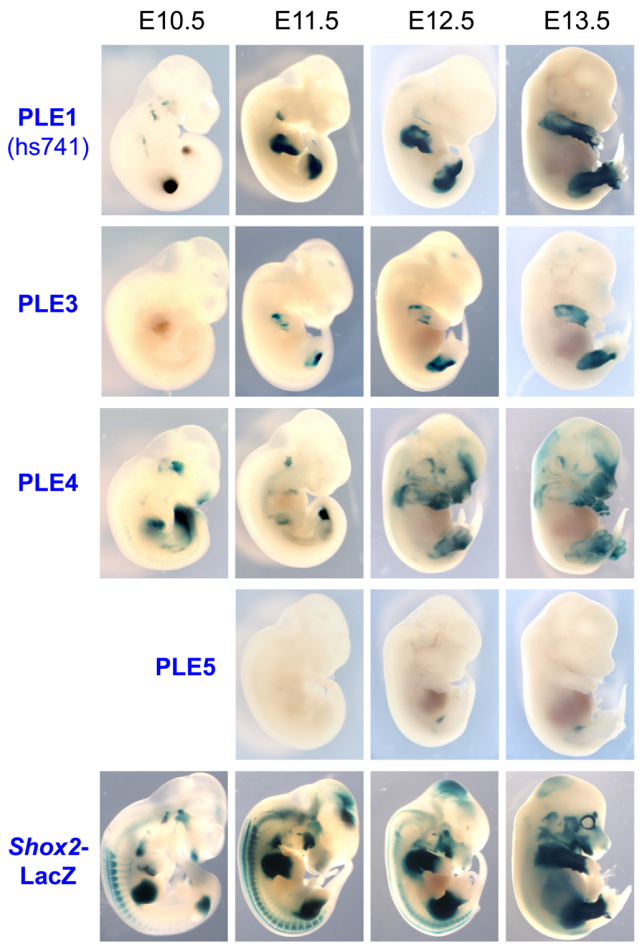

Figure S3. Spatio-temporal activity patterns of *Shox2*-contacting proximal limb enhancers (PLEs). Developmental time course (at E10.5-E13.5) of PLE activities compared to the *Shox2* expression pattern (*Shox2*-LacZ) in stable transgenic lines. Embryos from one representative transgenic line per element are shown (see **Fig. 2C and Table S4**). PLE1 represents the mouse ortholog of the hs741 enhancer[24,36]. To aid in visualization, embryos are depicted at progressively lower magnification at later stages.

