Identification of Transcriptional Enhancers in Development and Disease

By

Brandon J. Mannion

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Comparative Biochemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Dr. Len A. Pennacchio, Co-Chair
Professor Fenyong Liu, Co-Chair
Professor Lin He
Professor Gary H. Karpen

Spring 2023

Abstract

Identification of Transcriptional Enhancers in Development and Disease

by

Brandon J. Mannion

Doctor of Philosophy in Comparative Biochemistry

University of California, Berkeley

Dr. Len A. Pennacchio, Co-Chair
Professor Fenyong Liu, Co-Chair

Enhancers are non-coding DNA elements found throughout the genome that, in concert with transcription factors, coactivators, and general transcriptional machinery, activate cell-type specific gene expression. Initial studies on enhancers demonstrated these regulatory elements contained clusters of transcription factor binding sites to recruit endogenous transcription factors and drive elevated expression of a target gene[1–4]. These early works highlighted that their regulatory activity was maintained despite alterations in their orientation and/or positioning relative to the targeted gene. Nearly half a century later, enhancers are center stage in efforts to characterize the regulatory components and mechanisms behind development and disease. This dissertation is a study on mammalian enhancers, the genome-wide approaches for their identification, and their contributions in early developmental processes. Chapter 1 provides an overview of the enhancer properties uncovered from various experimental systems and how these properties are harnessed to predict and further dissect enhancer activity. Chapters 2 and 3 comprise two separate projects that involve 1) an extensive *in vivo* assessment of active enhancers that are hidden from canonical biochemical-based methods for enhancer identification and 2) the characterization of tissue-specific enhancers across the *Shox2* locus that regulate early heart, face, and limb development. Altogether these works demonstrate the critical roles of enhancers for normal organismal development and the ongoing challenge of mapping increasingly large datasets to insights on enhancer prediction and function.

To my mother, for her love and support to explore and continue onward.
And to my loving and patient partner, who encourages me through any finish line.

Table of Contents

List of Figures

Chapter 2

Chapter 3

List of Tables

Chapter 2

Chapter 3

Acknowledgements

First, I want to recognize and deeply thank my advisor, Len Pennacchio, for his guidance, mentorship, expertise, and support throughout all these years. I also thank Diane Dickel, Axel Visel, Marco Osterwalder, and all the past and present members of the Mammalian Functional Genomics Group for their help, humor, and efforts inside and outside of the lab. Each of their contributions and insights, individually and as a team, provided a research environment that was exciting and rewarding to be a part of everyday.

I also want to thank my other dissertation committee members Fenyong Liu, Lin He, and Gary Karpen for their time and feedback during my graduate career. Their support and flexibility despite a global pandemic and a myriad of other consequential events allowed me to continue onward on this challenging path. I also thank Phong Trang and the Comparative Biochemistry group for the discussions and assistance we provided each other.

To my friends nearby and afar, thank you for the laughter and memories that were both relaxing and reenergizing.

Finally, I thank my family and my partner's family for their love, unwavering support, and encouragement over these weeks, months, and years.

To my partner and wife, Helen: your love, patience, and optimism fueled me through the long days and late nights and reminded me to savor this chapter and the challenges within it. Thank you here and always.

Chapter 1 : Introduction

For a majority of the past half century, an understanding of organismal development, complexity, and its evolutionary relationship to others has centered around genes and the genomes in which these genes reside. Studies on the duplication, mutation, or relocation of genes have provided numerous insights in the fields of evolutionary biology, developmental biology and medicine. Advancements in sequencing technologies near the end of the twentieth century enabled the completion of several whole-genome sequencing projects for organisms that include *Drosophila melanogaster* (fruit fly), *Mus musculus* (house mouse), and humans[5–8]. With such information, genes, gene number, and genome size were harnessed to relate organisms to each other and to refine existing phylogenies[9]. Nevertheless, that protein-coding genes are remarkably similar between humans and chimpanzees was one finding that suggested the coding genome alone was insufficient to explain phenotypic differences and variation across the tree of life[10]. Only a few years after this insight, short stretches of DNA derived from the SV40 (Simian Virus 40) early genes upstream region (*i.e.,* a non-coding region) were shown to significantly increase gene expression in a mammalian cell system[2]. These so-called "transcriptional enhancer elements" activated β-globin gene expression independent of the enhancer's position or orientation in the reporter setup. Subsequent studies demonstrated these enhancer elements coordinate gene expression via the recruitment of transcriptional activators and repressors to transcription factor binding sites (TFBSs) within the enhancer region[4,11,12]. Variants, mutations, or rearrangements of enhancer sequence can lead to disruptions in target gene expression and corresponding disease phenotypes exemplified in abnormal limb development, blood disorders, altered heart function, and metabolic diseases[13–18]. However, perturbation of enhancer sequence is not often sufficient to substantially alter gene expression, as many have demonstrated the presence of additional enhancers within a locus (*e.g.,* "shadow enhancers") that can provide a buffering effect against such variation[19–23]. Nonetheless, analyses of sequence composition (*e.g.,* the presence and arrangement of transcription factor motifs) have provided useful frameworks upon which other approaches can be applied to refine enhancer identification[24].

Models on enhancer composition

At the level of an enhancer's DNA sequence, a continuum of models exist to consolidate how both the organization of TFBS and the interaction of corresponding binding transcription factors (TFs) contribute to enhancer regulatory activity[25]. The enhanceosome model, typified initially by studies on regulation of the interferon-beta (IFN-β) gene, suggests a strict arrangement of binding sites in order for participating transcription factors to cooperatively bind DNA and subsequently activate gene expression[26]. Opposite this rigidity in binding site arrangement is the billboard model, which proposes that the presence of transcription factor binding sites - regardless of their relative arrangements around each other - are sufficient for enhancer function[27]. Nevertheless, models focused primarily on enhancer composition cannot explain the activities of all enhancers thus far characterized, and instead must account for both the combinatorial interactions of transcription factors and their specific cellular contexts of activity[28–30]. As introduced above, the observation of separate, redundant enhancers within a locus that can buffer against mutations to yield stable target gene expression is an additional layer of complexity for the dissection of enhancer function[19,23,31,32]. This configuration of multiple enhancers per target gene appears commonly throughout mammalian genomes[23,33,34]. Beyond sequence composition, it is necessary to evaluate an enhancer's endogenous context, *i.e.,* its interactions with other regulatory elements,

transcription factors, cofactors, and transcriptional machinery as permitted by the local chromatin organization, to understand how these components together effect transcriptional output[35].

3D chromatin organization

With the advent and refinement of chromatin conformation capture assays[36–40], it is increasingly appreciated how three-dimensional chromatin architecture impacts gene regulation via the specification of particular enhancer-promoter interactions. This selectivity of interactions is partly mediated by factors that include CTCF and Cohesin, which coordinate the looping of DNA into so-called topologically associated domains (TADs) and other sub-domains[41,42]. Structural variants (*e.g.,* inversions or duplications) or CTCF boundary mutations that disrupt TAD structure can lead to an enhancer regulating a different gene or failing to regulate its target gene(s). Notable examples include the multiple limb malformations (*e.g.,* brachydactyly, digit shortening; polydactyly, extra digits) from limb-related gene dysregulation that are attributed to variant-altered TADs[43–45]. However, regulatory element interactions (*e.g.,* enhancer-promoter, promoter-promoter) that span across TADs have been observed in the context of both human and mouse cell differentiation processes[46,47]. Central to understanding enhancer function is the identification of its target gene(s) and, more broadly, the enhancer's contributions within a gene's regulatory landscape[48]. Whether an enhancer targets a closely flanking gene or skips over intervening genes is an event partly determined by the local three-dimensional chromatin configuration. In mammals, the genomic distance between an enhancer and its target gene promoter varies from several hundred to millions of basepairs[49–51]. Examples of long-range enhancer-promoter interactions include the ZRS (zone of polarizing activity regulatory sequence) *Shh* enhancer, which skips nearly 1 Mb of intervening sequence with other genes, and the human *SOX9* enhancers that are over 1.4 Mb away from *SOX9*[52,53]. While TADs can facilitate preferential regulatory element interactions within a given domain, it is yet to be resolved how the frequencies and concentrations of such interactions account for both the activation of transcription and its varying transcriptional output[35,54]. Despite these gaps, both 3D chromatin organization and chromatin states are commonly profiled to predict candidate enhancers and their potential regulatory activities[55–57].

Enhancer identification

Initial methods to identify and test putative enhancers utilized the presence of TFBSs or regions with high evolutionary conservation[58,59]. However, the presence of a TFBS alone does not translate to a functional region, as they are both numerous genome-wide and often overlapping with other unrelated motifs. Additionally, transcription factor characterization is incomplete as further work continues to discover and characterize both their expression patterns and interactions within their regulatory networks[60,61]. Genomic regions with high evolutionary conservation also facilitated candidate enhancer identification[62–64]. Still, not all candidate or functional enhancers are characterized by high evolutionary conservation[65,66]. Transcription factor binding typically occurs at open chromatin or nucleosome-free DNA[67–69]. As such, open chromatin is a widely used characteristic used for genome-wide identification of candidate regulatory elements[69–72]. Techniques such as ChIP-chip (chromatin immunoprecipitation with DNA microarray[73]) or ChIP-seq (chromatin immunoprecipitation with next-generation sequencing[74,75]) can provide genome-wide information on transcription factor binding or post-translational histone modifications. Binding by CBP/p300 histone acetyltransferases and other coactivators to putative enhancer regions is informative for their identification across multiple developing tissues[76,77]. The presence of chromatin-based modifications around candidate enhancer regions, *i.e.*, post-translational

histone modifications mediated by histone acetyltransferases or histone methyltransferases, that include H3K27ac (acetylation of histone 3 lysine 27) and H3K4me1 (monomethylation of histone 3 lysine 4) is also informative for genome-wide candidate enhancer identification[78–80]. Beyond these canonical enhancer-associated chromatin marks, other histone modifications are also associated with candidate or active enhancers[81–84]. It is an ongoing area of investigation whether or not these marks are causative or merely correlated with transcriptional regulatory activity[85–87].

Mapping of regulatory elements

The ENCODE (Encyclopedia of DNA Elements) consortium has contributed extensive mapping and characterization of regulatory elements throughout the human and mouse genomes[88–90]. The most recent iteration includes genome-wide chromatin (*e.g.,* histone modifications, open chromatin, methylation patterns) and transcriptomic (*e.g.,* RNA-seq) data from both human and mouse tissue types[90–93]. In its latest iteration, the ENCODE consortium generated an extensive chromatin catalog that spans early mouse development and that systematically confirms the dynamic tissue-specific chromatin states of enhancers[93]. From these and adjacent studies, chromatin accessibility (via DNase-seq), histone modifications (*e.g.,* H3K27ac via ChIP-seq), and presence of insulator-binding protein CTCF were used to enumerate over a million candidate cis-regulatory elements (cCRE) throughout the human and mouse genomes[90]. Recent efforts to annotate the human genome using DNase-based footprinting similarly estimate over a million putative regulatory elements[72,94]. In parallel with the generation of these large tissue-based chromatin atlases is the increasing development and use of techniques to resolve genome-wide chromatin marks and transcript expression at the single-cell level[95–98]. Similar to the ENCODE consortium's objectives to profile the chromatin landscapes of multiple tissues across different developmental stages, the Human Cell Atlas and others currently aim to catalog the expression profiles of all the cells in the human body[99]. Such atlases will only increase in their breadth and depth of cells and organisms sampled as the methods for single-cell based genomics continue to grow[100]. Several types of functional assays exist that harness these vast sequence and epigenomic data to dissect enhancer function in *in vitro* and *in vivo* contexts.

Approaches for candidate enhancer validation

The use of functional assays to validate candidate enhancer activity in a given chromatin context is a critical step for enhancer characterization and further dissection[101]. Massively parallel reporter assays (MPRAs) are the prevalent form of cell-based assays that enable the testing of hundreds to millions of genomic fragments for activity in either fluorescent- or transcript-based readouts[102,103]. MPRAs are often paired with CRISPR-Cas9-mediated screens to assess the impacts of local chromatin context and regulatory element mutation on target gene expression[104–107]. While MPRAs can provide a substantial amount of functional data per experiment (*e.g.,* span of genome covered, number of variants tested), these and other high-throughput experiments currently lack the organismal contexts (*i.e.,* the relevant cell-cell and other higher-order interactions within heterogeneous tissues) provided by the *in vivo* system. Additionally, that MPRAs and other cell-based assays cannot yet assess different cell types per experiment poses additional challenges in consideration of enhancers with regulatory activities in multiple separate tissues[108].

Thus, *in vivo* reporter assays are a critical tool to validate and characterize candidate enhancers [101,102]. In particular, my dissertation lab and other research groups utilize a mouse *in vivo* transgenic reporter assay to test whether or not DNA sequences identified as candidate enhancers are active

in an *in vivo* context[109]. Briefly, candidate enhancers are cloned adjacent to a minimal promoter and the *LacZ* reporter gene. These reporter constructs are microinjected into mouse pronuclei and the transgenic embryos are later harvested and scored for reproducible tissue- or cell type- specific candidate enhancer-reporter activity (*i.e.*, colorimetric β-galactosidase activity). Results from my dissertation lab's mouse *in vivo* studies are compiled and made publicly available through a widely used enhancer catalog, the VISTA Enhancer Browser (https://enhancer.lbl.gov)[110]. Presently, the VISTA Enhancer Browser includes over 3,000 candidate enhancers (VISTA enhancers) from human, mouse, and other vertebrate sequences that have been tested for enhancer-reporter activity in developing mouse tissue. While mouse *in vivo* experiments are costly and low-throughput compared to MPRAs, recent improvements that include the CRISPR-Cas9 mediated site-specific integration of reporter constructs have increased transgenic rates and have enabled testing of manyfold elements or clinical variants at scale[111]. Together with further functional characterization in the form of targeted enhancer deletions and/or variant knock-ins, the *in vivo* reporter system has contributed to our understanding of enhancers in a variety of developing tissues that include the brain, limb, and craniofacial structures[53,112,113]. Further, VISTA enhancers paired with tissue- and stage-matched mouse chromatin data demonstrated that the H3K27ac (ChIP-seq) is the most informative mark for validated enhancers[93]. Efforts are ongoing to develop experimental systems that combine the advantages of both MPRAs and *in vivo* systems, *i.e.,* high-throughput experiments with cell-/tissue-type heterogeneity and context. Alternatively, computational approaches that incorporate these functional genomics and sequencing data are in use to predict 1) whether a given sequence is a functional regulatory element or 2) the impacts of regulatory element sequence variation on gene expression[114–118]. While *in silico* approaches are increasing in power to predict subtle transcript expression changes based on these data, functional studies that utilize the *in vivo* system remain valuable for physiologically-relevant insights applicable both to enhancer biology and human disease.

Conclusion

Considering the above models, data, and available experimental systems, it is an intriguing question how DNA sequence, chromatin marks, 3D chromatin architecture, and other varying cellular and molecular contexts independently and combinatorially contribute to enhancer function. It is not yet practical to employ current assays, whether in cells or an *in vivo* system, to systematically address this question across the temporal (*e.g.*, developmental stage) and spatial (*e.g.*, cell- or tissue-type) landscapes known for enhancer gene-regulatory activities. As one approach to this question, I first looked into the relationship between time- and tissue-specific chromatin marks and active enhancers. While epigenomic data in the form of histone modifications (*e.g.*, H3K27ac, H3K4me1) and open chromatin are useful for the genome-wide prediction of cis-regulatory elements, it remains unclear how well these data correctly identify enhancers that are active at a particular developmental stage and in a specific tissue or cell-type. Using candidate cis-regulatory elements tested in a mouse *in vivo* dataset generated by my dissertation lab (VISTA enhancers), I found a majority of elements that drove reproducible enhancer-reporter activity indeed had enhancer-associated chromatin marks (*e.g.*, H3K27ac, H3K4me1, open chromatin) in the corresponding tissue. Interestingly, I found a portion of elements that drove enhancer-reporter activity and yet did not have any of these enhancer-associated chromatin marks in the relevant tissue. The following chapter (Chapter 2) comprises a systematic approach in which I perform both a thorough retrospective study on the VISTA enhancers and an unbiased tiling study across two developmental gene loci to resolve how these enhancer-associated chromatin features both mark

and miss active enhancers. That prevalent epigenomic data available cannot reliably identify all active enhancers has implications not only for current studies that utilize these resources, but also for future technologies that aim to predict and characterize enhancers throughout the genome.

Addendum

Apart from the main works that follow this chapter, I collaborated with other research groups on projects primarily centered around the roles of enhancers in regulating early heart or brain development. Altogether these collaborations highlight the continued value in pairing epigenetics-based approaches, clinical data, and mouse *in vivo* transgenic reporter assays to screen for enhancers and to uncover their contributions in disease and developmental processes. Below I highlight 4 publications I was also involved with during my graduate studies.

First, the Vedantham research group is interested in pacemaker cardiomyocyte cells and the gene regulatory programs that contribute to their origin and function. Pacemaker cells are a part of the sinoatrial node (SAN) tissue that together with the conduction system stimulates heart contractions and proper heart rhythms[119,120]. To identify potential cis-regulatory elements specific to the SAN pacemaker cells, Galang et al. compared differentially accessible regions (open chromatin via ATAC-seq) between isolated pacemaker cardiomyocytes (PCs) and right atrial cardiomyocytes (RACMs)[121]. They then showed differentially accessible regions observed in PCs were 1) associated with genes also differentially expressed between PCs and RACMs and 2) enriched for transcription factor motifs known to be involved in heart and PC development. From their list of 59 PC-specific accessible regions, I incorporated both evolutionary conservation and ENCODE mouse E11.5 H3K27ac ChIP-seq to prioritize the testing of 17 elements via the recently scaled mouse *in vivo* transgenic reporter assay[111]. From these tests, we found 4 with reproducible enhancer-reporter activity in mouse E11.5 heart, 2 of which showed specific *LacZ* staining in the SAN region. All elements with SAN-specific staining are adjacent to genes enriched in PC relative to RACM (*i.e., Hcn4*, *Rgs6*). Galang et al. further showed that mice without a differentially accessible region adjacent to *Isl1* (termed "Isl1 Locus SAN Enhancer", or ISE) had reduced Isl1 expression, reduced number of PCs, increased occurrences of arrhythmias, and other phenotypes suggestive of ISE as a PC-specific enhancer. Apart from the characterization of PC- and SAN-related enhancers, this works highlights an effective strategy to sift for candidate regulatory elements in rare cell types based on the mouse *in vivo* testing of regions that are differentially accessible (in one cell type vs. another) and that have additional enhancer-associated chromatin marks (*i.e.,* H3K27Ac).

Second, shortly after the completion of the first draft of the human genome in the early 2000s, comparisons between the human, mouse, and rat genomes revealed there to be over 480 stretches of DNA with 100% sequence conservation among the three[122]. Interestingly, a large proportion of these so-called ultraconserved elements (UCEs) did not overlap with annotated exons and also clustered around transcription factors or developmental genes, which is suggestive of a functional, regulatory role for this subset of UCEs. Subsequent mouse *in vivo* studies demonstrated several of these UCEs drive tissue-specific enhancer-reporter activity, oftentimes in the developing brain[62,64]. Additional characterization of four UCEs that are proximal to *Arx*, which encodes a transcription factor involved in brain development, revealed these to act as tissue-specific enhancers whereby single and double deletions of these UCEs resulted in reduced *Arx* expression and altered brain

morphologies[113]. However, still unclear were the mechanisms that contributed to the extreme conservation across these elements and, relatedly, whether and to what extent the function of these UCEs would change as a result of mutations across these elements. Prior studies in cells showed that only a few sites within larger candidate enhancers were measurably affected by introduced mutations, and oftentimes these did not yield substantial changes in expression[123–125]. Snetkova et al. applied different levels of mutagenesis (2%, 5%, 20% of base pairs mutated for a given enhancer) to systematically assess the mutational tolerance of 23 ultraconserved enhancers in a mouse *in vivo* context[126]. For this project, I helped evaluate the presence of tissue- and stage-specific enhancer-associated chromatin marks around UCEs to determine a subset of UCEs that could be assessed for enhancer-reporter activity in the mouse *in vivo* transgenic assay. I also provided research support in the form of timed embryo collections, *LacZ* staining, and subsequent blinded scoring of enhancer-reporter activity across the different elements tested. Interestingly, Snetkova et al. found a majority of the 23 ultraconserved enhancers were tolerant to increasing mutation loads applied throughout the sequence, *i.e.,* they maintained their tissue-specific enhancer-reporter activity even in some cases with 5% of base pairs mutated within a given enhancer. That most of these enhancers maintained their reference activity even with mutations focused on especially conserved sites with predicted conserved TF motifs was a surprise in light of both the deep conservation of UCEs and lack of natural human variation within UCEs[122,127]. These latter observations are suggestive of tight control over the mutational landscape of UCEs. To assess the phenotypic consequences (or lack thereof) of mutagenized UCEs in both an *in vivo* and endogenous genomic context, Snetkova et al. used CRISPR-Cas9 to generate knock-in mouse lines for three different mutagenized alleles of two ultraconserved enhancers. The two mutagenized alleles that were observed as a loss of enhancer-reporter activity in the transgenic assay (embryonic mice) were correspondingly associated with altered brain morphology (*e.g.,* abnormal hippocampus) in the stable knock-in mice. Similarly, the one mutagenized allele that did not differ in enhancer-reporter activity from the unchanged (reference) ultraconserved enhancer also showed in the corresponding mouse knock-in similar brain morphology as that in the reference allele. These results demonstrated the correspondence between findings in the initial mouse *in vivo* transgenic assay and further functional characterization via stable mouse knock-in lines. Overall, this study revealed a portion of ultraconserved elements with enhancer activities that are surprisingly tolerant to mutations, which suggests additional functions beyond developmental transcriptional regulation are required to understand their ultraconservation.

Third, whole exome sequencing (WES) has been a powerful, relatively low-cost approach to identify candidate causal coding variants for human diseases[128–130]. Whole-genome sequencing (WGS), while costly and also yet to be fully applied to catalog the vast diversity in human populations worldwide, is capable of resolving both single nucleotide variants (SNVs) with high specificity, small insertions/deletions (indels), and copy number variants (CNVs)[131–133]. CNVs, which are a subset of structural variants (SVs), are implicated in a variety of human diseases that involve the nervous system and neurodevelopment[131,134,135]. One area of study in the Turner research group centers around neurodevelopmental disorders, particularly autism, and the identification of potential causal coding and regulatory variants (either SNVs or CNVs). From an extensive whole-genome sequencing and metaanalysis of hundreds of autism families, Turner et al. identified *de novo* mutations (DNMs) both in coding and noncoding DNA that were enriched in probands (child with autism) relative to the unaffected sibling[136]. Moreover, relative to unaffected siblings, probands were found to accumulate more DNMs in genes associated with

autism. In the Turner group's latest study, Padhi et al. incorporated additional whole-genome sequencing of families with autism to identify *de novo* variants (DNVs) enriched in enhancers and to characterize hs737, a VISTA enhancer with midbrain and hindbrain enhancer-reporter activity that was observed to have DNVs in multiple affected children[137]. For this project I helped assess the enrichment of DNVs in VISTA enhancers. Of those VISTA enhancers that were found to have DNVs, I also provided guidance on the use of chromatin data to characterize their regulatory activities across multiple developmental stages and tissues. Additionally, I coordinated the design of three hs737 variants, each of which had one of the three clinical DNVs and their testing via the mouse *in vivo* transgenic assay. Due to the elevated number of DNVs found in this particular enhancer, we were interested in other mutations within hs737 that could alter its enhancer-reporter activity. I designed 20 hs737 variants, each with randomly distributed mutations (SNVs) throughout the enhancer, and also coordinated their testing. This mouse *in vivo* characterization, while not in the Padhi et al. study, highlights an increasingly accessible approach of pairing clinical data with *in vivo* enhancer-reporter activity screens to understand potentially pathogenic variants[111]. Padhi et al. show the DNV-enriched hs737 to have chromatin characteristics relevant for its regulatory activity in the developing brain. They then focus on *EBF3*, a putative target gene of hs737 to highlight both its involvement in regulating neurodevelopmental disorder (NDD)-associated genes and the NDD-related phenotypes that are elevated in persons with *EBF3* variants. Overall, this work highlights the power of applying whole-genome sequencing toward a clinical cohort to identify *de novo* mutations enriched in children with autism (probands) and, subsequently, to nominate potential causal variants for further study of their contributions to autism and NDD-related etiologies.

Finally, the Firulli research group is focused on the characterization of the gene regulatory networks involved in heart development, particularly how the basic helix loop helix (bHLH) Hand/Twist-family of transcription factors contribute to cardiogenesis. Both *Hand1* and *Hand2* are expressed in the early heart within shared and also non-overlapping subregions[138]. *Hand2* is involved in the development and differentiation of the epicardium, endocardium, and myocardium. Aside from its roles in cardiac morphogenesis, *Hand2* is also involved in both limb and craniofacial development[139,140]. Deletion of *Hand2* in mice results in embryonic lethality[141]. *Hand2* conditional knockout lines present multiple endocardial-related developmental abnormalities (*e.g.,* ventricular septal defects)[142,143]. In order to characterize the components of the *Hand2* gene regulatory network responsible for endocardial development, George et al. paired scRNA-seq of *Hand2* conditional knockouts (H2CKO) with Hand2 ChIP-seq data to identify in mice Hand2-associated genes and their corresponding candidate cis-regulatory elements[144]. From this analysis they focused on the following genes that had altered gene expression in the H2CKO endocardial cluster (via scRNA-seq) and whose surrounding loci had Hand2 ChIP-seq signal (*i.e.*, signifying Hand2-DNA interactions) at conserved non-coding sites: *Igf2*, *Igf2R*, *Ptn*, *Tmem108*, and *Klf2*. They used both Hand2 ChIP-seq signal and evolutionary conservation to select candidate enhancers for a selection of these genes. I helped in the coordination of cloning, mouse *in vivo* testing, and scoring of these candidate enhancers. Of 6 candidate enhancers tested, we found 3 to have enhancer-reporter activity in endocardial or endothelial cells. The subregional activities of these enhancers recapitulated those of their adjacent genes, namely *Igf2R* and *Klf2*, which are expressed in endocardium and endothelium during heart development[145,146]. George et al. further characterized one of the two tested *Klf2* enhancers (-50kb CNE) by examining changes in tissue expression in a mouse line generated from a -50kb CNE *LacZ* transgenic line crossed with the H2CKO line. -

*50*kb CNE *LacZ* expression was still observed in the vasculature. However, no expression was observed in the ventricular endocardium, which indicated the -50kb CNE *Klf2* enhancer to be Hand2-dependent for this ventricular endocardium activity. Further, they observed in a mouse line generated with the -50kb CNE *Klf2* enhancer deleted (CRISPR-Cas9), a reduction in *Klf2* expression within the ventricular endocardium. *Klf2* expression within this region was not completely lost and mice from this -50kb CNE deletion line were both viable and fertile, which suggests additional components that may regulate *Klf2*. Altogether these findings uncover several enhancers within the *Hand2* gene regulatory network that contribute to heart development and function. Additional candidate enhancers identified by the H2CKO analysis remain to be studied for their potential roles within this network. Finally, the study highlights the increasing utility of single-cell based approaches (*e.g.,* scRNA-seq) to distinguish cluster-specific features (*i.e.,* expression changes for a given cluster or cell-type) that can then be targeted for further functional characterization. As these approaches improve in their scalability and resolution, we can expect a rise in our understanding of gene regulation in development and disease to follow.

References

1. Moreau, P. *et al.* The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res.* **9**, 6047–6068 (1981).
2. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
3. Philipsen, S., Talbot, D., Fraser, P. & Grosveld, F. The beta-globin dominant control region: hypersensitive site 2. *EMBO J.* **9**, 2159–2167 (1990).
4. Ikuta, T. & Kan, Y. W. In vivo protein-DNA interactions at the beta-globin gene locus. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 10188–10192 (1991).
5. Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science* **287**, 2185–2195 (2000).
6. Mouse Genome Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
7. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
8. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
9. Wolf, Y. I., Rogozin, I. B., Grishin, N. V. & Koonin, E. V. Genome trees and the tree of life. *Trends Genet.* **18**, 472–479 (2002).
10. King, M.-C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees: Their macromolecules are so alike that regulatory mutations may account for their biological differences. *Science* **188**, 107–116 (1975).
11. Serfling, E., Jasin, M. & Schaffner, W. Enhancers and eukaryotic gene transcription. *Trends Genet.* **1**, 224–230 (1985).
12. Philipsen, S., Pruzina, S. & Grosveld, F. The minimal requirements for activity in transgenic mice of hypersensitive site 3 of the beta globin locus control region. *EMBO J.* **12**, 1077–1085 (1993).
13. Lettice, L. A. *et al.* Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 7548–7553 (2002).
14. Kioussis, D., Vanin, E., deLange, T., Flavell, R. A. & Grosveld, F. G. Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. *Nature* **306**, 662–666 (1983).
15. Driscoll, M. C., Dobkin, C. S. & Alter, B. P. Gamma delta beta-thalassemia due to a de novo mutation deleting the 5’ beta-globin gene activation-region hypersensitive sites. *Proceedings of the National Academy of Sciences* vol. 86 7470–7474 Preprint at https://doi.org/10.1073/pnas.86.19.7470 (1989).
16. Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum. Mol. Genet.* **21**, 3255–3263 (2012).
17. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371–375 (2014).
18. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **46**, 136–143 (2014).
19. Hong, J.-W., Hendrix, D. A. & Levine, M. S. Shadow enhancers as a source of evolutionary novelty. *Science* **321**, 1314 (2008).
20. Frankel, N. *et al.* Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**, 490–493 (2010).