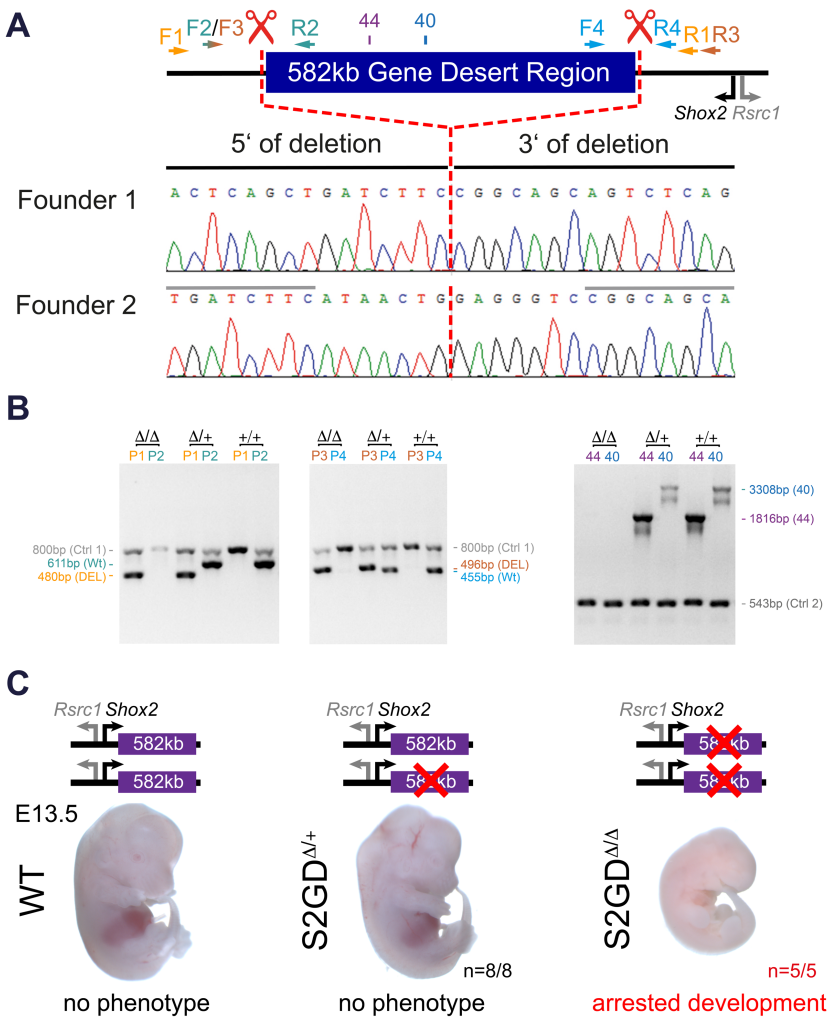

Figure S4. CRISPR/Cas9-mediated deletion of the *Shox2* gene desert causes embryonic lethality. (**A**) Validation of clean deletion breakpoints in gene desert knockout mouse lines. Red scissors indicate the CRISPR guide RNA locations flanking the deleted gene desert region. Sanger sequencing traces show the nearly identical deletion breakpoints (indicated by the red dashed line) in the two lines used in this study (**Table S5**). Location of primers (arrows) and amplicons (blocks) used for PCR (in B) are indicated. (**B**) PCR validation and genotyping used to detect the wild-type (+) and *Shox2* gene desert (S2GD) deletion (Δ) alleles. Amplicon sizes are indicated on the side. Primers (Ctrl-1 or Ctrl-2) amplifying an unrelated genomic region were used as positive controls. See **Table S6** for primer sequences and related PCR product sizes. P, product. (**C**) Homozygous gene desert deletion ($S2GD^{\Delta/\Delta}$) leads to arrested development and embryonic lethality, similar to the lethality pattern observed in *Shox2*-deficient embryos[22]. Heterozygous ($S2GD^{\Delta/+}$) genotypes develop into viable and fertile mice.