21. Ghiasvand, N. M. *et al.* Deletion of a remote enhancer near ATOH7 disrupts retinal neurogenesis, causing NCRNA disease. *Nat. Neurosci.* **14**, 578–586 (2011).
22. Antosova, B. *et al.* The Gene Regulatory Network of Lens Induction Is Wired through Meis-Dependent Shadow Enhancers of Pax6. *PLoS Genet.* **12**, e1006441 (2016).
23. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).
24. Hardison, R. C. & Taylor, J. Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet.* **13**, 469–483 (2012).
25. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**, 1170–1187 (2016).
26. Thanos, D. & Maniatis, T. Virus induction of human IFNβ gene expression requires the assembly of an enhanceosome. *Cell* **83**, 1091–1100 (1995).
27. Kulkarni, M. M. & Arnosti, D. N. Information display by transcriptional enhancers. *Development* **130**, 6569–6575 (2003).
28. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
29. Rao, S., Ahmad, K. & Ramachandran, S. Cooperative binding between distant transcription factors is a hallmark of active enhancers. *Mol. Cell* (2021) doi:10.1016/j.molcel.2021.02.014.
30. Jindal, G. A. & Farley, E. K. Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Dev. Cell* **56**, 575–587 (2021).
31. Perry, M. W., Boettiger, A. N., Bothma, J. P. & Levine, M. Shadow enhancers foster robustness of Drosophila gastrulation. *Curr. Biol.* **20**, 1562–1567 (2010).
32. Scholes, C., Biette, K. M., Harden, T. T. & DePace, A. H. Signal Integration by Shadow Enhancers and Enhancer Duplications Varies across the Drosophila Embryo. *Cell Rep.* **26**, 2407–2418.e5 (2019).
33. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
34. Wang, X. & Goldstein, D. B. Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. *Am. J. Hum. Genet.* **106**, 215–233 (2020).
35. Furlong, E. E. M. & Levine, M. Developmental enhancers and chromosome topology. *Science* **361**, 1341–1345 (2018).
36. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
37. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38**, 1341–1347 (2006).
38. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
39. Hughes, J. R. *et al.* Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* **46**, 205–212 (2014).
40. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.* **21**, 207–226 (2020).
41. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
42. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles

of chromatin looping. *Cell* **159**, 1665–1680 (2014).
43. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
44. Kragesteen, B. K. *et al.* Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. *Nat. Genet.* 1 (2018) doi:10.1038/s41588-018-0221-x.
45. Kraft, K. *et al.* Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations. *Nat. Cell Biol.* **21**, 305–310 (2019).
46. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369–1384.e19 (2016).
47. Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557–572.e24 (2017).
48. Bolt, C. C. & Duboule, D. The regulatory landscapes of developmental genes. *Development* **147**, (2020).
49. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
50. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
51. Spitz, F. Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles. *Semin. Cell Dev. Biol.* **57**, 57–67 (2016).
52. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735 (2003).
53. 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).
54. Schoenfelder, S. & Fraser, P. Long-range enhancer-promoter contacts in gene expression control. *Nat. Rev. Genet.* **20**, 437–455 (2019).
55. Robson, M. I., Ringel, A. R. & Mundlos, S. Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3D. *Mol. Cell* **74**, 1110–1122 (2019).
56. Millán-Zambrano, G., Burton, A., Bannister, A. J. & Schneider, R. Histone post-translational modifications - cause and consequence of genome function. *Nat. Rev. Genet.* (2022) doi:10.1038/s41576-022-00468-7.
57. Hafner, A. & Boettiger, A. The spatial organization of transcriptional control. *Nat. Rev. Genet.* 1–16 (2022) doi:10.1038/s41576-022-00526-0.
58. Berman, B. P. *et al.* Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 757–762 (2002).
59. Hallikas, O. *et al.* Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47–59 (2006).
60. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263 (2009).
61. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
62. Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
63. Prabhakar, S. *et al.* Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res.* **16**, 855–863 (2006).

64. Visel, A. *et al.* Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.* **40**, 158–160 (2008).
65. Blow, M. J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* **42**, 806–810 (2010).
66. Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116–120 (2012).
67. Felsenfeld, G., Boyes, J., Chung, J., Clark, D. & Studitsky, V. Chromatin structure and gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 9384–9388 (1996).
68. Felsenfeld, G. & Groudine, M. Controlling the double helix. *Nature* **421**, 448–453 (2003).
69. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
70. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
71. Vierstra, J. *et al.* Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**, 1007–1012 (2014).
72. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).
73. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
74. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* vol. 316 1497–1502 Preprint at https://doi.org/10.1126/science.1141319 (2007).
75. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
76. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
77. May, D. *et al.* Large-scale discovery of enhancers from human heart tissue. *Nat. Genet.* **44**, 89–93 (2011).
78. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
79. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
80. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–21936 (2010).
81. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
82. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
83. Pérez-Lluch, S. *et al.* Absence of canonical marks of active chromatin in developmentally regulated genes. *Nat. Genet.* **47**, 1158–1167 (2015).
84. Regadas, I. *et al.* A unique histone 3 lysine 14 chromatin signature underlies tissue-specific gene regulation. *Mol. Cell* (2021) doi:10.1016/j.molcel.2021.01.041.
85. Dorighi, K. M. *et al.* Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Mol. Cell* **66**, 568–576.e4 (2017).
86. Rickels, R. *et al.* Histone H3K4 monomethylation catalyzed by Trr and mammalian

COMPASS-like proteins at enhancers is dispensable for development and viability. *Nat. Genet.* **49**, 1647–1653 (2017).
87. Gandara, L. *et al.* Developmental phenomics suggests that H3K4 monomethylation confers multi-level phenotypic robustness. *Cell Rep.* **41**, 111832 (2022).
88. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
89. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
90. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
91. He, Y. *et al.* Spatiotemporal DNA methylome dynamics of the developing mouse fetus. *Nature* **583**, 752–759 (2020).
92. He, P. *et al.* The changing mouse embryo transcriptome at whole tissue and single-cell resolution. *Nature* **583**, 760–767 (2020).
93. Gorkin, D. U. *et al.* An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744–751 (2020).
94. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
95. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
96. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
97. Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
98. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309–1324.e18 (2018).
99. Elmentaite, R., Domínguez Conde, C., Yang, L. & Teichmann, S. A. Single-cell atlases: shared and tissue-specific cell types across human organs. *Nat. Rev. Genet.* (2022) doi:10.1038/s41576-022-00449-w.
100. Preissl, S., Gaulton, K. J. & Ren, B. Characterizing cis-regulatory elements using single-cell epigenomics. *Nat. Rev. Genet.* **24**, 21–43 (2023).
101. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics* (2020) doi:10.1038/s41576-019-0209-0.
102. Kvon, E. Z. Using transgenic reporter assays to functionally characterize enhancers in animals. *Genomics* **106**, 185–192 (2015).
103. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).
104. Canver, M. C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192–197 (2015).
105. Sanjana, N. E. *et al.* High-resolution interrogation of functional elements in the noncoding genome. *Science* **353**, 1545–1549 (2016).
106. Klann, T. S. *et al.* CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat. Biotechnol.* **35**, 561–568 (2017).
107. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**, 377–390.e19 (2019).

108. Sabarís, G., Laiker, I., Noon, E. P.-B. & Frankel, N. Actors with Multiple Roles: Pleiotropic Enhancers and the Paradigm of Enhancer Modularity. *Trends in Genetics* **0**, (2019).
109. Osterwalder, M. *et al.* Characterization of Mammalian In Vivo Enhancers Using Mouse Transgenesis and CRISPR Genome Editing. *Methods Mol. Biol.* **2403**, 147–186 (2022).
110. 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).
111. Kvon, E. Z. *et al.* Comprehensive In Vivo Interrogation Reveals Phenotypic Impact of Human Enhancer Variants. *Cell* **180**, 1262–1271.e15 (2020).
112. Attanasio, C. *et al.* Fine tuning of craniofacial morphology by distant-acting enhancers. *Science* **342**, 1241006 (2013).
113. Dickel, D. E. *et al.* Ultraconserved Enhancers Are Required for Normal Development. *Cell* **172**, 491–499.e15 (2018).
114. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
115. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
116. Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28**, 739–750 (2018).
117. Avsec, Ž. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021).
118. de Almeida, B. P., Reiter, F., Pagani, M. & Stark, A. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat. Genet.* **54**, 613–624 (2022).
119. 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).
120. Liang, X., Evans, S. M. & Sun, Y. Development of the cardiac pacemaker. *Cell. Mol. Life Sci.* **74**, 1247–1259 (2017).
121. Galang, G. *et al.* ATAC-Seq Reveals an Isl1 Enhancer That Regulates Sinoatrial Node Development and Function. *Circ. Res.* **127**, 1502–1518 (2020).
122. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
123. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
124. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).
125. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 3583 (2019).
126. Snetkova, V. *et al.* Ultraconserved enhancer function does not require perfect sequence conservation. *Nat. Genet.* **53**, 521–528 (2021).
127. Drake, J. A. *et al.* Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat. Genet.* **38**, 223–227 (2006).
128. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
129. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
130. Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through

exome sequencing of the UK Biobank. *Nat. Genet.* **53**, 942–948 (2021).
131. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
132. Belkadi, A. *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 5473–5478 (2015).
133. Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. A. & Gilissen, C. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Hum. Mutat.* **36**, 815–822 (2015).
134. Spielmann, M. & Klopocki, E. CNVs of noncoding cis-regulatory elements in human disease. *Curr. Opin. Genet. Dev.* **23**, 249–256 (2013).
135. Turner, T. N. *et al.* Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *Am. J. Hum. Genet.* **98**, 58–74 (2016).
136. Turner, T. N. *et al.* Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171**, 710–722.e12 (2017).
137. Padhi, E. M. *et al.* Coding and noncoding variants in EBF3 are involved in HADDS and simplex autism. *Hum. Genomics* **15**, 44 (2021).
138. George, R. M. & Firulli, A. B. Hand Factors in Cardiac Development. *Anat. Rec.* **302**, 101–107 (2019).
139. Charité, J. *et al.* Role of Dlx6 in regulation of an endothelin-1-dependent, dHAND branchial arch enhancer. *Genes Dev.* **15**, 3039–3049 (2001).
140. Osterwalder, M. *et al.* HAND2 targets define a network of transcriptional regulators that compartmentalize the early limb bud mesenchyme. *Dev. Cell* **31**, 345–357 (2014).
141. Srivastava, D. *et al.* Regulation of cardiac mesodermal and neural crest development by the bHLH transcription factor, dHAND. *Nat. Genet.* **16**, 154–160 (1997).
142. Barnes, R. M. *et al.* Hand2 loss-of-function in Hand1-expressing cells reveals distinct roles in epicardial and coronary vessel development. *Circ. Res.* **108**, 940–949 (2011).
143. VanDusen, N. J. *et al.* Hand2 is an essential regulator for two Notch-dependent functions within the embryonic endocardium. *Cell Rep.* **9**, 2071–2083 (2014).
144. George, R. M. *et al.* Single cell evaluation of endocardial HAND2 gene regulatory networks reveals critical HAND2 dependent pathways impacting cardiac morphogenesis. *Development* **150**, (2023).
145. Goddard, L. M. *et al.* Hemodynamic Forces Sculpt Developing Heart Valves through a KLF2-WNT9B Paracrine Signaling Axis. *Dev. Cell* **43**, 274–289.e5 (2017).
146. Wang, K. *et al.* Differential roles of insulin like growth factor 1 receptor and insulin receptor during embryonic heart development. *BMC Dev. Biol.* **19**, 5 (2019).

Chapter 2 : Uncovering hidden enhancers through unbiased *in vivo* testing

In this chapter I describe my main project, for which the motivation was established in Chapter 1. This work has been posted on bioRxiv and will be submitted for publication as follows: Brandon J. Mannion, Marco Osterwalder, Stella Tran, Ingrid Plajzer-Frick, Catherine S. Novak, Veena Afzal, Jennifer A. Akiyama, Ismael Sospedra, Sarah Barton, Erik Beckman, Tyler H. Garvin, Patrick Godfrey, Janeth Godoy, Riana D. Hunter, Momoe Kato, Michael Kosicki, Anne N. Kronshage, Elizabeth A. Lee, Eman M. Meky, Quan T. Pham, Kianna von Maydell, Yiwen Zhu, Javier Lopez-Rios, Diane E. Dickel, Axel Visel, and Len A. Pennacchio. Uncovering Hidden Enhancers Through Unbiased *In Vivo* Testing.

Uncovering Hidden Enhancers Through Unbiased *In Vivo* Testing

Abstract

Transcriptional enhancers are a predominant class of noncoding regulatory elements that activate cell type-specific gene expression. Tissue-specific enhancer-associated chromatin signatures have proven useful to identify candidate enhancer elements at a genome-wide scale, but their sensitivity for the comprehensive detection of all enhancers active in a given tissue in vivo remains unclear. Here we show that a substantial proportion of in vivo enhancers are hidden from discovery by conventional chromatin profiling methods applied to the tissues in which they are active. In an initial comparison of over 1,200 in vivo validated tissue-specific enhancers with tissue-matched mouse developmental chromatin data, 14% (n=286) of active enhancers did not show canonical enhancer-associated chromatin signatures in the tissue in which they are active. To assess the prevalence of enhancers not detectable by conventional chromatin profiling approaches in more detail, we used a high throughput transgenic enhancer reporter assay to systematically screen over 1.3 Mb of mouse genomic sequence at two critical developmental loci. In total, we assessed 281 consecutive ~5 kb regions for in vivo enhancer activity in mouse embryos. We observed 88 instances of reproducible tissue-specific enhancer activity, 26% of which occurred in the absence of canonical enhancer-associated chromatin signatures in the respective tissue. In combination, our retrospective and prospective studies assessed only a small fraction of the mouse genome (0.1%) and identified 309 enhancers that are hidden from proper tissue-specific genome annotation using current chromatin-based enhancer identification approaches. Our findings both suggest the existence of tens of thousands of enhancers throughout the genome that remain undetected by prevalent chromatin profiling approaches and underscore the growing utility of incorporating complementary and multimodal data for enhancer detection.

Introduction

The importance of distant-acting enhancers in the temporal and spatial control of human gene expression is well established[1–4]. Proper transcriptional regulation by enhancers, which are particularly enriched near developmentally important genes, enables normal organismal development and function[5–7]. The initial characterization of enhancers was enabled by pioneering molecular studies of individual loci such as locus control regions at the β-globin locus[8–10], the availability of initial noncoding comparative genomic information from species such mouse, rat, and pufferfish[11–13], and powerful genomic approaches including ChIP-chip and subsequent next-generation sequencing techniques[14–16].

Dedicated genomic efforts such as ENCODE have sought to systematically identify enhancers via suitable *in vitro* and *in vivo* approaches[17]. Remarkably, while the human genome contains only ~20,000 protein-coding genes, these studies identified on the order of one million putative enhancers[18]. For example, one ChIP-Seq study that examined the enhancer-associated mark H3K27ac on a panel of 12 tissues isolated from eight mouse developmental stages covering critical phases of mammalian prenatal development (embryonic day [E]10.5 to postnatal day [P]0) uncovered ~200,000 candidate enhancers[19]. Chromatin accessibility (mapped by ATAC-seq) and the histone modifications H3K4me1 and H3K27ac (mapped by ChIP-seq) are widely utilized as canonical enhancer-associated chromatin marks[16,20–22]. However, the accuracy and practical utility of these data sets critically depends on the correlation of the marks examined with true *in vivo* activity, which can be assessed in transgenic reporter assays[23]. For instance, enhancer validation efforts using *in vivo* mouse reporter assays revealed a substantial number of false-positives in these putative H3K27ac-derived putative enhancer datasets[19]. Conversely, it remains unknown if the use of these canonical enhancer-associated chromatin marks comprehensively captures all *in vivo* enhancers or misses substantial numbers of bona fide *in vivo* enhancers (*i.e.*, false negatives)[24].

To assess the prevalence and characteristics of enhancers potentially missed in current chromatin-based datasets, we first performed comparisons of pre-existing large functionally validated enhancer sets with comprehensive mouse embryonic tissue chromatin atlases. These retrospective analyses provided initial indications that many *in vivo* enhancers are missed by chromatin-based discovery strategies that rely on canonical enhancer-associated chromatin marks[19,25]. Next, we conducted a comprehensive prospective analysis in which we performed nearly 300 transgenic enhancer assays[26] for the unbiased tiling of over 1.3 Mb of the mouse genome, which uncovered dozens of hidden enhancers (*i.e.*, without detectable canonical enhancer-associated chromatin marks) in these regions.

Results

Many *in vivo* enhancers show no canonical enhancer marks

As an initial exploration of the comprehensiveness of chromatin-based enhancer mapping strategies, we used the VISTA Enhancer Browser database (https://enhancer.lbl.gov)[25] to retrospectively assess the relationship between enhancer-associated chromatin marks and validated enhancer activity *in vivo*. To date, this resource includes over 3,200 human and mouse elements that have been tested for enhancer-reporter activity, primarily at mouse embryonic day 11.5 (E11.5), a stage when multiple developing tissues (*e.g.,* limb, heart, brain, craniofacial structures) can be assessed through whole-mount imaging in mice and compared with their functional counterparts in humans. We focused on the 1,272 validated enhancers that drove reproducible expression in one or more of the following anatomical structures: forebrain (n=450 enhancers); midbrain (398); hindbrain (366); craniofacial region (262); limb (304); and heart (272). We compared these data to chromatin data (H3K27ac ChIP-seq, H3K4me1 ChIP-seq, and ATAC-seq,) from these same tissues collected from E11.5 mouse embryos (**Table S2.1**). For each of the six tissues, we examined the presence of canonical enhancer-associated chromatin signatures at each positive element's endogenous site (**Fig. 2.1a-b, Table S2.2**).

For example, for the 304 VISTA limb enhancers, we found that 116 (38%) do not have a limb-specific H3K27ac enhancer-associated mark (**Fig. 2.1c**). In addition, of these 116 limb enhancers lacking H3K27ac marks, 60 (20%) also lack an H3K4me1 mark. Finally, 45 of these limb enhancers (15% of VISTA limb-positive elements) are completely lacking any of the three enhancer-associated chromatin marks (H3K27ac ChIP-seq, H3K4me1 ChIP-seq, or ATAC-seq) in limb tissue. Across all six tissues examined, these "hidden" enhancers represent 9% to 25% of VISTA enhancers (**Fig. S2.3**). Overall, we found that 50% (1028) of tissue-specific VISTA enhancers have all three marks, 22% (461) have at least two marks, 13% (277) have only one of the three marks, and 14% (286) are hidden enhancers without any of the three marks in the corresponding tissue. The relative proportions of these chromatin mark categories are similar across the six considered tissues (**Fig. S2.3**). We observed hidden enhancers in all six developing tissues that we assessed at E11.5 both for their enhancer-associated marks and transgenic enhancer-reporter activity, which suggests their existence is a general phenomenon across other cell and tissue types.

Mouse *in vivo* tiling assay uncovers additional hidden enhancers

Since many of the enhancers reported in the literature and VISTA database were found through chromatin signature-guided enhancer discovery screens, retrospective intersections are likely to underestimate the proportion of enhancers lacking canonical chromatin signatures. To assess this phenomenon in a more unbiased manner, we selected two separate loci (*Gli3* and *Smad3/Smad6*) to test the enhancer activity of 281 overlapping elements regardless of their chromatin state. The *Gli3* gene encodes a transcription factor that is involved in pathways for the development of the limb, face, and nervous system[27–29]. Apart from *Gli3* itself, the flanking region included in the tiling is generally depleted of other genes and includes a gene desert that spans over 800 kb[30]. Dozens of regions (n=38) across the locus are predicted to be enhancers based on tissue-specific H3K27ac (**Table S2.4, Fig. 2.2a**), and prior limited candidate enhancer studies within this locus identified enhancers active in the limb and brain in E11.5 mouse embryos[31–33]. Additionally, we

performed unbiased tiling across a second locus that encompasses the *Smad3* and *Smad6* genes (**Fig. S2.4**). As with the *Gli3* locus, the *Smad3*/*Smad6* locus considered for tiling also includes several (n=86) H3K27ac-marked regions (**Table S2.4**). While *Smad3* is broadly expressed in all six of the tissues examined in this study at mouse E11.5, expression of *Smad6* is highest in mouse E11.5 heart (**Fig. S2.5**), consistent with its importance in cardiovascular development[34]. We designed elements ~5 kb in size with boundaries chosen to fully capture complete H3K27ac-enriched regions where possible and with overlaps to adjacent elements in order to tile across both loci. Altogether, we tested 281 of the sequences in a site-directed mouse *in vivo* transgenic assay[26,35] and assessed enhancer activity in six tissues, for a total number of 1,686 enhancer-tissue observations (**Fig. 2.2b-d**). Collectively, the tested elements span over 1.3 Mb (approximately 1 and 0.3 Mb of the *Gli3* and *Smad3/Smad6* loci, respectively) of the mouse genome.

We observed that 63 of 281 tested elements showed reproducible enhancer-reporter activity at mouse embryonic day 11.5 (E11.5) in at least one tissue (**Fig. 2.2b, Fig. S2.4**). A majority of elements tested around the *Gli3* loci showed reproducible LacZ activity in the developing brain, limb, and craniofacial regions (**Fig. S2.6)**, all tissues where *Gli3* is expressed[27–29]. Similarly, elements tested around the *Smad3/Smad6* loci show activity in a variety of developing tissues (**Table S2.2**), which likely reflects both the observed broad expression patterns of *Smad3* and the known tissue-specific roles of *Smad6* in cardiovascular development[34].

Similar to the retrospective VISTA study, we focused on six tissues (forebrain, midbrain, hindbrain, craniofacial structures, limb, heart) to assess the relationship between experimental enhancer data in transgenic reporter assays and chromatin data (H3K27ac, H3K4me1, ATAC-seq, **Table S2.2**). The 63 elements that showed reproducible *in vivo* enhancer-reporter activity in one or more developing tissue altogether represented 88 tissue-enhancer activities. We used these 88 tissue-enhancer activities to compare with stage- and tissue-matched chromatin data and observed that 23 (26%) were hidden, *i.e.,* they lack enhancer-associated chromatin marks in their active tissue. We observed that hidden enhancers from the unbiased tiling represent a larger proportion of tissue-specific enhancers (26%) relative to the retrospective VISTA enhancer comparison (14%) described above (**Fig. 2.3a**). We identified hidden enhancers in all 6 tissues under investigation, including forebrain (of 14 forebrain-enhancers, 2 were hidden) and hindbrain (of 16 hindbrain-enhancers, 7 were hidden) (**Fig. 2.3b**). Aside from a lack of the three enhancer-associated chromatin marks, we found a majority of these hidden enhancers also were without alternative histone marks examined by ENCODE (H3K27me3, H3K36me3, H3K4me2, H3K4me3, H3K9ac, H3K9me3; **Fig. 2.3b, Fig. S2.6**). Further, hidden enhancers across the *Gli3* locus were active in tissues that show Gli3 mRNA expression *in situ* and marked enhancers at the same developmental stage, which supports their role in regulating neighboring gene expression (**Fig. S2.7)**. Additionally, we generated mouse lines with over 800 kb of sequence (upstream of *Gli3* TSS) deleted. Notably, we did not observe any gross phenotypic changes in these deletion lines, which is similar to prior studies on single enhancer knockouts at this locus[33]. Altogether, our retrospective VISTA study and unbiased systematic experimental testing uncovered 309 tissue-specific hidden enhancers, supporting the existence of substantial numbers of missing enhancers genome-wide.

Hidden enhancers are indistinguishable from their marked counterparts

We next assessed the properties of hidden versus marked enhancers in an attempt to explain their functional differences. From the VISTA retrospective study and the unbiased tiling, both hidden and marked enhancers have similar levels of evolutionary conservation, *i.e.*, both categories have

elevated conservation scores (phastCons) relative to genomic background and show no significant difference in the level of conservation (**Fig. S2.8**). Within each tiling locus we did not find specific transcription factor binding sites that were enriched in hidden enhancers relative to marked enhancers (**Table S2.5**). Similarly, by functional enrichment analysis there are no significant biological processes or phenotypes that distinguish hidden enhancers from their marked counterparts (**Table S2.6**). To explore the potential contributions of transposable elements (TEs) within these enhancer regions[36], we enumerated the TEs within hidden and marked enhancers to assess whether particular TE families were enriched or depleted in either group. We found similar proportions of LINE, SINE, DNA transposon, and other repeat element families between hidden and marked enhancers (**Fig. S2.9**). These comparisons further confirm that hidden enhancers show all hallmarks of bona fide *in vivo* enhancers with canonical marks.

Some hidden enhancers can be identified from alternative chromatin data

Given the absence of canonical enhancer-associated chromatin marks in embryonic mouse tissue-derived data, we examined if complementary chromatin data types offer potential avenues for the discovery of these hidden enhancers. We first evaluated if hidden enhancer activity at E11.5 could be the outcome of residual LacZ reporter activity from enhancer activity that occurred at an earlier developmental stage. Of the 309 tissue-specific hidden enhancers assayed at E11.5, 173 (56%) have enhancer-associated chromatin marks at an earlier stage, *i.e.*, H3K27ac and/or H3K4me1 at embryonic day 10.5 (E10.5) (**Fig. 2.4a**).

Next, we examined if available single-cell chromatin data could resolve enhancer-associated chromatin marks around hidden enhancers that might have been missed from standard chromatin data derived from bulk tissue preparations. Of the six tissues for which we compared transgenic enhancer-reporter activity with the corresponding mouse tissue chromatin data from ENCODE, two tissues (forebrain, hindbrain) have single nucleus ATAC-seq (snATAC-seq) data across early mouse development[37,38]. Of the 44 hidden enhancers that are active in either the forebrain or the hindbrain, only 8 (18%) could be identified via available corresponding single cell data (**Fig. 2.4b**).

Since 236 (83%) of the 286 hidden enhancers identified in the VISTA retrospective study are the human orthologues of human-mouse conserved sequences tested in mouse transgenic enhancer assays, we also examined if available human tissue-matched epigenomic data would have predicted any of these hidden enhancers. For 112 human-derived hidden enhancers that did not have enhancer-associated chromatin marks either in earlier (E10.5) or in single-cell chromatin data, 49 were assessable with available tissue-matched, similar-staged human chromatin data from craniofacial, heart, and limb bud tissues. Only 10 (20%) showed enhancer marks in available human chromatin data. Altogether through these stage- and tissue-matched analyses, 118 (38%) of the originally identified hidden enhancers could not be identified despite at least two of the three complementary data types being available (**Fig. 2.4c**).

Finally, while our major focus has been on comparing experimentally validated enhancers to their chromatin profiles in their exact tissue of activity, we sought to explore if chromatin data from disparate sources (*i.e.*, without the previous constraints to precisely match developmental stage and tissue type to the tissue-specific enhancers of this study) could nonetheless be informative for the identification of hidden enhancers. We used a broad catalog of candidate cis-regulatory elements (cCREs) derived from chromatin-based profiling of various human and mouse cell lines and tissues[18]. Collectively, the human- and mouse-derived cCREs annotated for enhancer-like signatures cover over 14% of the mouse genome. Across both the pre-existing VISTA and tiling

studies, we found that 243 of 271 (89.7%) of the hidden enhancers showed enhancer-like signatures (ELS) in at least one data set from this comprehensive cCRE catalog (**Fig. S2.10**). However, a majority of elements (1225 of 1505; 81.4%) that were negative in our transgenic mouse assay also overlapped with enhancer-like cCREs. Additionally, hidden enhancers that are not marked by cCREs from this expanded search have similar levels of elevated evolutionary conservation as those with cCREs, which further supports their functional constraint (**Fig. S2.11)**. Altogether, intersection with this generalized collection of cCREs from differing cell types and developmental stages has no substantial predictive power beyond the use of tissue-specific chromatin data sets.

Discussion

In this study, we report the existence of hundreds of hidden enhancers in the human/mouse genome that lack canonical enhancer-associated marks in chromatin profiling data from the tissue in which they are active. This includes a retrospective analysis of over 1,200 *in vivo* validated tissue-specific enhancers in VISTA and a prospective tiling study of 281 candidate sequences, which implemented a recently scaled transgenic assay[26] to systematically test elements for mouse *in vivo* enhancer activity across over 1.3 Mb of a mammalian genome. In contrast to previous *in vitro* approaches or studies in humans, mice, and *Drosophila*[39–44], the present screen represents a comprehensive and systematic assessment of sizable genomic intervals across two mammalian loci for bona fide *in vivo* enhancer activity. We show that a majority of tissue-specific enhancers have corresponding enhancer-associated chromatin marks in corresponding tissue(s), which supports the continued use of these datasets for candidate enhancer identification. However, we also show that reliance on the current, prevalent applications of these chromatin-based assays to identify candidate enhancers misses a notable portion of so-called hidden enhancers that are active in the transgenic *in vivo* reporter assay but do not show any of the noted marks. Within the tiling study across two separate loci, we show that the tissue-specific enhancer-reporter activities of hidden enhancers are similar to those of their marked counterparts, which suggests these sequences contribute tissue-specific enhancer activity at their endogenous sites in ways similar to canonical enhancers. Although we deleted a large genomic interval at the *Gli3* locus that encompasses several reproducible tissue-specific enhancers (including the previously characterized mm1179[33]), we did not observe any changes in viability, limb morphology, or other related phenotypes. This is similar to prior observations at this locus and highlights the challenges in enhancer dissection when multiple enhancers can provide redundancy or a buffering effect to maintain gene expression. Future work to understand the roles of both marked and hidden enhancers within a locus may require additional genome engineering to account for this system. We also find that the levels of evolutionary conservation between both marked and hidden enhancers are similarly elevated relative to random genomic background, which is also supportive of their functional relevance or utility both within and outside their endogenous contexts. Apart from the absence of enhancer-associated chromatin marks, we could not identify sequence, genomic, or other epigenomic properties that could distinguish hidden enhancers from their marked counterparts.

We found many of these hidden enhancers can be identified by considering complementary data either from other time points, single cell chromatin measurements, or other species. While public single cell chromatin accessibility data are currently limited to a few tissues in mice[37,38], it is likely that additional developing tissues (*e.g.,* face, limb, heart) will soon be surveyed at single-cell resolution. As supported by our comparisons of hidden forebrain and hindbrain enhancers, these single-cell approaches should enable the resolution of both common and rare cell types in tissues and, subsequently, the identification of enhancers missed by bulk tissue-derived data. Additionally, human chromatin data within a similar developmental window are currently available only from a few tissues (*i.e.,* heart[45], face[46], limb bud[47]), and future characterization of other similarly staged human tissues should facilitate cross species comparisons of enhancer-associated chromatin marks and *in vivo* enhancer activity. Further, we utilized cCREs derived from various human and mouse cell types to recover a majority of the hidden enhancers that were previously missed by the available stage- and tissue-matched chromatin data. Although this raises the possibility that additional chromatin data from other sources can assist with the identification of these hidden enhancers, this is at the expense of poor specificity or selectivity for elements that do validate as

active enhancers in the mouse *in vivo* system. Moreover, these pooled data cannot provide insights on the specific cell type and/or developmental stage of candidate enhancers.

Across the two tiling loci we found over 80 instances of reproducible tissue-specific enhancer activity representing ~26% of which are hidden enhancers in their corresponding tissue. We focused on only six tissues from which we could compare between tissue-matched chromatin properties and mouse *in vivo* data at E11.5, yet there are vast numbers of other tissues and developmental time points relevant for enhancer identification[48]. With some estimates of hundreds of thousands to nearly one million candidate enhancers in mammalian genomes, one might speculate from our tiling study that there are tens of thousands of additional enhancers unaccounted for by current approaches[18]. As sequencing expands to cover the full range of human tissues, diversity, environmental perturbations, and as related technologies provide even higher resolution approaches to probe gene regulatory activity, we can expect to better understand and annotate the unique characteristics of hidden enhancers and their functional significance in transcriptional regulation.

Methods

Experimental model

All animal work was reviewed and approved by the Lawrence Berkeley National Laboratory Animal Welfare and Research Committee. All mice used in this study were housed at the Animal Care Facility (ACF) of 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 deletion mouse models were performed in the *Mus musculus* FVB strain.

ENCODE mouse chromatin and RNA-seq data

Processed mouse chromatin data[19] (ATAC-seq; ChIP-seq for H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K27me3, H3K9ac, H3K36me3, H3K9me3) and RNA-seq data[49] were downloaded from the ENCODE resource portal (https://www.encodeproject.org/). Details on the generation and processing of these data are available here: https://www.encodeproject.org/pipelines/. See Supplementary Table 2.1 for a listing of all the bulk tissue mouse data used for chromatin intersections or tissue expression analyses.

VISTA enhancers

Human and mouse candidate enhancers were tested in a mouse *in vivo* transgenic reporter assay, as previously described[26] (see also "Locus selection for tiling and mouse in vivo enhancer validation"). Candidate enhancers were assessed for reproducible enhancer-reporter activity in forebrain, midbrain, hindbrain, craniofacial structures (*e.g.,* branchial arches; nose; facial mesenchyme), limb, and heart. The genomic coordinates (assembly mm10) of these elements were downloaded from the VISTA Enhancer Browser (https://enhancer.lbl.gov/)[25]. Human elements were lifted over from hg38 to mm10 via the UCSC liftOver tool using *minMatch=0.1*[50].

Chromatin intersections and hidden enhancer identification

Mouse *in vivo* validated elements from both the VISTA Enhancer Browser and the tiling assay were intersected with tissue-specific ENCODE chromatin data via bedtools[51] (v2.29.0) to check for the presence or absence of enhancer-associated chromatin signatures (*e.g.*, tissue-specific mouse E11.5 peaks from H3K27ac ChIP-seq, H3K4me1 ChIP-seq, and/or ATAC-seq data) within each elements' genomic coordinates. Elements with reproducible enhancer-reporter activity (positive elements) but without any of the three enhancer-associated chromatin signatures in the relevant tissue(s) were designated as hidden enhancers. Positive elements with any (up to all) of the three enhancer-associated chromatin signatures were considered marked enhancers. Positive elements were also checked for overlap with other chromatin features available from mouse ENCODE: DNase-seq, H3K27me3 ChIP-seq, H3K36me3 ChIP-seq, H3K4me2 ChIP-seq, H3K4me3 ChIP-seq, H3K9ac ChIP-seq, and H3K9me3 ChIP-seq. Both mouse embryonic days 10.5 (E10.5) and 11.5 (E11.5) data were used for the above analyses.

Locus selection for tiling and mouse *in vivo* enhancer validation

Coordinates used for the *Gli3* locus are chr13:14,626,494-15,785,614 (mm10). Coordinates used for the *Smad3/Smad6* locus are chr9:63,685,831-64,099,907 (mm10). Tiling elements

approximately 5kb in size (with overlap to adjacent tiling elements) were cloned into the pCR4-Shh::lacZ-H11 vector (Addgene plasmid #139098), which includes the mouse Shh promoter, the *LacZ* gene for enzymatic, colorimetric readout, and flanking homology arms that enable site-specific integration at the H11 locus[26]. Each tiling element was PCR amplified from mouse BAC DNA template (CHORI). A mixture of the reporter construct, Cas9 protein (Integrated DNA Technologies, catalog #1081058), and sgRNAs were transferred by microinjection into the pronucleus of mouse embryos (FVB strain) and then transferred to the uterus of pseudopregnant females (CD-1 strain). Transgenic embryos were then collected at mouse embryonic day 11.5 (E11.5) for LacZ staining and the assessment of enhancer-reporter activity in several developing tissues (*e.g.,* forebrain, midbrain, hindbrain, craniofacial, heart, and limb). For more detailed steps and information on the workflow that spans cloning, mouse colony management, microinjection, and embryo staining, refer to the recently published protocol[35]. Transgenic embryos for each tiling element were assessed for reproducible enhancer-reporter activity in separate, independent embryos. Genomic coordinates, transgenic embryo images, and tissue annotations for each element are available on the VISTA Enhancer Browser (https://enhancer.lbl.gov).

Generation of large *Gli3* upstream deletion using CRISPR-Cas9

Mouse lines lacking the large genomic interval upstream of *Gli3* (within the *Gli3* TAD) were generated using CRISPR-Cas9 editing. Two pairs of guide RNAs (gRNAs) 5' and 3' to this large interval were designed using CHOPCHOP[52]. Cas9 protein (catalog #1081058), tracrRNA (catalog #1072533), and crRNAs were ordered from Integrated DNA Technologies. Upstream deletion mice were generated via injection of the CRISPR-Cas9 mixture (final concentrations: Cas9 protein at 20ng/μl; four gRNAs at 50ng/μl; tracrRNA at 50ng/μl; injection buffer: 10mM Tris, pH 7.5; 0.1mM EDTA) into the pronuclei of FVB embryos and then transferred to the uteri of CD-1 pseudopregnant females (similar to microinjection protocol described above). Founder (F0) mice were genotyped for the targeted deletion with PCR amplification of primers immediately adjacent to the deletion breakpoints. From this approach, multiple F0s with a large upstream deletion were obtained and subsequently maintained through outcrossing with wildtype FVB mice. Unless otherwise noted, all founder lines were fertile and displayed normal pre- and postnatal viability.

Evolutionary conservation

PhastCons scores were downloaded from the UCSC Genome Browser at https://hgdownload.cse.ucsc.edu/goldenPath/mm10/phastCons60way/. phastCons scores were calculated for each element (mean across region) and used to compare the levels of evolutionary conservation between different categories of tested elements (*e.g.,* hidden enhancers vs. marked enhancers). The Kolmogorov-Smirnov test was used to assess potential differences in phastCons distributions between the considered enhancer categories.

Additional epigenomic data

Publicly available single cell chromatin accessibility data from mouse E11.5 forebrain (GSE100033)[37] and mouse E11.5 cerebellum (https://apps.kaessmannlab.org/mouse_cereb_atac/)[38] were used to compare differences between bulk tissue and single cell assays in the resolution of enhancer-associated chromatin signatures, *i.e.,* if there were open chromatin regions absent in bulk chromatin data but detected in single cell data. Human chromatin data from approximately stage-matched limb bud (GSE42413)[47], heart (GSE137731)[45], and face (GSE97752)[46] were used to evaluate if hidden enhancers

from human sequence could be identified with these complementary data. The candidate cis-regulatory elements (cCRE) catalog was provided by Jill Moore and Zhiping Weng[18].

Transcription factor motif and functional enrichment analyses

HOMER[53] version 4.10 was used to assess enrichment of both known and *de novo* motifs within hidden enhancers, via *findMotifsGenome.pl* and the following parameters: -size given -len 8,9,10,12,14 -bg <background file = all positive VISTA enhancers>. GREAT[54] version 4.0.4 (http://great.stanford.edu/public/html/) was used to assess the enrichment of biological ontologies in hidden enhancers, via the basal plus extension setting (5,000bp upstream, 1,000bp downstream, distal up to 1Mbp).

Repeat element analysis

Repeat elements annotated across the mouse genome (by family, class, and name) were obtained from the RepeatMasker track via the UCSC Table Browser as a BED file. The number of repeat elements within each tested element's genomic interval was tallied by *bedtools intersect* [51] and used to compare the proportion of repeat element classes between enhancer mark categories (*e.g.,* marked enhancers vs. hidden enhancers). Human elements (tested in the mouse *in vivo* system) were lifted over from hg38 to mm10 via the UCSC liftOver tool using minMatch=0.1.

Figures

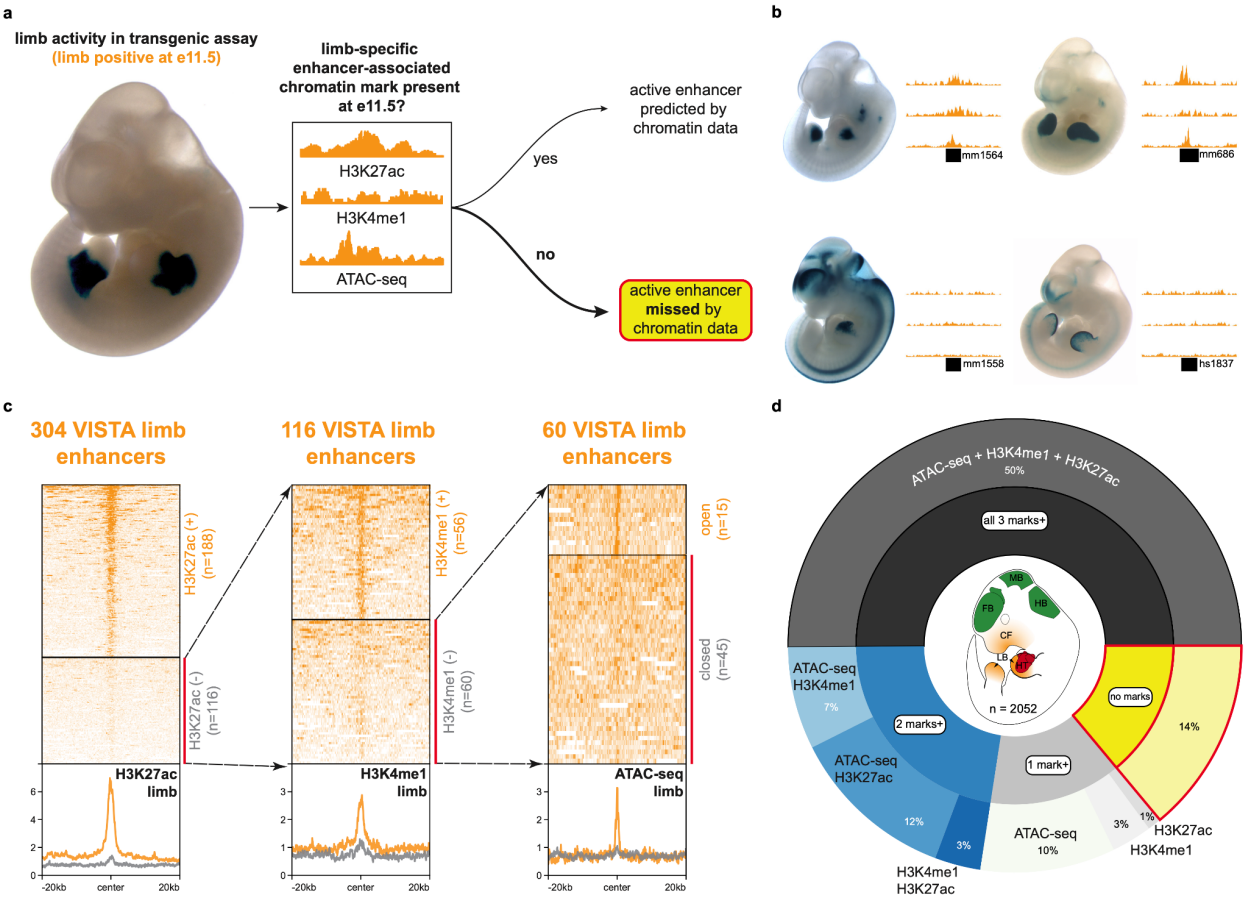


Figure 2.1. Mouse *in vivo* enhancers without canonical enhancer-associated chromatin marks. (a) Approach to retrospectively identify active enhancers without tissue-specific enhancer-associated chromatin marks. **(b)** Examples of active limb enhancers with (top row) and without (bottom row) enhancer-associated chromatin marks in stage-matched embryonic limb tissue. See **Fig. S2.1** for examples of active enhancers without these marks in other tissues. **(c)** Chromatin profiles of active limb enhancers with and without H3K27ac (ChIP-seq), H3K4me1 (ChIP-seq), or open chromatin (ATAC-seq). See **Fig. S2.2** for another example of chromatin mark filtering for forebrain enhancers. **(d)** Proportion of VISTA enhancers across six tissues (forebrain, midbrain, hindbrain, craniofacial structure, limb, heart) with and without enhancer-associated chromatin marks. For this study we focused on the VISTA enhancers with activity ("positive" elements) in the above six tissues (**Table S2.3**). Active enhancers without any of these chromatin marks are in yellow.

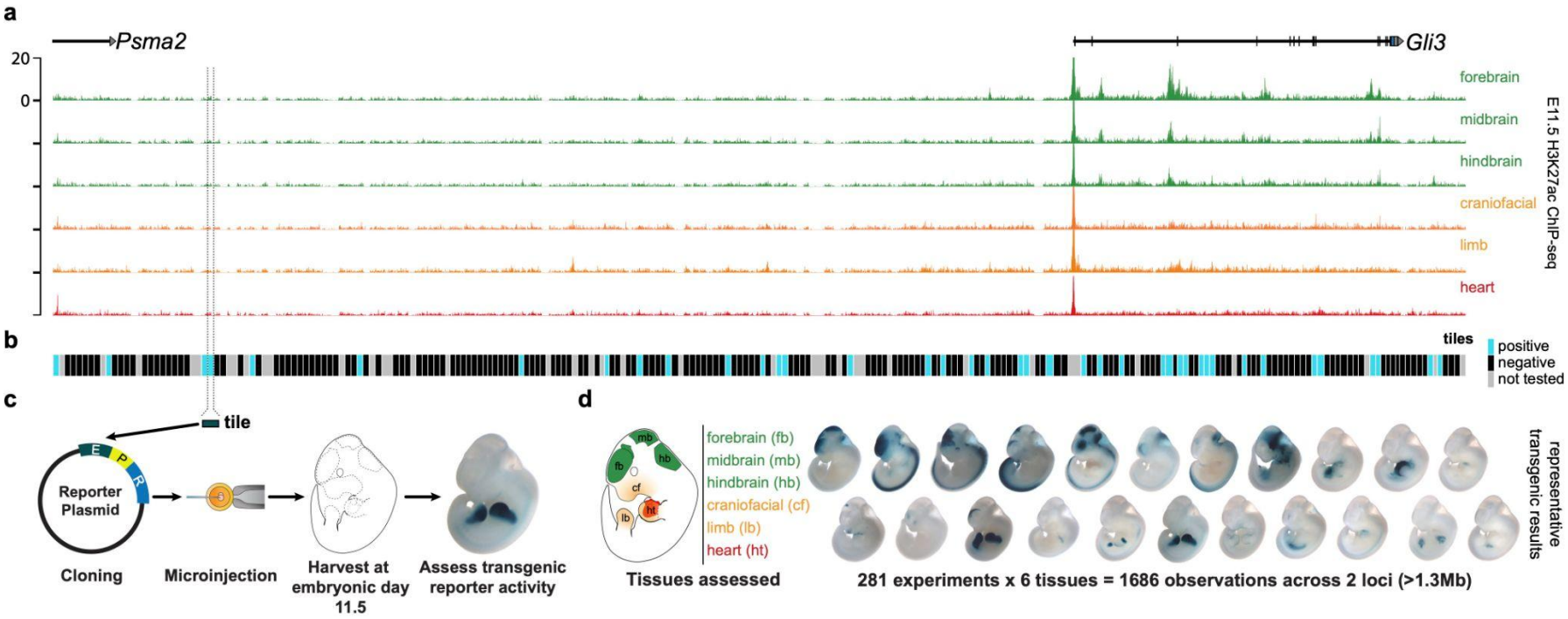

a
Psma2
Gli3
20
0
forebrain
midbrain
hindbrain
craniofacial
limb
heart
E11.5 H3K27ac ChIP-seq
b
tiles
positive
negative
not tested
c
tile
E
P
R
Reporter Plasmid
Cloning
Microinjection
Harvest at embryonic day 11.5
Assess transgenic reporter activity
d
mb
fb
hb
cf
ht
lb
forebrain (fb)
midbrain (mb)
hindbrain (hb)
craniofacial (cf)
limb (lb)
heart (ht)
Tissues assessed
representative transgenic results
281 experiments x 6 tissues = 1686 observations across 2 loci (>1.3Mb)

Figure 2.2. Systematic tiling for the unbiased identification of mouse in vivo enhancers. (a) *Gli3* locus with H3K27ac ChIP-seq data (ENCODE) for six tissues. **(b)** Elements for the unbiased tiling assay were ~5 kb in size and designed to overlap with adjacent elements. Elements that were tested and that had reproducible enhancer-reporter activity in the mouse *in vivo* transgenic assay are shaded blue. We observed 63 tested elements with tissue-specific enhancer-reporter activity at mouse embryonic day 11.5 (E11.5) with tissue-specific activity. Of these 63 enhancers, 36 show reproducible enhancer-reporter activity in multiple tissues at E11.5. Elements without reproducible activity are shaded black. Elements not successfully tested are shaded gray. **(c)** Approach for testing each tile in the mouse *in vivo* transgenic assay. E, enhancer; P, promoter; R, reporter. (**d**) Left: Depiction of tissues that were checked for reproducible enhancer-reporter activity. Right: Example transgenic results from tiling across the *Gli3* and *Smad3/Smad6* loci (see **Fig. S2.4** for the *Smad3/Smad6* locus).

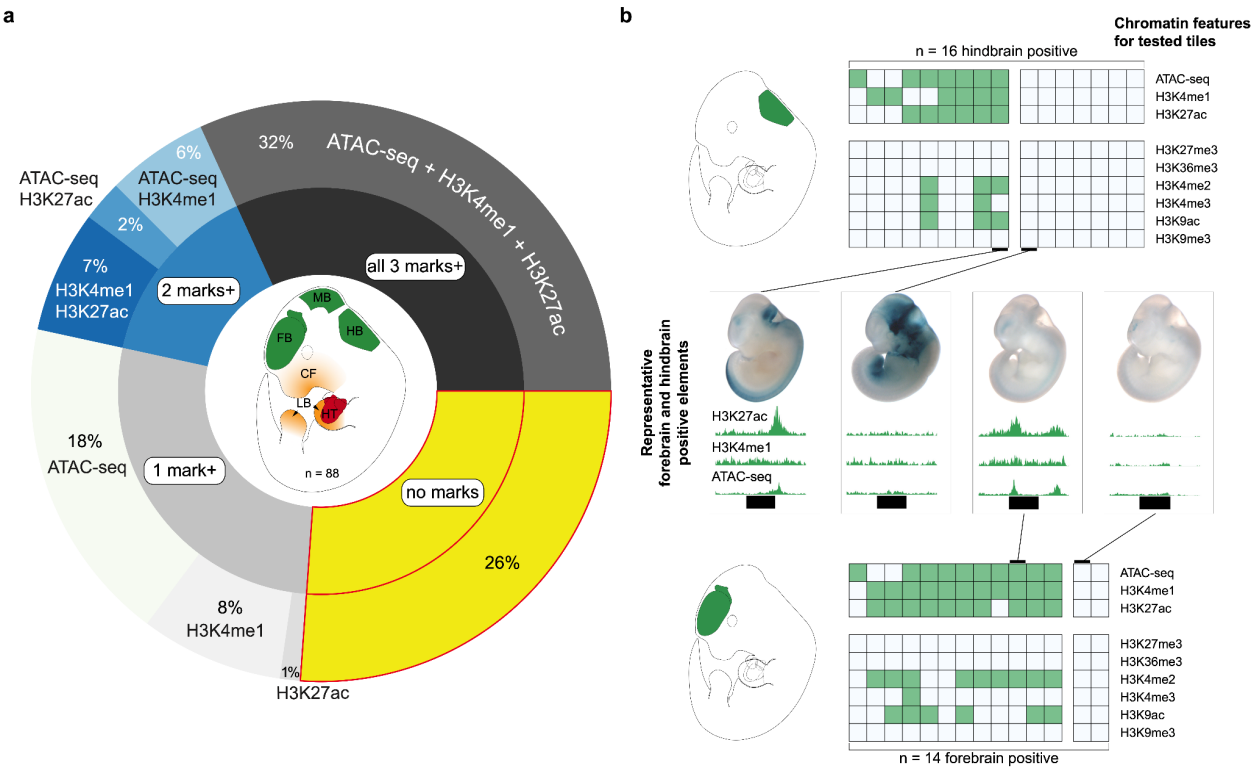


Figure 2.3. Active enhancers from unbiased tiling with and without enhancer-associated chromatin marks. (a) Proportion of active enhancers from unbiased tiling with and without enhancer-associated chromatin marks. Active enhancers without any of these chromatin marks are in yellow. **(b)** Example of active hindbrain (n=16) and active forebrain enhancers (n=14) from the tiling assay. Columns within the square table represent a tested element (a region with genomic coordinates), whereas rows within a single column represent the chromatin feature (shaded green if a peak in the given chromatin feature is present) for that particular element. The square tables are split into two main categories, those elements with at least one of the considered enhancer-associated chromatin marks present (left) and those without any of the three considered enhancer-associated chromatin marks (right). Additional chromatin data depicted show that a portion of hidden enhancers are not marked by any of the chromatin marks assayed by ENCODE (**Fig. S2.6**). Representative transgenic results (two enhancers for hindbrain; two for forebrain) are depicted as well as the chromatin profile for the relevant element.

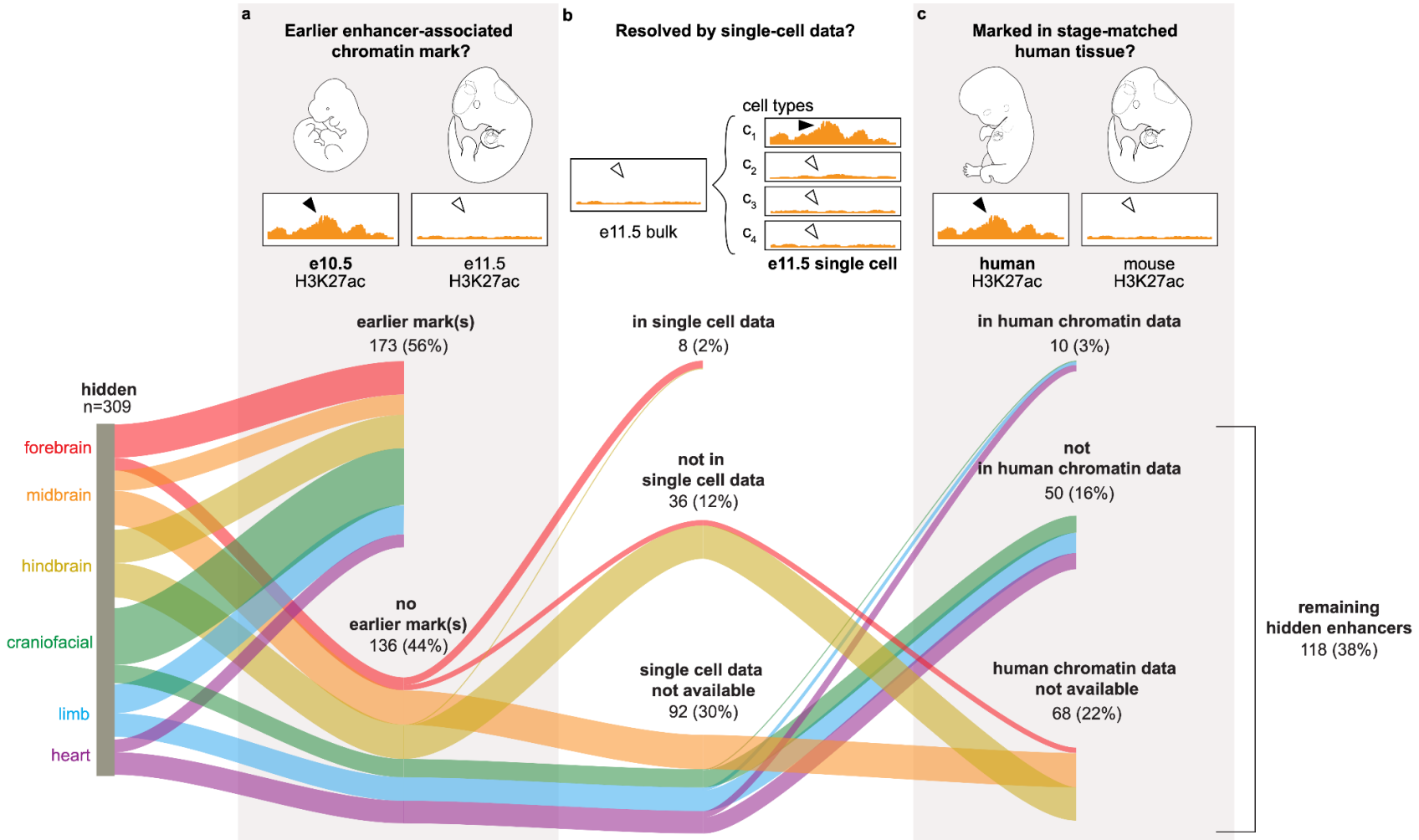

a
Earlier enhancer-associated chromatin mark?
e10.5 H3K27ac
e11.5 H3K27ac
b
Resolved by single-cell data?
cell types
e11.5 bulk
e11.5 single cell
c
Marked in stage-matched human tissue?
human H3K27ac
mouse H3K27ac
hidden n=309
forebrain
midbrain
hindbrain
craniofacial
limb
heart
earlier mark(s) 173 (56%)
no earlier mark(s) 136 (44%)
in single cell data 8 (2%)
not in single cell data 36 (12%)
single cell data not available 92 (30%)
in human chromatin data 10 (3%)
not in human chromatin data 50 (16%)
human chromatin data not available 68 (22%)
remaining hidden enhancers 118 (38%)

Figure 2.4. Hidden enhancers cannot be fully recovered from alternative chromatin data. (a) Hidden enhancers that are active at E11.5 are assessed for earlier enhancer-associated chromatin marks at E10.5. **(b)** Hidden enhancers based on bulk tissue chromatin data and that do not have earlier enhancer-associated chromatin marks are assessed for corresponding enhancer-associated chromatin marks from available single cell chromatin accessibility data. **(c)** Hidden enhancers neither recoverable from earlier enhancer-associated chromatin marks nor single cell chromatin accessibility data are assessed for enhancer-associated chromatin marks in available human chromatin data. Percentages represent the proportion of the filtered elements relative to the starting set of 309 hidden enhancers.