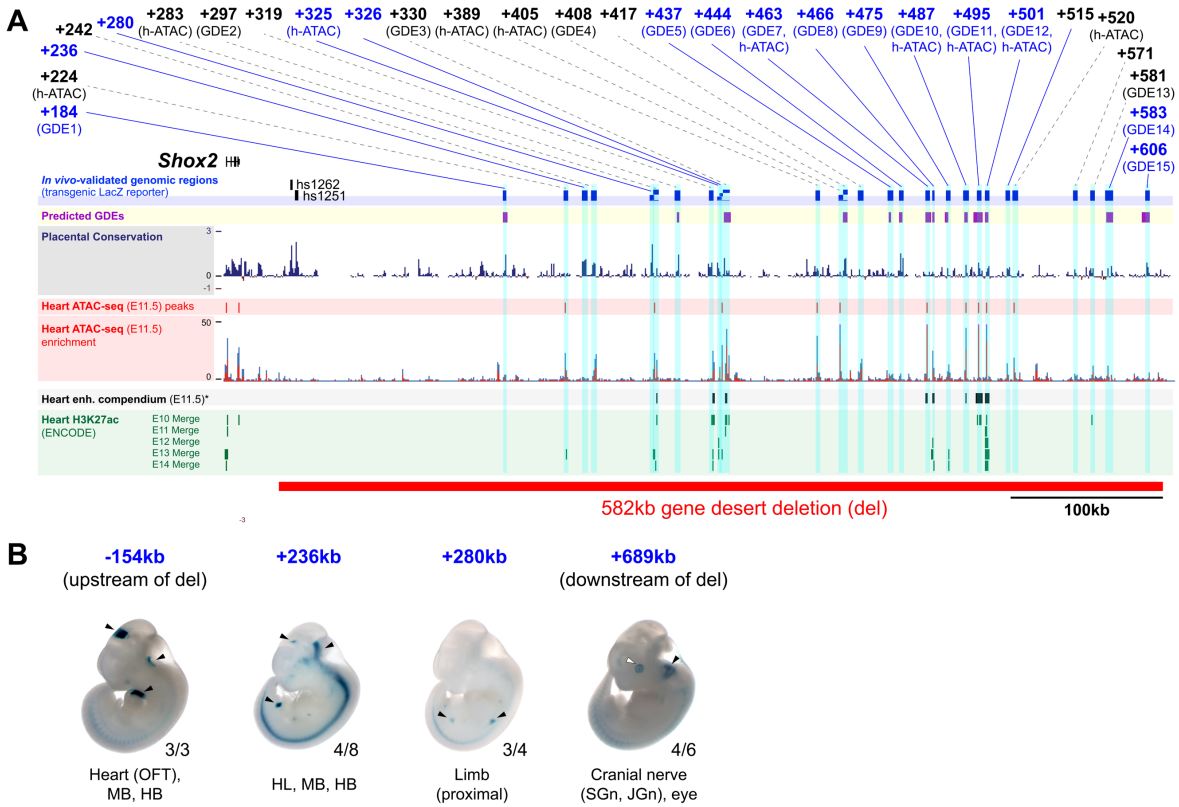

Figure S5. *Shox2* candidate enhancer regions validated in this study at E11.5. (**A**) UCSC browser window showing the *Shox2* gene and the deleted centromeric gene desert region. Candidate enhancer elements validated via Hsp68-*LacZ* reporter transgenesis in this study (blue bars**)** are listed and named based on the distance (in *kb*) to the *Shox2* transcriptional start site (TSS) (**Tables S2, S8**). IDs of elements with validated reproducible *LacZ* reporter activities in any tissue at E11.5 are marked blue, those without reproducible activities are shown in black. All tested sequences and transgenic results can be retrieved from the Vista Enhancer Browser repository (https://enhancer.lbl.gov). Regions were selected based on "stringent" enhancer predictions (GDE elements, **Fig. 1B**), heart ATAC-seq at E11.5 (h-ATAC, Fig. **5C**), cardiac enhancer potential predicted by integrative analysis[42] (asterisk) or heart H3K27ac ChIP-seq from ENCODE[28], and/or sequence conservation (PhyloP). (**B**) Reproducible *LacZ* reporter activities in transgenic embryos at E11.5 from additional candidate enhancer elements tested in this study are shown (also listed in **A**). Arrowheads mark reproducible enhancer activities. Numbers on the bottom right of each embryo denote reproducibility in indicated tissues. +, downstream; -, upstream of the *Shox2* TSS.

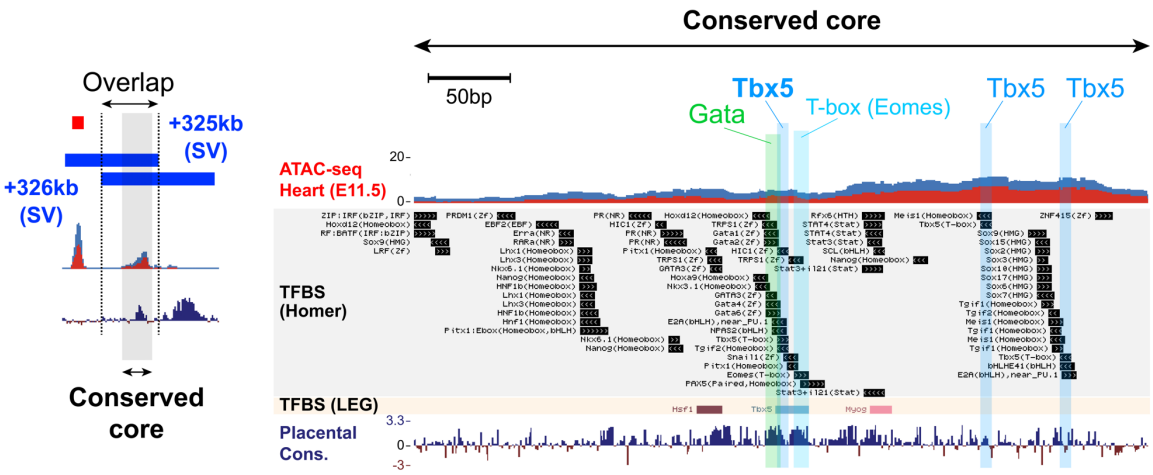

Figure S6. Tbx5 binding motifs within the *Shox2*-SV enhancer region. UCSC browser window showing the conserved core within the region of overlap of +326kb and +325kb SV enhancers (**Fig. 5C, E**). Predictions of transcription factor binding sites (TFBS) mapping to this region are derived from the HOMER database. The conserved Tbx5 motif is also part of the more stringent (and limb-specific) TF motif predictions of the LEG database[79]. The ATAC-seq track from embryonic hearts at E11.5 shows enrichment of reads in a stacked configuration (blue: replicate 1; red: replicate 2). The placental mammal base-wise conservation by PhyloP is shown. SV, sinus venosus.

Supplementary Tables

Table S1 (provided as separate EXCEL file). Developmental enhancer predictions within the *Shox2* TAD. *Sheet 1*: List of genomic elements within the extended *Shox2* TAD (chr3:65996078-67396078) that show significant ENCODE H3K27ac ChIP-seq enrichment in at least one tissue at any developmental stage and that were further filtered for ChromHMM[72] strong enhancers calls (as defined in Gorkin *et al.*, 2020). H3K27ac RPKM (Reads Per Kilobase of transcript, per Million mapped reads) values are shown for each developmental tissue and timepoint in a matrix format and are underlying the heatmap shown in **Figs. 1B and S1**. For genes present in this domain, RNA-seq read counts are listed as fragments per kilobase of exon per million reads mapped (FPKM) (red shaded). Blue shades mark predicted gene desert enhancers (GDEs). Mouse (mm10) coordinates (chrom, start, end) are given for each putative enhancer identified. Element IDs indicate distance to *Shox2* transcriptional start site (TSS). *Sheet 2*: The same matrix as in *sheet 1* but including q-values (-log10(q)) for each H3K27ac-enriched region as a measurement of statistical significance. *Sheet 3*: Metadata to be able to retrieve peak lists (for q-values) from the ENCODE Data Coordination Center (DCC, http://www.encodeproject.org/).

Table S2. Primers used for PCR amplification of predicted gene desert enhancer (GDE) elements for Hsp68-*LacZ* reporter assays. Distance (in *kb*) from the *Shox2* TSS is indicated in brackets for each element (-, upstream; +, downstream).

Element ID	*Forward primer*	*Reverse primer*	*Product Size (bp)*	*Genomic coordinates (mm10)*
GDE1 (+183)	TCCAACTAGCCACAATCCACTA	GGTTGACAAAGGTTCAGAAAGG	2486	chr3 66797249 66799734
GDE2 (+297)	TCCTCTCTGTGTTTCAGCTTTG	TGGGTGACTCAGGTAAACCTCT	3167	chr3 66682968 66686134
GDE3 (+330)	ACCATGGTAGGAAGTTCATTGG	GTTAGAGCTGTTGGGAAAATGC	4408	chr3 66650021 66654428
GDE4 (+408)	GCTATACGCCGTCAGCTTTAGT	ATGTGAATGAAGCACAAATTGC	3236	chr3 66572737 66575972
GDE5 (+437)	GATGTGGGGAAACTTCTGAAAC	TACAGACCCAGACAAGAGCAGA	4335	chr3 66542058 66546392
GDE6 (+444)	GATGCAGGCACGATATACAAAA	AGACCTTACACACGTGCACAAC	2962	chr3 66535471 66538432
GDE7 (+463)	CTGCGCTTTCTTCTTATCCCTA	CAGATCCACCTCTTCCTTCATC	3402	chr3 66518263 66521664
GDE8 (+466)	GGAATTGCTTTGTAGCTCTGCT	CAGGGAGGAAGCTTCTAGTTCA	1816	chr3 66514951 66516766
GDE9 (+475)	GACACCACCAAGAGTTCGTGTA	AATTACAATGTGTGGGGGAGAC	2824	chr3 66504602 66507425
GDE10 (+487)	TCTCTATGACCAAACGGGCTAT	GGATTTGGAAGAACAAGAGGTG	3109	chr3 66492941 66496049
GDE11 (+495)	CTGTGTATGCCTTTGCTCTCAG	CTCTGCTCATATTCTGCCTCCT	3067	chr3 66484145 66487211
GDE12 (+501)	CTGCTCTAATTCTGGGAGGTTG	TTATTGCTTGGTGAGAATGTGG	3041	chr3 66478804 66481844
GDE13 (+581)	TGTATTCCACAGCCTCCCTAGT	CCCAAGGTCTGGTTTAGAACTG	2511	chr3 66400179 66402689
GDE14 (+583)	TCCTACAGGCAAGACCTCTCTC	CATGGTCCAACTGGTATTGATG	1942	chr3 66397740 66399681
GDE15 (+606)	CATTGGTACTTGGGCTGAAAA	TTACAAAGCTCCTGACGCAGT	3139	chr3 66373217 66376355
GDE16 (+656)	CAGAGGTCCTGAACTCAATTCC	TCCTGCTGTGCATAGAACAACT	2839	chr3 66324764 66327602