Supplementary Materials

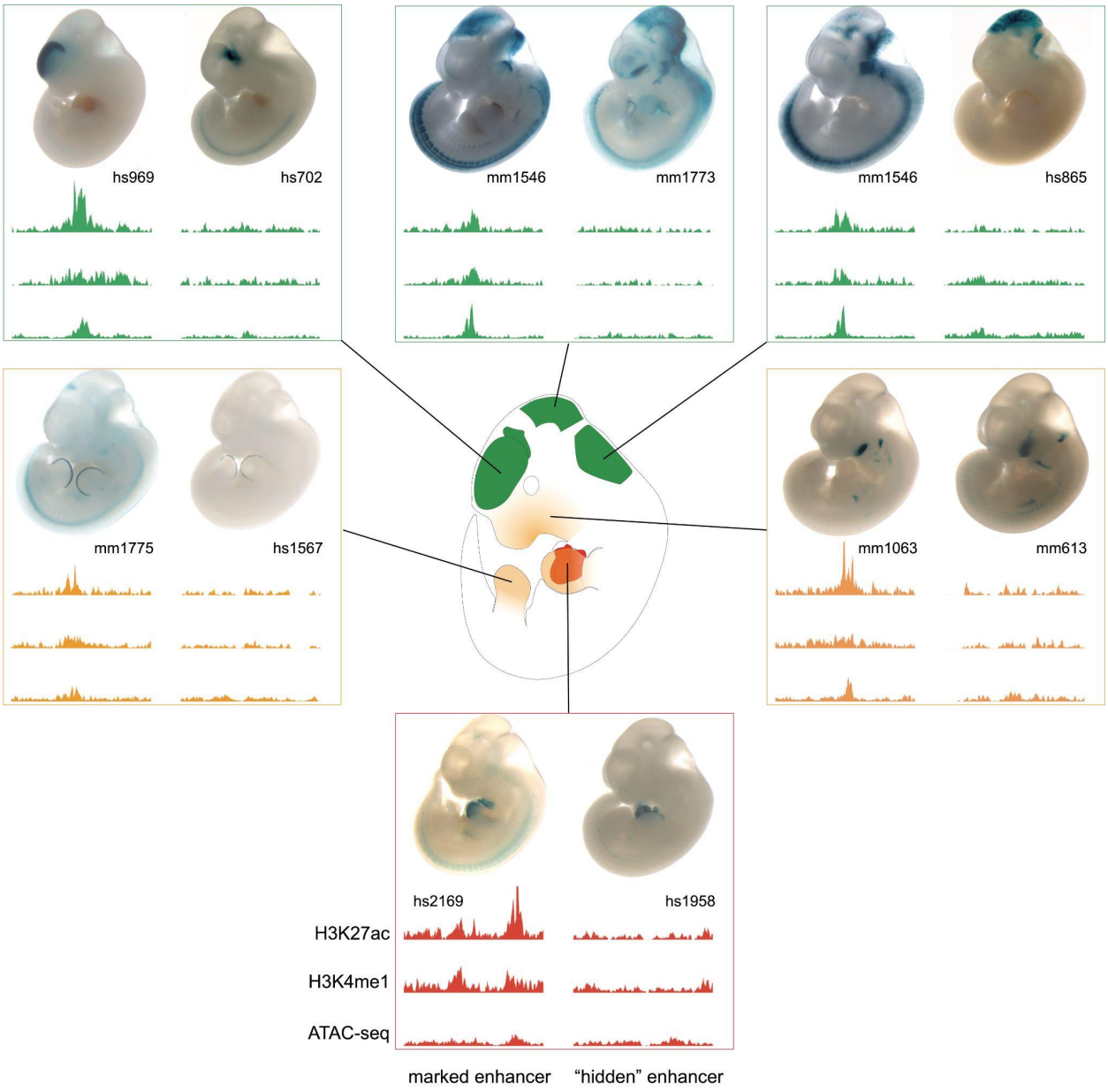


Figure S2.1. Mouse *in vivo* enhancers with and without canonical enhancer-associated chromatin marks. Representative transgenic result (mouse E11.5 embryos) displayed above tissue-specific chromatin profile for each tested element (VISTA ID provided). For each of the 6 considered tissues, an active enhancer with canonical enhancer-associated chromatin marks (left) is displayed alongside an active enhancer without canonical enhancer-associated chromatin marks (right). Mouse tissue- and stage-matched H3K27ac ChIP-seq, H3K4me1 ChIP-seq, ATAC-seq are from ENCODE[19].

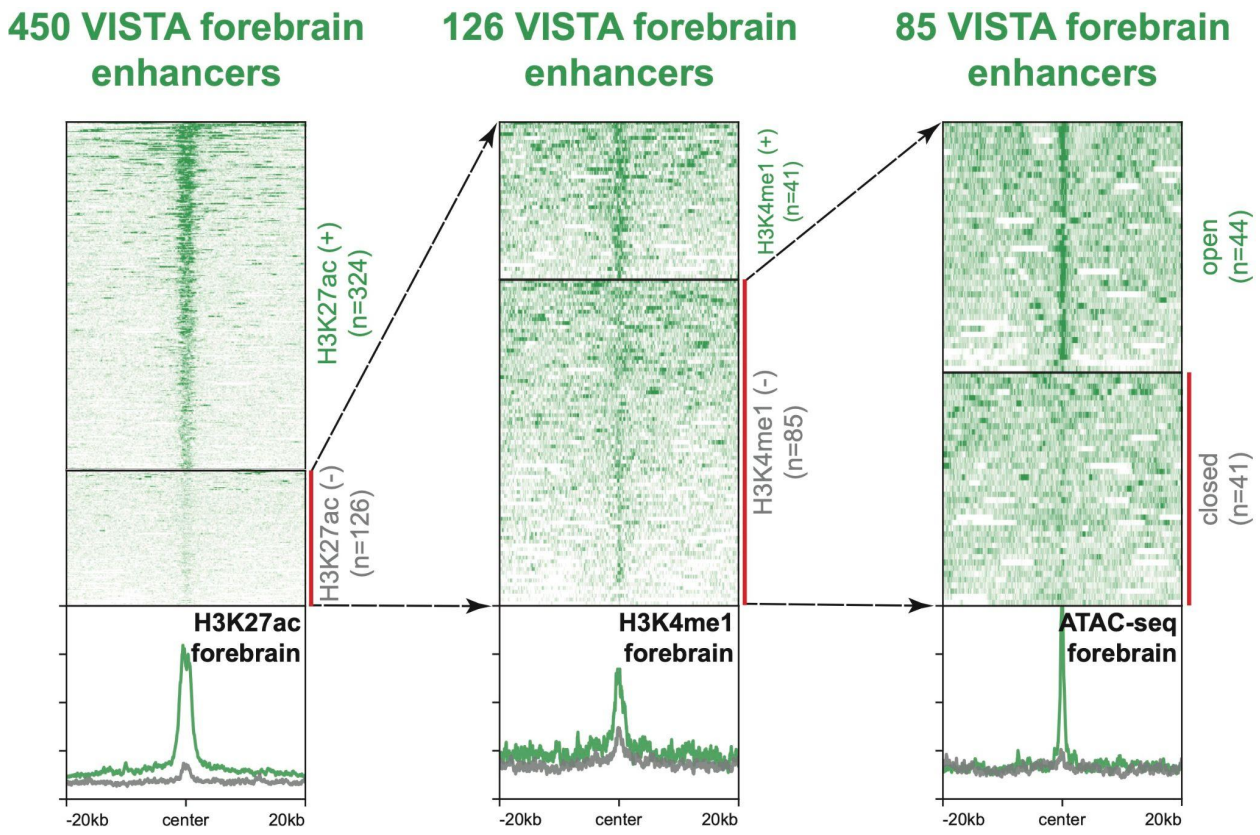


Figure S2.2. Chromatin profiles of active forebrain enhancers with and without H3K27ac, H3K4me1, and ATAC-seq (open chromatin). Forebrain enhancers from the VISTA Enhancer browser stratified across three canonical enhancer-associated chromatin marks. Processed mouse chromatin data are from ENCODE[19].

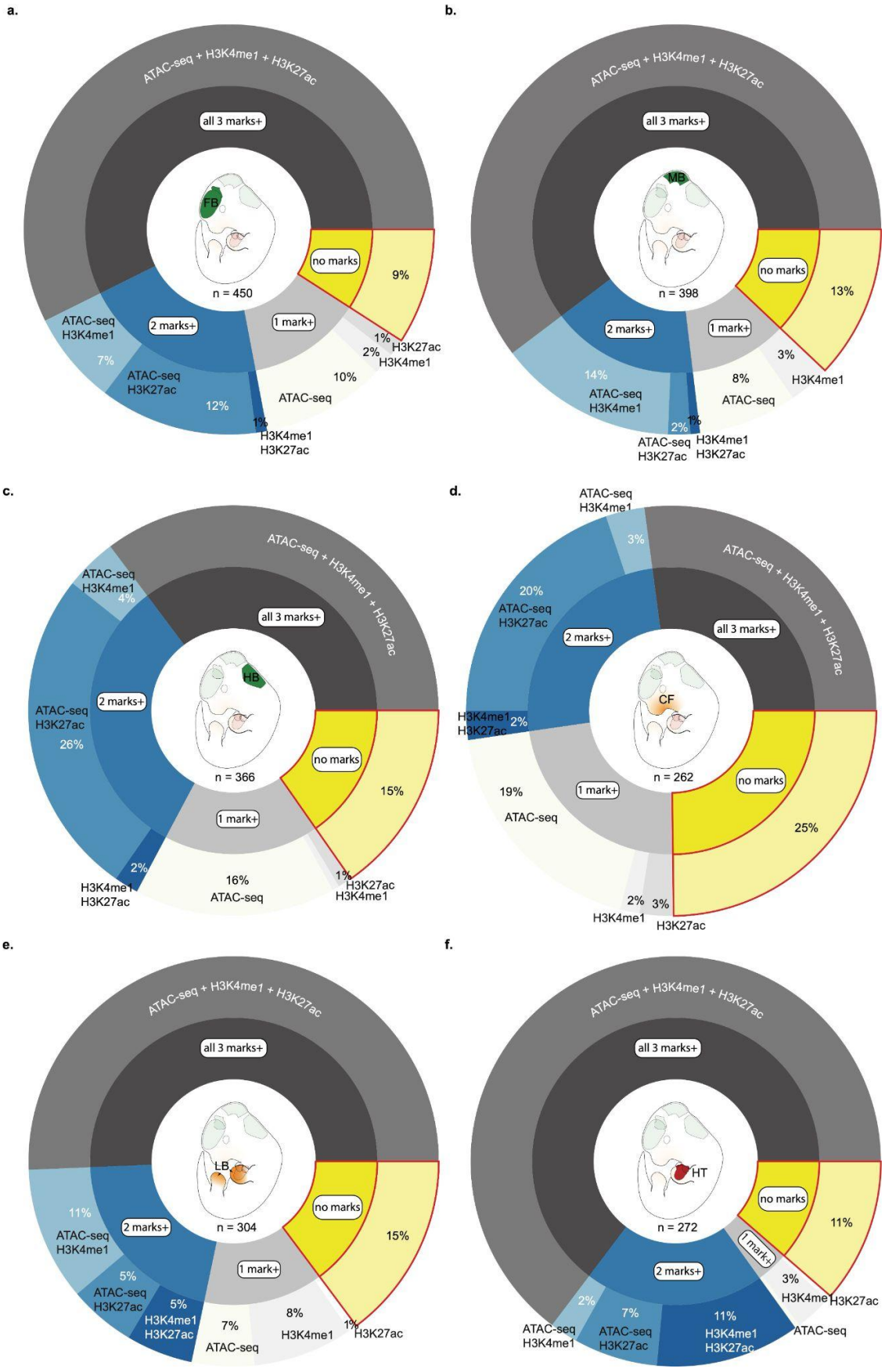

a.
ATAC-seq + H3K4me1 + H3K27ac
all 3 marks+
FB
n = 450
no marks
9%
2 marks+
1 mark+
ATAC-seq H3K4me1
7%
ATAC-seq H3K27ac
12%
1%
H3K4me1 H3K27ac
10%
ATAC-seq
1%
2%
H3K27ac H3K4me1
b.
ATAC-seq + H3K4me1 + H3K27ac
all 3 marks+
MB
n = 398
no marks
13%
2 marks+
1 mark+
14%
ATAC-seq H3K4me1
2%
ATAC-seq H3K27ac
1%
H3K4me1 H3K27ac
8%
ATAC-seq
3%
H3K4me1
c.
ATAC-seq + H3K4me1 + H3K27ac
all 3 marks+
HB
n = 366
no marks
15%
2 marks+
1 mark+
ATAC-seq H3K4me1
4%
ATAC-seq H3K27ac
26%
2%
H3K4me1 H3K27ac
16%
ATAC-seq
1%
H3K27ac H3K4me1
d.
ATAC-seq + H3K4me1 + H3K27ac
all 3 marks+
CF
n = 262
no marks
25%
2 marks+
1 mark+
ATAC-seq H3K4me1
3%
20%
ATAC-seq H3K27ac
H3K4me1 H3K27ac
2%
19%
ATAC-seq
2%
H3K4me1
3%
H3K27ac
e.
ATAC-seq + H3K4me1 + H3K27ac
all 3 marks+
LB
n = 304
no marks
15%
2 marks+
1 mark+
11%
ATAC-seq H3K4me1
5%
ATAC-seq H3K27ac
5%
H3K4me1 H3K27ac
7%
ATAC-seq
8%
H3K4me1
1%
H3K27ac
f.
ATAC-seq + H3K4me1 + H3K27ac
all 3 marks+
HT
n = 272
no marks
11%
1 mark+
2 marks+
3%
H3K4me
H3K27ac
2%
ATAC-seq H3K4me1
7%
ATAC-seq H3K27ac
11%
H3K4me1 H3K27ac
ATAC-seq

Figure S2.3. Proportions of VISTA enhancers with enhancer-associated chromatin signatures by tissue. Active enhancers across the six considered tissues with different combinations of canonical enhancer-associated chromatin marks. For every case there are active enhancers that do not have any of these considered marks. The tissues/regions are: **(a)** forebrain, **(b)** midbrain, **(c)** hindbrain, **(d)** craniofacial, **(e)** limb, and **(f)** heart.

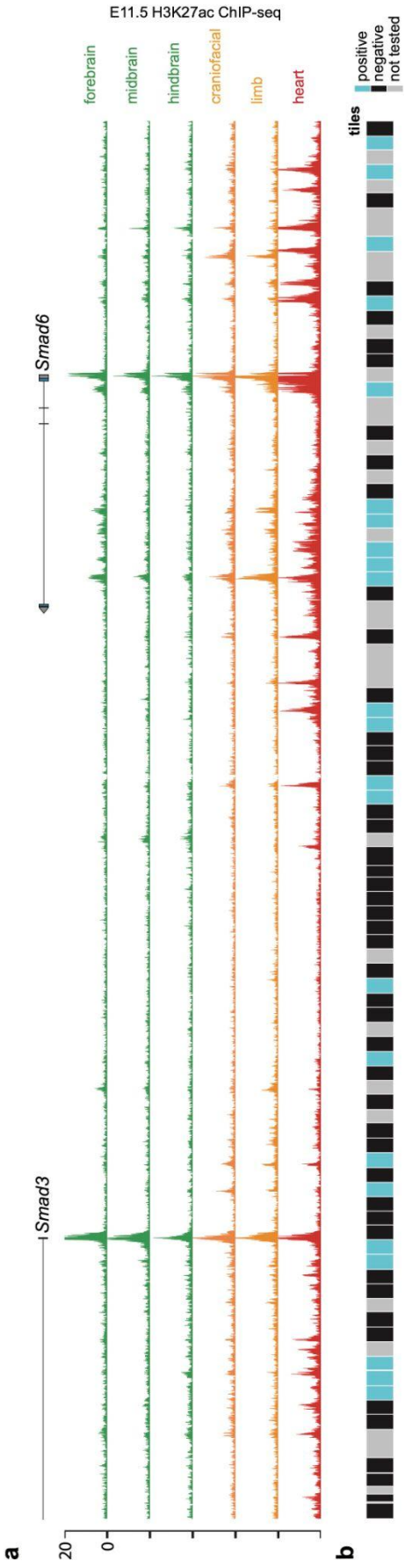
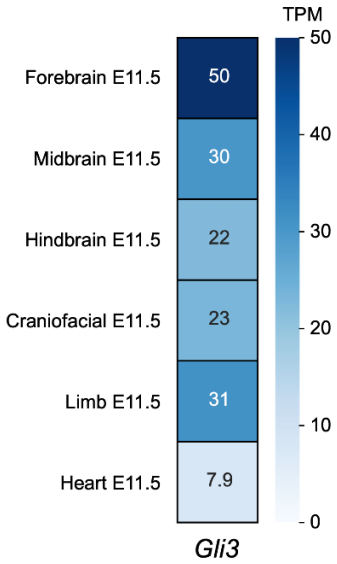
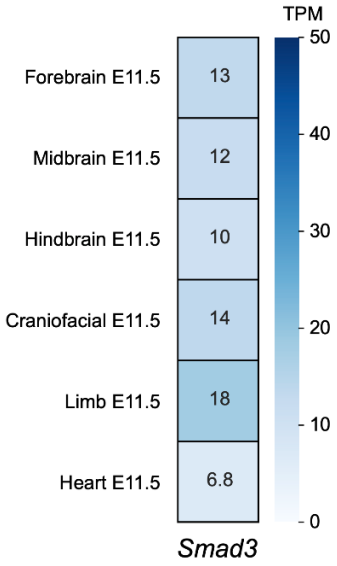

E11.5 H3K27ac ChIP-seq
forebrain
midbrain
hindbrain
craniofacial
limb
heart
Smad3
Smad6
20
0
a
b
tiles
positive
negative
not tested

Figure S2.4. Tiling a second locus for the unbiased identification of mouse in vivo enhancers. **(a)** *Smad3*/*Smad6* locus with mouse E11.5 H3K27ac ChIP-seq data (ENCODE) for six tissues. **(b)** Elements (~5kb in size and overlapping with adjacent elements) designed for the unbiased tiling assay. Elements that were tested and that had reproducible enhancer-reporter activity (in one or more tissues) in the mouse *in vivo* transgenic assay are shaded blue. Elements not tested are shaded gray.

a.


TPM
Forebrain E11.5
Midbrain E11.5
Hindbrain E11.5
Craniofacial E11.5
Limb E11.5
Heart E11.5
50
30
22
23
31
7.9
50
40
30
20
10
0
Gli3

b.


TPM
Forebrain E11.5
Midbrain E11.5
Hindbrain E11.5
Craniofacial E11.5
Limb E11.5
Heart E11.5
13
12
10
14
18
6.8
50
40
30
20
10
0
Smad3

c.

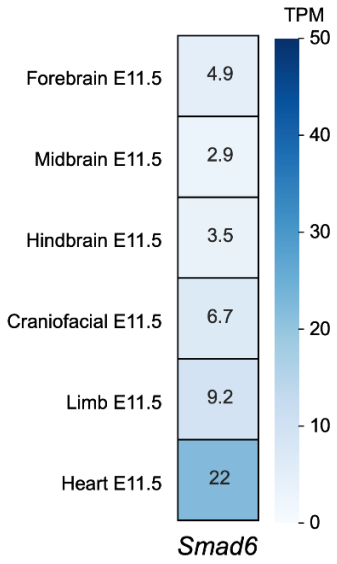

TPM
Forebrain E11.5
Midbrain E11.5
Hindbrain E11.5
Craniofacial E11.5
Limb E11.5
Heart E11.5
4.9
2.9
3.5
6.7
9.2
22
50
40
30
20
10
0
Smad6

Figure S2.5. Mouse E11.5 gene expression by tissue for the *Gli3*, *Smad3*, and *Smad6* genes. **(a)** Per tissue RNA-seq and mouse *in situ* data for **(a)** *Gli3*, **(b)** *Smad3,* and **(c)** *Smad6* genes. RNA-seq data are from mouse ENCODE[49]. TPM, transcripts per million.

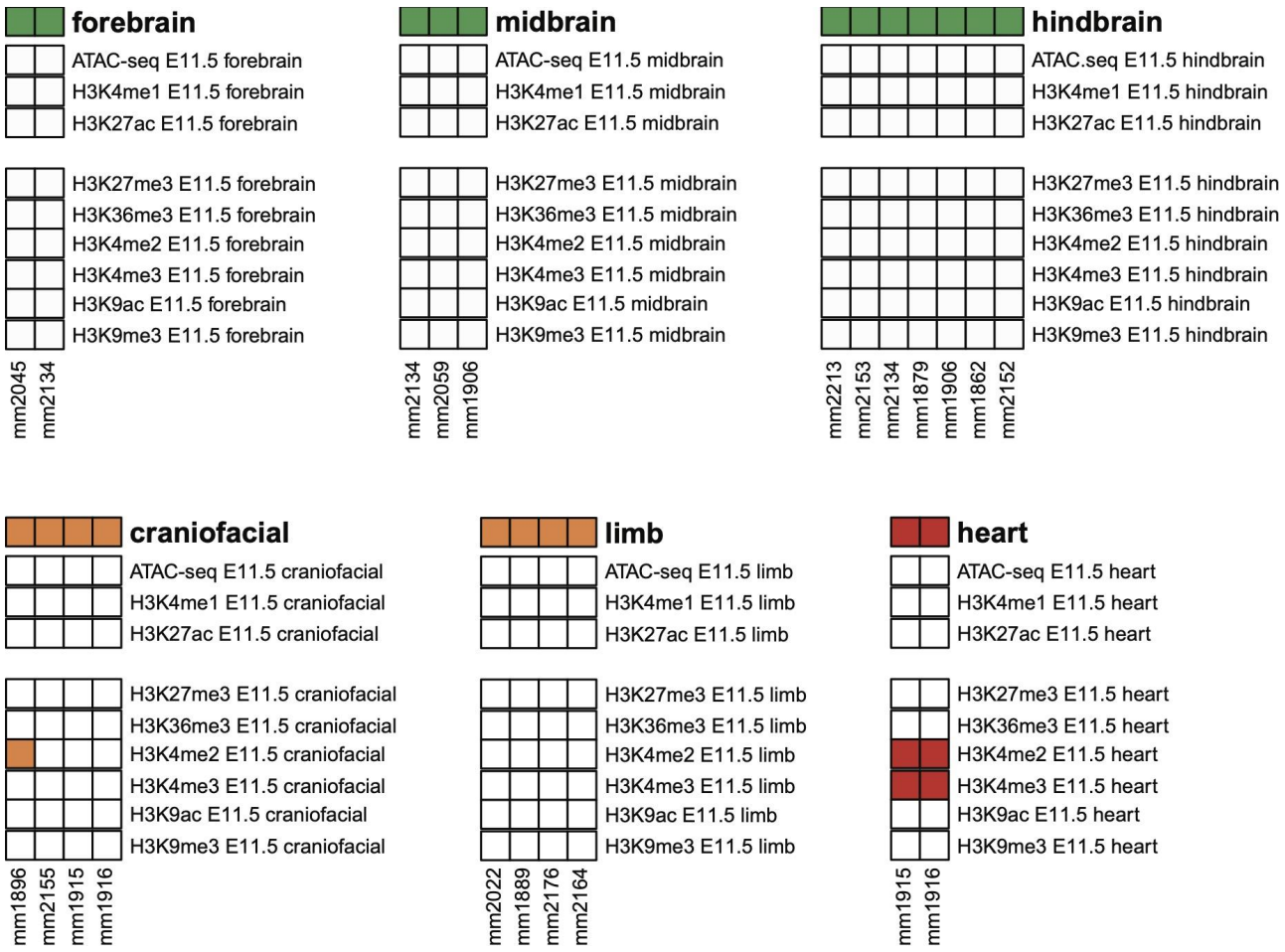


Figure S2.6. Hidden enhancers commonly lack other chromatin marks at their endogenous site. A majority of hidden enhancers identified from the unbiased tiling (across the *Gli3* and *Smad3/Smad6* loci) do not have other chromatin marks. Processed mouse chromatin data are from ENCODE[19].

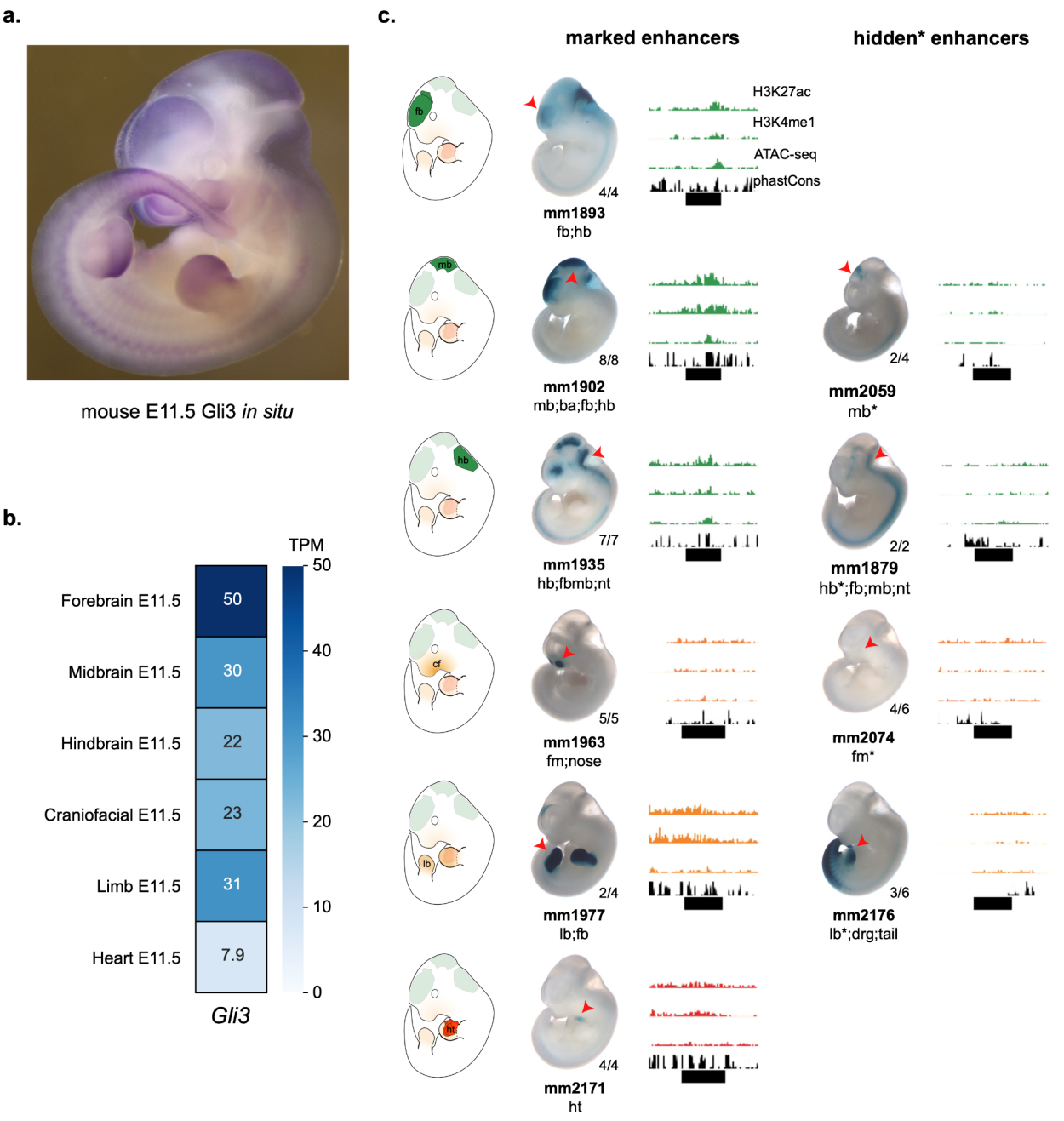


Figure S2.7. Hidden enhancers within the *Gli3* locus show similar tissue-specific reporter activities as their marked counterparts and correlate with Gli3 *in situ* expression data. (a) Whole-mount *in situ* for *Gli3* in E11.5 mouse. **(b)** ENCODE RNA-seq data from E11.5 mouse across six developmental tissues. **(c)** Examples of marked and hidden enhancers identified across the *Gli3* locus. Red ticks mark regions with reproducible tissue-specific enhancer reporter activity, summarized by one example transgenic embryo. Indicated with each representative transgenic result is the number of independent embryos with LacZ staining in the considered tissue over the total number of transgenic embryos obtained. Black bars underneath the chromatin and evolutionary conservation tracks represent the candidate element that was tested in the mouse *in vivo* transgenic reporter assay.

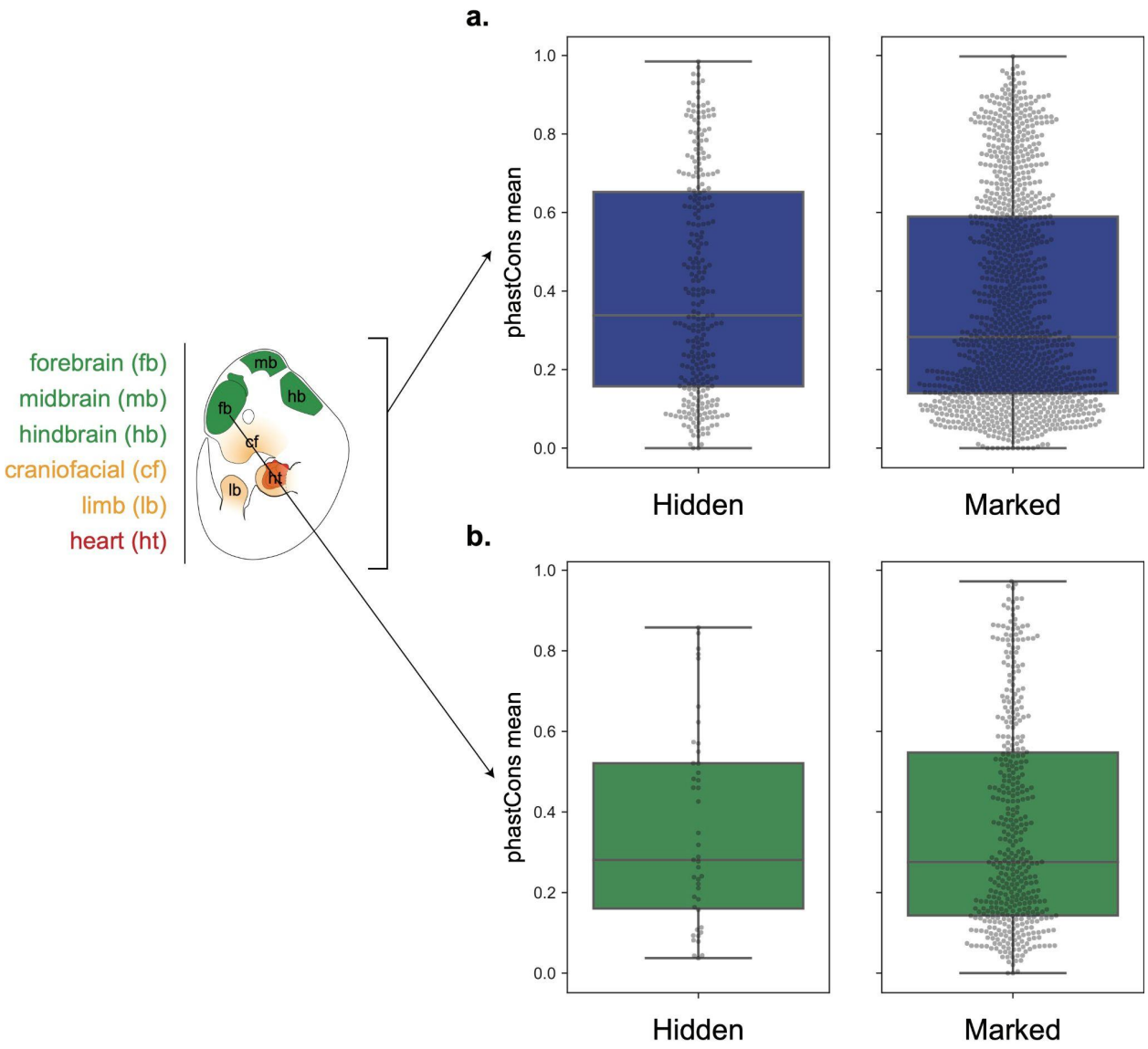


Figure S2.8. Similar levels of evolutionary conservation (phastCons) between hidden enhancers and marked enhancers. Hidden enhancers and marked enhancers have similar levels of evolutionary conservation (phastCons) for **(a)** all tissues considered together and also for each considered tissue, exemplified by **(b)** forebrain enhancers. Data not shown for the other five tissue types. No statistically significant difference via Kolmogorov-Smirnov comparison.

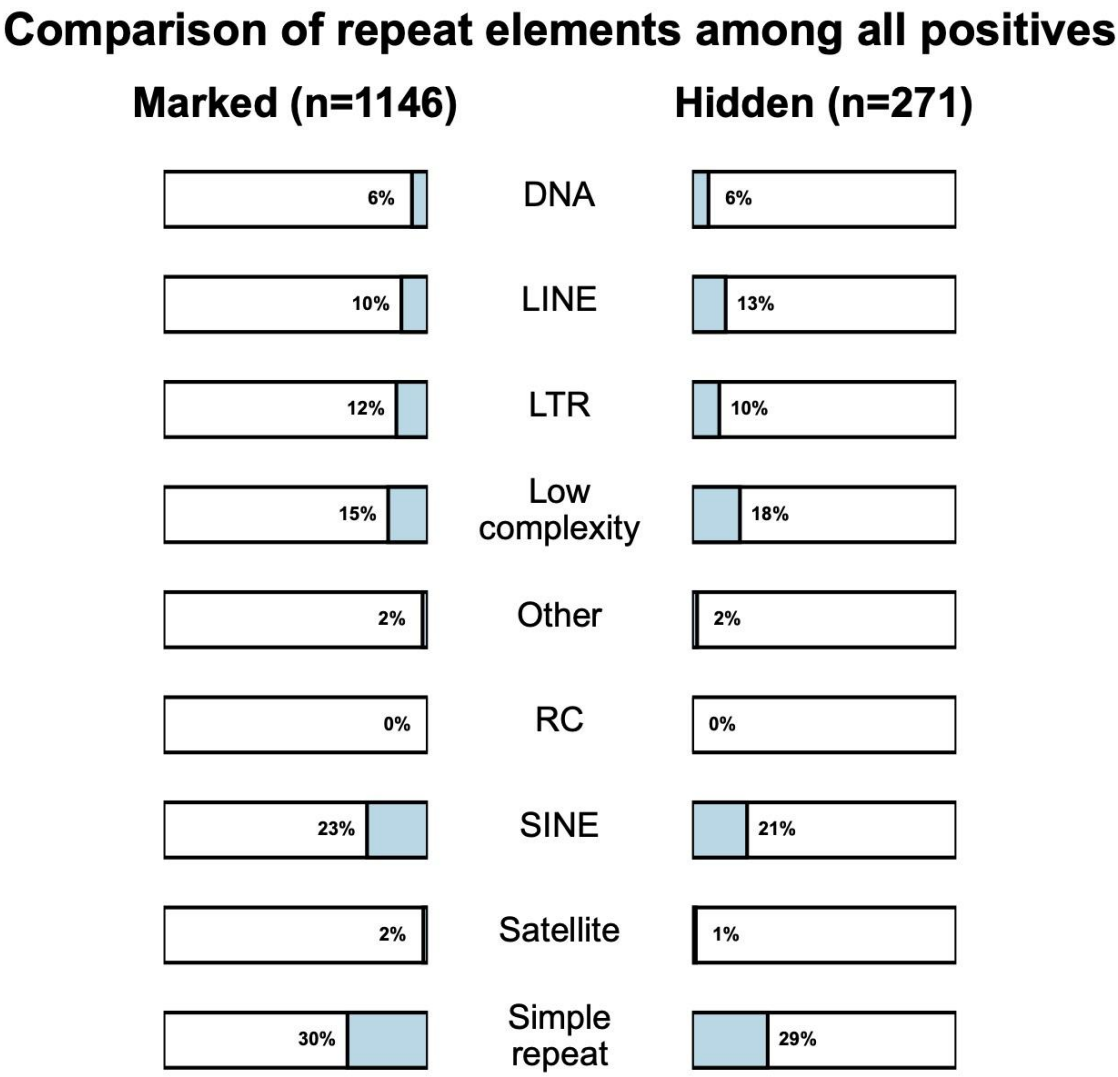


Figure S2.9. Similar proportions of transposable element families between marked and hidden enhancers. The proportions of transposable element classes among positive elements are comparable between marked and hidden enhancers (across all six tissues). All elements were evaluated for their repeat content from RepeatMasker mouse mm10 annotations.

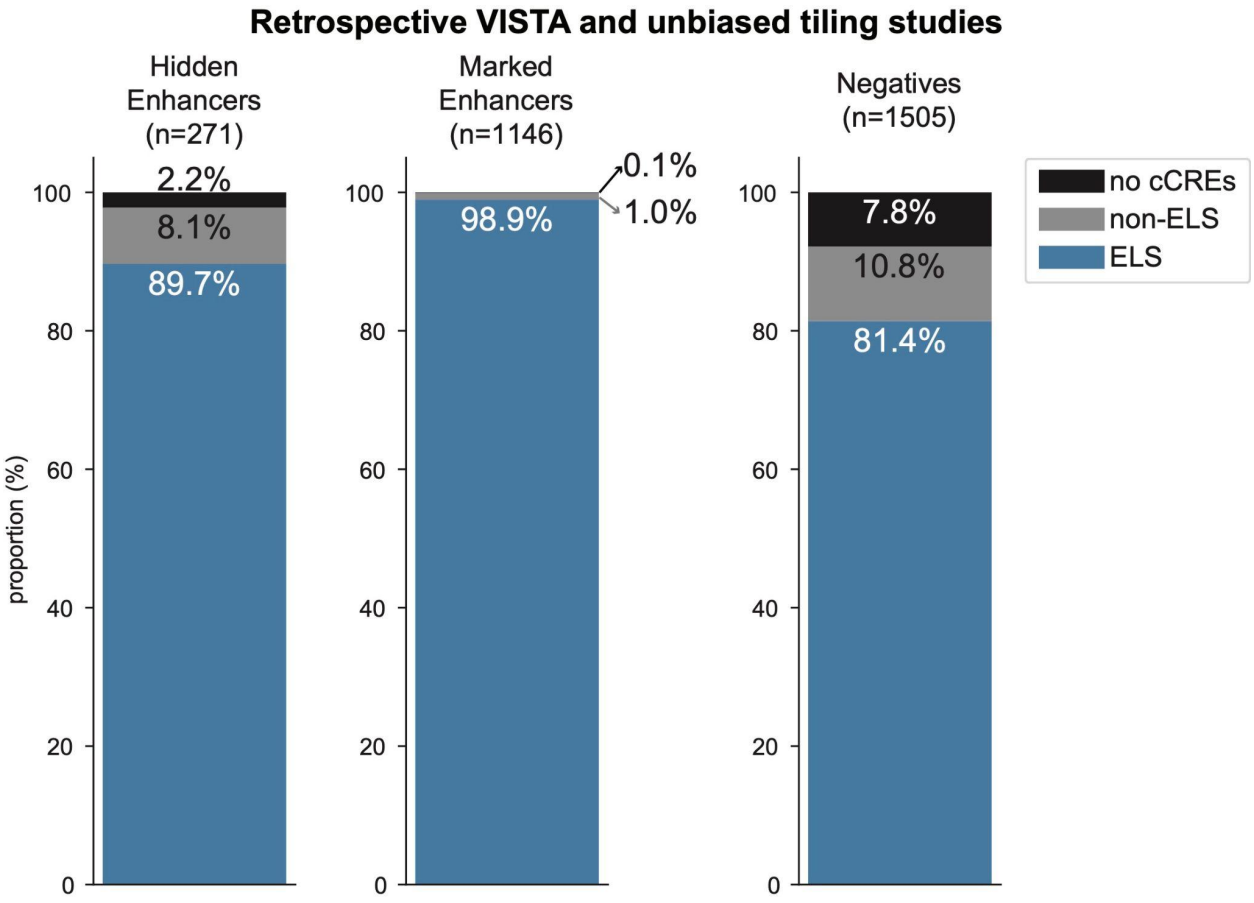


Figure S2.10. Majority of hidden enhancers identified from the retrospective VISTA and unbiased tiling studies contain candidate cis-regulatory elements (cCREs) that are derived from multiple tissue types and developmental stages. cCREs with enhancer-like signatures (ELS[18]) are present in the hidden enhancers identified across the retrospective VISTA and tiling studies. ELS cCREs are also present in a majority of active enhancers with canonical enhancer-associated chromatin marks (marked enhancers). A majority of negative elements (did not show reproducible tissue-specific enhancer-reporter activity at E11.5) also overlap with ELS cCREs.

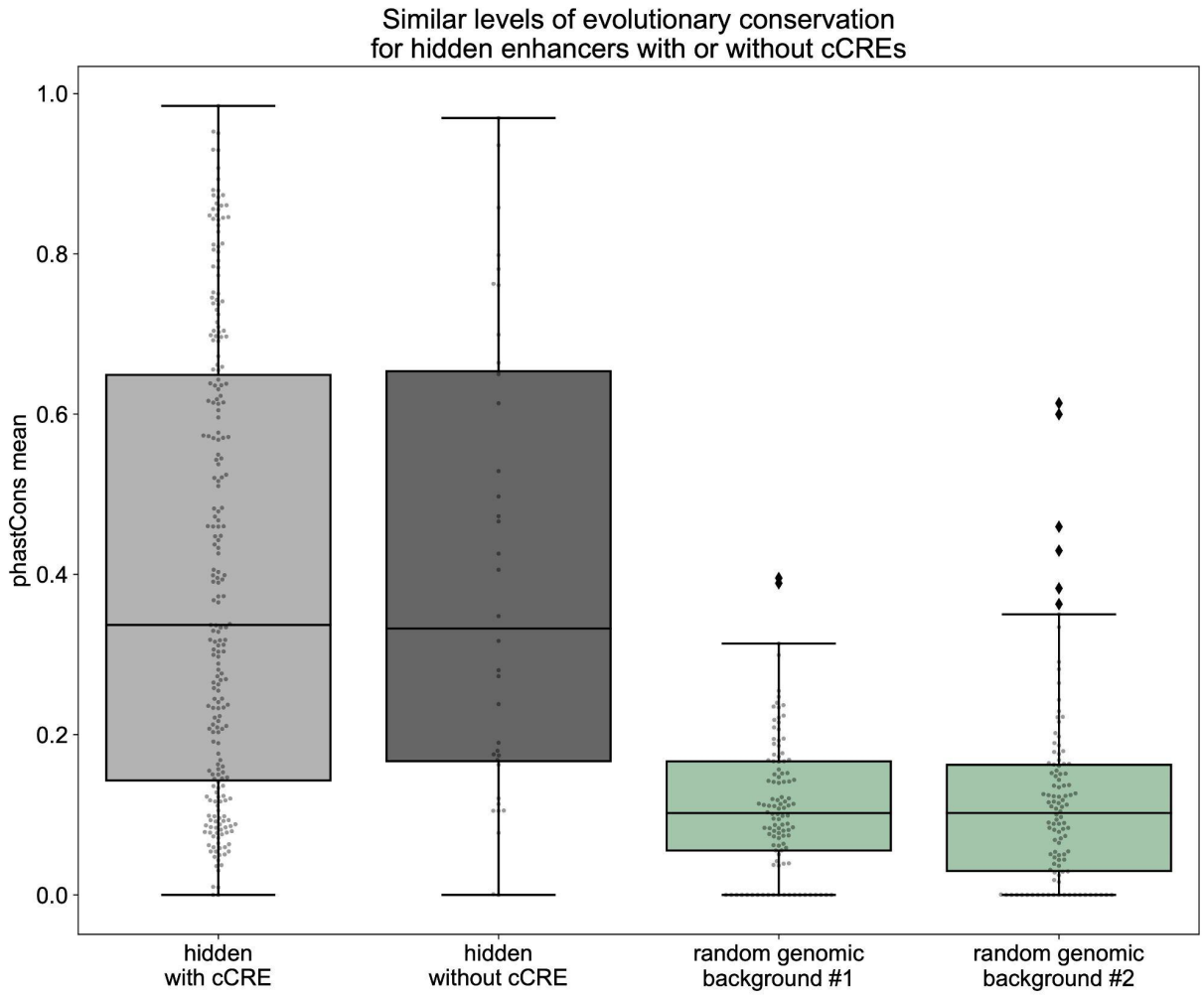


Figure S2.11. Hidden enhancers that do not overlap with candidate cis-regulatory elements (cCREs) have similar levels of evolutionary conservation (phastCons) as those that do. Among hidden enhancers identified across the retrospective VISTA and tiling studies, those that do not overlap with cCREs[18] have similar levels of elevated evolutionary conservation (compared to random genomic background) as those that do.

Table S2.1. ENCODE mouse chromatin and RNA-seq data.

Assay	Timepoint	Tissue	accession ID
ATAC-seq	E11.5	forebrain	ENCFF767MGH
ATAC-seq	E11.5	midbrain	ENCFF145KYY
ATAC-seq	E11.5	hindbrain	ENCFF740NUE
ATAC-seq	E11.5	e.f.p.	ENCFF994YJC
ATAC-seq	E11.5	limb	ENCFF549YKV
ATAC-seq	E11.5	heart	ENCFF071VIV
ChIP-seq_H3K4me1-mouse	E10.5	forebrain	ENCFF635WQB
ChIP-seq_H3K4me1-mouse	E10.5	midbrain	ENCFF060BRH
ChIP-seq_H3K4me1-mouse	E10.5	hindbrain	ENCFF105VCR
ChIP-seq_H3K4me1-mouse	E10.5	e.f.p.	ENCFF107SDV
ChIP-seq_H3K4me1-mouse	E10.5	limb	ENCFF761PEJ
ChIP-seq_H3K4me1-mouse	E10.5	heart	ENCFF677RSH
ChIP-seq_H3K4me1-mouse	E11.5	forebrain	ENCFF147OKD
ChIP-seq_H3K4me1-mouse	E11.5	midbrain	ENCFF202HIO
ChIP-seq_H3K4me1-mouse	E11.5	hindbrain	ENCFF098IGX
ChIP-seq_H3K4me1-mouse	E11.5	e.f.p.	ENCFF880FVO
ChIP-seq_H3K4me1-mouse	E11.5	limb	ENCFF255WOG
ChIP-seq_H3K4me1-mouse	E11.5	heart	ENCFF218FKJ
ChIP-seq_H3K4me2-mouse	E11.5	forebrain	ENCFF047OVD
ChIP-seq_H3K4me2-mouse	E11.5	midbrain	ENCFF132QFU
ChIP-seq_H3K4me2-mouse	E11.5	hindbrain	ENCFF835VQG
ChIP-seq_H3K4me2-mouse	E11.5	e.f.p.	ENCFF272ZCQ
ChIP-seq_H3K4me2-mouse	E11.5	limb	ENCFF703AUU
ChIP-seq_H3K4me2-mouse	E11.5	heart	ENCFF316YSJ
ChIP-seq_H3K4me3-mouse	E10.5	forebrain	ENCFF007CCC
ChIP-seq_H3K4me3-mouse	E10.5	midbrain	ENCFF390OXU
ChIP-seq_H3K4me3-mouse	E10.5	hindbrain	ENCFF086LDB
ChIP-seq_H3K4me3-mouse	E10.5	e.f.p.	ENCFF535KDR
ChIP-seq_H3K4me3-mouse	E10.5	limb	ENCFF708VXA
ChIP-seq_H3K4me3-mouse	E10.5	heart	ENCFF645KWB
ChIP-seq_H3K4me3-mouse	E11.5	forebrain	ENCFF635RVF
ChIP-seq_H3K4me3-mouse	E11.5	midbrain	ENCFF098WIS
ChIP-seq_H3K4me3-mouse	E11.5	hindbrain	ENCFF292XPT
ChIP-seq_H3K4me3-mouse	E11.5	e.f.p.	ENCFF902KTA
ChIP-seq_H3K4me3-mouse	E11.5	limb	ENCFF388OWQ
ChIP-seq_H3K4me3-mouse	E11.5	heart	ENCFF908XCE
ChIP-seq_H3K9ac-mouse	E11.5	forebrain	ENCFF246ERL

ChIP-seq_H3K9ac-mouse	E11.5	midbrain	ENCFF661ZXD
ChIP-seq_H3K9ac-mouse	E11.5	hindbrain	ENCFF741ESD
ChIP-seq_H3K9ac-mouse	E11.5	e.f.p.	ENCFF039EDR
ChIP-seq_H3K9ac-mouse	E11.5	limb	ENCFF146WOQ
ChIP-seq_H3K9ac-mouse	E11.5	heart	ENCFF028MGI
ChIP-seq_H3K9me3-mouse	E10.5	forebrain	ENCFF061KIC
ChIP-seq_H3K9me3-mouse	E10.5	midbrain	ENCFF844JAM
ChIP-seq_H3K9me3-mouse	E10.5	hindbrain	ENCFF338NFN
ChIP-seq_H3K9me3-mouse	E10.5	e.f.p.	ENCFF998SXT
ChIP-seq_H3K9me3-mouse	E10.5	limb	ENCFF829MJU
ChIP-seq_H3K9me3-mouse	E10.5	heart	ENCFF173HXP
ChIP-seq_H3K9me3-mouse	E11.5	forebrain	ENCFF110BTA
ChIP-seq_H3K9me3-mouse	E11.5	midbrain	ENCFF266QHJ
ChIP-seq_H3K9me3-mouse	E11.5	hindbrain	ENCFF431DLU
ChIP-seq_H3K9me3-mouse	E11.5	e.f.p.	ENCFF892BOY
ChIP-seq_H3K9me3-mouse	E11.5	limb	ENCFF649BGJ
ChIP-seq_H3K9me3-mouse	E11.5	heart	ENCFF565ETE
ChIP-seq_H3K27ac-mouse	E10.5	forebrain	ENCFF473BCV
ChIP-seq_H3K27ac-mouse	E10.5	midbrain	ENCFF656RYT
ChIP-seq_H3K27ac-mouse	E10.5	hindbrain	ENCFF567PMM
ChIP-seq_H3K27ac-mouse	E10.5	e.f.p.	ENCFF419UIR
ChIP-seq_H3K27ac-mouse	E10.5	limb	ENCFF205SAP
ChIP-seq_H3K27ac-mouse	E10.5	heart	ENCFF855NXH
ChIP-seq_H3K27ac-mouse	E11.5	forebrain	ENCFF759KHX
ChIP-seq_H3K27ac-mouse	E11.5	midbrain	ENCFF650WFB
ChIP-seq_H3K27ac-mouse	E11.5	hindbrain	ENCFF083MLY
ChIP-seq_H3K27ac-mouse	E11.5	e.f.p.	ENCFF680UPD
ChIP-seq_H3K27ac-mouse	E11.5	limb	ENCFF016BEF
ChIP-seq_H3K27ac-mouse	E11.5	heart	ENCFF236UMU
ChIP-seq_H3K27me3-mouse	E10.5	forebrain	ENCFF032HUD
ChIP-seq_H3K27me3-mouse	E10.5	midbrain	ENCFF504GUV
ChIP-seq_H3K27me3-mouse	E10.5	hindbrain	ENCFF443TNH
ChIP-seq_H3K27me3-mouse	E10.5	e.f.p.	ENCFF179QHY
ChIP-seq_H3K27me3-mouse	E10.5	limb	ENCFF736SSJ
ChIP-seq_H3K27me3-mouse	E10.5	heart	ENCFF110HRW
ChIP-seq_H3K27me3-mouse	E11.5	forebrain	ENCFF013NFN
ChIP-seq_H3K27me3-mouse	E11.5	midbrain	ENCFF927GUC
ChIP-seq_H3K27me3-mouse	E11.5	hindbrain	ENCFF641DRR
ChIP-seq_H3K27me3-mouse	E11.5	e.f.p.	ENCFF533OZG
ChIP-seq_H3K27me3-mouse	E11.5	limb	ENCFF601DCY

ChIP-seq_H3K27me3-mouse	E11.5	heart	ENCFF043FMD
ChIP-seq_H3K36me3-mouse	E10.5	forebrain	ENCFF247CKJ
ChIP-seq_H3K36me3-mouse	E10.5	midbrain	ENCFF455MXT
ChIP-seq_H3K36me3-mouse	E10.5	hindbrain	ENCFF371EQO
ChIP-seq_H3K36me3-mouse	E10.5	e.f.p.	ENCFF524LYP
ChIP-seq_H3K36me3-mouse	E10.5	limb	ENCFF662LXS
ChIP-seq_H3K36me3-mouse	E10.5	heart	ENCFF421CMA
ChIP-seq_H3K36me3-mouse	E11.5	forebrain	ENCFF636IYQ
ChIP-seq_H3K36me3-mouse	E11.5	midbrain	ENCFF074QQA
ChIP-seq_H3K36me3-mouse	E11.5	hindbrain	ENCFF519TDQ
ChIP-seq_H3K36me3-mouse	E11.5	e.f.p.	ENCFF208EZF
ChIP-seq_H3K36me3-mouse	E11.5	limb	ENCFF214JJB
ChIP-seq_H3K36me3-mouse	E11.5	heart	ENCFF271RAX
polyA plus RNA-seq	E11.5	forebrain	ENCFF042VCB
polyA plus RNA-seq	E11.5	midbrain	ENCFF184FWR
polyA plus RNA-seq	E11.5	hindbrain	ENCFF606UHO
polyA plus RNA-seq	E11.5	e.f.p.	ENCFF343KWN
polyA plus RNA-seq	E11.5	limb	ENCFF836WUM
polyA plus RNA-seq	E11.5	heart	ENCFF159DWP

“mixed sex embryo (11.5 days) strain B6NCrl C57BL/6” abbreviated as E11.5
“mixed sex embryo (10.5 days) strain B6NCrl C57BL/6” abbreviated as E10.5
“embryonic facial prominence” abbreviated as e.f.p.

Table S2.2. Enhancers from the VISTA retrospective study and the unbiased tiling used for chromatin intersections.

Note: this table comprises over 1,500 rows and over 80 columns. The table provided here is an excerpt, and a full version is available in the bioRxiv submission.

Study	Chromosome	Start	End	VISTA ID	Transgenic result	Tissue
VISTA	chr1	6729233	6730318	hs698	positive	drg;fb;hb;mb;nt;tri
VISTA	chr1	9648223	9650965	mm1546	positive	cn;drg;hb;mb;nt;tri
VISTA	chr1	12509176	12511893	mm1416	positive	fb;mb;nose
VISTA	chr1	19106737	19107712	hs865	positive	hb;mb
VISTA	chr1	19699462	19700530	hs217	positive	hb;mb
VISTA	chr1	20919958	20922631	hs2064	positive	hb;mb
VISTA	chr1	39441763	39444730	hs1933	positive	ht
VISTA	chr1	39945610	39950472	mm1333	positive	cn;drg;fb
VISTA	chr1	40942806	40944270	hs1212	positive	ht
VISTA	chr1	41165839	41167312	hs1093	positive	mb
VISTA	chr1	41274880	41276860	hs1112	positive	ba;fb;hb;lb;mb;other
VISTA	chr1	41388600	41391399	hs1555	positive	hb;lv;mb;nt
VISTA	chr1	41603695	41607210	hs1526	positive	fb
VISTA	chr1	41812941	41815356	hs1529	positive	fb
VISTA	chr1	41917498	41920605	hs1303	positive	fb
VISTA	chr1	41935262	41937041	hs1554	positive	ba
VISTA	chr1	41981433	41982259	hs401	positive	hb
VISTA	chr1	42242415	42244647	hs1131	positive	fb;mb
VISTA	chr1	42255267	42258543	hs1534	positive	fb;hb;mb
VISTA	chr1	42365768	42366767	hs702	positive	fb
		...				
tiling	chr13	14626494	14631743	mm2080	positive	lb;other
tiling	chr13	14635984	14641308	mm2079	negative	NA
tiling	chr13	14640829	14646066	mm2078	negative	NA
tiling	chr13	14645607	14651010	mm2178	negative	NA
tiling	chr13	14650658	14655894	mm2177	negative	NA
tiling	chr13	14655343	14660594	mm2077	negative	NA
tiling	chr13	14660219	14665547	mm2076	negative	NA
tiling	chr13	14669878	14675087	mm2074	positive	fm
tiling	chr13	14674606	14679912	mm2073	negative	NA
tiling	chr13	14679658	14684970	mm2072	negative	NA
tiling	chr13	14684752	14690058	mm2071	negative	NA
tiling	chr13	14689596	14694744	mm2070	negative	NA
tiling	chr13	14699365	14704573	mm2069	negative	NA
tiling	chr13	14704284	14709747	mm2068	negative	NA
tiling	chr13	14709364	14714468	mm2067	negative	NA
tiling	chr13	14714267	14719523	mm2066	negative	NA
tiling	chr13	14719097	14724249	mm2065	negative	NA
tiling	chr13	14723863	14729105	mm2064	negative	NA
tiling	chr13	14728859	14734326	mm2063	negative	NA
tiling	chr13	14733886	14739186	mm2062	negative	NA

Provided coordinates correspond to mouse genome, reference assembly GRCm38/mm10.
"Pre-existing VISTA data" abbreviated as VISTA.
ba: branchial arch; bv: blood vessels; cn: cranial nerve; drg: dorsal root ganglion;; fm: facial mesenchyme; fb: forebrain; gen: genital tubercle; ht: heart; hb: hindbrain; lb: limb; lv: liver; mb: midbrain; nt: neural tube; som: somite; tri: trigeminal V

Table S2.3. Overview of VISTA E11.5 enhancers by tissue.

Tissue	Total elements
forebrain*	450
midbrain*	398
hindbrain*	366
neural tube	256
dorsal root ganglion	84
somite	58
cranial nerve	62
trigeminal V	56
facial mesenchyme*	95
branchial arch*	165
nose*	85
ear	29
eye	90
limb*	304
heart*	272
liver	8
blood vessels	24
tail	32
genital tubercle	12

Enhancers active in branchial arch, facial mesenchyme, and/or nose are grouped as craniofacial enhancers. Rows with * asterisk: enhancers used in this study.

Table S2.4. Tissue-specific H3K27ac peak counts across the two loci tested by tiling for enhancer activity.

Provided coordinates correspond to mouse genome, reference assembly GRCm38/mm10.