Table S3. Viewpoints and primers used for 4C-Seq.

Viewpoint	**Genomic coordinates** (NlaIII fragment) (mm10)	**Primer sequence:** Illumina adapter sequences are shown in italics. Sequence specific to the viewpoint in bold.
Shox2	chr3:66,980,317-66,981,259*	Forward/Reading primer: *AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCGATCT* **CCAATTAAGAAAATATGTGGCATG** Reverse Primer: *CAAGCAGAAGACGGCATACGA***AGAATGTGAAGTTTGGTCCC**
PLE2	chr3:66,938,480-66,939,521	Forward/Reading primer: *AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCGATCT* **ACTGCTTAGTAAAGACTAATTATTCATG** Reverse Primer: *CAAGCAGAAGACGGCATACGA***AATGACATTATTATAAAATGCAATACTCT**
PLE4	chr3:66,573,586-66,574,775	Forward/Reading primer: *AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCGATCT* **GGCTGATTCTCCTGCATG** Reverse Primer: *CAAGCAGAAGACGGCATACGA***AGTTATAAAGATGATTAAGCTCTGATC**

*The *Shox2* viewpoint spans the 3' end of the first *Shox2* exon and the 5' end of intron 1.

Table S4. PCR primers and amplicons to test 4C-seq-predicted proximal limb enhancer elements (PLEs) via Hsp68-*LacZ* transgenesis. Distance (in *kb*) from the *Shox2* TSS is indicated in brackets for each element (-, upstream; +, downstream).

Element ID	*Forward primer*	*Reverse primer*	*Product Size (bp)*	*Genomic coordinates (mm10)*
PLE1 (-89)	TGGGCAAAGATCACAGAACA	GTGTGTGTGTGTGTGGTGGA	1674	chr3:67070163-67071836
PLE2 (+43)	GAAGGACCGCACAGCTTATC	GGTCCACATATGCCCAAGGA	2428	chr3:66937659-66940086
PLE3 (+237)	GAAGAGGGGGCAGATTGTGTTGACTG	TGCTTCTTCAAATATTGCTTTGCTAAT	10351*	chr3:66739935-66750285
PLE4 (+407)**	GTGAATGAAGCACAAATTGCAA	AAAGCCCATGTGTTCATCCCAG	3718	chr3:66572253-66575970
PLE5 (+568)	GGTCTATCTTGTTGCATGTTTTGTT	GGACAAACAGAGCTCAGAAGAGA	9473***	chr3:66409729-66419201

*A 9128bp ApaI/SalI sub-fragment (mm10: chr3:66740432-66749559) of the 10351bp PCR fragment was cloned into the pβlacz vector and used for *LacZ* transgenesis.
The PLE4 fragment contains the GDE4 element (Fig. 1B**) and an additional 486bp.
***A 8520bp ApaI/SalI sub-fragment (mm10: chr3:66409729-66418248) of the 9473bp PCR fragment was cloned into the pβlacz vector and used for *LacZ* transgenesis.