Chromosome	Start	End	Peak ID
chr9	63726550	63726767	smad3-6_forebrain_1
chr9	63750284	63750584	smad3-6_forebrain_2
chr9	63803081	63803952	smad3-6_forebrain_3
chr9	63958607	63961013	smad3-6_forebrain_4
chr9	63965116	63966263	smad3-6_forebrain_5
chr9	63973357	63976505	smad3-6_forebrain_6
chr9	63979405	63981266	smad3-6_forebrain_7
chr9	63998102	63998341	smad3-6_forebrain_8
chr9	64045612	64045866	smad3-6_forebrain_9
chr9	64057937	64058780	smad3-6_forebrain_10
chr9	64066676	64067379	smad3-6_forebrain_11
chr9	63750797	63751077	smad3-6_midbrain_1
chr9	63878861	63880288	smad3-6_midbrain_2
chr9	63922622	63922868	smad3-6_midbrain_3
chr9	63959436	63960939	smad3-6_midbrain_4
chr9	64057956	64058438	smad3-6_midbrain_5
chr9	64066639	64067229	smad3-6_midbrain_6
chr9	63715978	63716450	smad3-6_hindbrain_1
chr9	63879904	63880391	smad3-6_hindbrain_2
chr9	63959442	63961038	smad3-6_hindbrain_3
chr9	64045211	64045719	smad3-6_hindbrain_4
chr9	64049721	64050212	smad3-6_hindbrain_5
chr9	64058574	64058797	smad3-6_hindbrain_6
chr9	64066382	64067553	smad3-6_hindbrain_7
chr9	63697926	63698205	smad3-6_craniofacial_1
chr9	63716584	63717165	smad3-6_craniofacial_2
chr9	63726440	63726872	smad3-6_craniofacial_3
chr9	63749000	63749383	smad3-6_craniofacial_4
chr9	63771723	63772602	smad3-6_craniofacial_5
chr9	63779909	63781192	smad3-6_craniofacial_6
chr9	63896092	63896424	smad3-6_craniofacial_7
chr9	63941630	63943144	smad3-6_craniofacial_8
chr9	63958692	63961155	smad3-6_craniofacial_9
chr9	63970884	63974810	smad3-6_craniofacial_10
chr9	63979718	63981328	smad3-6_craniofacial_11

chr9	64045082	64045754	smad3-6_craniofacial_12
chr9	64049674	64050367	smad3-6_craniofacial_13
chr9	64057792	64059839	smad3-6_craniofacial_14
chr9	64066668	64067537	smad3-6_craniofacial_15
chr9	64084683	64086523	smad3-6_craniofacial_16
chr9	63697466	63697688	smad3-6_limb_1
chr9	63701812	63702071	smad3-6_limb_2
chr9	63737619	63738254	smad3-6_limb_3
chr9	63746338	63749661	smad3-6_limb_4
chr9	63772045	63772615	smad3-6_limb_5
chr9	63779875	63780875	smad3-6_limb_6
chr9	63786288	63786657	smad3-6_limb_7
chr9	63802860	63803888	smad3-6_limb_8
chr9	63836780	63837082	smad3-6_limb_9
chr9	63896107	63897026	smad3-6_limb_10
chr9	63927583	63927817	smad3-6_limb_11
chr9	63958079	63961434	smad3-6_limb_12
chr9	63968666	63968869	smad3-6_limb_13
chr9	63973468	63975045	smad3-6_limb_14
chr9	63979750	63981531	smad3-6_limb_15
chr9	63992882	63993250	smad3-6_limb_16
chr9	64045043	64045765	smad3-6_limb_17
chr9	64049057	64050976	smad3-6_limb_18
chr9	64057794	64060556	smad3-6_limb_19
chr9	64066919	64067251	smad3-6_limb_20
chr9	63672936	63673302	smad3-6_heart_1
chr9	63677225	63678748	smad3-6_heart_2
chr9	63685500	63687755	smad3-6_heart_3
chr9	63692186	63692615	smad3-6_heart_4
chr9	63696640	63699128	smad3-6_heart_5
chr9	63704106	63707455	smad3-6_heart_6
chr9	63711419	63727984	smad3-6_heart_7
chr9	63732879	63734277	smad3-6_heart_8
chr9	63738857	63739082	smad3-6_heart_9
chr9	63742046	63751162	smad3-6_heart_10
chr9	63779898	63780829	smad3-6_heart_11
chr9	63785228	63785454	smad3-6_heart_12
chr9	63791597	63791975	smad3-6_heart_13
chr9	63803047	63803598	smad3-6_heart_14
chr9	63877001	63878147	smad3-6_heart_15

chr9	63895686	63897427	smad3-6_heart_16
chr9	63916500	63930108	smad3-6_heart_17
chr9	63936830	63937650	smad3-6_heart_18
chr9	63941230	63943767	smad3-6_heart_19
chr9	64000345	64001042	smad3-6_heart_20
chr9	64029760	64031081	smad3-6_heart_21
chr9	64040117	64061530	smad3-6_heart_22
chr9	64064532	64071298	smad3-6_heart_23
chr9	64075022	64080975	smad3-6_heart_24
chr9	64083528	64089329	smad3-6_heart_25
chr9	64093616	64094963	smad3-6_heart_26
chr13	15394688	15395294	gli3_forebrain_1
chr13	15479312	15479633	gli3_forebrain_2
chr13	15485686	15487797	gli3_forebrain_3
chr13	15517024	15517455	gli3_forebrain_4
chr13	15555749	15557297	gli3_forebrain_5
chr13	15602024	15603254	gli3_forebrain_6
chr13	15617567	15625314	gli3_forebrain_7
chr13	15706611	15709033	gli3_forebrain_8
chr13	15759485	15759904	gli3_forebrain_9
chr13	15484727	15487585	gli3_midbrain_1
chr13	15541210	15545380	gli3_midbrain_2
chr13	15556376	15556912	gli3_midbrain_3
chr13	15602073	15604266	gli3_midbrain_4
chr13	15618344	15618697	gli3_midbrain_5
chr13	15624436	15625033	gli3_midbrain_6
chr13	15707230	15708344	gli3_midbrain_7
chr13	15759433	15759727	gli3_midbrain_8
chr13	15011716	15011955	gli3_hindbrain_1
chr13	15484878	15487084	gli3_hindbrain_2
chr13	15517258	15517632	gli3_hindbrain_3
chr13	15541229	15544343	gli3_hindbrain_4
chr13	15577227	15577683	gli3_hindbrain_5
chr13	15602649	15602892	gli3_hindbrain_6
chr13	15624043	15624829	gli3_hindbrain_7
chr13	15670212	15670747	gli3_hindbrain_8
chr13	15758976	15759947	gli3_hindbrain_9
chr13	15053111	15053383	gli3_craniofacial_1
chr13	15590182	15590460	gli3_craniofacial_2
chr13	15052610	15053924	gli3_limb_1

chr13	15212606	15213147	gli3_limb_2
chr13	15394910	15395396	gli3_limb_3
chr13	15487304	15487569	gli3_limb_4
chr13	15543578	15544169	gli3_limb_5
chr13	15549182	15551291	gli3_limb_6
chr13	15560581	15560861	gli3_limb_7
chr13	15577231	15577682	gli3_limb_8
chr13	15590323	15590809	gli3_limb_9
chr13	15665585	15665940	gli3_heart_1

Table S2.5. Summary of hidden enhancer transcription factor motif analysis.

Tissue	Motif Name	Consensus	Result
forebrain	n/a	n/a	n/a
midbrain	n/a	n/a	n/a
hindbrain	NeuroD1(bHLH)/Islet-NeuroD1-ChIP-Seq(GSE30298)/Homer	GCCATCTGTT	*not significant*
hindbrain	HOXA2(Homeobox)/mES-Hoxa2-ChIP-Seq(Donaldson_et_al.)/Homer	GYCATCMATCAT	*not significant*
craniofacial	n/a	n/a	n/a
limb	Foxh1(Forkhead)/hESC-FOXH1-ChIP-Seq(GSE29422)/Homer	NNTGTGGATTSS	*not significant*
limb	GABPA(ETS)/Jurkat-GABPa-ChIP-Seq(GSE17954)/Homer	RACCGGAAGT	*not significant*
limb	ETS:RUNX(ETS,Runt)/Jurkat-RUNX1-ChIP-Seq(GSE17954)/Homer	RCAGGATGTGGT	*not significant*
heart	n/a	n/a	n/a
all tissue	n/a	n/a	n/a

Table S2.6. Summary of hidden enhancer functional enrichment analysis.

Tissue	Result
forebrain	n/a
midbrain	n/a
hindbrain	n/a
craniofacial	n/a
limb	n/a
heart	n/a
all hidden enhancers	n/a

References

1. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735 (2003).
2. Sur, I. K. *et al.* Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* **338**, 1360–1363 (2012).
3. Gupta, R. M. *et al.* A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. *Cell* **170**, 522–533.e15 (2017).
4. 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).
5. Zeitlinger, J. & Stark, A. Developmental gene regulation in the era of genomics. *Dev. Biol.* **339**, 230–239 (2010).
6. Bolt, C. C. & Duboule, D. The regulatory landscapes of developmental genes. *Development* **147**, (2020).
7. Kvon, E. Z., Waymack, R., Elabd, M. G. & Wunderlich, Z. Enhancer redundancy in development and disease. *Nat. Rev. Genet.* (2021) doi:10.1038/s41576-020-00311-x.
8. Kioussis, D., Vanin, E., deLange, T., Flavell, R. A. & Grosveld, F. G. Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. *Nature* **306**, 662–666 (1983).
9. Philipsen, S., Talbot, D., Fraser, P. & Grosveld, F. The beta-globin dominant control region: hypersensitive site 2. *EMBO J.* **9**, 2159–2167 (1990).
10. Philipsen, S., Pruzina, S. & Grosveld, F. The minimal requirements for activity in transgenic mice of hypersensitive site 3 of the beta globin locus control region. *EMBO J.* **12**, 1077–1085 (1993).
11. Brenner, S. *et al.* Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. *Nature* **366**, 265–268 (1993).
12. Mouse Genome Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
13. Gibbs, R. A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
14. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).
15. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
16. 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).
17. ENCODE Project Consortium *et al.* Perspectives on ENCODE. *Nature* **583**, 693–698 (2020).
18. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
19. Gorkin, D. U. *et al.* An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744–751 (2020).
20. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).

21. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–21936 (2010).
22. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
23. Kvon, E. Z. Using transgenic reporter assays to functionally characterize enhancers in animals. *Genomics* **106**, 185–192 (2015).
24. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics* (2020) doi:10.1038/s41576-019-0209-0.
25. 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).
26. Kvon, E. Z. *et al.* Comprehensive In Vivo Interrogation Reveals Phenotypic Impact of Human Enhancer Variants. *Cell* **180**, 1262–1271.e15 (2020).
27. Biesecker, L. G. What you can learn from one gene: GLI3. *J. Med. Genet.* **43**, 465–469 (2006).
28. Blaess, S., Stephen, D. & Joyner, A. L. Gli3 coordinates three-dimensional patterning and growth of the tectum and cerebellum by integrating Shh and Fgf8 signaling. *Development* **135**, 2093–2103 (2008).
29. Lopez-Rios, J. *et al.* GLI3 constrains digit number by controlling both progenitor proliferation and BMP-dependent exit to chondrogenesis. *Dev. Cell* **22**, 837–848 (2012).
30. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).
31. Abbasi, A. A. *et al.* Human intronic enhancers control distinct sub-domains of Gli3 expression during mouse CNS and limb development. *BMC Dev. Biol.* **10**, 44 (2010).
32. Osterwalder, M. *et al.* HAND2 targets define a network of transcriptional regulators that compartmentalize the early limb bud mesenchyme. *Dev. Cell* **31**, 345–357 (2014).
33. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).
34. Galvin, K. M. *et al.* A role for smad6 in development and homeostasis of the cardiovascular system. *Nat. Genet.* **24**, 171–174 (2000).
35. Osterwalder, M. *et al.* Characterization of Mammalian In Vivo Enhancers Using Mouse Transgenesis and CRISPR Genome Editing. *Methods Mol. Biol.* **2403**, 147–186 (2022).
36. Fueyo, R., Judd, J., Feschotte, C. & Wysocka, J. Roles of transposable elements in the regulation of mammalian transcription. *Nat. Rev. Mol. Cell Biol.* 1–17 (2022) doi:10.1038/s41580-022-00457-y.
37. Preissl, S. *et al.* Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432–439 (2018).
38. Sarropoulos, I. *et al.* Developmental and evolutionary dynamics of cis-regulatory elements in mouse cerebellar cells. *Science* **373**, (2021).
39. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
40. Kvon, E. Z. *et al.* Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature* **512**, 91–95 (2014).
41. Dickel, D. E. *et al.* Function-based identification of mammalian enhancers using site-specific integration. *Nat. Methods* **11**, 566–571 (2014).

42. Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods* **14**, 629 (2017).
43. Gasperini, M. *et al.* CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *Am. J. Hum. Genet.* **101**, 192–205 (2017).
44. Muerdter, F. *et al.* Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods* **15**, 141–149 (2018).
45. VanOudenhove, J., Yankee, T. N., Wilderman, A. & Cotney, J. Epigenomic and Transcriptomic Dynamics During Human Heart Organogenesis. *Circ. Res.* **127**, e184–e209 (2020).
46. Wilderman, A., VanOudenhove, J., Kron, J., Noonan, J. P. & Cotney, J. High-Resolution Epigenomic Atlas of Human Embryonic Craniofacial Development. *Cell Rep.* **23**, 1581–1597 (2018).
47. Cotney, J. *et al.* The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell* **154**, 185–196 (2013).
48. Nord, A. S. *et al.* Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155**, 1521–1531 (2013).
49. He, P. *et al.* The changing mouse embryo transcriptome at whole tissue and single-cell resolution. *Nature* **583**, 760–767 (2020).
50. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–8 (2006).
51. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
52. Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M. & Valen, E. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res.* **42**, W401–7 (2014).
53. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
54. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).

Chapter 3 : A gene desert required for regulatory control of pleiotropic *Shox2* expression and embryonic survival

In this chapter I describe a second study separate from my main project (Chapter 2) that involves a collaborative effort to characterize the *cis*-regulatory landscape within the gene desert that flanks *Shox2*, which encodes a transcription factor essential for proper craniofacial, limb, and heart development. This work has been posted on bioRxiv and will be submitted for publication as follows: Samuel Abassah-Oppong, Brandon J. Mannion, Virginie Tissières, Eddie Rodríguez-Carballo, Anja Ljubojevic, Fabrice Darbellay, Tabitha A. Festa, Carly S. Sullivan, Guy Kelman, Riana D. Hunter, Catherine S. Novak, Ingrid Plajzer-Frick, Stella Tran, Jennifer A. Akiyama, Iros Barozzi, Guillaume Andrey, Javier Lopez-Rios, Diane E. Dickel, Axel Visel, Len A. Pennacchio, John Cobb, and Marco Osterwalder. A gene desert required for regulatory control of pleiotropic *Shox2* expression and embryonic survival.

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

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. 3.1A**). This *Shox2*-TAD harbors one additional protein-coding gene (*Rsrc1*), while three other genes (*Mlf1, Veph1, Ptx3)* are found in neighboring chromatin domains (**Fig. 3.1A, S3.1A**). 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. S3.1A**). 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. 3.1B, S3.1A**). 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). 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. 3.1B, S3.1 and Table S3.1**). 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. 3.1B**). 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.

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. 3.1C and Table S3.2**), 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. 3.1B-C**). However, while multiple GDEs showed elevated H3K27ac signatures in developing limbs, transgenic screens identified only one element (GDE6) able to drive reporter activity in forelimbs (**Fig. 3.1C**). 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. 3.1C**). 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. 3.1C**). *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. 3.1C**). 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. 3.1B-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.

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. 3.1B**), 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. 3.2A, S3.2A-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. 3.2B, S3.2C**). We conducted two independent 4C-seq experiments using a viewpoint directly adjacent to the *Shox2* promoter (**Fig. 3.2A-B and Table S3.3**). The two replicates displayed reproducible interaction profiles revealing discrete regions with high interaction frequencies with the *Shox2* promoter (**Fig. 3.2B)**. 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. 3.2A**), 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. 3.1C, S3.1, 3.2A, 3.2B**). 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. 3.2B-C and Table S3.4**)[24,25,36]. Finally, our 4C-seq analysis identified three *Shox2*-contacting gene desert modules (+237kb, +407kb and +568kb) with limb enhancer signatures (**Fig. 3.2A-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. 3.2B-C and Table S3.4**). 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. S3.3**). In a last step, using 4C-seq we assessed the 3D interaction profiles of selected individual enhancers (PLE2 and PLE4) (**Fig. S3.2C**). These experiments corroborated the specific interactions observed between both enhancers and the *Shox2* promoter (**Fig. S3.2C**). Interestingly, while PLE2 shows no interaction with other enhancers, PLE4 is establishing contacts with two other proximal limb enhancers (PLE1

and PLE3) (**Fig. S3.2C**). 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.

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. 3.2A, S3.4A and Tables S3.5, S3.6**). Heterozygous F1 mice with clean deletion breakpoints (*S2*GD$^{\Delta/+}$) (**Fig. S3.4A**) 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. S3.4C**), 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. 3.3A**), despite the loss of multiple PLEs (**Fig. 3.2C, S3.3**). 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. 3.3B and Table S3.7**). 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. 3.3C**). 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. 3.3C**). 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.

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 (**Fig. 3.4A-B**). Furthermore, *Shox2* expression in the medial nasal process (MNP) was severely downregulated at E10.5 and E11.5 (**Fig. 3.4A-B**). 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. 3.1B, 3.2A**). Importantly, GDE9 and GDE15 show activity patterns that closely overlap *Shox2* in the maxillary-mandibular (MXP-MDP) and MNP compartments, respectively (**Fig. 3.4C**). In addition, transgenic validation of other predicted GDEs identified multiple brain and cranial nerve activities (**Fig. 3.1B-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. 3.4A**). Hereby, the presence of multiple brain enhancers with overlapping activities in the diencephalon, midbrain and hindbrain, both inside and outside the gene desert (**Fig. S3.5**, 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. 3.4A**). Quantitative expression profiling in *S2*GD$^{\Delta/\Delta}$ embryonic hearts at E11.5 furthermore revealed severe downregulation of cardiac *Shox2* transcripts (**Fig. 3.4B**). 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. 3.1B-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 3.5A-B**). In accordance with the absence of *Shox2* transcripts in the SV at E10.5 (**Fig. 3.4A**), 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. 3.5A-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. S3.4C**), 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. 3.5C-D**). ATAC-seq peak calling analysis uncovered 10 elements within the deleted gene desert region which were significantly enriched for open chromatin (**Fig. 3.5C and Table S3.8**). 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. 3.1B**). As none of these elements drove reproducible *LacZ* reporter activity in cardiac regions at E11.5 (**Fig. 3.1C**), 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. 3.5C**). 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. 3.5B,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. 3.5E**). 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. 3.5E**). 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. S3.6**), 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.

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. 3.4A, 3.5A-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. 3.6C,E**), potentially acting as a 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] demonstrates the promise of such an approach.

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. 3.2 and S3.3**, 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. 3.1A** 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 S3.2 and S3.8**. 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 S3.4)** 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. S3.3**) 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. 3.2B**. 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 S3.3**. 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. 3.2B, S3.2C**). 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 S3.5** 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. S3.4A** and **Table S3.6** for genotyping strategy, primer sequences and PCR amplicons). Sanger sequencing was used to identify and confirm deletion breakpoints in F0 and F1 mice (**Fig. S3.4A**).

***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 S3.1**, *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 S3.1**, *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).

Figures

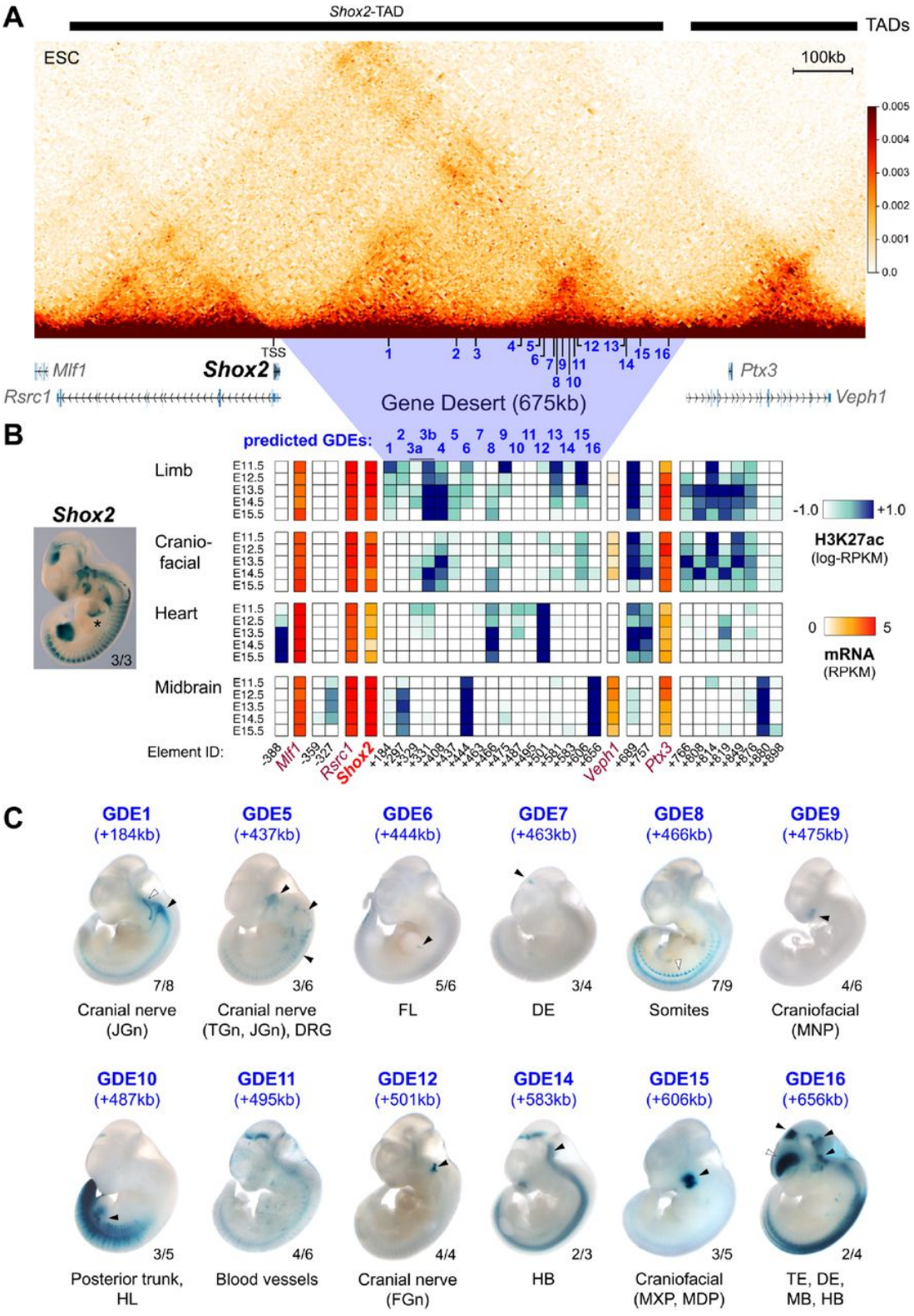

A
Shox2-TAD
TADs
ESC
100kb
0.005
0.004
0.003
0.002
0.001
0.0
TSS
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Mlf1
Shox2
Ptx3
Rsrc1
Veph1
Gene Desert (675kb)
B
predicted GDEs:
1
2
3a
3b
4
5
6
7
8
9
10
11
12
13
14
15
16
Shox2
3/3
Limb
Cranio-facial
Heart
Midbrain
E11.5
E12.5
E13.5
E14.5
E15.5
-1.0
+1.0
H3K27ac
(log-RPKM)
0
5
mRNA
(RPKM)
Element ID:
-388
Mlf1
-359
-327
Rsrc1
Shox2
+194
+297
+329
+331
+408
+437
+444
+463
+466
+475
+487
+495
+501
+583
+606
+656
Veph1
+689
+757
Ptx3
+766
+808
+814
+819
+849
+876
+890
+898
C
GDE1
(+184kb)
7/8
Cranial nerve
(JGn)
GDE5
(+437kb)
3/6
Cranial nerve
(TGn, JGn), DRG
GDE6
(+444kb)
5/6
FL
GDE7
(+463kb)
3/4
DE
GDE8
(+466kb)
7/9
Somites
GDE9
(+475kb)
4/6
Craniofacial
(MNP)
GDE10
(+487kb)
3/5
Posterior trunk,
HL
GDE11
(+495kb)
4/6
Blood vessels
GDE12
(+501kb)
4/4
Cranial nerve
(FGn)
GDE14
(+583kb)
2/3
HB
GDE15
(+606kb)
3/5
Craniofacial
(MXP, MDP)
GDE16
(+656kb)
2/4
TE, DE,
MB, HB

Figure 3.1. 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 S3.1 and methods**). The full set of tissues is shown in **Fig. S3.1**. 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.

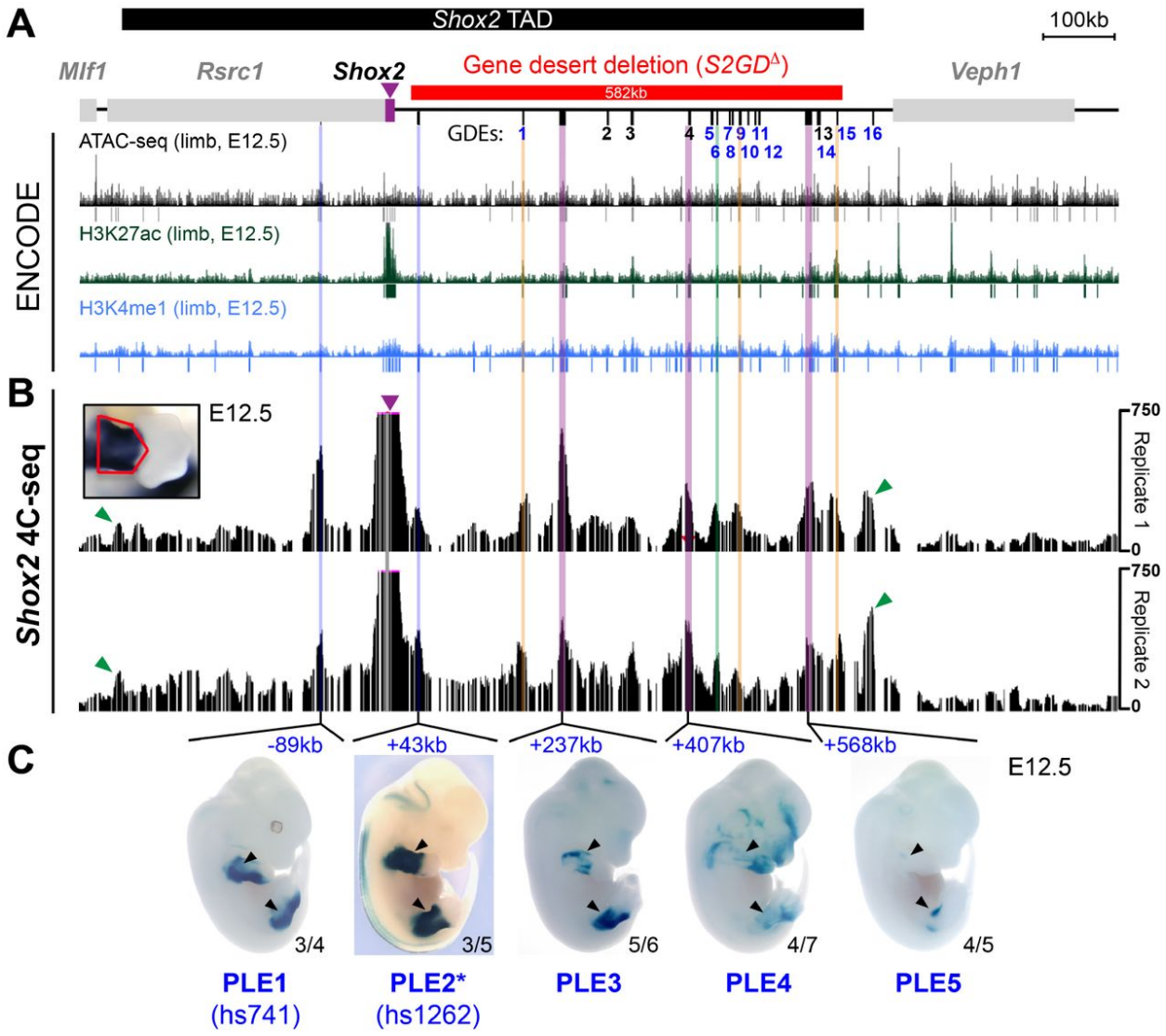


Figure 3.2. Chromatin conformation capture identifies the gene desert as a hub for *Shox2*-interacting limb enhancers. (**A**) Map of the 1.4 Mb extended *Shox2*-TAD (see **Fig. 3.1A**) and the locations of the predicted gene desert enhancers (GDEs) 1-16, in blue those with confirmed *in vivo* enhancer activities at E11.5 (**Fig. 3.1B, 3.1C and Table S3.1**). 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. S3.2A, 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. 3.1C** 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. 3.1C**). The embryos shown are representatives from stable transgenic *LacZ* reporter lines (see **Fig. S3.3**). 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).

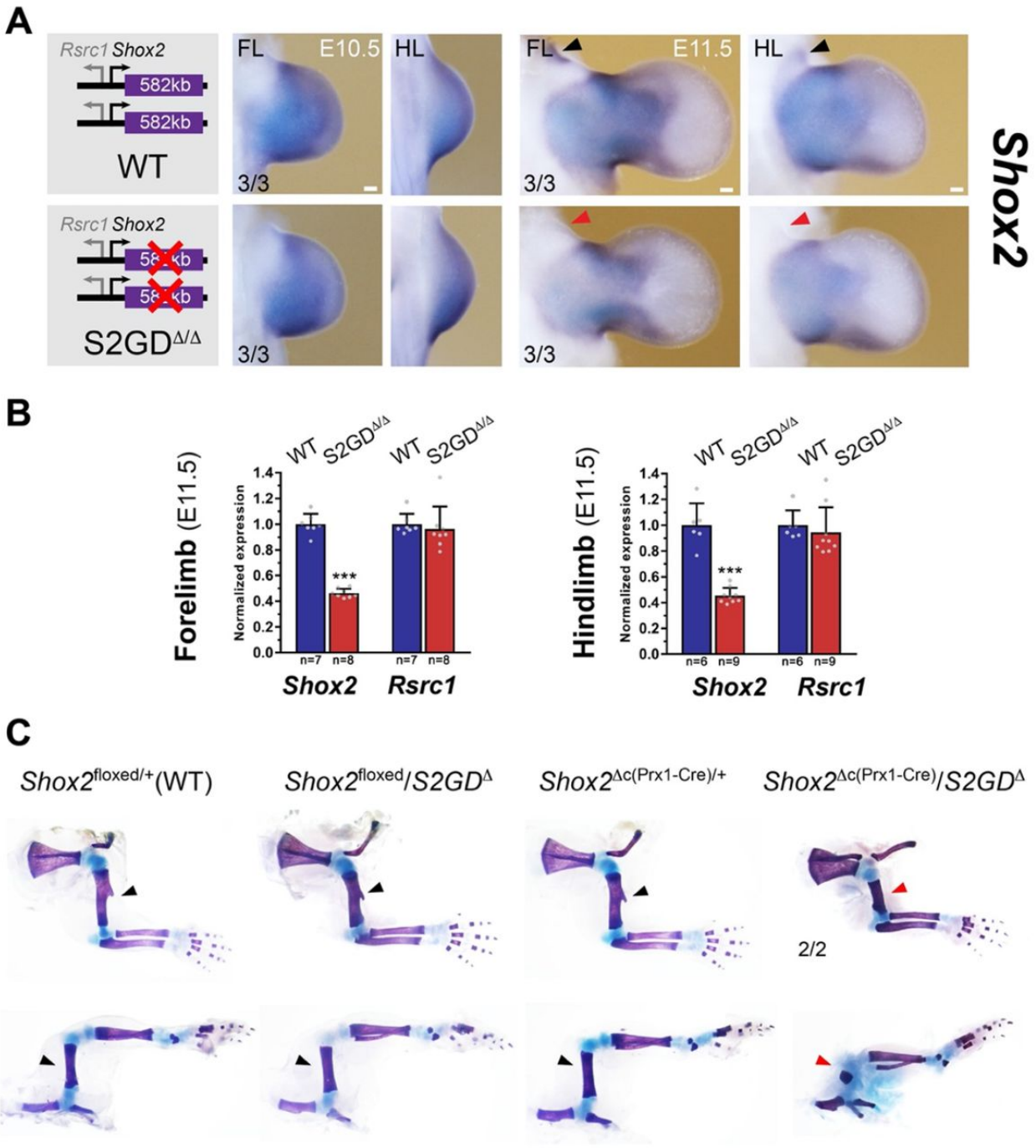


Figure 3.3. The gene desert controls quantitative *Shox2* expression in limbs as part of a resilient regulatory architecture. (**A**) ISH reveals spatial *Shox2* expression in fore- and hindlimb buds of embryos homozygous for the gene desert deletion (*S2*GD$^{\Delta/\Delta}$) at E10.5 and E11.5. Red 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.

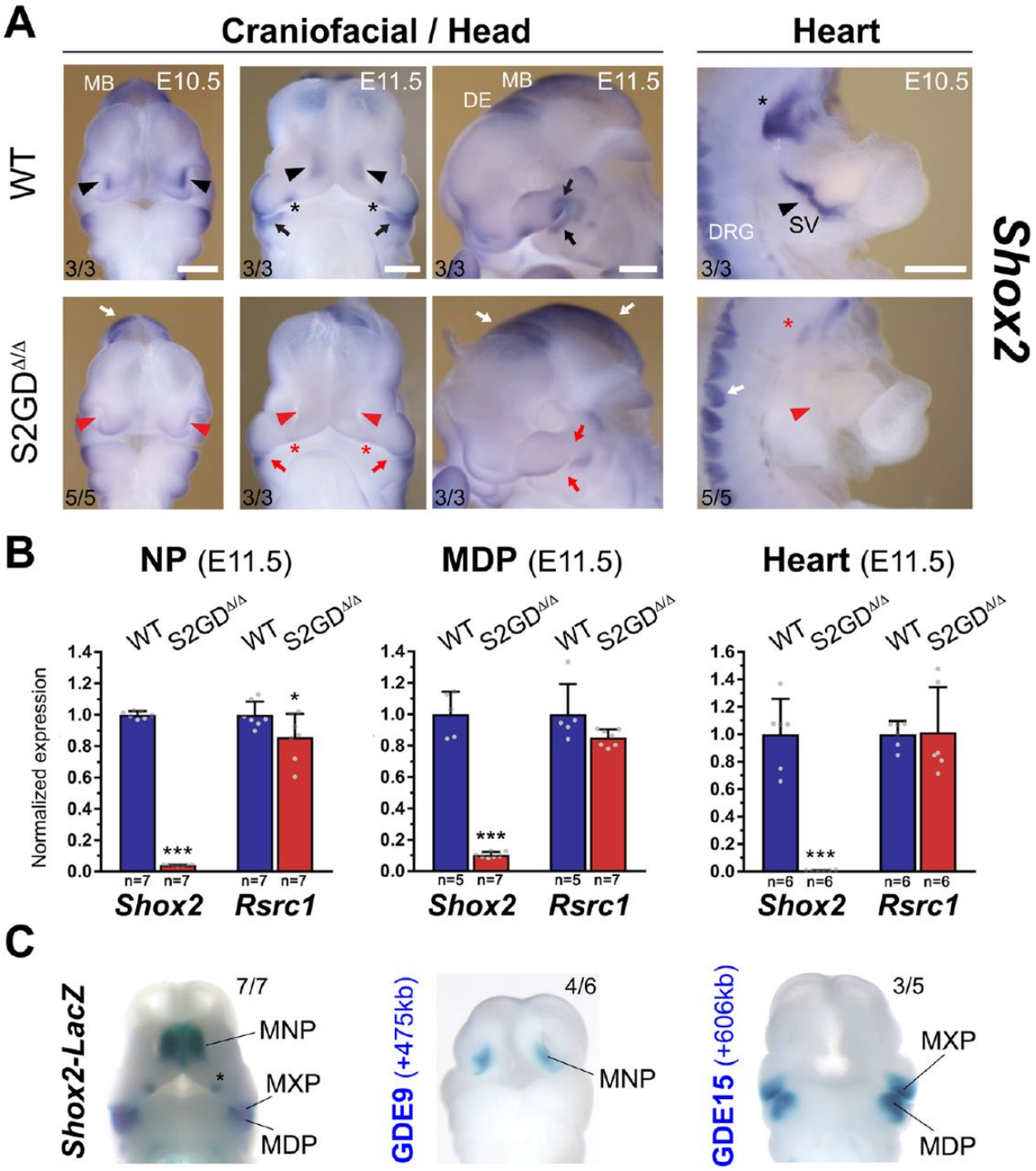


Figure 3.4. The gene desert controls pleiotropic *Shox2* expression with a predominant impact in craniofacial and cardiac domains. (**A**) RNA *in situ* hybridization (ISH) reveals severely reduced *Shox2* transcripts in craniofacial and cardiac expression domains in embryos homozygous for the gene desert 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. 3.1C**) 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.

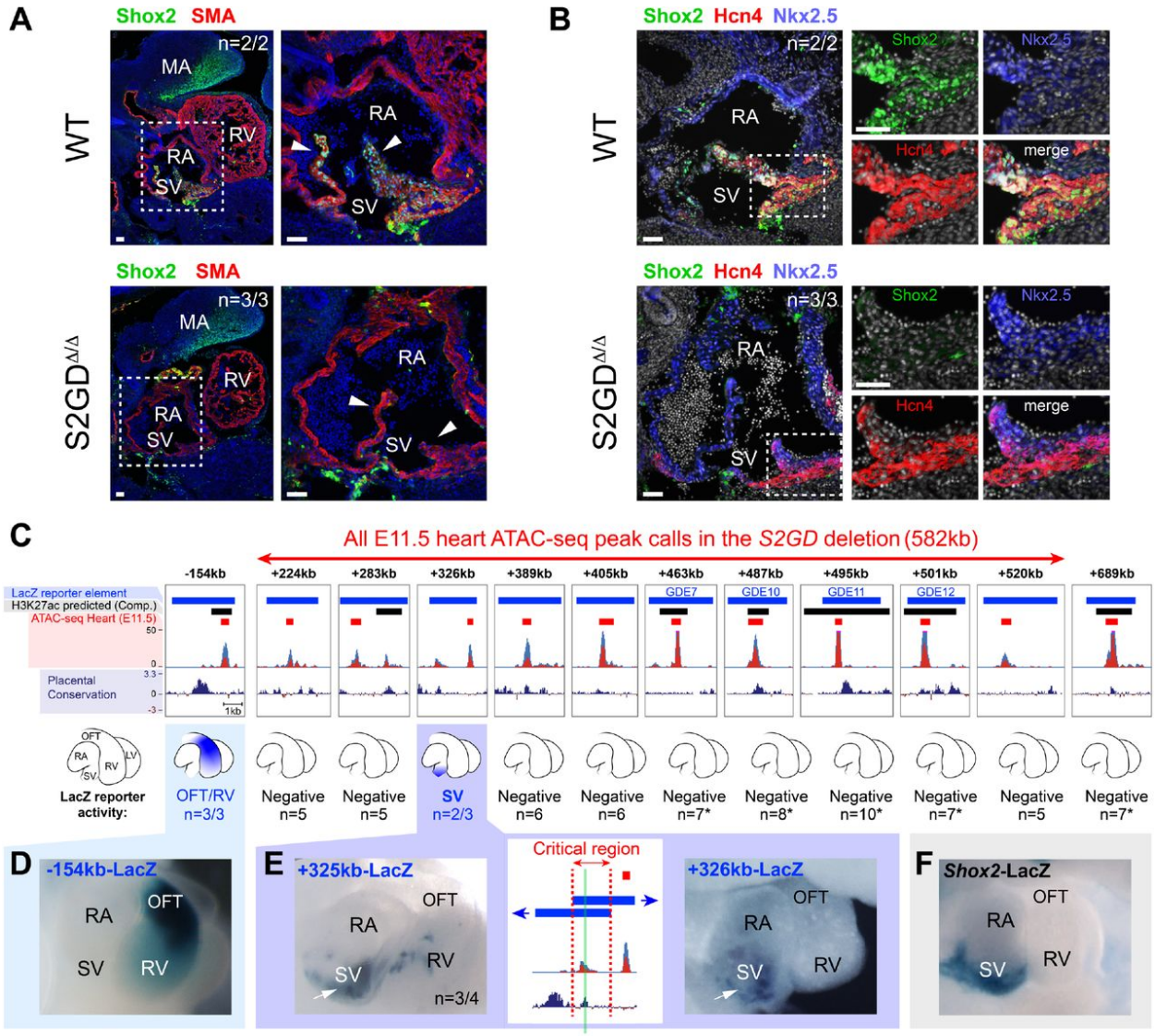


Figure 3.5. Identification of a sinus venosus (SV) gene desert enhancer implicated in critical regulation of *Shox2* in sinoatrial cardiac pacemaker cells. (**A**) Compared to abundant Shox2 protein (green) in the SV of wildtype (WT) hearts, Shox2 is depleted in myocardial cells of the SV including the 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. 3.1C, S3.5**). 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. S3.6**). RA, right atrium, RV, right ventricle, LV, left ventricle, OFT, outflow tract.

Supplementary Materials

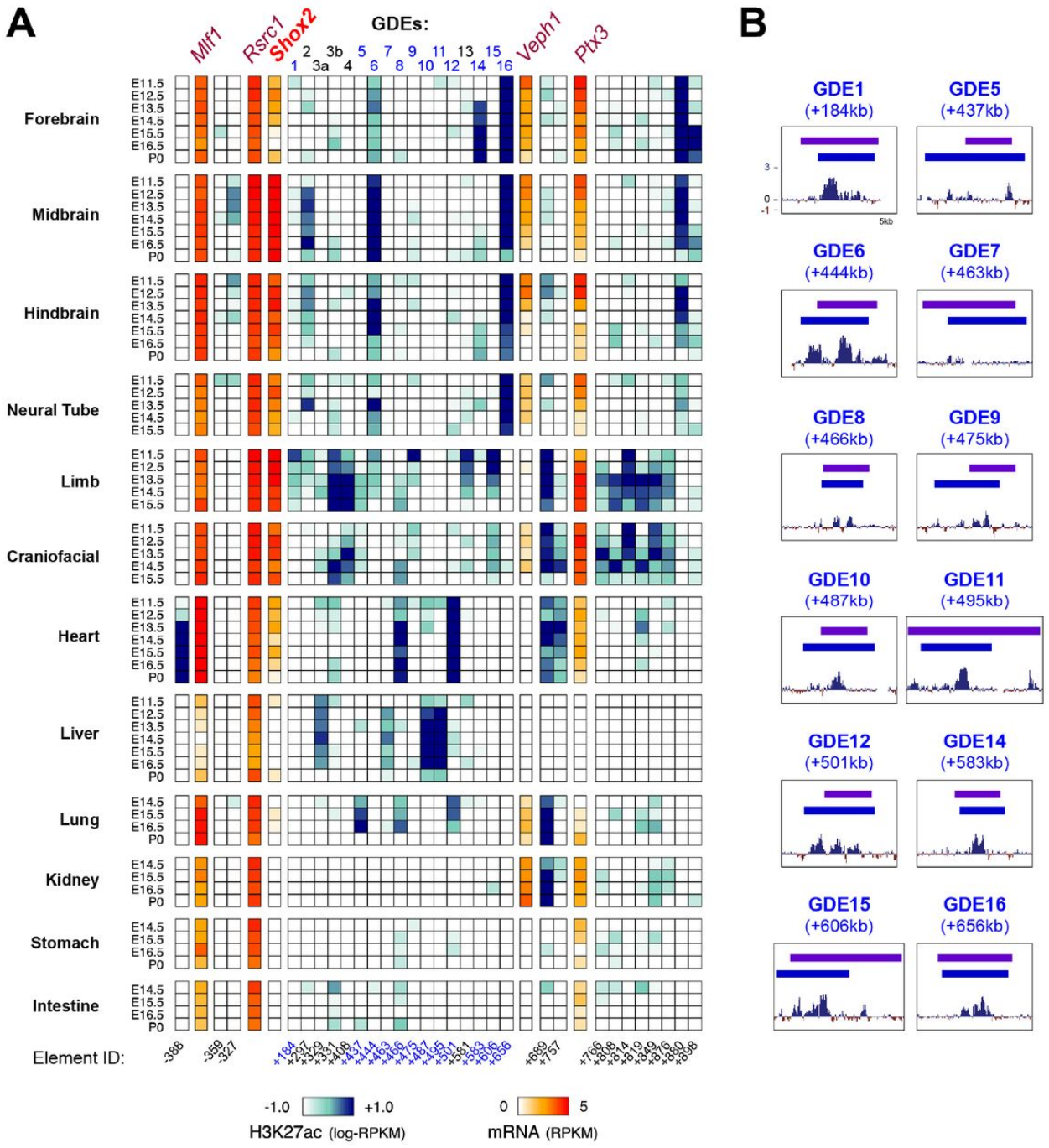


Figure S3.1. Prediction of spatio-temporal enhancer activities in the extended *Shox2* regulatory domain. (**A**) Complete heat map listing predicted enhancer elements (Element IDs) within the extended *Shox2* TAD region[10] across different time-points and tissues (see also **Fig. 3.1**), based on H3K27ac 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 S3.1**). Predictedgene desert enhancer elements (GDEs) are indicated and those with confirmed enhancer activities at E11.5 (**Fig. 3.1C**) 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. 3.1C and Table S3.2**). Distance (in *kb*) from the *Shox2* TSS is indicated for each Element ID (-, upstream; +, downstream).

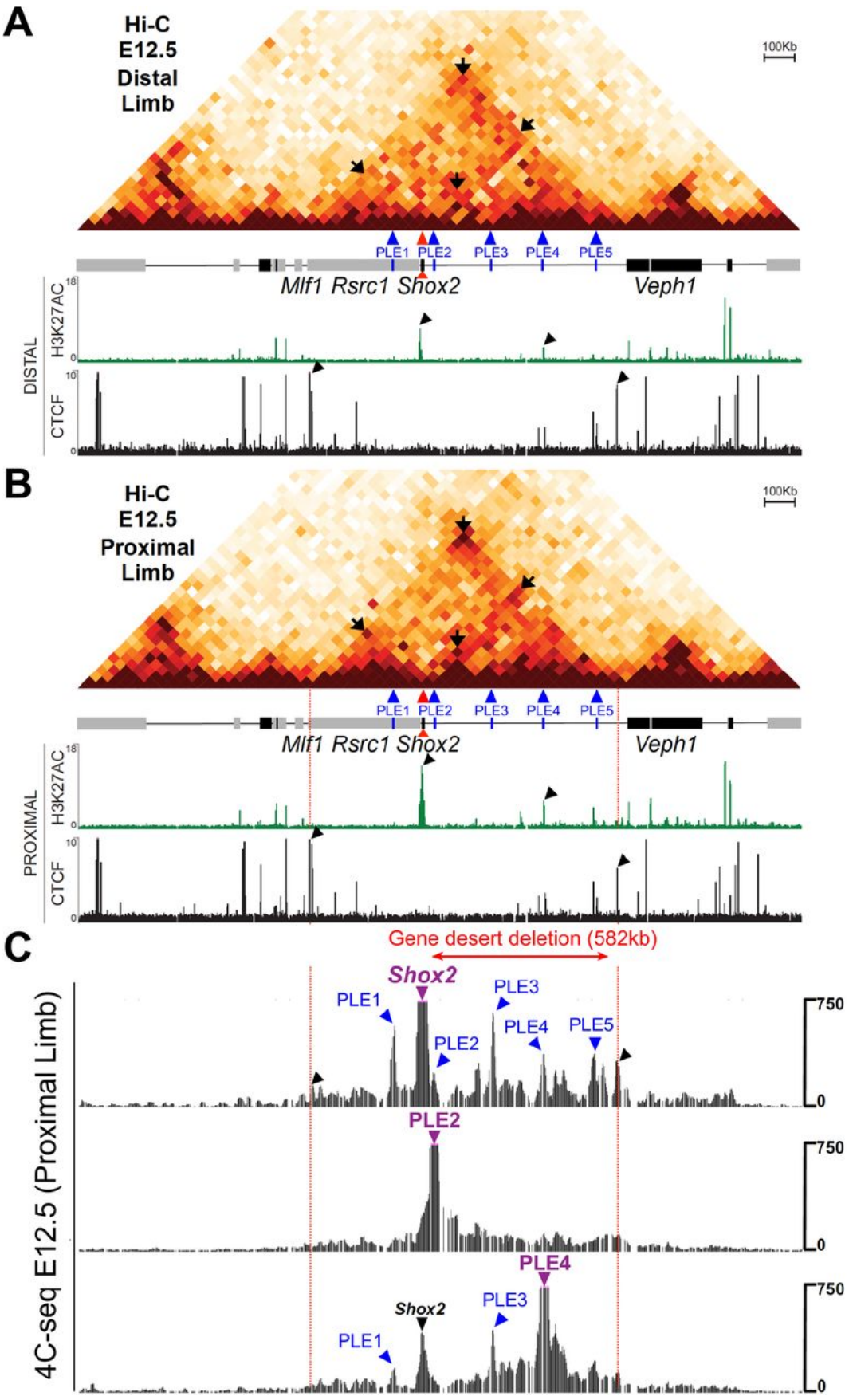

A
Hi-C
E12.5
Distal
Limb
100Kb
PLE1
PLE2
PLE3
PLE4
PLE5
Mlf1 Rsrc1 Shox2
Veph1
DISTAL
H3K27AC
CTCF
B
Hi-C
E12.5
Proximal
Limb
100Kb
PLE1
PLE2
PLE3
PLE4
PLE5
Mlf1 Rsrc1 Shox2
Veph1
PROXIMAL
H3K27AC
CTCF
C
Gene desert deletion (582kb)
4C-seq E12.5 (Proximal Limb)
Shox2
PLE1
PLE2
PLE3
PLE4
PLE5
750
0
PLE2
750
0
PLE4
PLE3
Shox2
PLE1
750
0

Figure S3.2. 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. 3.2**) 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. 3.2B**), PLE2 (hs1262/LHB-A) and PLE4 viewpoints (purple arrowheads) (**Table S3.3**). 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*.

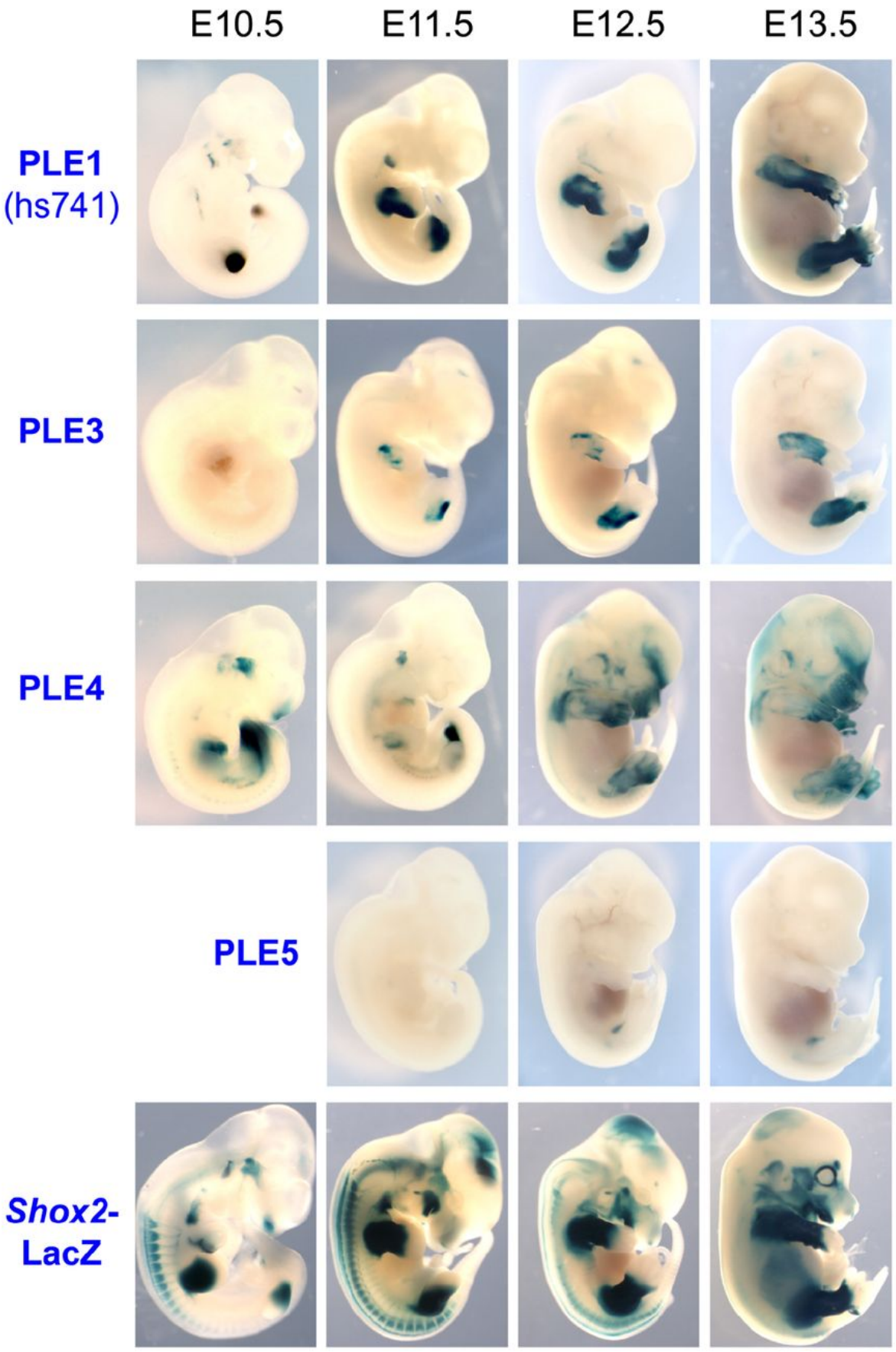

E10.5
E11.5
E12.5
E13.5
PLE1
(hs741)
PLE3
PLE4
PLE5
Shox2-
LacZ

Figure S3.3. 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. 3.2C and Table S3.4**). 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.

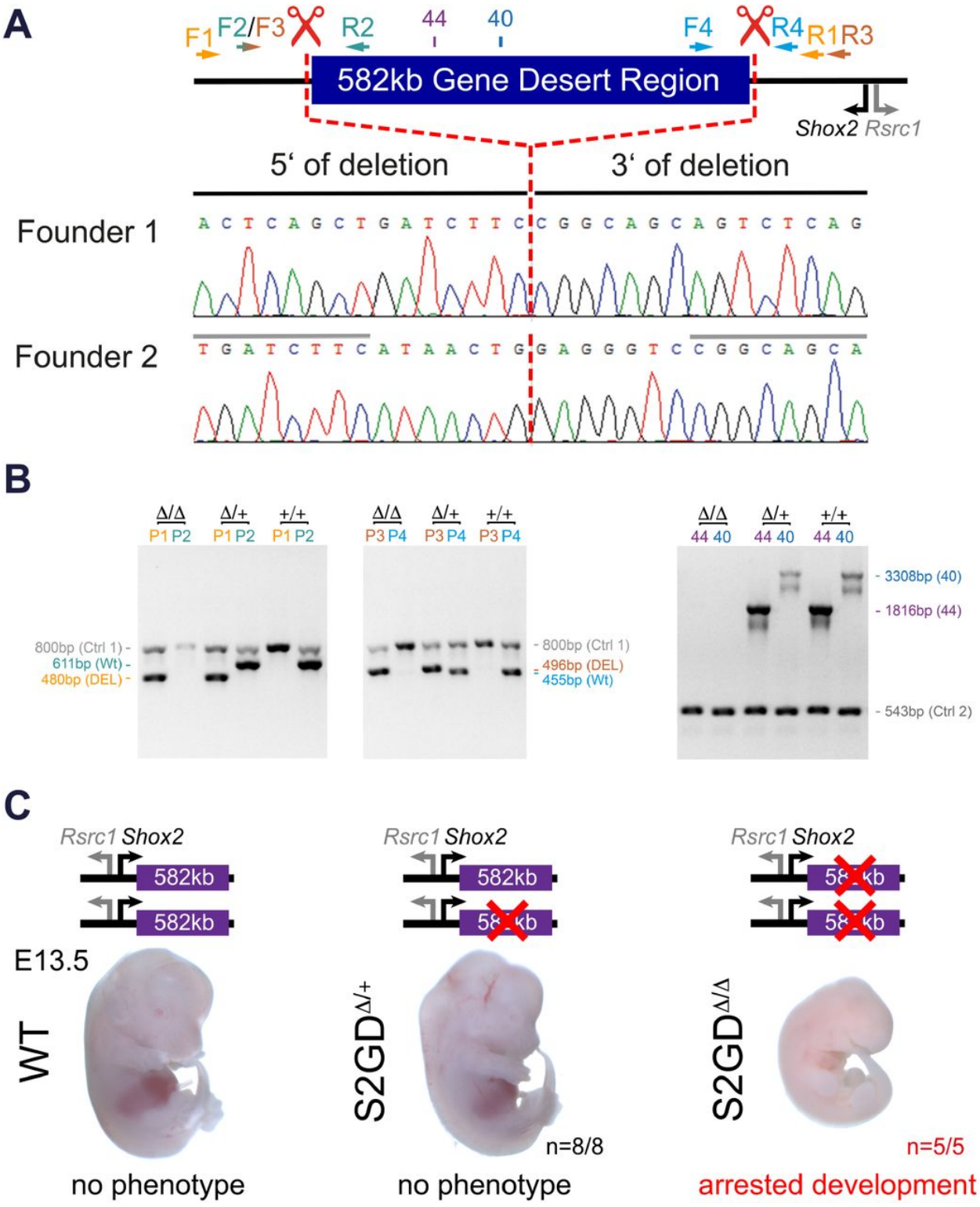


Figure S3.4. CRISPR/Cas9-mediated deletion of the *Shox2* gene desert causes embryonic lethality. (**A**) Validation of clean deletion breakpoints in gene desert knockout mouse lines. Red

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 S3.5**). 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 S3.6** 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.

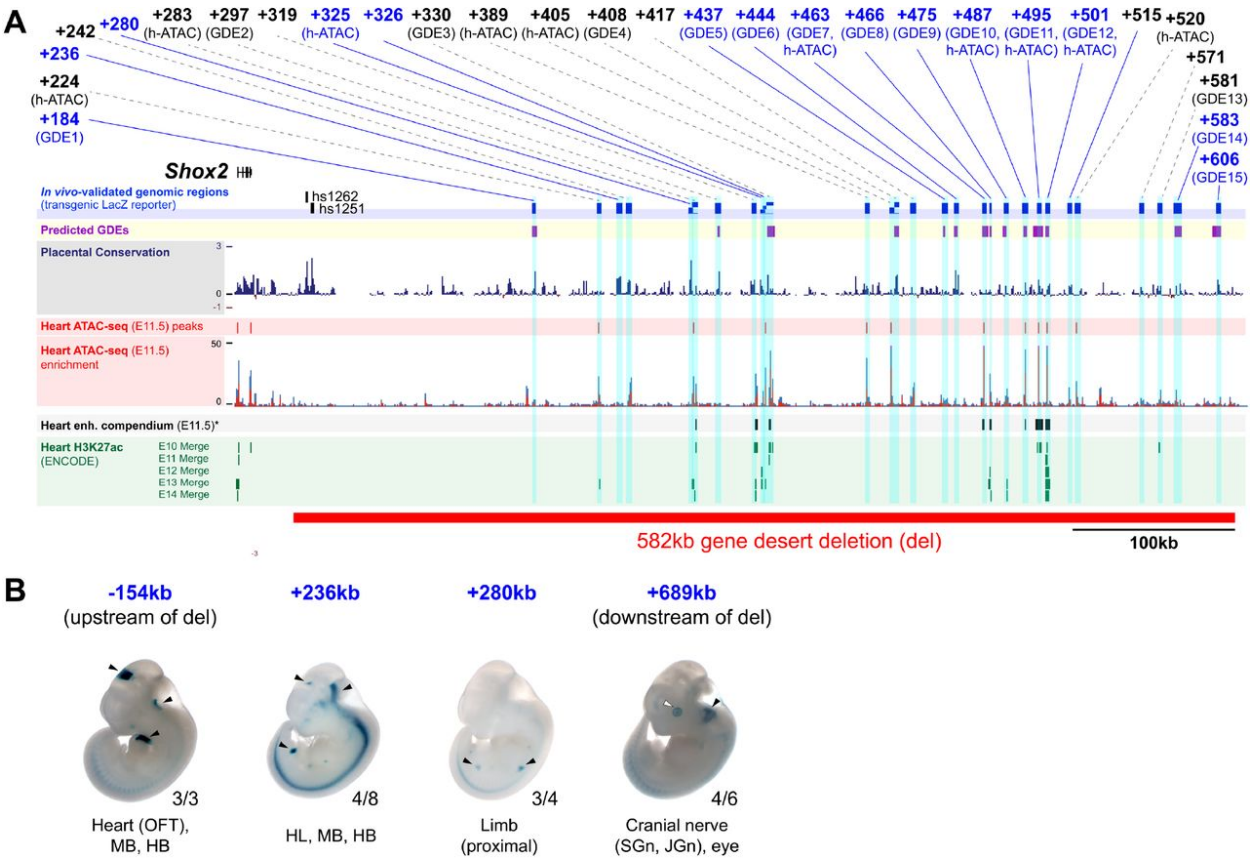


Figure S3.5. ***Shox2* candidate enhancer regions validated in this study at E11.5.** (**A**) UCSC browser window showing the *Shox2* gene and the deleted centromeric gene desert region. Candidate enhancer elements validated via Hsp68-*LacZ* reporter transgenesis in this study (blue bars**)** are listed and named based on the distance (in *kb*) to the *Shox2* transcriptional start site (TSS) (**Tables S3.2, S3.8**). 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. 3.1B**), heart ATAC-seq at E11.5 (h-ATAC, Fig. **3.5C**), 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.