Table S5. Targeted gene desert region and CRISPR sgRNA templates. Genomic coordinates of the CRISPR-deleted region are provided for each founder line. Use of unique sgRNAs resulted in the generation of two nearly identical founder lines (see **Fig. S4**).

Gene desert deletion strain	Genomic coordinates of deletion (mm10)	Deleted region (bp)	5' sgRNA target sequence	5' sgRNA target sequence
Founder 1	chr3 66365062 66947168	582107	TGATCTTCATAACTGCCATGGGG	TGAAGCACAAGGCTGGCGGGAGG
Founder 2	chr3 66365069 66947161	582093	Same as above	Same as above

Table S6. Primers used for screening and genotyping of CRISPR deletion mouse strains. PCR genotyping results using agarose gel electrophoresis are shown in **Fig. S4**. P, product. f, founder. N.A., not amplified.

Analyzed Region	*Primer name*	*Sequence*	*Product Size (bp)*
Desert deletion (P1)	**F1**	agcggagggatactttagcac	WT: 582587 (N.A.)
	R1	tgctgagagatgaaccctgat	KO: 480 (f1) / 494 (f2)
5' desert junction (P2)	**F2**	ccgcagagttctttgagagttt	WT: 611
	R2	gacccagcagtattcggagtta	KO: N.A.
Desert deletion (P3)	**F3**	ccgcagagttctttgagagttt	WT: 582603 (N.A.)
	R3	acaagagcatgtgttcaagtgg	KO: 496 (f1) / 510 (f2)
3' desert junction (P4)	**F4**	tgccctacagaagttaagcaca	WT: 455
	R4	tactgttgccatcactccattc	KO: N.A.
Region 44 (+466kb, GDE8)	**44 F**	ggaattgctttgtagctctgct	WT: 1816
	44 R	cagggaggaagcttctagttca	KO: N.A.
Region 40 (+389kb)	**40 F**	tctataacggagctgcacttga	WT: 3308
	40 R	ggcatttgtgagacatgagaaa	KO: N.A.
Control region 1 (Ctrl-1)	**Ctrl-1 F**	ccctagttctgtaaaccaggcta	WT/KO: 800
	Ctrl-1 R	tcatgtgtcttaggagagggttc	(Tbx3 locus)
Control region 2 (Ctrl-2)	**Ctrl-2 F**	agctggtagccttaaaataagccaa	WT/KO: 543
	Ctrl-2 R	gcctgaaagaggtcatcatcacc	(Gli3 locus)

Table S7. Primers used for SYBR Green Real-time PCR analysis.

Target Gene	qPCR primer	Sequence	*Product Size (bp)*
Shox2	Shox2_F	CCCGAGTACAGGTTTGGTTTC	119*
	Shox2_R	GAAGCTTGTAGAGTTGCACCC	
Rsrc1	Rsrc1_F	TGCAATTGGTCCTTGAAGCT	104*
	Rsrc1_R	GGTGGCTTGGTCTTCTTCTT	
Actb	Actb_F	ACACTGTGCCCATCTACGAGG	280*
	Actb_R	CATCACTATTGGCAACGAGCG	

*primer pair validated and used in a previous study[24].

Table S8 (provided as separate EXCEL file). List of all genomic elements analyzed in this study using Hsp68-*LacZ* transgenic reporter assays in mouse embryos at E11.5. Tested elements were selected based on the criteria summarized in **Fig. S5** and are named according to their distance (in *kb*) from the *Shox2* transcriptional start site (-, upstream; +, downstream). Corresponding UCSC browser coordinates (mm10), Vista Enhancer Browser IDs (https://enhancer.lbl.gov) and primer sequences (used for amplification of genomic regions cloned into the transgenic Hsp68-*LacZ* reporter construct) are shown. Rows of elements driving reproducible tissue-specific activities at E11.5 are marked blue.