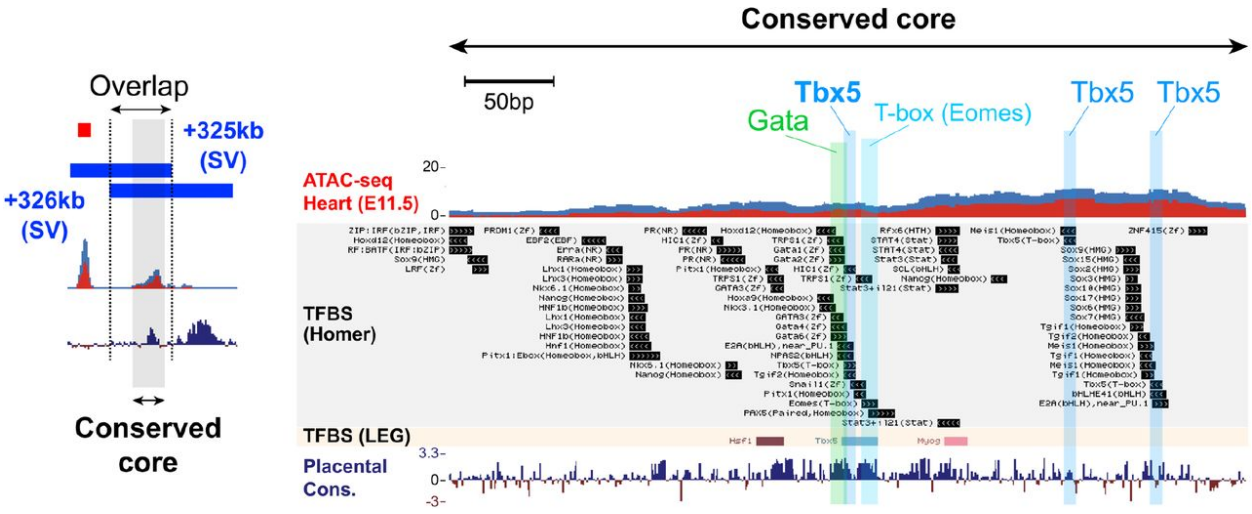


Figure S3.6. Tbx5 binding motifs within the *Shox2*-SV enhancer region. UCSC browser window showing the conserved core within the region of overlap of +326kb and +325kb SV enhancers (**Fig. 3.5C, E**). Predictions of transcription factor binding sites (TFBS) mapping to this region are derived from the 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.

Table S3.1. Developmental enhancer predictions within the Shox2 TAD.

Note: Table S3.1 sheet 1 comprises over 70 columns and Table S3.1 sheet 2 comprises over 130 columns. The two sheets provided below are excerpts, and a full version is available in the bioRxiv submission. *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. 3.1B and S3.1**. 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).

Element ID	chrom	start	end	id
Mlf1	chr3	67374097	67400003	rna_ENSMUSG00000048416.11_Mlf1
-388	chr3	67368500	67370500	ac_chr3:67368500-67370500
Gm17402	chr3	67365461	67375163	rna_ENSMUSG00000090408.1_Gm17402
-359	chr3	67340100	67342100	ac_chr3:67340100-67342100
-327	chr3	67307700	67309700	ac_chr3:67307700-67309700
Rsrc1	chr3	66981390	67358396	rna_ENSMUSG00000034544.13_Rsrc1
Shox2	chr3	66971727	66981771	rna_ENSMUSG00000027833.12_Shox2
+184	chr3	66796500	66799900	ac_chr3:66796500-66799900
+297	chr3	66683300	66685300	ac_chr3:66683300-66685300
+329	chr3	66652000	66654000	ac_chr3:66652000-66654000
+331	chr3	66649600	66651600	ac_chr3:66649600-66651600
+408	chr3	66572500	66575700	ac_chr3:66572500-66575700
+437	chr3	66543800	66545800	ac_chr3:66543800-66545800
+444	chr3	66536200	66538800	ac_chr3:66536200-66538800
+463	chr3	66517200	66521200	ac_chr3:66517200-66521200
+466	chr3	66515000	66517000	ac_chr3:66515000-66517000
+475	chr3	66506100	66508100	ac_chr3:66506100-66508100
+487	chr3	66493700	66495700	ac_chr3:66493700-66495700
+495	chr3	66483600	66489300	ac_chr3:66483600-66489300
+501	chr3	66479700	66481700	ac_chr3:66479700-66481700
+581	chr3	66399900	66401900	ac_chr3:66399900-66401900
+583	chr3	66397500	66399500	ac_chr3:66397500-66399500
+606	chr3	66373800	66378600	ac_chr3:66373800-66378600
+656	chr3	66324600	66327800	ac_chr3:66324600-66327800
+689	chr3	66291100	66294000	ac_chr3:66291100-66294000
+757	chr3	66224000	66226000	ac_chr3:66224000-66226000
Ptx3	chr3	66219910	66225805	rna_ENSMUSG00000027832.5_Ptx3
+766	chr3	66215000	66217000	ac_chr3:66215000-66217000
+808	chr3	66172300	66174300	ac_chr3:66172300-66174300
+814	chr3	66165800	66168800	ac_chr3:66165800-66168800
+819	chr3	66161700	66163700	ac_chr3:66161700-66163700
+849	chr3	66131300	66133300	ac_chr3:66131300-66133300
+876	chr3	66105100	66107100	ac_chr3:66105100-66107100
+880	chr3	66101100	66103100	ac_chr3:66101100-66103100
+898	chr3	66082800	66084800	ac_chr3:66082800-66084800
Veph1	chr3	66053558	66296837	rna_ENSMUSG00000027831.9_Veph1

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.

Element ID	chrom	start	end
-388	chr3	67368500	67370500
-359	chr3	67340100	67342100
-327	chr3	67307700	67309700
+184	chr3	66796500	66799900
+297	chr3	66683300	66685300
+329	chr3	66652000	66654000
+331	chr3	66649600	66651600
+408	chr3	66572500	66575700
+437	chr3	66543800	66545800
+444	chr3	66536200	66538800
+463	chr3	66517200	66521200
+466	chr3	66515000	66517000
+475	chr3	66506100	66508100
+487	chr3	66493700	66495700
+495	chr3	66483600	66489300
+501	chr3	66479700	66481700
+581	chr3	66399900	66401900
+583	chr3	66397500	66399500
+606	chr3	66373800	66378600
+656	chr3	66324600	66327800
+689	chr3	66291100	66294000
+757	chr3	66224000	66226000
+766	chr3	66215000	66217000
+808	chr3	66172300	66174300
+814	chr3	66165800	66168800
+819	chr3	66161700	66163700
+849	chr3	66131300	66133300
+876	chr3	66105100	66107100
+880	chr3	66101100	66103100
+898	chr3	66082800	66084800

Sheet 3: Metadata to be able to retrieve peak lists (for q-values) from the ENCODE Data Coordination Center (DCC, http://www.encodeproject.org/).

File_Accession	Experiment_Accession	Biorep	Description
ENCFF565NSZ	ENCSR672ZXY	1, 2	H3K27ac ChIP-seq on postnatal 0 day mouse midbrain
ENCFF323EFD	ENCSR123MLY	1, 2	H3K27ac ChIP-seq on embryonic 12.5 day mouse heart
ENCFF117LLJ	ENCSR863VHE	1, 2	H3K27ac ChIP-seq on embryonic 10.5 day mouse limb
ENCFF579HGD	ENCSR671NSS	1, 2	H3K27ac ChIP-seq on embryonic 13.5 day mouse midbrain
ENCFF384LAY	ENCSR382DRK	1, 2	H3K27ac ChIP-seq on embryonic 15.5 day mouse embryonic facial prominence
ENCFF659YSV	ENCSR289SWJ	1, 2	H3K27ac ChIP-seq on embryonic 13.5 day mouse neural tube
ENCFF468ZXG	ENCSR553IWV	1, 2	H3K27ac ChIP-seq on embryonic 16.5 day mouse midbrain
ENCFF897EEM	ENCSR275KPI	1, 2	H3K27ac ChIP-seq on embryonic 11.5 day mouse forebrain
ENCFF084IYL	ENCSR320EEW	1, 2	H3K27ac ChIP-seq on embryonic 14.5 day mouse forebrain
ENCFF223IHN	ENCSR546ANT	1, 2	H3K27ac ChIP-seq on embryonic 16.5 day mouse stomach
ENCFF290MLR	ENCSR616TJM	1, 2	H3K27ac ChIP-seq on postnatal 0 day mouse liver
ENCFF974RAJ	ENCSR136GMT	1, 2	H3K27ac ChIP-seq on embryonic 12.5 day mouse liver
ENCFF676TSV	ENCSR094TTT	1, 2	H3K27ac ChIP-seq on postnatal 0 day mouse forebrain
ENCFF203QTV	ENCSR129LAP	1, 2	H3K27ac ChIP-seq on embryonic 11.5 day mouse hindbrain
ENCFF378DZY	ENCSR599GVS	1, 2	H3K27ac ChIP-seq on embryonic 15.5 day mouse intestine
ENCFF834HBK	ENCSR797EYS	1, 2	H3K27ac ChIP-seq on embryonic 16.5 day mouse hindbrain
ENCFF272TNO	ENCSR401GRX	1, 2	H3K27ac ChIP-seq on embryonic 11.5 day mouse embryonic facial prominence
ENCFF083AAZ	ENCSR711SVB	1, 2	H3K27ac ChIP-seq on embryonic 15.5 day mouse kidney
ENCFF515NHP	ENCSR057SHA	1, 2	H3K27ac ChIP-seq on embryonic 14.5 day mouse kidney
ENCFF948YFR	ENCSR897WBY	1, 2	H3K27ac ChIP-seq on embryonic 11.5 day mouse limb
ENCFF700WUD	ENCSR140UEX	1, 2	H3K27ac ChIP-seq on embryonic 16.5 day mouse lung
ENCFF546IUI	ENCSR344HHI	1, 2	H3K27ac ChIP-seq on embryonic 13.5 day mouse hindbrain
ENCFF093QGY	ENCSR452WYC	1, 2	H3K27ac ChIP-seq on embryonic 14.5 day mouse lung
ENCFF764WNW	ENCSR332JYZ	1, 2	H3K27ac ChIP-seq on postnatal 0 day mouse hindbrain
ENCFF751ZHP	ENCSR594JGI	1, 2	H3K27ac ChIP-seq on embryonic 10.5 day mouse hindbrain
ENCFF014HMN	ENCSR825ZJV	1, 2	H3K27ac ChIP-seq on embryonic 10.5 day mouse forebrain
ENCFF566DFK	ENCSR088UKA	1, 2	H3K27ac ChIP-seq on embryonic 11.5 day mouse midbrain
ENCFF399KKD	ENCSR479LFP	1, 2	H3K27ac ChIP-seq on embryonic 15.5 day mouse liver
ENCFF998WFE	ENCSR316CNR	1, 2	H3K27ac ChIP-seq on embryonic 14.5 day mouse stomach
ENCFF386RYT	ENCSR357JII	1, 2	H3K27ac ChIP-seq on embryonic 16.5 day mouse kidney
ENCFF003VMR	ENCSR151APL	1, 2	H3K27ac ChIP-seq on embryonic 10.5 day mouse embryonic facial prominence
ENCFF467HTN	ENCSR905FFU	1, 2	H3K27ac ChIP-seq on embryonic 13.5 day mouse limb
ENCFF514SMO	ENCSR639DND	1, 2	H3K27ac ChIP-seq on embryonic 16.5 day mouse intestine
ENCFF531BZD	ENCSR481SGM	1, 2	H3K27ac ChIP-seq on embryonic 14.5 day mouse embryonic facial prominence
ENCFF153DSU	ENCSR846PJO	1, 2	H3K27ac ChIP-seq on embryonic 16.5 day mouse heart
ENCFF378OWA	ENCSR428OEK	1, 2	H3K27ac ChIP-seq on embryonic 16.5 day mouse forebrain
ENCFF889PZP	ENCSR737QWV	1, 2	H3K27ac ChIP-seq on embryonic 12.5 day mouse limb
ENCFF434LSI	ENCSR891SAW	1, 2	H3K27ac ChIP-seq on embryonic 12.5 day mouse neural tube
ENCFF954URD	ENCSR222IHX	1, 2	H3K27ac ChIP-seq on embryonic 11.5 day mouse heart
ENCFF241YYJ	ENCSR058DOA	1, 2	H3K27ac ChIP-seq on embryonic 11.5 day mouse liver
ENCFF877YAM	ENCSR066XFL	1, 2	H3K27ac ChIP-seq on embryonic 15.5 day mouse hindbrain
ENCFF213XYX	ENCSR252ONR	1, 2	H3K27ac ChIP-seq on embryonic 12.5 day mouse midbrain
ENCFF025JLK	ENCSR346FJG	1, 2	H3K27ac ChIP-seq on postnatal 0 day mouse stomach
ENCFF399WYR	ENCSR895BMP	1, 2	H3K27ac ChIP-seq on embryonic 15.5 day mouse lung
ENCFF627DHT	ENCSR966AIB	1, 2	H3K27ac ChIP-seq on embryonic 12.5 day mouse forebrain
ENCFF583IBI	ENCSR699XHY	1, 2	H3K27ac ChIP-seq on embryonic 13.5 day mouse heart
ENCFF157AZQ	ENCSR054JHZ	1, 2	H3K27ac ChIP-seq on embryonic 14.5 day mouse hindbrain
ENCFF026QRB	ENCSR311YPF	1, 2	H3K27ac ChIP-seq on embryonic 13.5 day mouse forebrain
ENCFF196WCO	ENCSR175KBJ	1, 2	H3K27ac ChIP-seq on embryonic 13.5 day mouse liver
ENCFF464KTV	ENCSR241BSK	1, 2	H3K27ac ChIP-seq on embryonic 15.5 day mouse neural tube
ENCFF573BYB	ENCSR265NBM	1, 2	H3K27ac ChIP-seq on embryonic 14.5 day mouse neural tube
ENCFF366XXD	ENCSR691NQH	1, 2	H3K27ac ChIP-seq on embryonic 15.5 day mouse forebrain
ENCFF399AMW	ENCSR988BRP	1, 2	H3K27ac ChIP-seq on embryonic 15.5 day mouse limb
ENCFF291VFI	ENCSR254AHA	1, 2	H3K27ac ChIP-seq on embryonic 14.5 day mouse midbrain
ENCFF720NTZ	ENCSR642VYW	1, 2	H3K27ac ChIP-seq on postnatal 0 day mouse intestine
ENCFF814SUK	ENCSR582SPN	1, 2	H3K27ac ChIP-seq on embryonic 10.5 day mouse heart
ENCFF281MKX	ENCSR989LUY	1, 2	H3K27ac ChIP-seq on embryonic 10.5 day mouse midbrain
ENCFF247VPI	ENCSR420MUV	1, 2	H3K27ac ChIP-seq on embryonic 13.5 day mouse embryonic facial prominence
ENCFF384FJW	ENCSR075SNV	1, 2	H3K27ac ChIP-seq on embryonic 14.5 day mouse liver
ENCFF872MVE	ENCSR140YPL	1, 2	H3K27ac ChIP-seq on postnatal 0 day mouse kidney
ENCFF409CQX	ENCSR784TLR	1, 2	H3K27ac ChIP-seq on embryonic 12.5 day mouse hindbrain
ENCFF160HCA	ENCSR929SEW	1, 2	H3K27ac ChIP-seq on embryonic 15.5 day mouse stomach
ENCFF470UWO	ENCSR802RET	1, 2	H3K27ac ChIP-seq on embryonic 16.5 day mouse liver
ENCFF447XAK	ENCSR428GHF	1, 2	H3K27ac ChIP-seq on embryonic 15.5 day mouse midbrain
ENCFF153SIZ	ENCSR360ANE	1, 2	H3K27ac ChIP-seq on embryonic 14.5 day mouse heart
ENCFF754NCW	ENCSR813SCQ	1, 2	H3K27ac ChIP-seq on embryonic 12.5 day mouse embryonic facial prominence
ENCFF463XJT	ENCSR675HDX	1, 2	H3K27ac ChIP-seq on postnatal 0 day mouse heart
ENCFF956JDU	ENCSR531RZS	1, 2	H3K27ac ChIP-seq on embryonic 11.5 day mouse neural tube
ENCFF309CWW	ENCSR574VME	1, 2	H3K27ac ChIP-seq on embryonic 15.5 day mouse heart
ENCFF583HBA	ENCSR884MYD	1, 2	H3K27ac ChIP-seq on postnatal 0 day mouse lung
ENCFF354CRO	ENCSR021ALF	1, 2	H3K27ac ChIP-seq on embryonic 14.5 day mouse limb
ENCFF354MWY	ENCSR424END	1, 2	H3K27ac ChIP-seq on embryonic 14.5 day mouse intestine

Table S3.2. 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)*
DE1 (+183) (mm1849)	TCCAACTAGCCACAATCCACTA	GGTTGACAAAGGTTCAGAAAGG	2486	chr3 66797249 66799734
DE2 (+297) (mm1852)	TCCTCTCTGTGTTTCAGCTTTG	TGGGTGACTCAGGTAAACCTCT	3167	chr3 66682968 66686134
DE3 (+330) (mm1853)	ACCATGGTAGGAAGTTCATTGG	GTTAGAGCTGTTGGGAAAATGC	4408	chr3 66650021 66654428
DE4 (+408) (mm1837)	GCTATACGCCGTCAGCTTTAGT	ATGTGAATGAAGCACAAATTGC	3236	chr3 66572737 66575972
DE5 (+437) (mm2108)	GATGTGGGGAAACTTCTGAAAC	TACAGACCCAGACAAGAGCAGA	4335	chr3 66542058 66546392
DE6 (+444) (mm1845)	GATGCAGGCACGATATACAAAA	AGACCTTACACACGTGCACAAC	2962	chr3 66535471 66538432
DE7 (+463) (mm2103)	CTGCGCTTTCTTCTTATCCCTA	CAGATCCACCTCTTCCTTCATC	3402	chr3 66518263 66521664
DE8 (+466) (mm1838)	GGAATTGCTTTGTAGCTCTGCT	CAGGGAGGAAGCTTCTAGTTCA	1816	chr3 66514951 66516766
DE9 (+475) (mm1846)	GACACCACCAAGAGTTCGTGTA	AATTACAATGTGTGGGGGAGAC	2824	chr3 66504602 66507425
DE10 (+487) (mm2109)	TCTCTATGACCAAACGGGCTAT	GGATTTGGAAGAACAAGAGGTG	3109	chr3 66492941 66496049
DE11 (+495) (mm2110)	CTGTGTATGCCTTTGCTCTCAG	CTCTGCTCATATTCTGCCTCCT	3067	chr3 66484145 66487211
DE12 (+501) (mm1839)	CTGCTCTAATTCTGGGAGGTTG	TTATTGCTTGGTGAGAATGTGG	3041	chr3 66478804 66481844
DE13 (+581) (mm1842)	TGTATTCCACAGCCTCCCTAGT	CCCAAGGTCTGGTTTAGAACTG	2511	chr3 66400179 66402689
DE14 (+583) (mm2111)	TCCTACAGGCAAGACCTCTCTC	CATGGTCCAACTGGTATTGATG	1942	chr3 66397740 66399681
DE15 (+606) (mm1843)	CATTGGTACTTGGGCTGAAAA	TTACAAAGCTCCTGACGCAGT	3139	chr3 66373217 66376355
DE16 (+656) (mm2112)	CAGAGGTCCTGAACTCAATTCC	TCCTGCTGTGCATAGAACAACT	2839	chr3 66324764 66327602

Table S3.3. 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 S3.4. 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 DE4 element (Fig. 3.2A**) 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 S3.5. 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 (**Fig. S3.4**).

Mouse allele	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	TGATCTTCATAACTGCCATGGGG	TGAAGCACAAGGCTGGCGGGAGG

Table S3.6. Primers used for screening and genotyping of CRISPR deletion mouse strains. PCR genotyping results using agarose gel electrophoresis are shown in **Fig. S3.4**. 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 S3.7. 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 S3.8. 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. S3.5 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.

Distance to Shox2 TSS (kb)	GDE element ID	Vista ID	*Coordinates (mm10)*	Forward primer	Reverse primer
-154	-	mm1848	chr3 67134438 67137518	TGCCAATTTGCAATTGTATCAC	GCAGCACTTTTCTTTCATCACA
+184	**GDE1**	mm1849	chr3 66797249 66799734	TCCAACTAGCCACAATCCACTA	GGTTGACAAAGGTTCAGAAAGG
+224	-	mm2113	chr3 66756856 66759558	GGGTACTGTGGTTGTCTTGTCA	TGGTTGTAGTACGAACAAAGTTGG
+236	-	mm1850	chr3 66743678 66747875	GAGGCACTAGGAACCAAAAAGA	AAGCAGACCTGAAAAGCAGAAG
+242	-	mm1851	chr3 66737978 66741264	ACTGACTTCTGCAGTGGCATTA	TGTAGGCAAGTGTGGGAGACTA
+280	-	mm2107	chr3 66699740 66703287	TCCGAAGGTCCTGAACTTAAAA	CAATGTTCACTTCCAACAGCAT
+283	-	mm2104	chr3 66697202 66700755	ACCCATCTCATTTTCCAACATC	AGCAGACATCTTGCCTATGGAT
+297	**GDE2**	mm1852	chr3 66682968 66686134	TCCTCTCTGTGTTTCAGCTTTG	TGGGTGACTCAGGTAAACCTCT
+319	-	mm2105	chr3 66660575 66664037	GGTCAGGAATTCAGAGGTCAAC	ATACATCTGGGTTTGTCCATCC
+325	-	mm2106	chr3 66655648 66658676	GCCATTATGGTCTTGAAGGAAG	ACTGACCCTTCACAGACTGGTT
+326	-	mm2114	chr3 66654695 66657196	ATCAGCTCAGCTTTGGTTAAGG	GAATTCCTGATGCACTCTTTCC
+330	**GDE3**	mm1853	chr3 66650021 66654428	ACCATGGTAGGAAGTTCATTGG	GTTAGAGCTGTTGGGAAAATGC
+389	-	mm2099	chr3 66590716 66594023	TCTATAACGGAGCTGCACTTGA	GGCATTTGTGAGACATGAGAAA
+405	-	mm2102	chr3 66575695 66578731	GCAGAAACCATACACCATCAGA	TCTCTCCCAAAACATGACTGAA
+408	**GDE4**	mm1837	chr3 66572737 66575972	GCTATACGCCGTCAGCTTTAGT	ATGTGAATGAAGCACAAATTGC
+417	-	mm2101	chr3 66562316 66565856	CTGCCATAACATTTGTGCTGTT	AATGCTTGTTTCCCAGAAGGTA
+437	**GDE5**	mm2108	chr3 66542058 66546392	GATGTGGGGAAACTTCTGAAAC	TACAGACCCAGACAAGAGCAGA
+444	**GDE6**	mm1845	chr3 66535471 66538432	GATGCAGGCACGATATACAAAA	AGACCTTACACACGTGCACAAC
+463	**GDE7**	mm2103	chr3 66518263 66521664	CTGCGCTTTCTTCTTATCCCTA	CAGATCCACCTCTTCCTTCATC
+466	**GDE8**	mm1838	chr3 66514951 66516766	GGAATTGCTTTGTAGCTCTGCT	CAGGGAGGAAGCTTCTAGTTCA
+475	**GDE9**	mm1846	chr3 66504602 66507425	GACACCACCAAGAGTTCGTGTA	AATTACAATGTGTGGGGGAGAC
+487	**GDE10**	mm2109	chr3 66492941 66496049	TCTCTATGACCAAACGGGCTAT	GGATTTGGAAGAACAAGAGGTG
+495	**GDE11**	mm2110	chr3 66484145 66487211	CTGTGTATGCCTTTGCTCTCAG	CTCTGCTCATATTCTGCCTCCT
+501	**GDE12**	mm1839	chr3 66478804 66481844	CTGCTCTAATTCTGGGAGGTTG	TTATTGCTTGGTGAGAATGTGG
+515	-	mm2100	chr3 66465121 66468456	GGTTGACACAAGTAACCAGCAA	GCAAGCACTCTACCCCATAATC
+520	-	mm2115	chr3 66459872 66463718	GTATGTTGTGGGCTTTCTCCTC	ATGAATCCCATGTAAGCAAACC
+571	-	mm1841	chr3 66409836 66412436	GGTTTCAGTCAAAAGAGCCTGT	GCCCAAAAAGTCTTGATACTGG
+581	**GDE13**	mm1842	chr3 66400179 66402689	TGTATTCCACAGCCTCCCTAGT	CCCAAGGTCTGGTTTAGAACTG
+583	**GDE14**	mm2111	chr3 66397740 66399681	TCCTACAGGCAAGACCTCTCTC	CATGGTCCAACTGGTATTGATG
+606	**GDE15**	mm1843	chr3 66373217 66376355	CATTGGTACTTGGGCTGAAAA	TTACAAAGCTCCTGACGCAGT
+656	**GDE16**	mm2112	chr3 66324764 66327602	CAGAGGTCCTGAACTCAATTCC	TCCTGCTGTGCATAGAACAACT
+689	-	mm1847	chr3 66291332 66293883	TACAGCAGACCTTTTCTGTCCA	GAGGGAGACTTGAGTGGTCATC

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.* cell115. 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**, p(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).

Chapter 4 : Conclusion

In parallel with past and ongoing studies on transcriptional regulation is a longstanding interest to consolidate the data and insights from these works to define the sequence, biochemical, and other molecular properties that are relevant for enhancer identification[1,2]. Enhancers are one of many components that coordinate and effect transcriptional activity, and their presence and proper functioning is intimately connected with both developmental- and disease-related processes[3,4]. Chapter 1 provided an overview of enhancers, their roles in gene regulation, and the genome-wide properties (*i.e.,* enhancer-associated chromatin marks) commonly used for their identification. In Chapter 2, I reported that while these commonly used enhancer-associated chromatin marks successfully identify *in vivo* validated enhancers that are active in the corresponding developmental tissue, there are a notable portion of active enhancers that are still missed by this approach. I used a large tiling study across two developmental loci (*e.g., Gli3*) to further assess in an unbiased manner these so-called hidden enhancers[5]. Indeed, this unbiased tiling study revealed additional *in vivo* active enhancers that did not have any of the enhancer-associated chromatin marks. In terms of sequence conservation, transcription factor motifs, and other related properties, these hidden enhancers were not distinguishable from marked enhancers. Additional data in the form of chromatin marks from an earlier developmental stage, related species, or single-cell based approaches could be used to identify some of these hidden enhancers. This work demonstrated in both the retrospective and tiling studies the technical limitations of current epigenomic data for identification of mouse *in vivo* validated enhancers. Ongoing developments in sequencing, imaging, and related chromatin profiling technologies (*e.g.,* single-cell chromatin accessibility) that better resolve the chromatin dynamics and transcriptional activities associated with these enhancers have the potential to greatly improve upon such approaches for enhancer identification and their subsequent characterization[6,7]. Chapter 3 reported an extensive study on the gene regulatory landscape that flanks the *Shox2* gene desert. Through the use of both enhancer-associated chromatin marks and also chromatin conformation capture, we identified multiple enhancers active in *Shox2* relevant tissues that include the limb, craniofacial structures, and heart[8]. These findings uncover additional components of the *Shox2* regulatory landscape that can be further explored in both developmental- and disease-related contexts. Altogether, the works presented here (including those in the addendum) demonstrate the value of harnessing enhancer-associated chromatin properties to sift through both the expansive genome and dynamic epigenome to find candidate enhancers for investigation of their gene regulatory activities[5,8–12]. Though we can expect additional layers of gene regulatory complexity to be revealed as sequencing and related computational approaches are continually applied to refine the full human genome, to represent the diversity of human populations, and to finely resolve the multitude of cell types and their transcriptional activities, we can look forward to the insights these layers will provide toward our understanding of transcriptional enhancers and their contributions in disease and development.

References

1. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
2. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics* (2020).
3. Rickels, R. & Shilatifard, A. Enhancer Logic and Mechanics in Development and Disease. *Trends Cell Biol.* **28**, 608–630 (2018).
4. Kvon, E. Z., Waymack, R., Elabd, M. G. & Wunderlich, Z. Enhancer redundancy in development and disease. *Nat. Rev. Genet.* (2021).
5. Mannion, B. J. *et al.* Uncovering Hidden Enhancers Through Unbiased In Vivo Testing. *bioRxiv* 2022.05.29.493901 (2022).
6. Hafner, A. & Boettiger, A. The spatial organization of transcriptional control. *Nat. Rev. Genet.* 1–16 (2022).
7. Preissl, S., Gaulton, K. J. & Ren, B. Characterizing cis-regulatory elements using single-cell epigenomics. *Nat. Rev. Genet.* **24**, 21–43 (2023).
8. Abassah-Oppong, S. *et al.* A gene desert required for regulatory control of pleiotropic Shox2 expression and embryonic survival. *bioRxiv* 2020.11.22.393173 (2020).
9. Galang, G. *et al.* ATAC-Seq Reveals an Isl1 Enhancer That Regulates Sinoatrial Node Development and Function. *Circ. Res.* **127**, 1502–1518 (2020).
10. Snetkova, V. *et al.* Ultraconserved enhancer function does not require perfect sequence conservation. *Nat. Genet.* **53**, 521–528 (2021).
11. Padhi, E. M. *et al.* Coding and noncoding variants in EBF3 are involved in HADDS and simplex autism. *Hum. Genomics* **15**, 44 (2021).
12. George, R. M. *et al.* Single cell evaluation of endocardial HAND2 gene regulatory networks reveals critical HAND2 dependent pathways impacting cardiac morphogenesis. *Development* **150**, (2023).